The Springer Series on Demographic Methods and Population Analysis 31

David A. Swanson
Jeff Tayman

Subnational Population Estimates

Springer

Subnational Population Estimates

THE SPRINGER SERIES ON

DEMOGRAPHIC METHODS AND POPULATION ANALYSIS

Series Editor

KENNETH C. LAND

Duke University

In recent decades, there has been a rapid development of demographic models and methods and an explosive growth in the range of applications of population analysis. This series seeks to provide a publication outlet both for high-quality textual and expository books on modern techniques of demographic analysis and for works that present exemplary applications of such techniques to various aspects of population analysis.

Topics appropriate for the series include:

- General demographic methods
- Techniques of standardization
- Life table models and methods
- Multistate and multiregional life tables, analyses and projections
- Demographic aspects of biostatistics and epidemiology
- Stable population theory and its extensions
- Methods of indirect estimation
- Stochastic population models
- Event history analysis, duration analysis, and hazard regression models
- Demographic projection methods and population forecasts
- Techniques of applied demographic analysis, regional and local population estimates and projections
- Methods of estimation and projection for business and health care applications
- Methods and estimates for unique populations such as schools and students

Volumes in the series are of interest to researchers, professionals, and students in demography, sociology, economics, statistics, geography and regional science, public health and health care management, epidemiology, biostatistics, actuarial science, business, and related fields.

For further volumes:
http://www.springer.com/series/6449

David A. Swanson • Jeff Tayman

Subnational Population Estimates

Springer

David A. Swanson
Department of Sociology &
The Center for Sustainable
Suburban Development
University of California Riverside
Riverside, CA 92521, USA

Jeff Tayman
Department of Economics
University of California San Diego
Gilman Drive 9500-0508
La Jolla, CA 92093, USA

ISSN 1389-6784
ISBN 978-90-481-8953-3 ISBN 978-90-481-8954-0 (eBook)
DOI 10.1007/978-90-481-8954-0
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2012937295

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In many areas of applied demography one of the most difficult tasks, in either gaining mastery of the content of an area or of teaching courses in the area is finding a set of materials that adequately covers the field without having to access an inordinate number of partial sources, none of which alone provides a sufficient overview for mastering the area. This has clearly been the case in the area of small area population estimation where, although there are useful overviews of some components of the processes and principals involved (see for example, Siegel and Swanson 2004), the need to examine both the academic and the pragmatic aspects of the methods and principles for completing, evaluating and knowledgeably using a set of estimates, have simply not been available in a single source.

For the first time a book which is both a comprehensive and rigorous scholarly work as well as a user oriented and pragmatic methodological source has become available with the publication of this text by Swanson and Tayman. In fact, I believe that it will become for those who do small area population estimates what Shryock and Siegel and Siegel and Swanson have provided for basic demographic methods– the source for learning how to approach, complete and evaluate small area population estimates in a variety of settings and considering a wide range of theoretical and pragmatic factors.

Its authors' credentials for doing such a work are flawless. Together they have more than a half century of experience in making, evaluating, and presenting estimates made in their capacities as demographers for state and local governments, public sector utilities, for private businesses and corporations, and as academic demographers estimating population change under challenging environmental and socioeconomic conditions. They not only know this area as practitioners but are also at the forefront of the academic literature in this area publishing widely and frequently on the methods, the evaluation and the use of small area population estimates.The result for the reader is a truly comprehensive volume on small area population estimation. The volume defines what an estimate is, what factors must be considered in selecting the estimation method to be used, and what data are needed to complete different types of estimates and where they can be found. It examines the methods available to complete estimates (e.g., extrapolation,

housing unit, regression-base, censal-ratio, component, sample-based, geographic based). It presents methods and measures for evaluating estimates, and examines how estimates fit into comprehensive programs for population analysis. This work also provides a broad range of quantitative examples of the applications of estimation methods and their evaluation and use under alternative circumstances.

When examined in its totality, however, the work is not only likely to become a basic text for the applied demographer but will also be of substantial utility for the general demographer attempting to operationalize variable parameters in an analysis, estimating values for periods where complete data are lacking and evaluating patterns of estimation errors. It will be a useful addition to both applied and basic demographers' collections of readings. Although its length will likely be daunting to some it is structured to allow one with basic knowledge of estimation processes to use it as a reference source for examining specific, as well as generic, issues impacting population estimation processes.

In sum, the volume you are now reading is one of those seminal pieces of work by a set of experienced authors. The work reflects the wide background and depth of experience and knowledge of its authors and there is no doubt that it is the most useful single source available on small area population estimates. Not only current but I believe students and other scholars a generation from now will thank the authors for leaving a record of the application of their basic and applied knowledge of small area estimates. I know that I have found my basic text for the next time I teach either a course in applied demographic methods or one on general demographic methods.

Reference

Siegel, J. S. & Swanson, D. A. (2004). *The methods and materials of demography, Second Edition*. Amsterdam: Elsevier Academic Press.

Steve H. Murdock
Allyn R. and Gladys M. Cline Professor of Sociology
Department of Sociology
Rice University
Houston, Texas 77005-1892, USA

Acknowledgements

The authors are grateful to Jack Baker, Eddie Hunsinger, Jerry McKibben, and Lynn Wombold for substantive suggestions and to Fred Cavanaugh, Chuck Gossman, Theresa Lowe, Don Pittenger, Lucky Tedrow, Ravi Verma, Paul Voss, the late John Walker, Meyer Zitter, and especially, Bob Schmitt, for planting ideas about population estimation methods many years ago.

Although indirect, we are deeply grateful to the late Professor Calvin F. Schmidt who put together the Washington State Census Board in 1945 while he was a Professor of Sociology at the University of Washington. The estimation methods pioneered by the Board have had an impact far beyond the borders of Washington, as have the many graduate students trained under its auspices.

We also are grateful to Jim Dannemiller, Beth Jarosz, Stefan Rayer, and Stan Smith, and Yi Zhao for providing data. In addition to providing data, Stan Smith, has provided a wealth of ideas over the years that has benefited not only us, but the entire field of population estimation (and projections).

We thank Steve Murdock for writing the preface. It would be hard to find a person more qualified, given his contributions to applied demography and his knowledge of the major vehicles for developing demographic information: enumeration; estimation; and projection. We are grateful for his kind words.

We would like to acknowledge the San Diego Association of Governments and other Associations and Councils of Governments whose innovative work on small area estimates goes largely without recognition. We also acknowledge Lawrence Weisser, from whom we both learned the value of field work and the importance of "ground truth," an important aspect of estimates work that also is largely without recognition.

Above all, we express our gratitude to our wives—Rita and Melinda—for their love, encouragement, and understanding as we worked on this project.

Contents

Chapter 1
Introduction

Although subject to flaws, the most complete and reliable source of information on a population is taken from a census (Bryan 2004a, 2004b; Swanson and Walashek 2011). However, a complete enumeration of a population is costly and not all populations have been subject to a census. Even in countries such as the United States, where census counts have been mandated since 1790, their high costs only allow them to be done once every ten years. This means that data can become outdated and that a substitute is needed – a set of population estimates. The development of methods of population estimation roughly corresponds to the development of censuses and vital statistics registries. For example, in the late 18th century, the French mathematician, Laplace, was using what we would today call a censal-ratio method in combination with recorded births and a population sample to estimate the population of France (Stigler 1986: 163-164). However, methodological development really only took off in the late 1930s and early 1940s, fueled in large part by the need for low-cost and timely information generated by the great depression of the 1930s and World War II. (Bryan 2004a; Eldridge 1947; Hauser and Tepping 1944; Shryock 1938; Shryock and Lawrence 1949; US Census Bureau 1945, 1949). In the United States, the Census Bureau played a major role in this effort, but it was not alone. During the early 1940s, the Washington State Census Board, for example, developed a comprehensive program of annual population determinations based on estimation methods that are still used today (Swanson and Pol 2005). Around this same time, demographers also began developing estimation methods for what were then called "underdeveloped countries," (Brass et al. 1968, Chandrasekaran and Deming 1949; Davis 1951; Popoff and Judson 2004) and the use of sample surveys as a substitute for complete census counts took hold (Bryan 2004a; Featherman 2004).

Today, population estimates are ubiquitous. They are done around the world by a host of governmental and non-governmental entities, as well as individual consultants (Bryan 2004b; Siegel 2002; Swanson and Pol 2005). The widespread availability of data, methods, and technology has made it possible for many people not only to develop estimates, but to do so more quickly and

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_1,

less expensively than has ever been done before. This trend is not likely to abate, but it carries with it a cost in that estimates may both be made and used with little or no understanding of the issues involved, what constitutes good estimates, and how to identify them. This book is designed to provide guidance on these issues and advice on how both to make and identify good estimates. Before we proceed, though, it is good to talk about what estimates are – and what they are not.

1.1 What is a Population Estimate?

A population *estimate* is the determination of the size or the characteristics of a population at a current or past date in the absence of census data for the same date. An estimate generally makes use of historical census data and data correlated with the population (s) in question, such as vital records (e.g., births and deaths) and administrative records (e.g., school enrollments, covered employment, automobile registrations, housing permits). However, there are ways in which an estimate can be done that do not rely directly upon either vital or administrative records, but rather on mathematical models or sample surveys.

The term population estimate is frequently used in the public domain to refer to the determination of the size or the characteristics of a population at a future date. However, most demographers prefer to use the term *projection* when talking about the possible size and characteristics of a population in the future. In developing a portrait of a given population in the future, it is not uncommon for a series of projections to be made that incorporate a range of plausible assumptions (e.g., expected trends in fertility, mortality, and migration). However, when one of these projections is selected as representing the most likely future, it then becomes the *forecast* for the population in question.

As opposed to a projection or a forecast, then, a population estimate is concerned with either the present or the past, but not the future (Smith, Tayman, and Swanson 2001: 3-4). In regard to this temporal dimension, we find it useful to make three distinctions in terms of estimates that provide a means of organizing techniques that we discuss in this book: (1) pre-censal; (2) inter-censal; and (3) post-censal. This temporal classification is useful because different methods are typically employed in the development of inter-censal, post-censal, and pre-censal estimates (Bryan 2004b). It also serves to keep an important principle in mind. Namely, that one should make full use of census information, vital statistics, and other relevant administrative records in developing estimates relative to the cost and resources required to make them usable. In turn, this principle serves to guide the selection of estimation methods. For example, since the data from the two censuses that bound an inter-censal estimate date contain information that both implicitly and explicitly bound the estimate itself, the principle

suggests that if these data are readily available, then an interpolation method is more likely to produce more accurate estimates than an extrapolative method and, as we discuss shortly, be of higher utility.

In using the term "pre-censal," we are referring to a time period prior to the initiation of census counts for the population in question. Since we are focusing on populations for which good census and administrative records data are available, this implies that the estimates are generally for a period in the distant past, to include pre-history. As it implies, we are referring to a period between census counts when we use the term "inter-censal." As such, we are referring to estimates for a time in the past, but not one that precedes the availability of census counts for the population in question. We view the term "post censal" as one that refers to a current point in time or the very recent past.

An estimate can be prepared for a nation or a subnational area such as a state, county, city, town, or census tract. An estimate also can be prepared for groups of subnational areas, groups of nations, or even the world as a whole. As the title and the examples reveal we focus on subnational estimates in this book, but virtually all of the methods we describe could be used at the national level. The major issue distinguishing national from subnational estimates is the fact that there is no domestic migration to account for at a national level.

The principal demographic characteristics for which an estimate is made include age and gender. However, in multiracial and multi-ethnic countries such as the United States and Canada, an estimate might be done not only by age and gender, but also by race and ethnicity. An estimate also can be made of social and economic subgroups of the population, households, and families.

1.2 How are Estimates Done?

Demographers and statisticians have developed a population estimation toolkit that contains a range of methods designed to meet different information needs at varying levels of accuracy and cost. The methods can be roughly placed into three categories: (1) analytical and statistical models that use data symptomatic of population and its changes; (2) mathematical models that use historical census data; and (3) sample surveys. Methods falling into the first category have generally been developed by and for applied demographers, most of whom work for national, state, and local governments. Methods falling into the second category have generally developed by and for academic demographers, most of whom work at universities and research institutes. The methods falling into the third category have generally been developed by and for statisticians and survey research scientists, but they also are widely used by demographers. Not surprisingly, there also are techniques that combine methods from two or even all three categories.

1.3 What makes a Good Estimate?

Without question, an estimate should be accurate, but accuracy is not the only criterion by which an estimate should be judged. Following the argument presented by Swanson and Tayman (1995), we suggest that attention be focused on the broader concept of utility. As alluded to earlier, there are many methods that in principle can be used to estimate a population, and improvements are a regular feature of these methods. Further, there is a wide range of decision-making situations in which population estimates are used. It follows, therefore, that no method should be universally judged to be superior to others and, by the same token, neither should any method be judged universally inferior to all others. We suggest instead, that relative to a given use, utility is gained by selecting a method that provides a sufficient amount of information for the purpose(s) at hand, while keeping cost and time to a minimum. In the case of an estimate, the sufficiency of the information provided is judged on the ability of using it to make good decisions. So, if an estimate is produced at minimal cost but provides timely information sufficient to make good decisions, then it has high utility. If an estimate does not meet these conditions then it has low utility. This follows the principle we described earlier that an estimate should make full use of census information, vital statistics, and other relevant administrative records, given the time and resources required to use them. An important underlying component of sufficiency is "transparency." That is, the ability of a decision-maker to understand how an estimate was done so that he or she can determine if the assumptions, methods, and data are reasonable.

1.4 Who makes Population Estimates?

Following World War II, many agencies responded to the demand for timely and low-cost population information. The United Nations started publishing estimates in its annual *Demographic Yearbook* in 1948 and in 1947 the US Census Bureau began publishing them regularly in its series, *Current Population Reports*. Since that time, many national and subnational statistical offices, as well as private vendors and consultants have also issued population estimates (Bryan 2004a, 2004b; Swanson and Pol 2005).

Today, National Statistical offices (e.g., The US Census Bureau), a number of sub-national governmental offices, non-profit organizations, and private sector firms publish population estimates (as well as projections and forecasts) on a regular basis. International agencies such as the United Nations also provide estimates. The US Census Bureau makes its estimates available at no cost on its website (http://www.census.gov/population). Statistics Canada, however, charges for virtually all of its products. The estimates done by state demographic centers are usually available for a nominal fee, and following the lead of the Bureau, many of the centers have websites. Private sector firms also make estimates, some of which are available on websites, but for a fee.

Although the Census Bureau produces sub-state estimates, this largely remains the domain of state demographic centers, local governmental entities, and the private sector. For their part, the state demographic centers have rarely ventured below the county and city level (e.g., they do not routinely make estimates for census tracts, although one notable exception is the state of Washington's demographic center). This terrain is claimed mainly by local government entities and the private sector. However, given the rapid advent of GIS (Geographic Information Systems) and other technological advances, there is much talk in the air of the Bureau and more than a few state demographic centers developing sub-county estimates. However, while it is possible that the Bureau and state demographic centers may extend their interest to lower levels of geography, it is highly unlikely that this would occur in the opposite direction - the Washington State Demographic Office is not likely to start doing estimates for the entire United States.

There is coordination between the Bureau of the Census and state demographic centers, accomplished mainly through the Federal State Cooperative Program for Population Estimates (FSCPE). Although some informal cooperation existed by the early 1960s, it was not until the latter part of that same decade that the Census Bureau and State agencies agreed, among other things, to establish close working relationships in the preparation of State population estimates and to facilitate the flow of technical information on population estimates between States. The FSCPE has remained operational ever since (Bryan 2004b: 525).

The situation in Canada is similar to that found in the United States. Starting in the 1940s with national estimates, Statistics Canada now prepares estimates for the country as a whole, its provinces and territories, and statistically-defined areas such as census divisions (Statistics Canada 1987: 2). Estimates done by Statistics Canada can be found at http://www.statcan.ca.

There are provincial demographic centers that prepare estimates specific to their own provinces and sub-areas such as counties. These include centers in Nova Scotia, Quebec, Ontario, Manitoba, and British Columbia; the provinces without centers use the estimates done by Statistics Canada. However, Canada does not have a federal-provincial program similar to the FSCPE found in the United States. As is the case in the United States, there also are firms in the private sector active in preparing estimates for Canada and its subareas.

There also are academics and others who make population estimates, most of which are pre-censal and inter-censal (Brass et al. 1968; Coale and Zelnick 1963; Nordycke 1989; Lee 1985; Schmitt 1977; Reher and Schofield 1993; Wrigley and Schofield 1981).

1.5 Why make Population Estimates?

So far, we have discussed what estimates are (and are not), how they are made, and who makes them. At this point you may be asking yourself, why are they made? The principal uses of population estimates and other demographic estimates relate

to government or private planning, particularly in regard to resource allocation (Siegel 2002: 398-399; Statistics Canada 1987: 2-3; Swanson 1980). Demographic estimates may be used directly or as the basis for preparing other more specialized types of information. These include, for example, estimates of the number of people of working age in a given labor market area, birth and death rates, the incidence of AIDS cases, the demand for assisted care centers, customers, households, and so on. The users include local, state and national governments, business firms, university research centers, and non-profit organizations. In addition to the uses in the field of planning, there are important uses in demographic analysis and related types of scientific studies.

As mentioned earlier, population estimates are widely used to make decisions. The following examples provide an idea of some of the real world uses of population estimates. As it turns out, each of the examples illustrates estimates that had high utility. That is, they were produced at minimal cost and provided timely information sufficient to make good decisions.

1.5.1 Political Redistricting in Florida

Using post-censal estimates, a team of demographers from Florida State University assisted Palm Beach County, Florida in the design of new voting districts following an amendment to the county's charter approved by voters in 1988 (Serow et al. 1997). Because of strong evidence of population change in Palm Beach County, the Board of County Commissioners decided that it would be best to use current population estimates rather than 1980 census data in designing the new voting districts. The Board then selected the team from Florida State University to prepare the estimates and to recommend voting district boundaries. In the first part of their work, the team used two standard techniques, the housing unit method and a regression method, to develop estimates of very small pieces of geography within the county (827 Traffic Analysis Zones). In the second part, the team developed five alternative plans that were presented and discussed in public forums and with the Board. From these meetings, a consensus emerged in favor of one of the plans and with slight modifications, it was approved by the Board in 1989.

1.5.2 The Country Mart Store, Omaha, Nebraska

Louis Pol (1988) describes the case of the Country Mart Store, a locally owned and operated grocery store in Omaha that was facing a possible reduction of customers in its market area. Until 1986, the store's owner, Wilbur Fast, had witnessed strong sales since purchasing it in 1982. By 1986, however, Mr. Fast was concerned and sought advice from the School of Business at the University of Nebraska, Omaha. Two sets of population estimates for 1986 using different data sources: (1) records

assembled by the Country Mart Store on customer locations; and (2) population estimates produced by a national vendor for zip codes covering the store's market area. The two sets were in agreement with one another, which indicated that the resulting single estimate was both valid and reliable. This 1986 estimate was compared with census and other data from 1980, and it was found that the population had actually increased since 1980. However, there also were indications that it was changing in terms of average age and household structure, factors important to retailers. Based on the analysis, Mr. Fast was advised on a marketing strategy that he then employed.

1.5.3 How many Visitors are in Hawai'i?

In 2002, tourists spent nearly 10 billion dollars in the state (Hawai'i Department of Business, Economic Development, and Tourism 2004) and the state depends heavily on tourism. In recognition of this dependence, the state initiated a series of data collection and estimation efforts many years ago, largely under the guidance of Robert C. Schmitt, the Hawaii State Statistician. Today, Hawai'i remains the only governmental unit in the country to systematically and regularly estimate its "de facto" population. In 2002, the state estimated that 6.45 million people visited Hawai'i, staying an average of nearly 9.4 days, and spending an average of $155 daily (Department of Business, Economic Development, and Tourism 2004). The state uses this information in the development of its budgets and the tourism industry uses it for planning.

1.5.4 The Impact of European Contact on Native Hawaiians

Although controversies exist over the exact numbers and cases, there is agreement that European contact led to population decline for many cultures, Hawai'i among them. Schmitt (1968) reports estimates of the Native Hawaiian population for the period 1778-79 from 100,000 to 400,000. Schmitt (1968) himself developed annual estimates that indicated a substantial population decline subsequent to European contact, from 130,300 in 1832 to 93,500 by 1848. By the time of the government of Hawai'i put together a reasonable (although one that is acknowledged to have undercounted the population) census in 1850, only 84,165 people were counted (Schmitt 1968). In successive census counts, accuracy improved as the counts continued to decline through 1878, when only 55,800 were counted (Schmitt 1968). Following this census and the annexation of Hawai'i by the United States in 1896, enumerations indicate the population started to increase, but it was assisted by migration of people into Hawai'i (Nordycke 1989; Schmitt 1968, 1977).

Now that you have an idea of what population estimates are, how they are made, by whom and why they are made - as well as some idea of their limitations, let me

turn to what is in store for you in the remainder of this book, which is largely concerned with how they are made.

1.6 About this Book

The book is applied in nature. It is primarily designed as a guide for developing estimates of populations in small areas. Its major focus is on developing estimates within countries that have advanced public information systems, including regularly conducted census counts and a wide range of administrative record systems. Similarly, the book can be used by people working in the private sector who have access not only to the public information systems, but also proprietary data. Thus, the book is primarily aimed at people responsible for making population estimates in state and local government, the private sector, and non-profit organizations. This includes not only demographers, but land use planners, transportation planners, and market researchers,

This book also is intended to serve as a guide for those learning how to make population estimates. In this category are students and those already in full-time jobs, who have to teach themselves about estimates. It also can serve as a useful tool to persons who may not make estimates themselves, but have an interest in knowing how they are done. This latter group can encompass a wide range, from planners to survey statisticians.

In terms of classroom instruction, this book should prove adequate as a primary textbook in a course that is largely or exclusively focused on population estimates. However, an instructor will need to develop exercises because the book contains none. In addition, an instructor may need to provide some supplementary material. In the many courses in which population estimation is covered in a short module, the book should be useful as supplementary reading or reference.

The book is neither an "easy read" nor a highly mathematical treatise. It assumes that the reader has at least an undergraduate degree, with the typical reader being either in a planning position or a graduate student in a social science field, including planning. A chapter on basic demographic concepts and measures has been included for two reasons. First, for those who have had no formal demographic training, it serves as the foundation for the chapters that come later and second, having it in the book makes the book relatively self-contained.

Although the book really is designed to be used for population estimates at the sub-national level, it can be used to develop estimates for higher levels of geography given that small area estimates can be aggregated upward to higher levels of geography. However, if the real aim is to produce estimates at a high level of geography such as a state or province, there are more efficient methods available and, moreover, readily available estimates (Bryan 2004b).

The book is not all encompassing. There are many topics within the field of population estimation that it does not cover. For example, it does not cover methods used to estimate the foreign-born population, a topic covered in depth by Judson

and Swanson (2011). Rather, it focuses on more general needs and what generally works in terms of making population estimates for subnational areas in countries that have well-developed statistical information systems. The book is aimed specifically at geographical units that correspond to states/provinces, counties, census tracts and other small pieces of geography. The main thrust is on developing estimates of the total population, but methods for estimating the characteristics (e.g., age, race, sex) of populations in small areas also are discussed.

The book need not be read in the order that the chapters are presented. However, for the most part the later chapters assume knowledge of what is covered earlier. Similarly, the earlier chapters generally present more simple methods than do the later chapters. The book consists of three major sections, followed by a glossary and a subject and name index. In Section I of this book, this chapter is followed by chapters 2 through 4, which cover fundamentals of basic concepts, data sources and basic demographic measures. The material in chapters 2 through 4 is as much a review as it is a tutorial and in either case, the intent is to provide just enough to get someone through the material in the 2nd section of this book, which covers basic estimation methods.

Section II covers the methods used in population estimation and is really the heart and soul of the book. It is organized into seven chapters, starting with Chapter 5, which provides an overview of population estimation methods. Chapter 6 covers extrapolative methods, chapter 7, housing unit methods, chapter 8, regression methods, Chapter 9, censal-ratio methods and Chapter 10, component methods. Chapter 11 discusses sample-based methods, to include "SPREE" and "synthetic" methods. However, Chapter 11 also touches on related methods, those used by survey statisticians to extend the coverage of sample surveys to small areas by "borrowing strength" from other information. These methods are not described in detail since the focus of the book is on developing population estimates. We believe, however, that it is important for demographers and others who may not be familiar with these methods to at least have an idea of what they are. In Chapter 12, structural models are described, as well as methods based on administrative records and special types of samples. We discuss the methods described in Section II largely in terms of "post-censal" since they are largely used to develop estimates for a current point in time or the very near past. They also can be used for "inter-censal" estimates, but are not well-suited, if at all, for developing "Pre-censal" estimates. That is, estimates in the distant past.

Section III deals first with special cases and adjustments (Chapter 13), then with methods of evaluation (Chapter 14), and then guidelines for developing estimates (Chapter 15). Chapter 16 is devoted to the development of estimates for types of populations either not counted at all in a census or ones that are difficult for the census to count, to include daytime and seasonal and visitor populations, the homeless, and populations impacted by disasters. Chapter 17 continues this theme, but takes it back in time. It focuses on the estimation of inter-censal populations, but also includes some discussion on the estimation of pre-censal populations. The book concludes with Chapter 18, which looks at the future of population estimation methods.

We decided against discussing any software in the book for three reasons. First, software technology has been undergoing a period of rapid change as this volume was being prepared, and was likely to be outdated as we wrote. The second reason is that we believed the reader could implement any demographic method electronically using standard, readily available, spreadsheet and statistical software with only limited training and experience on computers. Third, we felt that for the present purpose it was more important to convey the logic of the methods rather than present a device for accomplishing the result without thorough training as to its purpose and interpretation. The book concludes with a glossary and a subject and author index. References for the citations in each chapter are found at the end of the respective chapter.

References

Brass, W., A. Coale, P. Demeny, D. Heisel, F. Lorimer, A. Romaniuk, and E. van deWalk. (1968). *The Demography of Tropical Africa*. Princeton, NJ: Princeton University Press.

Bryan, T. (2004a). "Basic Sources of Statistics." pp. 9–39 in J. Siegel and D. Swanson (eds.) *The Methods and Materials of Demography 2nd edition*. New York, NY: Elsevier Academic Press

Bryan, T. (2004b). "Population Estimates." pp.523–560 in J. Siegel and D. Swanson (eds.) *The methods and materials of demography 2nd edition*. New York, NY: Elsevier Academic Press

Chandrasekaran, C. and W. E. Deming. (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44: 101–115.

Coale, A. and M. Zelnick. (1963). *New Estimates of Fertility and Population in the United States: A Study of Annual White Births from 1855 to 1960 and of Completeness of Enumeration in the Censuses from 1880 to 1960*. Princeton, NJ. Princeton University Press.

Davis, K. (1951). *The Population of India and Pakistan*. Princeton, NJ: Princeton University Press.

Eldridge, H. T. (1947). "Problems and Methods of Estimating Post-censal Population." *Social Forces* 24: 41–46.

Featherman, D. (2004). "Foreword." pp. xi – xv in J. House, F. T. Juster, R. Kahn, H. Schuman, and E. Singer (eds.) *A Telescope on Society: Survey Research and Social Science at the University of Michigan and Beyond*. Ann Arbor, MI: The University of Michigan Press.

Hauser, P. and B, J. Tepping. (1944). "Evaluation of Census Wartime Population Estimates and of Predictions of Postwar Population Prospects for Metropolitan Areas." *American Sociological Review* 9: 473–480.

Hawai'i Department of Business, Economic Development, and Tourism. (2004). *Annual Visitor Research Report 2002*. Honolulu, HI: Hawai'i Department of Business, Economic Development, and Tourism. (http://www.hawaii.gov/dbedt/02vrr/).

Judson, D. and D. A. Swanson. (2011). *Estimating Characteristics of the Foreign Born by Legal Status: An Evaluation of Data and Methods*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Lee, R. 1985. (1985). "Inverse Projection and Back Projection: A Critical Appraisal and Comparative Results for England, 1539–1871," *Population Studies* 39: 233–248.

Nordycke, E. (1989). *The Peopling of Hawai'i, 2nd Edition*. Honolulu, HI; University of Hawai'i Press.

Pol, L. (1988). "Determining the Demographics of a Market Area." pp. 43–45 in T. Merrick and S. Tordella, Demographics: People and Markets. *Population Bulletin* 43 (1). Washington, D.C: Population Reference Bureau.

Popoff, C. and D. Judson. (2004). "Some Methods of Estimation for Statistically Underdeveloped Areas." pp. 603–641 in J. Siegel and D. Swanson (eds.) *The Methods and Materials of Demography 2nd Edition*. New York, NY: Elsevier Academic Press

Reher, D. and R. Schofield (Eds.). (1993). *Old and New Methods in Historical Demography*. Oxford, England: Clarendon Press.

Schmitt, R. (1968). *Demographic Statistics of Hawaii, 1778–1965*. Honolulu, HI: University of Hawai'i Press.

Schmitt, R. (1977). *Historical Statistics of Hawaii*. Honolulu, HI: The University of Hawaii Press.

Serow, W., E. W. Terrie, R. Weller, and R. Wichmann. (1997). "The Use of Inter-censal Population Estimates in Political Redistricting." pp. 33–54 in H. J. Kintner, T. Merrick, P. Morrison, and P. Voss (eds.). (1997). *Demographics: a casebook for business and government*. Santa Monica, CA: RAND

Shryock, H. (1938). "Methods of Estimating Post-censal Population." *American Journal of Public Health* 28: 1042–1047.

Shryock, H. and N. Lawrence. (1949). "The Current Status of State and Local Population Estimates in the Census Bureau." *Journal of the American Statistical Association* 44: 157–173.

Siegel, J. (2002). *Applied Demography: Applications to Business, Government, Law, and Public Policy*. San Diego, CA; Academic Press.

Smith, S., J. Tayman, and D. A. Swanson. (2001). *State and Local Population Projections: Methodology and Analysis*. New York, NY: Kluwer Academic/Plenum Publishers.

Statistics Canada. (1987). *Population Estimation Methods, Canada*. Ottawa, ON: Statistics Canada.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.

Swanson, D. A. (1980). "Allocation Accuracy in Population Estimates: an Overlooked Criterion with Fiscal Implications." pp. 13–22 in *Small area population estimates – methods and their accuracy*, Small Area Statistics Papers, Series GE-41, no. 7. Washington DC: US Census Bureau.

Swanson, D. A. and L. Pol. (2005). "Contemporary Developments in Applied Demography." *Journal of Applied Sociology* 21: 26–56. 2005

Swanson, D. A. and G. E. Stephan. (2004). "A Demography Time Line." pp. 779–786 in J. Siegel and D. A. Swanson (eds.) *The Methods and Materials of Demography 2nd Edition*. Elsevier Academic Press. New York, NY.

Swanson, D. A., and J. Tayman. (1995). "Between a Rock and a Hard Place: The Evaluation of Demographic Forecasts." *Population Research and Policy Review* 14: 233–249.

Swanson, D. A. and P. Walashek. (2011). *CEMAF as a Census Method: A Proposal for a Re-designed Census and an Independent US Census Bureau*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

US Census Bureau. (1945). "Suggested Procedures for Estimating the Current Population of Counties" *Population Special Reports, Series P-47. No. 4*.

US Census Bureau. (1949). "Illustrative Examples of Two Methods of Estimating the Current Population of Small Areas." *Current Population Reports Series P-25 No. 20*.

Wrigley, E., and R. Schofield. (1981). *The Population History of England, 1541 – 1871: A Reconstruction*. Cambridge, MA: Harvard University Press.

Chapter 2
Basic Concepts

Creating, interpreting, and evaluating population estimates involves demographic, geographic, and statistical methods and data. This chapter introduces the major demographic concepts of size, distribution, characteristics, and the components of population change along with geographic concepts including Geographic Information Systems (GIS), density, center of population, concentration and clustering, distance, accessibility, and spatial interaction. We conclude this chapter with material on the concepts of descriptive and inferential statistics, and regression techniques.

2.1 Demographic

Demography is the scientific study of population (Swanson and Siegel 2004: 1). It focuses on five general topics: population size, population distribution across geographic areas, population composition (e.g., age, sex, and race), population change, and the determinants and consequences of population growth. Our focus is on the first four of these topics, and the Glossary contains definitions for many other demographic concepts related to population estimates taken directly from (Swanson and Stephen 2004).

2.1.1 Size

Population estimates start with the same basic consideration as a census: What is the size of a population? The concept of population size refers to the number of people residing in a specific area at a specific time (the de jure approach). According to the latest census, The City of San Diego had population of 1,307,402 on April 1, 2010, whereas the City of Del Mar had a population of only 4,161. These were the largest and smallest cities in San Diego County in terms of population size. However, in the censuses of many countries the concept of population size refers to

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_2,

the number of people actually present in a given area at a given time (the de facto approach). Under this approach, all tourists, business travelers, and seasonal residents present in Miami on census day would be counted along with usual residents who are also in town that day. Usual residents of Miami who were out of town would not be counted. De facto population estimates have many uses including dealing with potential traffic congestion and long commuting times, disaster and relief activities to understand the number of people that may be affected if a disaster was to occur, and defining the at-risk population for crime and arrest rates.

The de jure concept is more ambiguous in that it comprises all of the people who "belong" to a given area by virtue of legal residence, usual residence, or some similar criterion (Wilmoth 2004: 65). However, the de jure concept is used as the census definition of population in the United States, Canada, and most other developed countries and, as such, becomes the dominant concept in population estimates. Not surprisingly, the dominant focus of this book is on the estimation of de jure populations, although there is some discussion of methods for estimating de facto populations in Chapter 16.

By the same token, regardless of which of the two concepts used, virtually all census data refer to the populations of given geographic areas and, as such, most estimates, whether of de jure or de facto populations, are done for given geographic areas. However, the concept of a population need not be linked to a geographic area. For example, a population could refer to all the dependents of employees working for a multi-national corporation or the potential customers of an insurance company. As such, estimates of these populations are not confined to for these given geographic areas. This book, however, is focused on the methods used to estimate populations in given geographic areas.

2.1.2 Distribution

The distribution of a population refers to its geographic location. As is the case with the concept of population size, there are two major ways in which geographic areas can be identified. The first is the administrative approach, where areas are defined according to administrative or political criteria. Examples include states, counties, and cities. For many purposes these are the most important types of geographic areas that can be defined and as such, they are used by most censuses in reporting census population data, including the United States (Plane 2004). However, administrative areas also have several limitations. Their boundaries may not account for important economic, cultural, and social considerations. For example, Gary, Indiana is administratively distinct from the city of Chicago, Illinois, but it is economically, culturally, and socially linked to it. Another problem is that administrative boundaries may not remain constant over time –the annexations of a city are a case in point– and changing boundaries make it difficult not only to make comparisons over time, but to produce consistent estimates

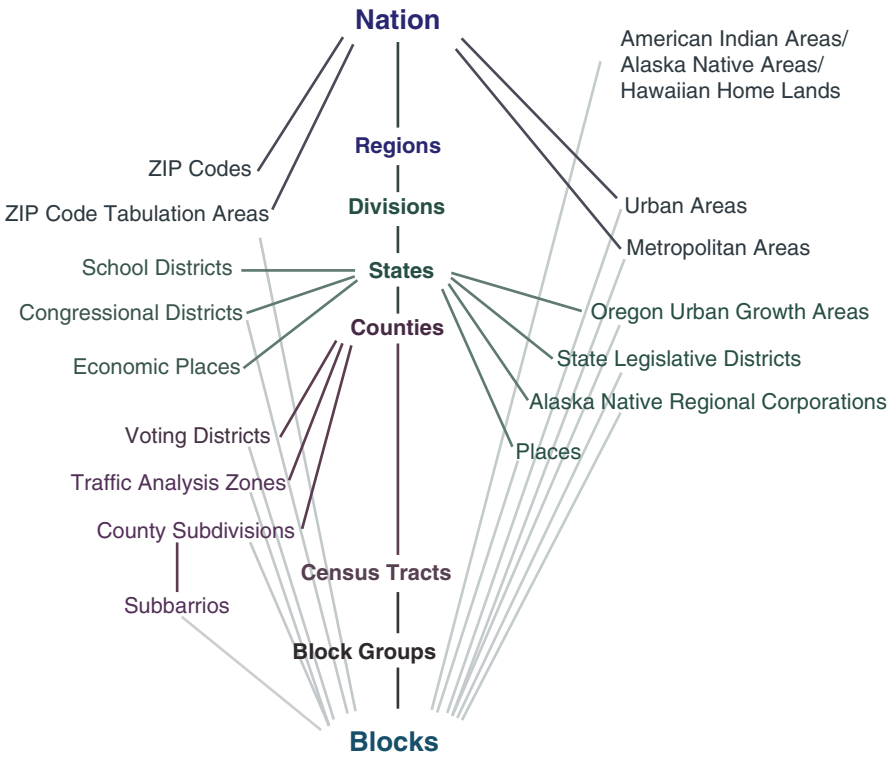


Fig. 2.1 Geographic Hierarchy for the 2000 Decennial Census
Source: Census 2000 Basics IV Geographic Areas. http://www.census.gov/mso/www/c2000basics/chapter4.htm

A way to get around of the limitations imposed by administrative definitions is to define geographic areas specifically for purposes of identifying areas that are economically, socially, and culturally linked that also are consistent over time. These so-called statistically defined areas are used in many countries, including the United States (Plane 2004, SANDAG 2010a).

In the United States, important statistical areas are based on geography used in the census–census blocks, block groups, and census tracts. *Blocks* are basically city blocks. They are small areas bounded on all sides by visible features such as streets or railroad tracks or by invisible boundaries such as city or township limits; they are the smallest geographic unit for which data are tabulated. *Block groups* are clusters of blocks and generally contain 250 to 550 housing units; block groups do not cross census tract boundaries. *Census tracts* are small, relatively permanent areas defined for all metropolitan areas and other densely populated counties. They do not cross county boundaries and generally contain between 2,500 and 8,000 persons and are designed to be relatively homogeneous with respect to population characteristics, living conditions, and economic status. Figure 2.1 shows a hierarchy of geographic

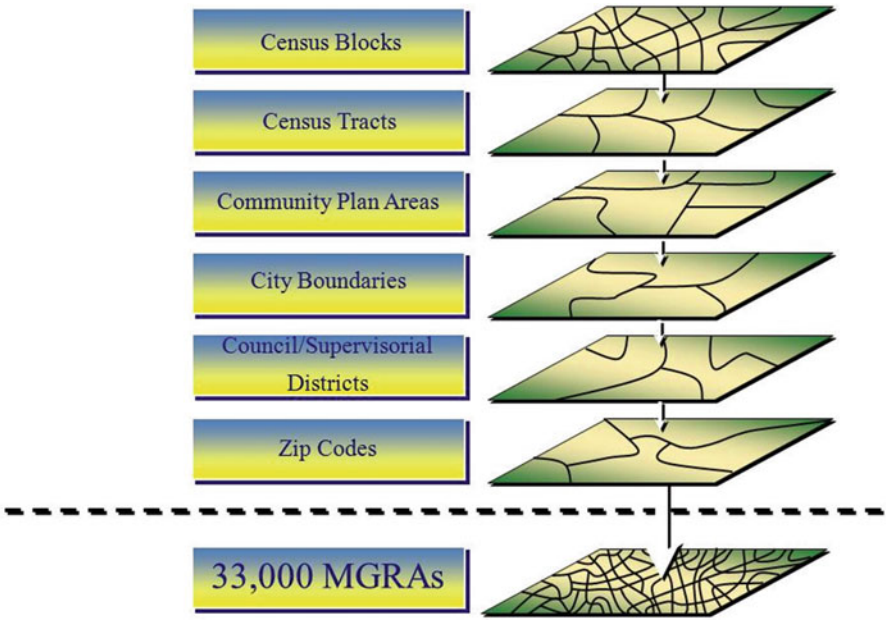


Fig. 2.2 Master Geographic Reference Areas in San Diego County
Source: San Diego Association of Governments (February 2010), 2050 Regional Growth Forecast

areas built from the 2000 census geography. Geographic areas may work in a hierarchical fashion, with smaller areas nesting in larger ones (e.g., census tracts within counties), while others like Metropolitan Areas are given only as subsets of the nation.

Geographic boundaries can also be defined according to other criteria. In the United States, for example, one can obtain estimates for Postal ZIP code areas. You may recall from Chapter 1 the example of the Country Mart Store, which defined its market area using ZIP codes. In another example from Chapter 1, the voting districts for Palm Beach County, Florida, were constructed using data from 827 traffic analysis zones. As the Palm Beach County example suggests, it is not uncommon to produce estimates and other forms of population data for a combination of administrative and statistical areas. Figure 2.2 shows an example of such a system used in San Diego, California known as the Master Geographic Reference Area (MGRA). It combines census geography, political boundaries, and zip codes into a spatially detailed spatial system that supports a wide range of uses.

2.1.3 Composition

Composition refers to the characteristics of the population. For population estimates, the most commonly used characteristics are age, sex, race, and Hispanic Origin. For many purposes, age is the most important demographic characteristic

because it has such a large impact on so many aspects of life, for individuals as well as for society as a whole. The age structure of a population affects its birth, death, and migration rates, and the demand for public education, health care, and nursing home care. It also impacts the housing market, the labor market, and the marriage market. No other characteristic is more valuable for a wide variety of planning and analytical purposes than the age composition of the population (Smith, Tayman, and Swanson 2001: 23). Sex composition also is important for many purposes. It is often used in combination with age to show a population's age-sex structure (Hobbs 2004).

The age-sex structure is often illustrated using *population pyramids* (Hobbs 2004: 161-166). Population pyramids are graphic representations showing the number (or proportion) of the population. The basic pyramid form consists of bars, representing age groups in ascending order from the lowest to the highest, pyramided horizontally on one another (see Figure 2.3). The bars for males are given on the left of a central vertical axis and the bars for females on the right of the axis. The characteristics of pyramids (e.g., the length of a bar to others, the steepness and regularity of its slope) for different populations quickly reveal any differences in the proportion of the sexes, the proportion of the population in any particular age class or classes, and the general age structure of the population (Hobbs 2004: 163).

Figure 2.3 shows pyramids for four populations with different age–sex structures. The pyramid for Uganda has a very broad base and narrows very rapidly. This pyramid illustrates the case of an age–sex structure with a very large proportion of children, a very small proportion of elderly persons, and a low median age. It reflects a "young" population with relatively high fertility rates. The pyramid for Sweden is very different. It has a relatively narrow base and a middle section of nearly the same dimensions. It illustrates the case of an age–sex structure with a very small proportion of children, a very large proportion of elderly persons, and a high median age. It reflects an "old" population and relatively low fertility rates. The pyramids for Argentina and China illustrate age-sex structures intermediate between those for Uganda and Sweden, with China showing the impact of its "one-child" policy in its youngest ages (0-14).

Race and ethnicity are two other widely used demographic characteristics. In the 2000 census the Census Bureau used five broadly defined racial categories: African American; American Indian or Alaska Native; Asian; Native Hawaiian or other Pacific Islander; and White (McKibben 2004). The 2000 census incorporated several changes in the collection of racial data. One important change is that the Census Bureau for the first time allowed respondents to list themselves as belonging to more than one racial category; prior to that time, respondents could list only a single category (McKibben 2004). In addition to race, the US Census uses an ethnic dimension, with two categories: Hispanic; and non-Hispanic (McKibben 2004). It should be noted that "Hispanic" is *not* a racial category; that is, people are classified both by race and by Hispanic origin. Composition also can refer to other characteristics such as employment status, income, education, and occupation (O'Hare, Pollard, and Ritualo 2004).

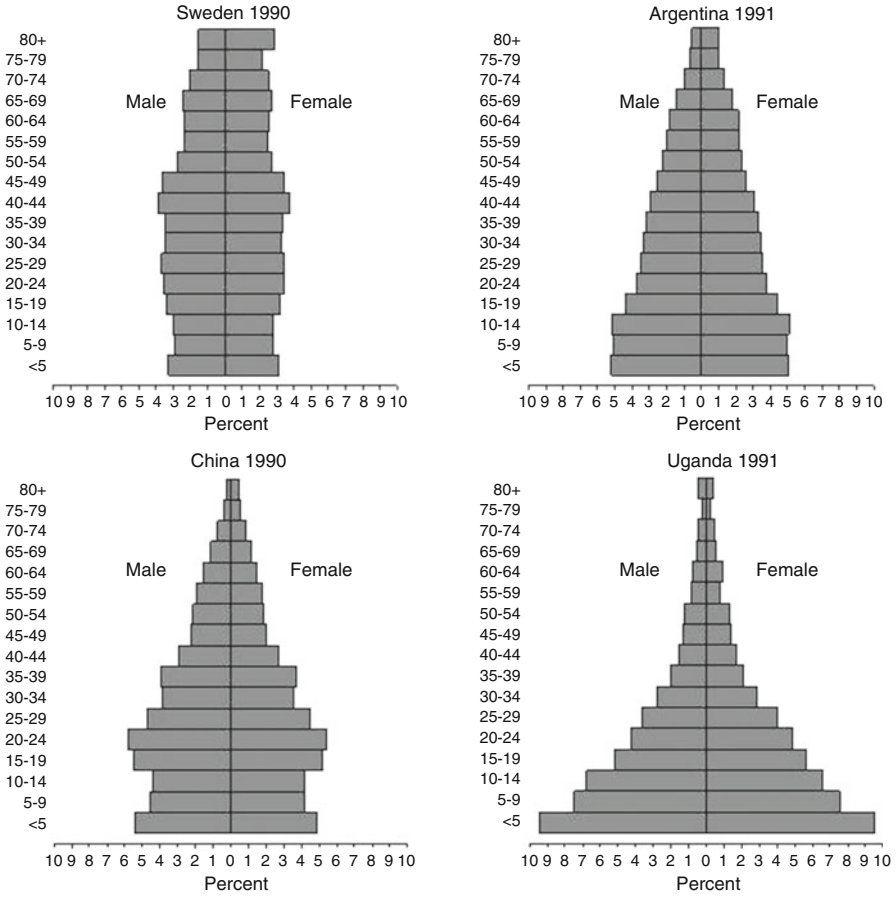


Fig. 2.3 Percent Distribution by Age and Sex of the Population of Sweden, China, Argentina, and Uganda Around 1990
Source: Hobbs (2004: 164)

Hispanics and race groups often have different demographic characteristics and patterns of growth that influence population estimation. For example, between 2000 and 2010 in Texas, the percentage change in the Hispanic population is more than double the overall population, similar to the percent change in American Indians and Alaskan Natives (see Table 2.1). Consequently, the Hispanic share of the total population increased from 32.0% in 2000 to 37.6% in 2010. Asians are the fastest growing race group, increasing by 71.5%, and almost 700,000 people in Texas identify themselves as belonging to 2 or more race groups in 2010. Non-Hispanic Whites, another widely use distinction, grew slowly during the first decade of the 21st century, causing its share of the total population to drop from 52.4% to 45.3%. Non-Hispanic Whites have the oldest age structure with a median age of 41.3 years

Table 2.1 Population by Race and Hispanic Origin in Texas, 2000 and 2010

			Change	
	2000	2010	Number	Percent
Total	20,851,820	25,145,561	4,293,741	20.6%
White	14,799,505	17,701,552	2,902,047	19.6%
Black or African American	2,404,566	2,979,598	575,032	23.9%
American Indian and Alaska Native	118,362	170,972	52,610	44.4%
Asian	562,319	964,596	402,277	71.5%
Native Hawaiian and Other Pac. Is.	14,434	21,656	7,222	50.0%
Other Races	2,438,001	2,628,186	190,185	7.8%
Two or More Races	514,633	679,001	164,368	31.9%
Hispanic Origin	6,669,666	9,460,921	2,791,255	41.8%
Non-Hispanic White	10,933,313	11,397,345	464,032	4.2%

Sources: US Census Bureau, Census 2000 and 2010, http://factfinder2.census.gov

Table 2.2 Median Age by Race and Hispanic Origin, Texas, 2010

Total	33.6
White	36.0
Black or African American	31.6
American Indian and Alaska Native	30.7
Asian	33.9
Native Hawaiian and Other Pac. Is.	27.6
Other Races	26.1
Two or More Races	20.9
Hispanic Origin	27.0
Non-Hispanic White	41.3

Source: US Census Bureau, 2010 Census, http://txsdc.utsa.edu/Resources/Decennial/2010/SF1/profiles/Texas_2010_SF1_Profile.pdf

in 2010, more than 14 years older than Hispanics (see Table 2.2). Native Hawaiian and Other Pacific Islanders and Other races have relatively low median ages (27.6 and 26.1) and younger age structures.

2.1.4 Change

Population change is measured as the difference in population size between two points in time (Perz 2004). A point in time can correspond to the date of a census or to the date of a population estimate. Measures of population change always refer to a specific population and a specific period of time; in most instances, they refer to a specific geographic area as well. Population change can also be measured for various subgroups of the population (e.g., females, Asians, or teenagers), different geographic areas (e.g., counties, cities), and different time periods (e.g., 1980-1990). In other words, population change can refer to changes in size, distribution, or composition, or to any combination of the three.

2.1.4.1 Components of Population Change

There are only three components of population change: births, deaths, and migration. A population grows through the addition of births and in-migrants, and declines through the subtraction of deaths and out-migrants. Understanding these three demographic processes is essential to understanding the nature and causes of population change. Fertility is the reproductive performance of a woman, man, couple, or group; it also is a general term for the incidence of births in a population or group (Swanson and Stephan 2004: 760). Although fertility rates are generally low in the US and other developed countries, they can vary substantially from place to place and from one race, ethnic or socioeconomic group within a given country. In 2003, the total fertility rate (average number of children per woman) for states ranged from 1.7 in Vermont to 2.7 in Utah (National Vital Statistics Reports 2010). Mortality is a general term for the incidence of deaths in a population or group (Swanson and Stephan 2004: 767). While mortality rates do not vary greatly within high income countries there are differences between race, ethnic and socioeconomic groups. In 2006, there was a 17.5 year difference in life expectancy (average number of years of remaining life) at birth between Black Males (69.4 years) and Asian Females (86.9 years) (LA County Department of Health 2010).

Migration is a general term for the incidence of movement by individuals, groups, or populations seeking to make permanent changes of residence (Swanson and Stephan 2004: 766). It refers to changes in usual place of reference and excludes short-term temporary movements such as commuting, visiting friends or relatives, or taking a business trip. The migration literature uses several terms to describe migration. Gross migration refers to the total number of migrants into or out of an area (e.g. 200 in-migrants and 300 out-migrants). Net migration is the difference between the two (e.g., a net outflow of 100); it shows the net effect of migration on the change in population. Internal or domestic migration refers to changes of residence within a county, while foreign or international migration refers to changes of residence from one county to another. People leaving a country are emigrants and those entering a country are immigrants. The migration level can vary considerably from place to place within the United States and can undergo large sudden changes. In San Bernardino County, California, for example, net migration for the years 2006, 2008, and 2010 was 7.548, -17,214, and -3,167 (State of California 2011).

2.1.4.2 Fundamental Demographic Equation

The overall change in a population is formalized in the fundamental demographic equation:

$$P_l - P_b = B - D + IM - OM$$

where P_l is the population at the end of the time period; P_b is the population at the beginning of the time period; and B, D, IM, OM are the number of births, deaths,

Table 2.3 Cumulative Estimates of the Components of Resident Population Change for Counties of Arizona: April 1, 2000 to July 1, 2009

Geographic Area	Total Population Change[a]	Natural Increase	Vital Events		Net Migration		
			Births	Deaths	Total	International[b]	Domestic
Arizona	1,465,171	464,238	875,726	411,488	986,764	272,410	714,354
Apache	1,168	6,829	11,465	4,636	−5,366	184	−5,550
Cochise	11,786	6,069	16,474	10,405	6,453	2,076	4,377
Coconino	13,531	12,722	18,473	5,751	1,515	2,022	−507
Gila	869	205	6,319	6,114	987	496	491
Graham	3,556	2,123	4,690	2,567	1,562	183	1,379
Greenlee	−506	479	997	518	−1,001	71	−1,072
La Paz	297	104	2,082	1,978	299	650	−351
Maricopa	950,964	338,001	564,289	226,288	632,032	215,566	416,466
Mohave	39,793	−636	20,655	21,291	41,241	2,777	38,464
Navajo	15,507	9,464	16,808	7,344	6,583	628	5,955
Pima	176,458	47,933	121,594	73,661	100,945	28,620	72,325
Pinal	161,242	18,224	35,399	17,175	131,833	4,890	126,943
Santa Cruz	5,390	4,996	7,233	2,237	631	2,313	−1,682
Yavapai	48,170	−1,423	19,235	20,658	50,085	3,125	46,960
Yuma	36,946	19,148	30,013	10,865	18,965	8,809	10,156

[a] Total population change includes a residual. This residual represents the change in population that cannot be attributed to any specific demographic component. See State and County Terms and Definitions at http://www.census.gov/popest/topics/terms/states.html.

[b] Net international migration includes the international migration of both native and foreign-born populations. Specifically, it includes: (a) the net international migration of the foreign born, (b) the net migration between the United States and Puerto Rico, (c) the net migration of natives to and from the United States, and (d) the net movement of the Armed Forces population between the United States and overseas.

Note: The April 1, 2000 estimates base reflects changes to the Census 2000 population resulting from legal boundary updates, other geographic program changes, and Count Question Resolution actions. All geographic boundaries for the 2009 population estimates series are defined as of January 1, 2009.

Source: US Census Bureau, Population Division (CO-EST2009 09 04-04), Release Date: March 2010

in-migrants, and out-migrants during the time period.[1] The difference between births and deaths (B − D) is called natural change coming from the population itself. It may be either positive (natural increase) or negative (natural decrease) depending on whether births exceed deaths or deaths exceed births. The difference between IM and OM reflects the change in population due to migration and can be either positive or negative depending on whether in-migrants exceed out-migrants or out-migrants exceed in-migrants. The fundamental demographic equation has a wide range of uses, including the development of estimates of population and net migration (Smith, Tayman & Swanson 2001: 30) and estimates of net census undercount (Robinson et al. 1993).

Table 2.3 shows natural increase and estimates of net domestic and foreign migration for counties in Arizona from 2000 to 2009. Domestic migration accounted for 49% of the population change in Arizona, followed by natural increase (32%) and

foreign migration (19%). There is substantial variability in the components of change among counties. Thirteen counties showed natural increase, while deaths slightly exceeded births in Mohave and Yavapai Counties. All counties showed positive growth due to foreign migration, but one-third of the counties lost population as the result of domestic migration. In counties with natural increase and positive total migration, the share of growth due to natural increase ranged from 11% to 94%.

2.2 Geographic

Population data are used to support private-sector marketing, business decision making, and public planning and policy making. For many purposes information on the size and characteristics of the population of a state or even a county is not sufficient. There is an increasing demand for population data for smaller scale areas that define more precisely where people live. Today population estimates are done for a wide range of subregional geographic areas including cities, census tracts, block, and parcels. Spatially intensive population estimates rely heavily on geographic methods to analyze, manage, create, and disseminate information. In this section, we discuss geographic information systems and some major geographic concepts.

2.2.1 Geographic Information Systems (GIS)

Geographic information systems (GIS) work with geographically (geo-) referenced data that are identified by coordinates that represent the position on the earth. GIS represent a unique combination computer hardware and software that are used to manipulate geo-referenced data. These systems provide four main capabilities: (1) input; (2) data management; (3) data manipulation and analysis; and (4) output (Aronoff 1989: 39). Many associate GIS with map making, but GIS has revolutionized our ability to create and display a wide range of graphical displays. GIS is much more than a map making utility. It has become a valuable tool for exploring spatial patterns, relationships, and predictive modeling (Bryan and George 2004).

GIS has many capabilities (e.g., Bryan and George 2004; Fotheringham, Brunsdon, and Charlton 2000: Chapter 3) and we illustrate a few that are relevant to population estimation. Say you had an address list of 800,000 electric meters used to analyze housing trends. How could you get a count of the meters for the 600 census tracts within the region? "Admatch" is a GIS procedure that can assign an address to any spatial location (see Figure 2.4). The address is matched to a street-based file that contains address ranges, geographical coordinates and or codes that define geographic areas. Once the matching is completed the addresses can be accumulated as desired. Tiger/Line street files developed by the Census Bureau are available for public use. Matching to street files approximate the location by interpolating within the address range on each side of the street. More precise locations can be obtained by matching to parcel file addresses (discussed in Chapter 3) (Tayman 1999).

Fig. 2.4 Admatch: Assign Activities to any Spatial Location
Source: Generated by the authors

Another important use of GIS is data retrieval and the ability to quickly create data profiles for any area, any shape, and any size (Hodges 1995; Sharkova 2000; Tordella 1987). What if you were asked to compute the current size and characteristics of the population at various distances from a store site? Figure 2.5 shows a travel time contour map that identifies various driving distances from a location in Carlsbad, CA; a city about 25 miles north of San Diego. The GIS technology used to create this map could prepare custom reports for each travel time contour.

Our last example deals with global positioning systems (GPS). GPS involves 24 satellites in low earth orbit (12,000 miles) that continuously beam their locations and temporal positions toward the earth (Bryan and George 2004). With GPS very precise coordinates can be found for any place on earth. The Census Bureau used GPS to verify Tiger/Line files and to determine the precise location of housing units for the 2010 census. Traditional travel diaries are known to undercount trips and GPS can help obtain more accurate information about travel times and volumes (Kreitz, Doherty, and Rindsfuser 2002). Figure 2.6 shows the traffic flows in 2001 for individual vehicles in a retail area in San Diego. This information was used to study travel patterns throughout the day, calibrate travel models, and develop daytime (de facto) population estimates.

2.2.2 Density

Population density is a simple concept that relates the size of a population to areal size of a particular geographic area where it is located. Density is usually computed

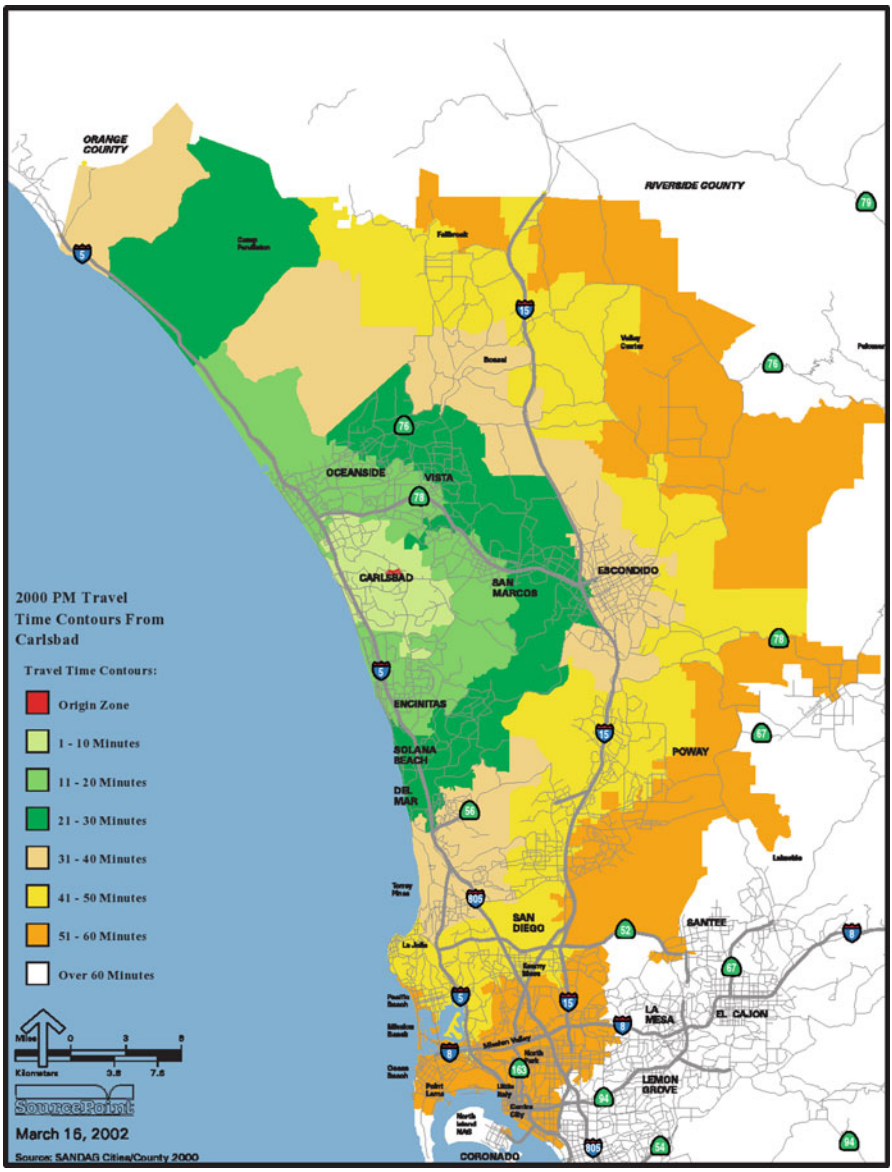


Fig. 2.5 Travel time distances from a Location
Source: San Diego Association of Governments (2002), Travel Time Contour Map

as population per square mile (or square kilometer in the metric system) or per acre of land area rather than gross area including land and water (Plane 2004). Population densities (per square mile) vary considerably across counties in the US (see Figure 2.7). High densities in 2009 are in the Northeast Corridor, on both coasts

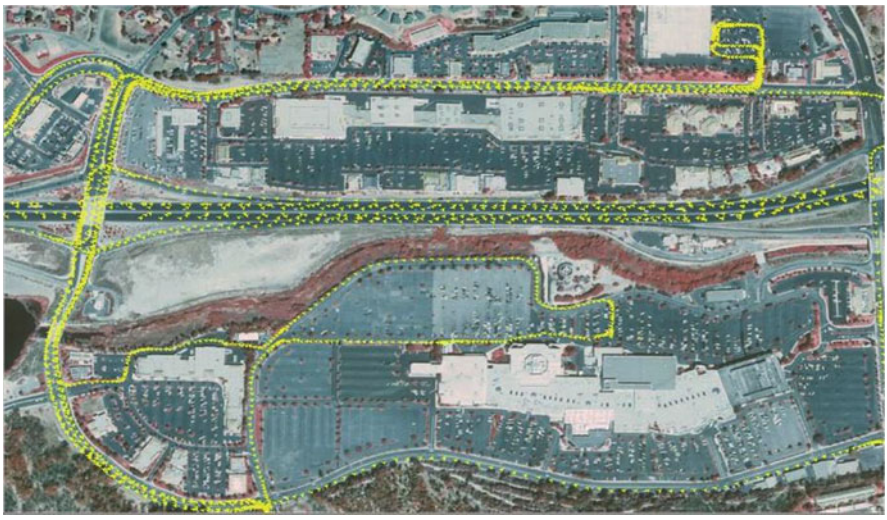


Fig. 2.6 GPS Location of Traffic Flows in San Diego, California
Source: California Department of Transportation (June 2002), 2000-2001 California Statewide Household Travel Survey. Sacramento, CA

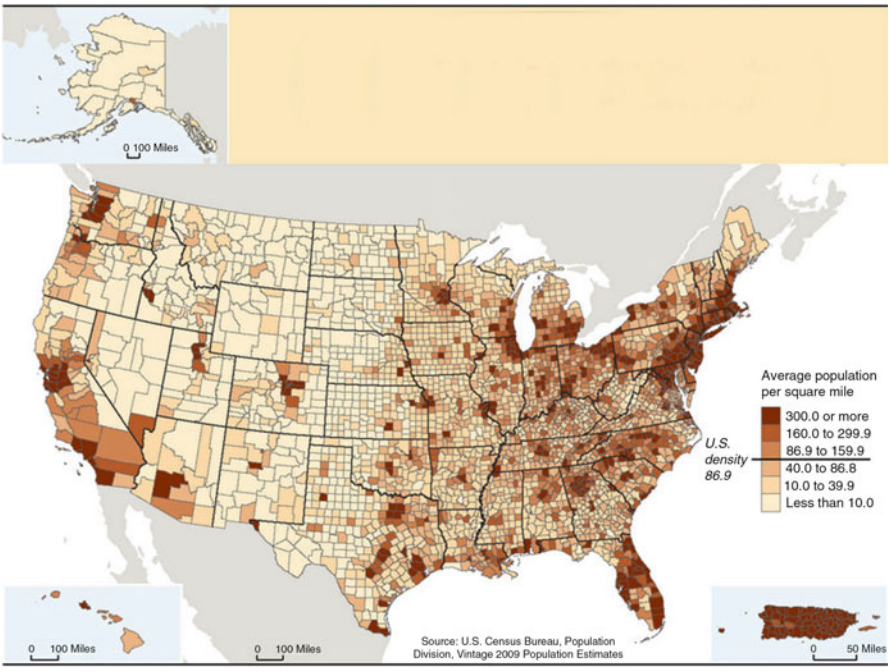


Fig. 2.7 Population Density for Counties and Puerto Rico Municipios: July 1, 2009
Source: US Census Bureau Population Division, Vintage 2009 Population Estimates. http://www.census.gov/popest/gallery/maps/County-Density-09.html

in Florida, in the metropolitan areas of the west coast, and around the great lakes. Low densities occur in many parts of the west and in the US breadbasket. The very low densities in Alaska are also evident. The overall density in the US is 86.9 persons per square mile and ranges from 0.039 in Yukon-Koyukuk Census Area to 71,505.7 in New York County.

Places with relatively low population densities may result from small population and/or large areas of undeveloped or uninhabitable land. More refined measures of population density would use the amount of settled area or settled area with residential land uses in the denominator. In 2008, San Diego County had a population density of 1.5 persons per acre based on the total land area of the County (SANDAG 2010b). When restricting the land area to that containing residential activity, the density increases to 9.4.

2.2.3 Center of Population

The mean point of the population distributed over an area is its center of population. It is the point where the area would balance with each individual having an equal weight and exerting influence on the central point proportionate to their distance from that point. The mean center of population is influenced by the distance of a person from it (Plane 2004). Population change farther away from the center of population will influence the mean point more so than population change near the center. The mean center of California's population in 1880 was in the San Francisco Bay area and has moved south to near Bakersfield by the year 2000 (NOAA 2004). So, population change in Riverside or San Bernardino counties would have more influence on the center of population than changes within 15 miles of Bakersfield. The mean center of the US population has been moving steadily westward since the first US census was taken in 1790 (see Figure 2.8). The center was near the top of Chesapeake Bay in 1790 and by 2010 it was near Plato Missouri. Since 1940, the center of population has moved south.

2.2.4 Spatial Distribution

Spatial distribution is the specific location or arrangement of events in space or time, or the arrangement of activities across the earth's surface. Several related terms have been used to characterize spatial distribution. Concentration is the degree to which population is focused or dispersed in geographic areas. Centralization is to form a center or the concentration of population in geographic space, while decentralization is the dispersal of population across geographic space. These concepts are

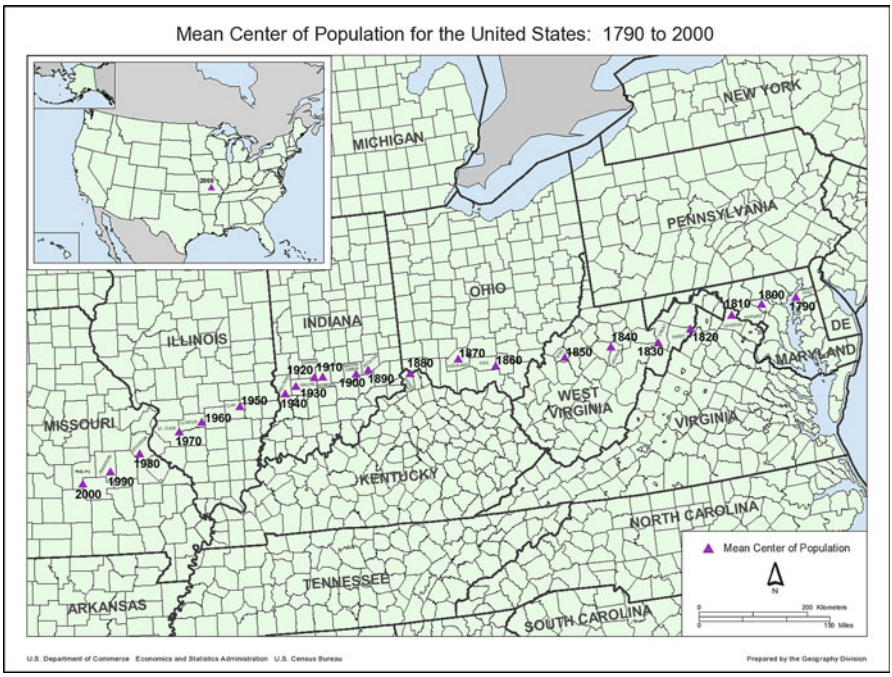


Fig. 2.8 Mean Center of the Population of the United States 1790 to 2010.
Source: US Department of Commerce, Economics and Statistics Administration, US Census Bureau. http://www.census.gov/geo/www/cenpop/meanctr.pdf

illustrated in Fig. 2.9. Housing units in San Diego County were concentrated along the South/Central coast and bay areas in 1940. The decentralization of development is clearly shown. By 1960, development had pushed east and pockets had developed along the north coast and north inland. These areas had further dispersed by 1980. The decentralization of housing continued and by 2000 development had spread beyond the urbanized area forming rural communities.

A cluster is a group of the same or similar items gathered closely together. A business cluster, for example, is a geographic cluster of interconnected businesses, suppliers, and associated activities in a particular field. Areal groupings of high income households, race, and age-restricted housing communities are examples of population clusters. Population clusters have been formally defined. For example, the Netherlands defines a distinct population cluster:

> as the population living in neighboring buildings that form a continuous built up area with a clearly recognizable street formation and certain land-use categories do not split up the population. Different population clusters of which the residential areas are separated no more than 200 meters of each other are considered to form a population cluster. An exception is made when residential areas are separated more than 200 meters by a canal or river, but are connected directly by a bridge or a tunnel. (Van Leeuwen 2007).

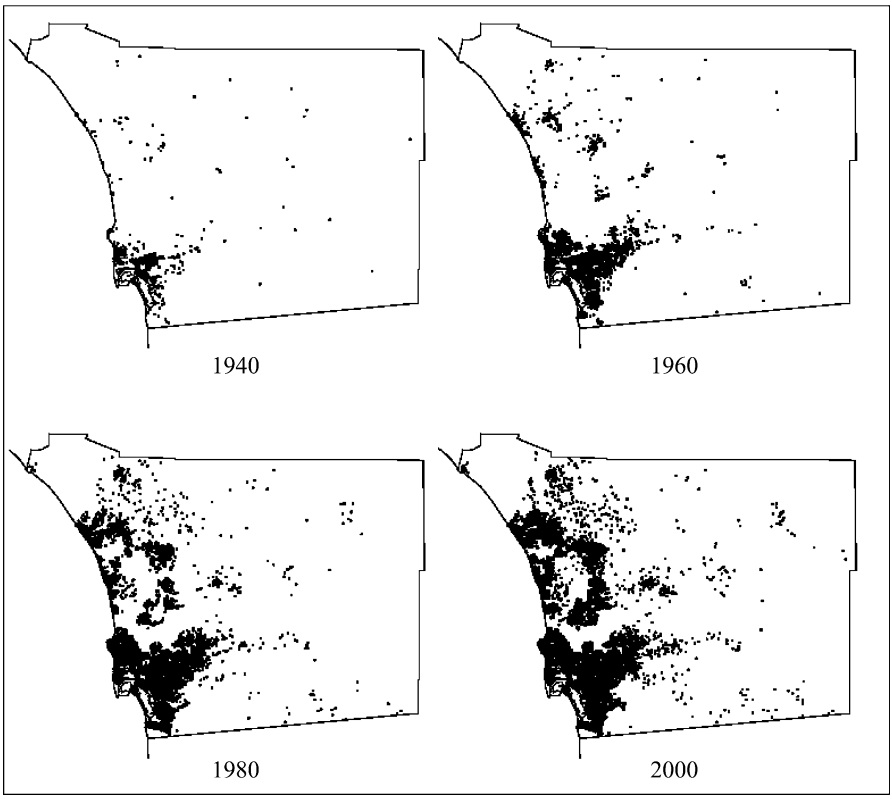


Fig. 2.9 Spatial Distribution of Housing Units, San Diego County, 1940-2000.
Source: Smith, Tayman, and Swanson (2001: 369)

2.2.5 Distance, Accessibility, and Spatial Interaction

A common and long-standing way to measure the dispersion of the population is the distance (Bachi 1985). Distance can be calculated between individual sites or locations, but it often measured from aggregate population grouped by geographic areas. In this case, the population is assumed to be concentrated at the geographic center of the area. Distance is often expressed in terms of miles or kilometers, but for some applications distance is represented by time or cost of travel between two areas (e.g., Putman 1983: 7).

For many applications such as site location or infrastructure demand it is useful to measure the accessibility of locations to a particular population distribution. Accessibility can be viewed as the ease of interaction between two or more locations. From a population perspective, accessibility is the proximity of a mass of persons to a particular location (Plane 2004). Accessibility over space is primarily influenced by the capacity of the transportation system relative to the travel demand. So in most cities, accessibility to the downtown core is much

quicker and easier during off-peak travel than during rush hour. The interstate highway system, for example, contributed significantly to the decentralization of activities across American because it increased the accessibility to formerly far flung areas.

Spatial interaction is the flow of products, people, services, or information among places, in response to localized supply and demand and is influenced by the accessibility between locations. It is a movement of people, freight, or information between an origin and a destination and a transport demand / supply relationship expressed over a geographical space (Rodriguez, Comtois and Slack 2009: Chapter 5.) Spatial interactions cover a wide variety of movements such as journey to work, migration, tourism, the usage of public facilities, the transmission of information or capital, the market areas of retailing activities, international trade, and freight distribution.

The basic assumption concerning many spatial interaction models is that flows are a function of the attributes of the locations of origin, the attributes of the locations of destination, and the accessibility between the concerned origins and the destinations. The gravity model is the most common formulation of the spatial interaction method (Lowry 1964, Putman 1994, Wegener 1994). It is named as such because it uses a similar formulation to Newton's formulation of gravity. Accordingly, the attraction between two objects is proportional to their mass and inversely proportional to their respective distance. A typical gravity model for population is based on the location of jobs, the ability of a location to accommodate additional growth, and the accessibility between these locations typically measured by travel time or cost.

2.3 Statistical

Statistical methods are widely used in population estimation. Regression techniques underlie ratio-correlation and other methods discussed in Chapters 6 and 8. They are also used for evaluating estimation model inputs deriving key parameters for these models, and evaluating the quality and validity of the resultant estimates (see Chapter 14). In this section, we discuss major concepts in descriptive and inferential statistics with specific attention given to regression modeling.

2.3.1 Descriptive Statistics

Descriptive statistics aim to summarize the main features of a distribution of data without employing a probabilistic formulation. Suppose you are analyzing estimation errors for cities within California and compute statistical measures that quantify the errors. These statistics describe the performance of the estimates for cities within California, but are not used to make generalizations about estimate errors in other states or other levels of geography. There are a variety of statistical measures

and graphical devices used to summarize and describe a distribution of data (e.g., Jaeger 1983; Langley 1970; Levin and Rubin 1998) and they can be generally grouped into three categories: central tendency (typicality), variability (dispersion), and distribution shape.

2.3.1.1 Central Tendency

Central tendency indicates the location or center of a distribution. The most common measures of central tendency are the mean, median, and mode (Swanson 2012). The average is represented by the arithmetic mean, which is the most widely used measure of central tendency. The mean is very familiar and has a number of desirable statistical properties, including that it uses all of the information in the distribution (Swanson, Tayman, and Barr 2000). This advantage is also a major drawback, as the mean can be influenced by extreme values. In this circumstance the mean may not represent the typical value.

The median, on the other hand, is the center point of the distribution and it not impacted by extreme values. The median is a resistant statistic and the mean is not, but the median ignores most of the information has other less desirable statistical properties than the mean. Other resistant measures of central tendency have been offered (e.g., M-estimators, the trimmed mean) to address the shortcomings of both the mean and median (Tayman and Swanson 1999).

The mode is the most frequently occurring observation. Unlike the mean and median, the may not be a unique mode for a distribution of data. If every observation is unique, there would be no mode or there could be more than one mode. Examples of data the often contain multiple modes are morning and evening commute-time travel volumes and household utility usage. The mode is, however, the only measure of central tendency applicable to nominal (unordered categories) data.

2.3.1.2 Variability

Although measures of central tendency provide useful information, they do not provide a complete picture of the data. Variability describes the deviation or spread of the observations from their center (Swanson 2012). Homogeneity and heterogeneity are also terms used to describe variability in data. For example, we may wish to compare estimate errors for counties in Florida during the decades 1990s and 2000s. The average error in both cases is 8.5%. With this information, you would note the average errors are identical and might conclude that there was no improvement in your methods. The next day your assistant tells you that for the 1990s the range of errors was 0.02% to 45%, and for the 2000s it was 5.6% to 20.2%. This new information shows that the methods had improved by reducing outlying errors and their variation.

The simplest measure of variability is the range, which is the difference between the highest and lowest values. It has the disadvantages of both the mean and median in that it ignores most of the information in the data set and is influenced by outliers. To combat the latter situation, measures like the interquartile range and quartile deviation have been developed (Blalock 1972: 79). The most widely used measures of variation are the variance and the standard deviation (square root of the variance). Both of these measures reflect the variability of the scores about the mean of the distribution. In the above example, the standard deviations for the population estimates of the 1990s and 2000s have values of 34.2 and 5.8.

Like the mean the variance and standard deviation are influenced by outliers, but they also present a dilemma in interpretation. What is a big one and what is a small one? The value of the variance is determined not only by the variability in the data, but also by the size of the mean. If you, for example, multiplied each observation in the distribution by 2 both the mean and standard deviation would double, but the variability of the new distribution would not change. What is needed is a measure that will provide an indication of the magnitude of the variation relative to magnitude of the mean. The coefficient of variance is such as measure of relative variation and is useful in comparing groups with respect to their homogeneity (e.g., Ikeda 2008).

2.3.1.3 Distribution Shape

We have discussed statistical descriptions under the general categories of measures of central tendency and variability. It is also useful to describe the general form or shape of the distribution. First, a distribution may be described by its number of relative maximums or modality. Strictly speaking, a distribution has only a single mode when an observation occurs most frequently. It is common to find a distribution described as bimodal or multimodal when there are two or more humps in the curve, even though there may be a single distinct mode.

Another distribution characteristic is its symmetry or conversely its skewness. A distribution is symmetrical if it can be divided into two mirror-image halves (see Figure 2.10). In a symmetrical distribution, the mean and median will be equal. If there is only one hump in the distribution the mode will be equal as well. A symmetrical distribution can also be multimodal in that case the mode would not equal the mean and median (Winkler and Hays 1975: 156). A non-symmetric or skewed distribution indicates the length of one of the tails of the distribution, relative to the center, is disproportionate to the other. A right or positively skewed distribution the bulk of the distribution falls into the lower values of the distribution with relatively few observations at the higher values. In a unimodel distribution, the mean exceeds the median, which exceeds the mode. Population estimate error distributions are often right-skewed (Tayman and Swanson 1999). On the other hand, in a left or negatively skewed distribution the long tail occurs among the lower values of the distribution. Here the mode exceeds the median, which exceeds the mean.

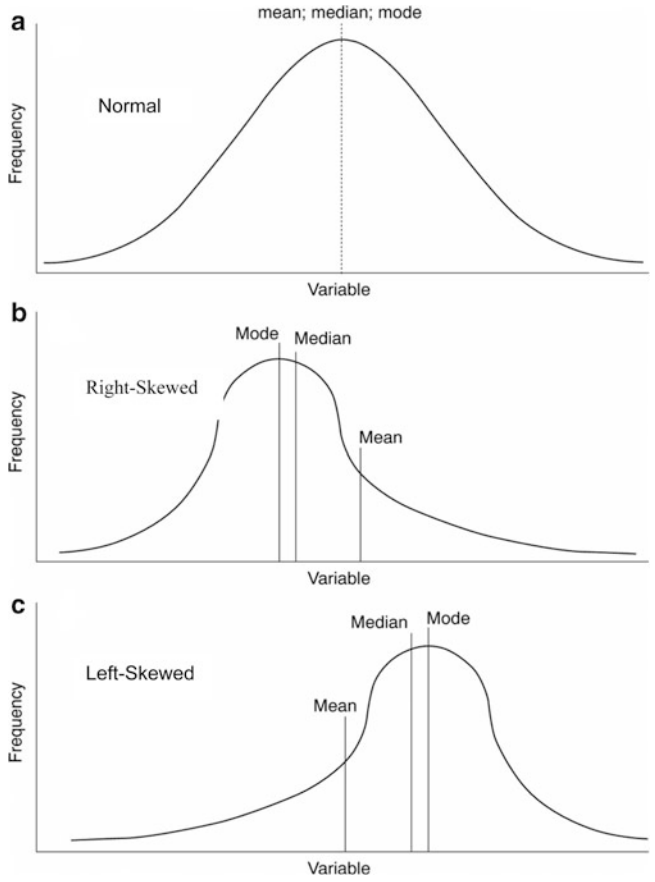


Fig. 2.10 Symmetrical, Right- and Left-Skewed Distributions
Source: http://www.google.com/images

Symmetrical curves are often associated with the bell-shaped normal distribution whose distribution shape is determined solely by its mean and standard deviation. But not all bell-shaped symmetrical distribution are normal (Blalock 1972: 98). Unimodal symmetrical curves may be more peaked (leptokurtic) or more flat (platykurtic) than the normal curve. Equations for these curves involve summarizing measures in addition to the mean and standard deviation.

2.3.2 Inferential Statistics

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. To legitimately use inferential statistics the immediate data must be either in fact or conceptually a random sample taken from the entire data set to which one wants to infer (Swanson 2012). Both of these perspectives

are employed, illustrated or otherwise discussed throughout this book (i.e., in Chapter 6, Chapter 7, Chapter 8, Chapter 9, Chapter 11, and Chapter 12).

As an example of the perspective where the immediate data are "in fact", a random sample, we can use inferential statistics to infer from a random sample of registered voters to the results of a forthcoming election. As another example where the immediate data are "in fact," a random sample, consider estimation errors for a sample of counties within the US and compute statistical measures that quantify the errors. Assuming a properly conducted sample, these statistics could be used to make generalizations about estimate errors for all counties in the US.

As an example of an immediate data set which "in concept" is a random sample, consider estimation errors for all 39 counties of the State of Washington as measured against the 2010 census. While in fact this is the entire population of interest if we are interested in estimation errors for the state of Washington's counties, the specific data in hand can be viewed as the probabilistic manifestation a process in which there is infinite number of outcomes. This perspective views our specific set of estimation errors as a random sample from the "super population" of infinite possible outcomes (Hartley and Sielken 1975; Sampath 2005).

Statistical inference is an important part of the toolkit used to develop population estimates. It is as a set of procedures designed to support generalizations by providing probabilistic evidence of their validity. However, one should use this set of procedures with an understanding of what it can and cannot do. This note of caution applies in particular to hypothesis testing, which we will shortly discuss. In using hypothesis tests, one needs to be cognizant of the important distinction between a substantive difference and a statistical difference (Swanson 2012). Keeping this distinction in mind will aid in avoiding the pitfalls associated with the practice of treating hypothesis testing as a ritual that in and of itself provides the answers to questions (Ziliak and McCloskey 2008).

2.3.2.1 Sampling Methods

Sampling is the process of selecting specific elements (the sample) from a population. A population is a complete group, whether people, houses, firms, electric light bulbs or geographic areas. Characteristics of the population are known as parameters, while characteristics of samples are known as statistics. Parameters are generally fixed and unknown. If they were known we would not need to sample. Statistics, on the other hand, vary from one sample to the next. The idea behind sampling is to use the sample elements to develop statistics that are used to estimate and make inferences about population parameters.

There are two general sample selection methods probability or random and non-probability or non-random (Warwick and Lininger 1975: 72). Probability sampling is where elements are chosen by chance procedures with known probabilities of selection. Simple random sampling (SRS) is basic selection process, where each element has the same probability of being selected (Kish 1965: 21). Modifications

to SRS include the use of stratification (selection from subpopulations), clustering (selection from groups of elements), and systematic selection (interval selection from lists). In non-probability samples elements are not selected by chance procedures or with known probabilities of selection. The most common types of non-probability samples are haphazard collection, judgment sampling, quota sampling, expert sampling, and purposeful sampling. Inferences are frequently made from non-probability samples, but they depend heavily on broad assumptions about the distribution of the survey variables in the population (Kish 1965: 19). On the other hand, inferences based on probability sampling can be made entirely from statistical methods, without assumptions about the population distribution.

The sampling distribution underlies the process of statistical inference. A sampling distribution is a probability distribution of a given statistic based on a random sample of size n. It may be considered as the distribution of a statistic for all possible samples from the same population of a given size. Imagine a taking a random sample of 50 households in Pacific Beach, a community of San Diego, and getting information of the number of persons permanently living in each house. The average of these values would be a statistic known as persons per household (PPH). If you took another sample of 50, the PPH value would likely be different from the first sample. Imagine you repeated this over and over again and infinite number of times. This distribution of these sample PPH values would be known as the sampling distribution of the sample mean. By the same token, there is a sampling distribution for any sample characteristic (e.g., mode, median, regression coefficient). A sampling distribution is never obtained by empirical means, as typically only a single sample of size n is selected. However, the sampling distribution provides a way to make inferences about population parameters on the basis of random samples in terms of the probability that a sample's statistic will arise from chance from a certain population (Winkler and Hays 1975: 305).

Under a random and unbiased sample, the mean of the sampling distribution is equal to the population parameter being estimated. The variability in a sampling distribution is known as sampling error and is a function of the variance of the variable in the population and the size of the sample. In general, increases in sample size will decrease the sampling error. As the sample size gets larger, the sampling distribution will cluster closer and closer to its mean or the population parameter. To clarify, we cannot be certain what the outcome of a single sample will be, but as sample size increases the probability increases that our single sample statistic will be closer to the parameter being estimated.

Table 2.4 illustrates these ideas based on a population of 100 households. Our aim is to estimate the PPH, which is 2.12 for the entire population. We simulated four sampling distributions with sample sizes of 2, 5, 10, and 20 by taking 153 random samples of each size and calculating the sample PHH for each sample. The means for each sampling distribution range from 2.06 to 2.18 close to the population PPH. They are not exact because our simulation has a relatively few number of random samples. For sample sizes of 2, there is considerable variability in the sampling distribution, which decreases as the sample size increases. For sample sizes of 20, all but one sample mean falls between two adjacent groups (1.1 to 3.0).

Table 2.4 Simulated Sampling Distributions

PPH	n = 2	n = 5	n = 10	n = 20
0	4	0	0	0
0.1 to 1.0	41	20	5	1
1.1 to 2.0	34	67	61	70
2.1 to 3.0	47	42	70	82
3.1 to 4.0	18	21	17	0
4.1 to 5.0	6	3	0	0
5.1+	3	0	0	0
Sampling Distribution Mean	2.15	2.06	2.18	2.08
Population PPH	2.12			

Source: Generated by the authors.

To calculate probabilities using a sampling distribution, we need to know its distribution shape. If the population is normally distributed, the sampling distribution will also be normally distributed, regardless of the sample size (Blalock 1972: 178). Otherwise, the Central Limit theorem states that as the sample size becomes large the sampling distribution approaches normality, regardless of the shape of the population distribution (Gnedenko 1967: 302-310). What is the appropriate number of observations to use the central limit approximation? Some believe that at least 30 observations are needed (Blalock 1972: 185), while others argue that as few as 10 can be considered normal for practical purposes (Winkler and Hays 1975: 316). Common practice is to use the Student's t distribution when the sample size is below 30.

2.3.2.2 Confidence Intervals

There are basically two kinds of estimates for population parameters: point estimation and interval estimation (Blalock 1972: 201). Point estimates provide the best single value to estimate a population parameter. According to the 2010 American Community Survey, the median household income was $59,923 in San Diego County. To ascertain the accuracy of this estimate, we would like to predict that the parameter is somewhere within a given interval on either side of the point estimate. In other words, we would like to develop a confidence interval such as we are 90% confident that the median household income in San Diego County lies between $58,848 and $60,998.[2] The 90% figure is known as the confidence level. Other common confidence levels are 95% and 99%, but there is nothing sacred about these levels.

The interval range (difference between the upper and lower limits) is a function of the sampling error and the confidence level (Hahn and Meeker 1991: 54). It does not reflect any other uncertainties in the estimation process such as non-response survey bias, questionnaire wording, or other data collection issues (Swanson 2012). The sampling error, measured by the standard error, takes into account the sample standard deviation and sample size. The standard error is an elegant measure in that

it takes into account all of the uncertainty associated with statistical inference and simultaneously links an empirical set of data to the theory that allows one to do statistical inference (Swanson 2012).

The confidence level is determined by the area under the sampling distribution curve corresponding to the level. In general, a larger sample, smaller sample standard deviation, and lower level of confidence will result in narrower confidence intervals. Methods have been developed for putting confidence intervals around population estimates (e.g., Espenshade and Tayman 1982; Kintner and Swanson 1993; Swanson 2008: 165-189; Swanson 1989).

2.3.2.3 Hypothesis Testing

In estimation, the question being answered is, "What is the value of the population parameter?" In hypothesis testing it is, "is it reasonable to believe that the value of the population parameter is a certain value?" That is, hypothesis testing is used to make comparisons (Swanson 2012). In regression analysis, for example, a common hypothesis test is whether the population slope is different from zero (Swanson 2004, 2012). Hypothesis testing begins with an assumption, called the null hypothesis, made about a population parameter. Sample data are used to determine the difference between the hypothesized value and sample statistic. Smaller differences increase the likelihood that the hypothesized value is correct. Larger differences decrease the likelihood. Because we are dealing with a sample, we cannot make a decision about the hypothesized parameter by simply examining the difference between it and the sample statistic. A formal statistical test of a hypothesis provides an objective framework for making this decision.

The outcome of a hypothesis test is a decision. In the case of the regression coefficient, a decision to continue to believe that its value is zero or the null hypothesis, or a decision to discard that belief in favor of an alternate hypothesis that the coefficient is not zero. This decision cannot be made with absolute certainly if for no other reason than we have sampling error. There are four possible outcomes in a hypothesis test:

1. Accept the hypothesis when it is true (correct decision);
2. Reject the hypothesis when it is true (incorrect decision, Type 1 error (α));
3. Reject the hypotheses when it is false (correct decision); and
4. Accept the hypothesis when it is false (incorrect decision, Type 2 error (β)).

A type one error (α), also known as the significance level, represents the risk or probability of making a Type I error. Similarly (β) represents the probability of making a Type II error. The significance level is typically set at values of 0.10, 0.05, and 0.01 (Lehmann 1986: 69) and is most often considered to the exclusion of β. Once the sample size is fixed, α and β vary inversely, so using small values for α could lead to high probabilities of making a Type II error. The compliment of a

Type II error is known as power or ability to reject a false hypothesis (Kraemer and Thiemann 1987). Ideally, the sample size should be set to yield adequate values for the significance level and power.

Using a test statistic and the appropriate sampling distribution (e.g., normal, t, F or Chi-square), one computes the likelihood or probability of getting this result assuming the null hypothesis is true, or the p-value. One accepts the null hypothesis if the p value is $\geq \alpha$ and rejects it if the p value is $< \alpha$. Rejection of the null hypotheses is known as a statistically significant result. It means that it not likely that the sample statistic would come from a population with the hypothesized parameter. Significance tests are sensitive to sample size. In vary large samples even substantively trivial differences will be statistically significant and in small samples potentially substantive differences will turn out to be not statistically significant (Henkel 1976).[3]

2.4 Regression

Regression analysis includes techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent and one or more independent variables (Swanson 2012). Regression with one independent variable is simple regression, and with more than one independent variable is multiple regression. Regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables changes, while the other independent variables are held fixed. Regression analysis is widely used for estimation, prediction, and forecasting. It also helps us understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. Regression is very flexible. It can handle variables with different measurement scales and incorporate variable transformations to study non-linear relationships (Draper and Smith 1981: Chapter 5; Hosmer and Lemeshow 1989).

Regression analysis begins with an equation representing straight line (Draper and Smith 1981: 9):

$$\begin{aligned} &\text{Population} \quad Y = \alpha + \beta X + \varepsilon; \text{ and} \\ &\text{Sample} \quad\quad\; y = a + bx + e, \end{aligned}$$

where a and b are constants, y is the dependent variable, x is the independent variable, and e is the error term. Constants a and b and the error term e are sample estimates of the corresponding population parameters α, β, and ε. The intercept a represents the point where the line crosses the y-axis at $x = 0$ and b is the slope that indicates the magnitude of change in Y for a one unit change in X. For example, an equation predicting the number of people based on changes in employment has a slope of 1.25. That means an increase in one job results in an increase of

1.25 people or a decrease in one job results in the loss of 1.25 people. A positive slope indicates a direct relationship between the dependent and independent variables (i.e., they change in the same direction). A negative slope indicates an inverse relationship in which the variables change in different directions. The error term (e) accounts for the fact that the independent variable will not perfectly predict the dependent variable; that is, there will be scatter about the regression line. The error term contains both measurement error in Y and the effects of other influences of Y not brought into the equation (Blalock 1972 367).

Several crucial assumptions underlie the regression model and there are procedures for testing and verifying their validity (e.g., Chatterjee and Hadi 2006; Draper and Smith 1981: Chapter 3; Stock and Watson 2003: 103-107). Some of the key assumptions are: 1) the independent variable is measured without error and is not related to the error term; 2) the errors are not correlated; and 3) the variance of the error is constant for each observation. The slope and intercept of the regression line are estimated using ordinary least squares (OLS), which minimizes the squared errors between the predicted and observed values. Minimizing the squared errors does not necessary mean these errors are small. Moreover, it is important to keep in mind that this minimization is relative to using the mean as an estimator, which in some cases, may not be an optimal estimator in the absence of the additional information associated with a regression model (Swanson 2004).

There are several ways to evaluate the fit of the regression line. One is the standard error of the estimate, which measures the scatter of the observed values around the regression line. The standard error gives a first handle on how well the fitted equation fits the sample data. But what is a 'big' and what is a 'small' standard error depends on the context, and it is sensitive to the units of measurement of the dependent variable. A more standardized statistic is the r-squared. R-squared ranges from zero to 1 and shows the proportion of the variance in Y accounted for by X. There is also a standard error associated with the slope, which can be used to make inferences about the population slope in the form of confidence intervals or hypothesis tests.

The extension to regression with more than one independent variable is relatively straightforward. OLS is used to estimate a slope and standard error for each independent variable, but now the slopes are known as partial regression coefficients (Blalock 1972: 431). They represent the slope that would be obtained by controlling or taking into account the remaining independent variables in the regression equation. Multiple regression does introduce another complexity; the relationship between the independent variables. Multicollinearity is a statistical phenomenon in which two or more independent variables in a model are highly correlated. In this situation the coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual independent variables. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the dependent variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others. Belsley, Kuh, and Welsch (1980) discuss ways to identify multicollinearity and potential remedies.

Endnotes

1. The IM and OM terms include both domestic and foreign migrants. If information is only available on net migration the IM and OM terms would be replaced by ± NM.
2. Strictly speaking this interval either does or does not contain the median household income for all households in San Diego County because the parameter is a fixed value. In the long run, we know that 90% of the infinite number of intervals that could be computed would contain the population parameter, which is the basis for our inference.
3. Significance testing is the cornerstone of research and the social sciences, but it is not without critics. Ziliak and McCloskey (2008) point out that "insignificance" does not mean unimportant, and propose that the scientific community should abandon usage of the test altogether, as it can cause false hypotheses to be accepted and true hypotheses to be rejected.

References

Aronoff, S. (1989). *Geographic information systems: A management perspective*. Ottawa: WDL Publications.

Bachi, R. (1958). "Statistical analysis of geographic series." *Bulletin of the International Statistical Institute* 32(2), 229–240.

Belsley, D., Kuh, A. E. & Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Blalock, H. M. (1972). *Social statistics*. New York: McGraw Hill.

Bryan, K. N. and R. George (2004). Geographic Information Systems. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 733–750). New York: Elsevier Academic Press.

Chatterjee, S. and Hadi, A. S. (2006). *Regression analysis by example*. New York: John Wiley & Sons.

Draper, N. and Smith, H. (1981). *Applied regression analysis, second edition*. New York: John Wiley & Sons.

Espenshade, T. J. and Tayman, J. (1982). "Confidence intervals for post-censal population estimates." *Demography* 19(2), 191–210.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative geography: Perspectives on spatial data analysis*. London: Sage Publications.

Gnedenko, B. V. (1989). *The theory of probability, Fourth Edition*. New York: Chelsea Publishing Company.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners*. New York: John Wiley & Sons.

Hartley, H. and R. Sielken, Jr. 1975. A "Super-Population Viewpoint" for Finite Population Sampling. *Biometrics* 31 (2): 411–422

Henkel, R. E. (1976). *Tests of significance*. Beverly Hills: Sage Publications.

Hobbs, F. B. (2004). Age and sex composition. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 125–173). New York: Elsevier Academic Press.

Hodges, K. (1995). An evaluation of geometric data retrieval methods. Paper presented at the annual meeting of the Population Association of America. San Francisco, CA.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression*. New York: John Wiley & Sons.

Ikeda, M. (2008). Developing guidelines based on CVs for when one-year estimates can be used instead of three-year estimates in the American Community Survey (ACS) for areas with populations of 65,000 or more. Washington, D.C: US Census Bureau, Statistical Research Division.

Jaeger, R. M. (1983). *Statistics: A spectator sport*. Beverly Hills: Sage Publications.
Kintner, H. J. and D. Swanson, D. A. (1993). "Measurement errors in census counts and estimates of inter-censal net migration." *Journal of Social and Economic Measurement* **19** (2), 97–120.
Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.
Kraemer, H. C. and Thiemann, S. (1987). *Statistical power analysis in research*. Beverly Hills: Sage Publications.
Kreitz, M. S., Doherty, T., & Rindsfuser, G. (2002). Collection of spatial behavior data and their use in activity scheduling models. Paper presented at the annual Meeting of the Transportation Research Board. Washington, DC
Langley, R. (1970). *Practical statistic: Simply explained*. New York: Dover Publications.
Lehmann, E. L. (1986). *Testing statistical hypothesis*. New York: John Wily & Sons.
Levin, R. S. and Rubin, D. S. (1998). *Statistics for management, Seventh Edition*. Englewood Cliffs: Prentice-Hall.
Los Angeles County Department of Public Health (2010). Life expectancy in Los Angeles County: How long do we live and why? A cities and communities report. Los Angeles, CA.
Lowry, I. S. (1964). *A model of metropolis*. Santa Monica: The Rand Corporation.
Martin, J. A., Hamilton, B. E., Sutton, P. D., et al. (2010). Births: Final data for 2007. National Vital Statistics Report, Vol. 58, No. 24. Hyattsville: National Center for Health Statistics.
McKibben, J. N. (2004). Racial and ethnic composition. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 175–189). New York: Elsevier Academic Press.
National Oceanic and Atmospheric Association, US Census Bureau, & American Congress on Surveying and Mapping. (2004). Center of population project. Washington, DC (http://www.ngs.noaa.gov/INFO/COP/).
O'Hare, W. P., Pollard, K. M. & Ritualo, A. (2004). Educational and economic characteristics. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp.211–215). New York: Elsevier Academic Press.
Perz, S. G. (2004). Population Change. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 253–263). New York: Elsevier Academic Press.
Plane, D. A. (2004). Population distribution-Geographic areas. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 81–104). New York: Elsevier Academic Press.
Putman, S. H. (1983). *Integrated urban models*. London: Pion Limited.
Putman, S. H. (1994). Integrated transportation and land use models: An overview of progress with DRAM and EMPAL, with suggestions for further research. Paper presented at the annual meeting of the Transportation Research Board. Washington, DC.
Robinson, J. G., B. Ahmed, P. Das Gupta, and K. Woodrow. 1993. Estimates of Population Coverage in the 1990 United States Census Based on Demographic Analysis. *Journal of the American Statistical System* 88: 1061–1071.
Rodrigue, J. P., Comtois, C. & Slack, B. (2009). *The geography of transport systems*. New York: Routledge.
Sampath, S. 2005. *Sampling Theory and Methods*. Harrow, England: Alpha Science International Ltd.
SANDAG (2010a). "Subregional and Major Statistical Areas in San Diego County." San Diego, CA: San Diego Association of Governments (http://gis.sandag.org/boundary/viewer.htm).
SANDAG (2010b). "2050 Regional Growth Forecast." San Diego, CA: San Diego Association of Governments (http://profilewarehouse.sandag.org/profiles/fcst/reg999fcst.pdf)
Sharkova, I. V. (2000). With or without GIS? Evaluating accuracy, timeliness, and costs of population estimates for user-defined areas. Paper presented at the annual meeting of the Population Association of America. Los Angeles, CA.
Smith, S. K., J. Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

State of California. (2011). Population estimates and components of change by county, July 1, 1999–2010 with 2010 Census benchmark. Sacramento, CA.

Stock, J. H. and Watson, M. H. (2003). *Introduction to econometrics*. Boston: Addison Wesley.

Swanson, D. A. 2004. "Advancing Methodological Knowledge within State and Local Demography: A Case Study." *Population Research and Policy Review* 23: 379–398.

Swanson, D. A. 2012. *Learning Statistics: A Manual for Sociology Students*. San Diego, CA: Cognella Academic Publishing.

Swanson, D. A. and Siegel, J. S. (2004). Introduction. in J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 1–8). New York: Elsevier Academic Press.

Swanson, D. A. and Stephen, G. E. (2004). Glossary. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 571–778). New York: Elsevier Academic Press.

Swanson, D. A., Tayman, J., & Barr, C. F. (2000). "A note on the measurement of accuracy for subnational demographic estimates." *Demography* 37(2), 193–201.

Swanson, D. A. (1989). "Confidence intervals for post-censal population estimates: A case study for local areas." *Survey Methodology* 15, 271–280.

Swanson, D. A. (2008). Measuring uncertainty in population data generated by the cohort-component method: A report on research in progress In S. H. Murdock & D. A. Swanson (Eds.) *Applied Demography in the 21st Century* (pp. 165–184). Dordrecht, Heidelberg, London, and New York: Springer.

Tayman, J. (1999). Post-censal population estimates for census blocks: The San Diego experience. Paper presented at the US Bureau of the Census Population Estimates Conference. Washington, DC.

Tayman, J. and Swanson. D. A. (1999). "On the validity of the MAPE as a measure of forecast accuracy." *Population Research and Policy Review* 18(4), 299–322.

Tordella, S. J. (1987). Geometric data retrieval methods in commercial sector demography. In D. A. Swanson and J. W. Wicks (Eds.), Is*sues in Applied Demography: Proceedings of the 1986 National Conference* (pp. 51–55). Bowling Green, OH.

Van Leeuwen, N. (2007). Delineating population clusters by polygons and research of a grid approach. Nordic Forum for Geo-Statistics Seminar 2007. Helsinki, Finland.

Warwick, D. P. and Lininger, C. A. (1975). *The sample survey: Theory and practice*. New York: McGraw Hill.

Wegener, M. (1994). "Operation urban models: State of the art." *Journal of the American Planning Association* 60, 17–30.

Wilmoth, J. (2004). Population size. In J. S. Siegel, & D. A. Swanson (Eds.) *Methods and Materials of Demography, Second Edition* (pp. 65–80). New York: Elsevier Academic Press.

Winkler, R. L. and Hays, W. L. (1975). *Statistics: Probability, inference and decision, Second Edition*. New York: Holt, Rinehart, and Winston.

Ziliak, S. T. and McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: The University of Michigan Press.

Chapter 3
Data Sources

Demographic data are collected, produced, and distributed by a variety of federal, state, and local government agencies and private companies (e.g. Bryan 2004a; Murdock and Ellis 1991: Chapter 3). These data are available as printed publications, unpublished reports, and electronic data files, with greater access and availability of data on the Internet. Table 3.1 shows the Web site addresses for commonly used public data sites. Data are often replicated in secondary sources such as professional journals, textbooks, and statistical abstracts. We will discuss the most important sources of demographic data used for population estimation in the United States.

3.1 Choice of Data

An important factor determining the choice of the method for preparing a population estimate is the type and quality of data available for this purpose (Bryan 2004b: 526). There two general categories of data used in population estimates: (1) "direct" data and (2) "indirect" or symptomatic data. The classification depends on the specific kind of data and their use in a given method. Direct data are obtained from censuses, surveys (e.g., ACS), from vital registration systems and reflect an unequivocal connection to population and population change. Indirect data, on the other hand, produce estimates based on information indirectly related to, or symptomatic of, the population characteristic being estimated. Examples of indirect data are school enrollment, income tax returns, voter registration, employment, electrical hook-ups, and housing. Data of a given type may be direct for one kind of estimate and indirect for another. For example, data on vital events are direct when used to estimate natural change in a population. They represent indirect data when used in the censal ratio method to estimate total population (see Chapter 9). Both direct and indirect data are often used in combination when estimating population. If data is lacking or there is insufficient time or resources to collect it, extrapolation methods can be used to estimate population (See Chapter 6).

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_3,

Table 3.1 Web Site Addresses of Major US Public Sector Data Providers

Bureau of Economic Analysis	http://www.bea.gov
Bureau of Labor Statistics	http://www.bls.gov
Bureau of Transportation Statistics	http://www.bts.gov
Census Bureau	http://www.census.gov
Department of Homeland Security	http://www.dhs.gov
Federal-State Cooperative Program for Population Estimates	http://www.census.gov/population/www/coop/fscpe.html
Geologic Survey	http://www.usgs.gov
National Association of Regional Councils	http://www.narc.org
National Center for Education Statistics	http://nces.ed.gov/
National Center of Health Statistics	http://www.cdc.gov/nchs
State Census Data Centers	http://www.census.gov/sdc

The usefulness of indirect data for population estimation depends on the extent to which factors other than population influence them. Changes in school attendance may result from changes in the laws relating to attendance and the availability of school facilities, as well as from changes in the number of children of school age. Additionally, the prevalence of private schools and home schooling can compromise the efficacy of enrollment data. Employment, housing construction, and public utility customers change with economic conditions (e.g., unemployment rate) as well as with population. The usefulness of indirect data as symptomatic indicators of population change will vary with the particular situation and these data should be carefully evaluated and understood prior to their use in population estimation. Important criteria to evaluate are relevance to population changes, completeness, bias, internal consistency, and temporal closeness to the estimation time point.

3.2 Decennial Census

The decennial census is by far the most important and comprehensive source of demographic data in the United States. Every 10 years, the federal government attempts to count the entire population of the country. The results determine each state's representation in Congress and are used by state legislatures and local governments to redraw electoral boundaries. Census data are basis for the distribution of billions of dollars in federal and state funds each year through a variety of revenue-sharing and grant-in-aid programs. Businesses and government agencies use the census for planning, budgeting, marketing, and policy-making purposes. Scholars and the media use them to analyze social, economic, and political issues, and they are the basis for population estimates. Excellent discussions of the decennial census and related issues can be found in Alonso and Starr (1987); Anderson (1988); Anderson and Fienberg (1999); Edmonston and Schultze (1995); and Swanson and Walashek (2011).

Table 3.2 Types of Data Collected in the 2000 and 2010 Censuses of Population and Housing

All households (2000 and 2010 data):

Population: Name, relationship to householder, sex, age, date of birth, race, and Hispanic origin, additional people on census day not included (2010), person live or stay somewhere else (2010)

Housing: Number of people in household, telephone number, tenure (ownership status), vacancy status.

Sample households (2000 long-form data):

Population: Same as short form, plus marital status, school enrollment, educational attainment, ethnic origin (ancestry), language spoken at home, place of birth, citizenship status, year of entry into the United States, place of residence five years ago, disability status, living with grandchildren, military service, employment status, employment history, place of work, transportation to work, occupation, industry, and income.

Housing: Same as short form, plus type of housing unit, year built, length of residence in current unit, number of rooms, number of bedrooms, plumbing facilities, kitchen facilities, telephone in unit, type of heating fuel, number of motor vehicles, size of lot, presence of home business, annual costs of utilities, monthly rent or mortgage payment, second mortgage, real estate taxes, property insurance, and value of property.

Other Characteristics: School enrollment, educational attainment, ancestry/ethnic origin, state or country of birth, citizenship and year of entry, language spoken at home, ability to speak English, residence 5 years ago, veteran status/period served, disability, grandparents as caregivers, children ever born, current employment status, hours worked per week, place of employment, travel time to work, means of travel to work, persons in car pool, industry/employer type, occupation/class of worker, self-employment, weeks worked last year, total income by source

Source: Adapted from Smith, Tayman, and Swanson (2001: 37)

The United States conducted its first census in 1790 and since then a census has been conducted every 10 years without interruption. The first census compiled a list of household heads and counted people in five demographic categories. More and more questions were added over the following decades, covering social, economic, and housing characteristics as well as demographic characteristics (Bryan 2004a:16-18. The practice of collecting a limited amount of data from all households (so-called *short-form data*) and a larger amount from a sample of households (*long-form data*) was begun in 1940 and continued up to the year 2000. In 2000, about five of six households received the short form of the census and one of six received the long form. The long form asks the same questions as the short form, plus a number of others. The 2010 census only included a 10 question short form sent to every household. The sample long-form data was replaced by the American Community Survey discussed later in this chapter. Table 3.2 describes the types of data collected on the short forms of the 2000 and 2010 censuses and the long form of the 2000 census.

The long-form provides the most commonly used and most comprehensive in terms of demographic and geographic detail data on gross migration. Since 1940 the census has included a question about previous residence five years ago asked since 1940; in 1950, the question asked about previous residence one-year prior. Migration data is reported down to the city and census tract level, but only for in-migrants. The Census Bureau has tabulated in- and out-migrants by age, sex, and race for states and counties. There are several problems with the migration data

collected in the decennial census (Smith et al. 2001: 113-114). First, they do not pick up multiple moves over the five year period and consequently, understate the full extent of mobility during that time (DaVanzo and Morrison 1981; Long and Boertlein 1990). Second, the long-form sample size can create data reliability problems for small places, especially when the migration is broken into demographic subgroups. Third, emigration to foreign countries is not covered. Finally, the migration data are only available every ten years, and is typically released three to five years after the data is collected.

The decennial census is based on self-enumeration. Households are mailed census forms in late March and are asked to fill them out and return them by mail; in some rural areas the forms are delivered by a census enumerator. The Census Bureau follows a number of procedures designed to maximize response rates and to collect information from non-responding households. In spite of these procedures, the data collected are incomplete and sometimes incorrect. Post-enumeration surveys and demographic analyses are used to measure the extent and nature of census errors and to develop estimates of the net undercount (or, in some instances, the net overcount) (US Census Bureau 2011).

The Census Bureau tabulates aggregate census results for a variety of geographic areas discussed in Chapter 2. Not all types of data are tabulated for all levels of geography, however. Short-form data is tabulated for each census block in the United States. Long-form data are tabulated only down to the block group level. Public Use Microdata Sample (PUMS) files are compiled for areas with 100,000 or more residents, providing individual records (stripped of identifying information) for use in more detailed analyses.

Government officials and other interested parties have been concerned about the accuracy of census results since the first census (Choldin 1994). Census errors are caused by missed households, refusal to respond, recording errors, sampling errors, geographic assignment errors, duplication errors, coding and data-processing errors, and the incorrect imputation of missing data. Although these errors can cause census counts to be either too high or too low for any given geographic area or population subgroup, in most instances they lead to net undercounts of the "true" population. Nationally, the net census undercount was estimated as 5.4% in 1940, 4.1% in 1950, 3.1% in 1960, 2.7% in 1970, 1.2% in 1980, and 1.8% in 1990; and 1.2% in 2000 (Farley 2008).

The net undercount differs by geographic area and population subgroups. For example, the net undercount is much greater for blacks, Hispanics, and American Indians than for non-Hispanic whites (Anderson and Fienberg 2000). Because of the increase in the net undercount between 1980 and 1990 and the large differences found among population subgroups, many concerns about the accuracy of the decennial census were voiced after the 1990 census. The Census Bureau responded by developing plans to use statistical sampling to account for non-responses and to adjust for the net differential undercount, but those plans encountered strong political opposition in Congress. A Supreme Court decision in 1999 prohibited the use of adjustments based on sampling for the reapportionment of Congress after the 2000 Census, but left unresolved several broader issues related to the use of statistical adjustments.[1]

3.3 Vital Events

Data on events such as births, deaths, marriages, and divorces are called vital statistics. The US vital registration system is somewhat unusual in that states collect vital events certificates and are paid to transmit them to the federal government. As early as 1639, the Massachusetts Bay Colony began reporting births, deaths, and marriages as part of its administrative/legal system and by 1933 all states had vital registration systems that met the federal standards (adoption of standard certificates and at least a 90% registration rate); the Alaskan territory was admitted in 1950 and the territory of Hawaii for deaths in 1917 and births in 1929 (Bryan 2004a: 26). The federal government sets standards for the collection and reporting of the data, compiles summaries from data collected by each state, and publishes a variety of reports based on these data. The quality of vital statistics data is generally very good in the United States and other developed countries.

In 1960, the National Office of Vital Statistics was reorganized and became part of the National Center for Health Statistics (NCHS), which today is a branch of the Centers for Disease Control (CDC). Annual and monthly reports on births, deaths, marriages, and divorces are available from the NCHS, along with annual life-tables for the nation.[2] State life tables are typically constructed every 10-year to use the census to calculate mortality rates; they are prepared by NCHS and by vital statistics agencies in many states.[3] It should be noted that some of the concepts and definitions used by the NCHS do not precisely match those used by the Census Bureau and adjustment might be needed when combining Census and NCHS data (Hahn et al. 1992; Miniño et al. 2010; Sink 1997).

Data from the NCHS are available only at the national and state levels; vital statistics data for local areas must be obtained elsewhere. Most states tabulate data at the county (or county-equivalent) level, but few go beyond that to develop regular data series for subcounty areas (Bogue 1998). Although individual records generally contain the information needed to allocate them to different types of subcounty areas (e.g., cities, census tracts), actually doing so requires a substantial effort. In addition, there are often errors in geocoding birth and death records at the subcounty level (Flotow and Burson 1996). Analysts needing vital statistics data for subcounty areas may have to develop those data themselves.

3.4 Surveys

The decennial census and vital statistics reports are valuable sources of demographic data. However, the census is conducted only once every 10 years, and vital statistics data cover only a small portion of the variables of interest to demographers. Sample surveys can be used to collect data on a variety of topics at various times between censuses. The Census Bureau conducts a variety of large on-going surveys, some of which are discussed below.

3.4.1 Current Population Survey

One of the most important sample surveys in the United States is the Current Population Survey (CPS), a monthly survey of about 50,000 households conducted by the Census Bureau for the US Bureau of Labor Statistics. Started in the early 1940s, this survey originally focused on the collection of labor force and unemployment data and is the source for the monthly updates of employed residents and the unemployment rate. It has since been expanded to include a variety of topics including occupation, industry, education, income, veteran status, marital status, living arrangements, fertility, and migration, as well as demographic data on age, sex, race, and ethnicity. Data from the CPS are currently tabulated at the national, regional, and state levels and for large metropolitan areas.

3.4.2 American Housing Survey

The American Housing Survey (AHS) is conducted for the Department of Housing and Urban Development (HUD). It collects data on the nation's housing, including apartments, single-family homes, mobile homes, vacant housing units, household characteristics, income, housing, and neighborhood quality, housing costs, equipment and fuels, size of housing unit, and recent movers. National data are collected in odd numbered years, and data for each of 47 selected Metropolitan Areas are collected currently about every six years. The national sample covers an average 55,000 housing units. Each metropolitan area sample covers 4,100 or more housing units. The AHS returns to the same housing units year after year to gather data; therefore, this survey is ideal for analyzing the flow of households through housing. AHS metropolitan data are available for subareas of 100,000 or more that are often defined as groups of census tracts.

3.4.3 Construction and Building Permits Survey

The Current Construction Survey (CCS) provides regional statistics on starts and completions of new single- and multi-family housing units and sales of new single-family houses. New residential buildings currently authorized by a building permit or started in areas not requiring a building permit. Data collected include start date, completion date, sales date, sales price (single-family houses only), and physical characteristics of each housing unit, such as square footage and number of bedrooms.

The Building Permits Survey (BPS) provides current data on new residential construction and additions, alterations, and renovations to existing residential buildings

from all places issuing building permits for private residential structures. Over 98 percent of all privately-owned residential buildings constructed are in permit-issuing places. Data collected on permits issued for private projects include number of buildings, number of housing units, and permit valuation by size of structure. The BPS provides housing permit information by county and place within the United States.

Most of the permit-issuing jurisdictions are municipalities; the remainder are counties, townships, or unincorporated towns. For the municipalities, and townships or towns, the area subject to building permit requirements to which the figures pertain is normally that of the governmental jurisdictions. A small number of municipalities have authority to issue building or zoning permits for areas extending beyond their corporate limits. In such cases, the data relate to the entire area within which the permit-issuing authority is exercised. Similarly, a small number of townships issue permits for only a part of the township and the data normally covers only the area subject to the township's permit system.

These surveys are used by The Conference Board for developing the index of leading economic indicators, the Bureau of Economic analysis for developing national income and products accounts, and The Federal Reserve Board for analyzing national and regional economic conditions. The Departments of Housing and Urban Development use the data to evaluate housing programs. Financial institutions use these statistics to estimate mortgage demand. Private businesses use them for market planning, material use, and investment analysis.

3.4.4 American Community Survey

The American Community Survey (ACS) is a relatively new survey conducted by the Census Bureau with a great deal of potential as a demographic data source.[4] The American Community Survey covers a broad range of topics about social, economic, demographic, and housing characteristics of the US population and it replaced the long form of the decennial census in 2010. The ACS was started in four sites in 1996 and has been expanded every year since that time; it was fully implemented in 2005 with three million households—drawn from all counties or county equivalents in the United States—contacted each year. Starting in 2005, the ACS has produced annual estimates of demographic, housing, social, and economic characteristics for every state, as well as for all cities, counties, metropolitan areas, and population subgroups of 65,000 or more. In 2007, the estimates for areas between 20,000 and 65,000 were released based on the accumulation of survey observations from 2005 to 2007. The first ACS estimates for areas less than 20,000, including block groups and census tracts were released in 2010 based on the accumulation of survey observations from 2005 to 2009. Estimates for areas smaller than 65,000 will be released annually based on successive accumulations of surveys over the preceding three and five year periods.

The American Community Survey (ACS) is designed to provide accurate and timely demographic and economic indicators on a "continuous measurement" basis for federal, state, and local governments, and businesses. A major goal of the ACS is to monitor change over time, but even if the ACS works perfectly, it will show implausible changes for some groups and areas (Hogan 2008; Swanson and Hough 2007). This orientation is a major departure from the traditional decennial census and from the major surveys managed by the Census Bureau, which provide "snapshots" at single points in time rather than continuous measurement observations. The statistical reliability of the ACS samples was intended to match that of the decennial census long form, but because the sample size of the ACS is smaller than originally expected, ACS estimates are less precise than the comparable estimates from Census 2000 and prior decennial census years (Rohanna and Tayman 2006).

There are important differences, aside from sample size, between the ACS and the decennial census that limit their comparability (US Census Bureau 2009a). The census uses the "usual residence" concept, while the ACS uses a length of stay of more than 2 months duration, which can affect the demographic characteristic of places with substantial seasonal populations (Van Auken et al. 2006). In the ACS, characteristics such as income are averages derived from successive monthly samples, as opposed to the point-in-time or interval-of-time reference of the census (Salvo and Lobo 2002). In contrast to the five year residency question in the decennial long form, the ACS asks about residency one year ago. Gross migration flows for states are currently available, but not for counties, and the state flows do not contain any demographic characteristics.

The ACS places a greater demand or burden on end users compared to decennial data. For any given area there are now three ACS numbers to choose from (1-year, 3-year, and 5-year accumulations) for any characteristic. These choices often show substantial variation (Swanson 2010). The interpretation of data accumulated over time intervals is more ambiguous than the decennial data that references an April 1 time point. How for example does one assess the reliability and validity of multiyear accumulations for areas with rapidly changing population or trends calculated from this kind of information? The ACS is very explicit in identifying the error in its estimates by including with its data releases $\pm$ values that represent 90% confidence limits. On the one hand, this information is welcomed, as the error inherent in long-from data was largely kept under wraps or required application of formulas buried deep in appendices. On the other hand, many if not most, users of census data are not concerned or do not have the background to understand and correctly interpret the statistical properties of ACS data (Rohanna and Tayman 2006).

These and other unresolved issues, such as controlling to estimates with a different residency rule and their own error, may limit the ACS's adequacy as a replacement for the decennial long form (GAO 2004).[5] While there are many challenges facing the ACS, strategies for using the ACS have been and are continuing to be developed (e.g., Citro and Kalton 2007; Gage 2006). The ACS holds great potential. Rather than waiting for 10 years for refreshed data from each

decennial census, local data can be made available each year. There is much work to be done, and it is important to recognize that the Census Bureau and data users are experiencing the growing pains of the ACS (Scardamalia 2006).

3.5 Adminstrative Records

Administrative records are records kept by agencies of federal, state, and local governments for purposes of registration, licensing, and program administration. Although not always designed explicitly to do so, these records provide valuable information or symptomatic indicators on specific demographic events or subgroups of the population; although, they do contain both random and systematic biases that can affect their efficacy for demographic analysis (Judson et al. 2001). We have already discussed vital statistics, one type of administrative record that is very valuable for demographic analysis (including the production of population estimates). Other types include Medicare, Internal Revenue Service (IRS), Department of Homeland Security (DHS), utility meters, drivers' licenses, building permits, school enrollment, voter registration, and property tax records. All these data sources can be used for various types of demographic analyses. We discuss several of these data sources in this section that are used for population estimation.

3.5.1 Internal Revenue Service

Internal Revenue Service (IRS) records can be used to estimate migration. By matching the addresses listed on annual income tax returns and adjusting for the number of exemptions claimed on each return, the IRS is able to create an annual set of state-to-state and county-to-county migration flows. These data have several advantages over decennial census data. They are available every year instead of every 10 years, they cover one-year intervals rather than five-year intervals, and they are available on a timelier basis (within one to two years instead of three to five years).

IRS migration data have several limitations, however. Not everyone files an income tax return. In particular, people with low incomes are not required to file. People moving to or from abroad are also likely to be missed. The address listed on a tax return may be that of a bank, law office, accounting firm, or post office box rather than the home address of the filer. This may lead to an inaccurate distribution of the population at the local level. The methodology assumes that people listed as exemptions on a tax return actually live (and move) with the filer; this may not be true (e.g., college students living away from home). Finally, IRS migration data provide no information on the characteristics of migrants and are not available below the county level. For further discussion of the strengths and weaknesses of IRS migration data, see Engels and Healy (1981); Isserman et al. (1982); and Wetrogan and Long (1990).

3.5.2 Department of Homeland Security

Formerly, the Immigration and Naturalization Service (INS), located in the Department of Justice, was the major source of international migration statistics in the United States. That function now resides in the Department of Homeland security. The INS began collecting immigration data in 1892. Before that time, immigration statistics were collected by the Department of State and the Treasury Department, beginning in 1820. Incomplete data on emigrants were collected for a number of years, but those collection efforts were discontinued in the late 1950s (Bryan 2004a: 30).

The DHS produces annual statistics on the number of legal immigrants by type, country of origin, place of intended residence, age, sex, marital status, occupation, and several other characteristics. Data are available for the nation, states, and metropolitan areas. Some DHS data are based on the year in which a person was granted legal immigrant status, which is not necessarily the same as the year in which that person entered the United States. This distinction had a particularly large impact on immigration statistics for 1989-1992, when many aliens who were residing in the country illegally were granted permanent resident status under the provisions of the Immigration Reform and Control Act of 1986 (Immigration and Naturalization Service 1999).

3.5.3 Other Administrative Records

The vast majority of people in the US ages 65 and older are enrolled in Medicare and Medicare information is often used to estimate the population in that age group (Bryan 2004b). School enrollment is another widely used symptomatic indicator of population change. School enrollment is used to estimate migration in the component method (Chapter 10), as well as an independent variable in ratio-correlation models (Chapter 8). Employment at the place of work often used in ratio-correlation models and is found in most economic-demographic models of migration (Smith et al. 2001: Chapter 9). The US Bureau of Labor Statistics produces employment, unemployment and wage data for states and counties. Drivers' license address change is another source of data to estimate migration; it is used in California's population estimation program. Drivers from other states are required to turn in their license when they apply for a California license and similarly for California drivers moving to other states. These data provide an annual estimate of gross migration for persons of driving and have been found to be of sufficient accuracy for use in population estimation (Johnson and Lovelady 1995).

The housing unit method (Chapter 7) requires information to update housing stock and household changes since the last census. One source of information for making these updates is building permits discussed earlier in this chapter. However,

not all building permits get built and there is a lag between when the permit is issued and when the unit is constructed (Smith and Lewis 1980). More refined indicators of housing unit change are certificates of occupancy, which are available from local planning or building departments. These agencies also often have information on housing demolitions and unit conversions (e.g., a simple family home replaced by eight condominiums), which can further improve housing unit estimates. Some applications of the housing unit method use electric utility data to estimate households. Households can be estimated directly from utility data, bypassing the intermediate steps of estimating housing stock and occupancy rate, but there is not always a one-to-one correspondence between the number of meters and the number of households (Smith 1986; Tayman 1994).

The geographic specificity of population estimates has evolved over time. Early on, population estimates were made mainly for states and counties. Estimates now are routinely made for cities and many subcity areas, including census tracts and block groups, but there is a growing demand for estimates for even smaller areas such as assessor's parcels, block faces, and street segments (Rynerson and Tayman 1998; Swanson and Pol 2005). A primary source of data to support these estimates is the parcel file. Parcels are individual house- or business-lots that are tracked by a tax assessor for ownership and taxation purposes. They contain information such as address, assessors' parcel number, unit counts, land and structure value, and land use and zoning codes. Some issues with parcel files include having data only for taxable, private lots; the resources required to reconcile parcel and census housing counts; the need for other sources (e.g., areal imagery) to augment and maintain parcel information; and the GIS expertise and technology to manage and manipulate parcel information (Jarosz 2008). Parcels do offer a consistent and up-to-date source on housing changes, a geography that planners and other users can relate too, and a very accurate and detailed spatial location of housing activity.

In closing, we note that we have largely confined our discussion to administrative records in the United States. However, we note that there are similar records in other countries in which subnational estimation is widely used (Bryan 2004a). These countries include Argentina, Australia, Canada, England, India, New Zealand, and others in which adequate census counts and vital records are available and population registries are not in place. However, as noted in Chapter 16, even in countries with population registers (e.g., Finland), estimation methods must be used if one is interested in De Facto populations.

Endnotes

1. *US House of Representatives v. Department of Commerce*, 525 US 316 (1999)
2. The Social Security Administration also produces national life tables several times each decade (e.g., Bell and Miller 2005).
3. NCHS did not produce state life tables after the 2000 census.

4. As of October, 2011, The URL for the ACS homepage is http://www.census.gov/acs/www/
5. ACS estimates are controlled to post-censal county level population and housing unit estimates. Estimates of person characteristics are based on the person weight and estimates of family, household, and housing unit characteristics are based on the housing unit weight (US Census Bureau 2009b: Chapter 11). A detailed discussion and evaluation of ACS weighting procedures is found in Citro and Kalton 2007: Chapters 5 and 6 and Appendix B).

References

Alonso, W., & Starr, P. (Eds.). (1987). *The politics of numbers*. New York: Russell Sage Foundation.

Anderson, M. J. (1988). *The American census: A social history*. New Haven: Yale University Press.

Anderson, M. J., & Fienberg, S. E. (1999). *Who counts? The politics of census taking in contemporary America*. New York: Russell Sage Foundation.

Anderson, M. J., & Fienberg, S. E. (2000). Race and ethnicity and the controversy over the US census. *Current Sociology*, 48(3), 87–110.

Bell, F. C., & Miller, M. L. (2005). Life tables for the United States social security area 1900–2100 Actuarial Study No. 120. Washington, DC: Social Security Administration.

Bogue, D. J. (1998). Techniques for indirect estimation of total, marital, and extra-marital fertility for small areas and special populations. Meeting of the Federal-State Cooperative Program for Population Projections. Chicago, IL.

Bryan, T. (2004a). Basic sources of statistics. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography*, *Second Edition* (pp. 9–42). New York: Elsevier Academic Press.

Bryan, T. (2004b). Population estimates. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography*, *Second Edition* (pp. 523–560). New York: Elsevier Academic Press.

Choldin, H. M. (1994). *Looking for the last percent: The controversy over census undercounts*. Brunswick: Rutgers University Press.

Citro, C.,F., & Kalton, G. (Eds.). (2007). *Using the American Community Survey: Benefits and challenges*. Washington, DC: The National Academies Press.

DaVanzo, J., & Morrison, P. M. (1981). Return and other sequences of migration in the United States. *Demography*, 18, 85–101.

Edmonston, B., & Schultze, C., L. (1995). *Modernizing the US census: Panel on census requirements in the year 2000 and beyond*. Washington, DC: National Academy Press.

Engels, R. A., & Healy, M. K. (1981). Measuring interstate migration flows: An origin-destination network based on Internal Revenue Service records. *Environment and Planning A*, 13, 1345–1360.

Farley, R. (2008). A brief history of the United States census: 1970 to 2000. National Poverty Center Workshop: Analyzing Poverty and Socioeconomic Trends using the American Community Survey. Ann Arbor, MI.

Flotow, M., & Burson, R. (1996). Allocation errors of birth and death records to subcounty geography. Meeting of the Population Association of America. New Orleans, LA.

Gage, L. (2006). Comparison of Census 2000 and American Community Survey 1999–2001 estimates: San Francisco and Tulare Counties, California. *Population Research and Policy Review*, 25, 243–256.

GAO. (2004). *American community survey: Key unresolved issues*. Washington, DC.

Hahn, R., Mulinare, J., & Teutsch, S. (1992). Inconsistencies in coding of race and ethnicity between births and deaths in US infants: A new look at infant mortality, 1983 through 1985. *Journal of the American Medical Association*, 267, 259–262.

Hogan, H. (2008). Measuring population change using the American Community Survey. In S. H. Murdock, & D. A. Swanson (Eds.), *Applied Demography in the 21st Century* (pp. 13–30). New York: Springer.

Immigration and Naturalization Service. (1999). 1996 statistical yearbook of the Immigration and Naturalization Service. Washington, DC: US Department of Justice.

Isserman, A. M., Plane, D. A., & McMillen, D. B. (1982). Internal migration in the United States: An evaluation of federal data. *Review of Public Data Use*, 10, 285–311.

Jarosz, B. (2008). Using assessor parcel data to maintain housing unit counts for small area population estimates. In S. H. Murdock, & D. A. Swanson (Eds.), *Applied Demography in the 21st Century* (pp. 89–101). Dordrecht, Heidelberg, London, and New York: Springer.

Johnson, H., & Lovelady, R. (1995). Migration between California and other states; 1985–1994. Sacramento, CA: California Department of Finance.

Judson, D. H., Popoff, C. M. & Batutis, M. J. (2001). An evaluation of the accuracy of US Census Bureau county population estimates. *Statistics in Transition*, 5(2), 205–235.

Long, J. F., & Boertlein, C. G. (1990). Comparing migration measures having different intervals. Current Population Reports, Series P-23, No. 166. Washington, DC

Miniño, A., M., Xu, J. Q., & Kochanek, K. D. (2010). Deaths: Preliminary Data for 2008. National Vital Statistics Reports (Vol. 59). Hyattsville: National Center for Health Statistics.

Murdock, S. H., & Ellis, D. R. (1991). *Applied Demography: An introduction to basic concepts, methods, and data*. Boulder: Westview Press.

Rohanna, K., & Tayman, J. (2006). Census data for transportation planning, analysis, and implementation. *Journal of Economic and Social Measurement*, 31, 167–183.

Rynerson, C., & Tayman, J. (1998). An evaluation of address-level administrative records used to prepare small area population estimates. Meeting of the Population Association of America. Chicago, IL.

Salvo, J. J., & Lobo, P. A. (2002). The American Community Survey: Quality response by mode of data collection in the Bronx test site. Paper presented at the Joint Statistical Meetings of the American Statistical Association. New York, NY.

Scardamalia, R. (2006). The American Community Survey: general commentary on the findings from external evaluations. *Population Research and Policy Review*, 25, 293–303.

Sink, L. (1997). Race and ethnicity classification consistency between the Census Bureau and the National Center for Health Statistics. Washington, DC: US Bureau of the Census.

Smith, S. K. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, 81, 287–296.

Smith, S. K., & Lewis, B. (1980). Some new techniques for applying the housing unit method. *Demography*, 17, 323–339.

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

Swanson, D. A., & Hough, G. (2007). An evaluation of persons per household (PPH) data generated by the American Community Survey: A demographic perspective. Southern Demographic Association. Birmingham, AL.

Swanson, D. A. (2010). The American Community Survey and the 2010 Census. Paper presented at the MPO Modeling Conference. San Diego, CA.

Swanson, D. A., & Pol, L. G. (2005). Applied demography in the United States and implications for practice elsewhere. International Population Conference, International Union for the Scientific Study of Population. Tours, France.

Swanson, D. A., & Walashek, P. J. (2011). *CMAF as a census method: A proposal for a redesigned census and an independent Census Bureau*. Dordrecht, Heidelberg, London, and New York: Springer.

Tayman, J. (1994). Estimating population, housing and employment for micro-geographic areas. In K. V. Rao, & J. W. Wicks, W. (Eds.), *Studies in Applied Demography: Proceedings from the International Conference on Applied Demography* (pp. 101–108). Population and Society Research Center, Bowling Green, OH.

US Census Bureau. (2009a). A compass for understanding and using American Community Survey data: What researchers need to know. Washington, DC: US Department of Commerce.
US Census Bureau. (2009b). Data and methodology: American community survey. Washington, DC: US Department of Commerce.
US Census Bureau. (2011). Post-enumeration surveys, (http://www.census.gov/coverage_measurement/post-enumeration_surveys/).
Van Auken, P. M., Hammer, R. B., Voss, P. R., & Veroff, D. L. (2006). The American Community Survey in counties with "seasonal" populations. P*opulation Research and Policy Review*, 25, 275–292.
Wetrogan, S. J., & Long, J. F. (1990). Creating annual state-to-state migration flows with demographic data. Washington, DC: US Bureau of the Census.

Chapter 4
Basic Measures

Demographic analysis and population estimation requires the use of quantitative measures and graphical techniques. This chapter discusses commonly used measures in demography, geography, and statistics (e.g., Barber 1988: Chapter 3; Freedman, Pisani, and Purves 2007; Siegel and Swanson 2004; Smith, Tayman, and Swanson 2001: Chapter 2). We also present graphical techniques for presenting and analyzing tabular and spatial data (e.g., Jacoby 1997, 1998; Krygier and Wood 2011; Tufte 1990, 1997, 2001; Tyner 2010).

Measuring population change and computing ratios and rates are fundamental operations in demography covered in this chapter. Chapter 2 described sources of direct information on migration. Here we discuss indirect methods for estimating net migration in aggregate and by age group. The geographic measures we cover deal with quantifying the distribution of activities across space, comparing the change or differences in spatial distributions, and measuring the distance and accessibility of spatial activities. We also present measures that describe location points, variability, and shape of data distributions, inferential procedures for computing confidence intervals and performing hypothesis tests, and an example of regression analysis. The chapter concludes with a discussion of selected graphical and mapping techniques.

4.1 Demographic

4.1.1 Change

Population change is measured as the difference in population size between two points in time (i.e., two specific dates). A point in time can correspond to the date of a census or to the date of a population estimate. It can refer to changes in size, distribution, or composition, or to any combination of the three. Since censuses are typically more accurate than estimates, measures of change based on censuses will

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_4,

generally be more accurate than measures based on estimates. The measures of population change discussed below above are simple and straightforward. However, they are not always easy to implement properly because of changes in geographic boundaries, changes in the accuracy of the base data, and changes in definitions.

The geographic boundaries of states and most counties have been constant for a long time. Other geographic areas (e.g., cities, zip codes, census tracts), however, have experienced sudden (and sometimes large) boundary changes. Consistent measures of population change are possible only if geographic boundaries are held constant over time. Changes in the accuracy of the base data (e.g., differential census undercount rates or allocation errors) also affect the measurement of population change. Finally, changes in the definition or interpretation of demographic concepts can also affect the measurement of change. Take race, for example. Since respondents were allowed to list only one racial category in the 1990 census and but could list multiple categories in 2000, apparent changes in race groups between 1990 and 2000 may have been caused by changes in reporting practices as well as by changes in the actual population.

Population change can be expressed in either absolute or percentage terms. Absolute change (AC) is computed by subtracting the population at the earlier date from the population at the later date. A negative sign indicates a population loss. Percentage change (PC) is computed by dividing the absolute change by the population at the earlier date and multiplying by 100:

$$\begin{aligned} AC &= P_l - P_b \text{ and} \\ PC &= AC/P_b{}^*100, \end{aligned}$$

where P_l is the population at the later date, P_b is the population at the earlier date.

Population change can also be expressed in terms of an average annual change. The average annual absolute change (AAAC) can be computed simply by dividing the total change by the number of years between the two dates:

$$AAAC = (P_l - P_b)/y,$$

where y is the number of years between the two dates.

For some purposes it is helpful to view annual population change in relative rather than absolute terms, or as annual percent changes (i.e., growth rates) rather than as annual absolute changes. Average annual growth rates can be calculated in two slightly different ways. The first is based on a geometric model:

$$r(geom) = (P_l/P_b)^{(1/y)} - 1,$$

where r is average annual geometric growth rate and the other terms are defined as before. The geometric growth rate calculated in this manner is based on compounding in discrete intervals (i.e., at specific dates). In this example, growth

is compounded once a year. Since population growth occurs continuously, it is useful to compute the average annual growth rate from an exponential model based on continuous compounding:

$$r(\text{expon}) = [\ln(P_l/P_b)]/y,$$

where ln is the natural logarithm. Geometric rates are always slightly larger than exponential rates because they are calculated at discrete intervals rather than continuously. The differences between geometric and exponential growth rates will widen as the annual average growth rate gets larger or more negative in the case of declining areas.

Examining the doubling time can give a more intuitive sense of the long-term impact of growth than simply viewing the average annual growth rate. The exact formula for the doubling time based on annual compounding is:

$$DT = \ln(2)/\ln(1 + r(\text{geom})).$$

A simple and accurate approximation for doubling time, known as the rule of 70, is given by (Keyfitz 1977:4):

$$DT = 70/100\,{}^{*}r(\text{geom}).^{1}$$

Table 4.1 shows the measures of change between 2000 and 2010 in San Diego County. Hispanics and non-Hispanic Asians are the fastest growing ethnic groups in San Diego County. For the first time since the race/ethnic data were collected, the County experienced a decrease in the non-Hispanic White population. The differences between r(geom) and r(expon) are the largest for the Hispanic and non-Hispanic Asian groups and are close in value for the other race groups and total population. For declining populations, the doubling time is negative, indicating the years require to half the population.

4.1.2 Ratio, Proportion, Percentage, and Rate

Demographic analysis requires the use of statistical measures. Two types can be identified. Absolute measures focus on single numbers such as population size, births, deaths, natural increase, or net migration. Relative measures focus on the relationship between two numbers; they are typically expressed as ratios, proportions, percentages, rates, or probabilities. All the relative measures are similar to each other, but each has a distinct meaning.

A ratio is simply one number divided by another. These could be any two numbers, but do not need to have any particular relationship to each other. To be useful, of course, the comparison of the two numbers should provide some type of

Table 4.1 Measures of Population Change by Race and Hispanic Origin, San Diego County, 2000 to 2010

Race/Ethnic Group	2000	2010	AC	PC	AAAC
Non-Hispanic	2,062,868	2,103,965	41,097	2.0%	4,109.7
White	1,548,833	1,500,047	−48,786	−3.1%	−4,878.6
Black of African American	154,487	146,600	−7,887	−5.1%	−788.7
American Indian & Alaskan Native	15,253	14,098	−1,155	−7.6%	−115.5
Asian	245,297	328,058	82,761	33.7%	8,276.1
Native Hawaiian & Other Pac. Is.	12,164	13,504	1,340	11.0%	134.0
Other races and 2 or more races	86,834	101,658	14,824	17.1%	1,482.4
Hispanic	750,965	991,348	240,383	32.0%	24,038.3
Total Population	2,813,833	3,095,313	281,480	10.0%	28,148.0

			Doubling Time[a]	
Race/Ethnic Group	r(geom)[b]	r(expon)[b]	Rule of 70	Exact
Non-Hispanic	0.20	0.20	354.5	351.7
White	−0.32	−0.32	−219.1	−216.2
Black of African American	−0.52	−0.52	−133.9	−131.9
American Indian & Alaskan Native	−0.78	−0.79	−89.2	−87.7
Asian	2.95	2.91	23.7	24.2
Native Hawaiian & Other Pac. Is.	1.05	1.05	66.6	66.7
Other races and 2 or more races	1.59	1.58	44.1	44.3
Hispanic	2.82	2.78	24.9	25.3
Total Population	0.96	0.95	73.1	73.0

[a]Negative values indicate years required to halve the population
[b]For ease of expression the average annual growth rate is multiplied by 100.
Sources: US Census Bureau, 2000 and 2010 censuses

meaningful information. A commonly used ratio in demography is the sex ratio, which is the number of males divided by the number of females (it is often multiplied by 100 for purposes of exposition). Almost universally, more males are born than females causing sex ratios to exceed 100 in the younger ages (see Table 4.2). The predominance of females in the older ages is also evident by sex ratio values considerably below 100. In San Diego, the large sex ratios in ages 15 to 24 are indicative of the military population, especially in ages 18 to 21 shown at the bottom of the table. Alachua County is home of the University of Florida. The sex ratios in the college-age population show predominance of woman attending that university.

A proportion is a special type of ratio in which the numerator is a subset of the denominator. In San Diego County in 2010, there are 991,348 Hispanics and a total population of 3,095,313. The proportion Hispanic is 0.321 (991,348 / 3,095,313). If we multiply a proportion by 100, we get a percentage. So, Hispanics account for 32.1% of San Diego County population in 2010.

A rate is also a special type of ratio. Strictly speaking, a rate is the number of events occurring during a given time period divided by the population at risk to the occurrence of those events.[2] For example, the death rate is the number of deaths divided by the population exposed to the risk of dying and the birth rate is the number of births divided by the population exposed to the risk of giving birth.

Table 4.2 Sex Ratio by Age, San Diego County and Alachua County, 2010

Age	San Diego	Alachua
Under 5	104	105
5 to 9	105	104
10 to14	105	111
15 to 19	111	91
20 to 24	122	98
25 to 29	110	106
30 to 34	105	121
35 to 39	102	104
40 to 44	101	102
45 to 49	100	92
50 to 54	98	87
55 to 59	94	93
60 to 64	92	90
65 to 69	88	87
70 to 74	82	78
75 to 79	78	73
80 to 84	70	53
85+	57	61
18 to 21	119	88

Sources: US Census Bureau, 2000 and 2010 censuses

Although the concept of a rate is clear, it is often difficult or impossible to develop an exact measure of the population at risk to the occurrence of a particular event (Smith, Tayman, and Swanson 2001: 33). This problem is usually solved by using the mid-year population as an approximation of the population at risk. This solution is based on the assumption that births, deaths, and migration occur evenly throughout the year, so that the mid-year population is a measure of the average population during the year.

Crude rates divide the event by the total population. In crude rates the denominator is only a rough approximation of the population at risk of the occurrence of an event. For example, males have a greater risk of dying than females, older people have a greater risk of dying than younger people, and only females of childbearing age can give birth. A commonly used strategy for refining crude rates is to develop rates for specific age-sex groups (race and ethnic groups can be used as well).

In addition to the distinction between crude and age-specific rates, a distinction can also be made between central rates and probabilities (Siegel 2002: 13). In a central rate, the denominator is the mid-year population. In a probability, the denominator is the population at the beginning of the time period, which is thought to correspond more closely to the population at risk to the occurrence of an event during the time period. In reality, the distinction between central rates and probabilities is somewhat fuzzy because of the movement of migrants into and out of the area. It is difficult (if not impossible) to construct true probabilities and central rates are widely used to approximate true probabilities for a variety of demographic measures.

Table 4.3 Selected Fertility Measures, San Diego County, 2010

Age	Female Pop	Births	ASFR[a]	Pop Both Sexes	
15 to 19	106,787	3,596	33.7	0 to 4	203,423
20 to 24	122,109	10,306	84.4	All Ages	3,095,313
25 to 29	119,659	11,620	97.1		
30 to 34	107,537	11,356	105.6		
35 to 39	104,621	6,180	59.1		
40 to 44	104,268	1,604	15.4		
Total	664,981	44,662	395.2		
CBR	14.4				
GFR	67.2				
TFR	1,976.2				
CW R (0-4)	0.306				

[a]Per 1,000 woman
Sources:
US Census Bureau, 2010 census
California Department of Public Health, Birth Records

4.1.2.1 Fertility Rates

A number of measures have been developed to reflect the fertility behavior of a population. We will describe several of the most commonly used. Discussions of these and other fertility measures can be found in Estee (2004), Pullum (2004), Dharmalingam (2004), and Smith (1992). Table 4.3 shows the fertility measures discussed below for San Diego County in 2010.

The simplest fertility measure is the crude birth rate (CBR), which is calculated by dividing the number of births during a year by the midyear population. It is generally multiplied by 1,000 to reflect the number of births per 1,000 persons:

$$\mathrm{CBR} = (\mathrm{B}/\mathrm{P})^{*}1{,}000,$$

where B is the number of births during the year and P is the midyear population. The CBR is limited because it does not account for differences in demographic characteristics. Births occur only to females, primarily those between 15 and 44. The age-sex structure of a population thus has a major impact on its fertility behavior. Other fertility measures have been developed to account for differences in age and sex characteristics.

The general fertility rate (GFR) relates the number of births to the number of females in their prime childbearing years. It is calculated by dividing the number of births by the number of females 15–44:

$$\mathrm{GFR} = (\mathrm{B}/\mathrm{F}_{15-44})^{*}1{,}000.$$

The GFR (sometimes simply called the *fertility rate*) provides a more refined measure than the CBR because it relates the number of births to the population most likely to give birth. It has a major shortcoming, however. The age distribution of persons *within* the 15-44 age group differs from one population to another and changes over time.

A third measure accounts for these differences by focusing on birth rates for each individual age group. The age-specific birth rate (ASBR) is calculated by dividing the number of births to females in a given age group by the number of females in that age group:

$$_{n}\mathrm{ASBR}_{x} = (_{n}\mathrm{B}_{x}/_{n}\mathrm{F}_{x})^{*}1{,}000,$$

where, x is the youngest age in the age interval, n is the number of years in the age interval, $_{n}\mathrm{B}_{x}$ is the number of births to females between the ages of x and x+n, and $_{n}\mathrm{F}_{x}$ is the number of females between the ages of x and x+n at mid-year.

All of this age detail, while valuable, makes it difficult to evaluate changes in fertility behavior over time and to compare differences among regions. The total fertility rate (TFR) summarizes the entire array of ASBRs and facilitates such comparisons. The TFR is the sum of all the individual ASBRs and is calculated as:

$$\mathrm{TFR} = \Sigma\mathrm{ASBR}_{x}, (\text{single year age groups})$$
$$\mathrm{TFR} = 5\Sigma_{5}\mathrm{ASBR}_{x}.(\text{5-year age groups})$$

The TFR can be interpreted as the number of children a hypothetical cohort of 1,000 women would have during their lifetimes if none died and if their fertility behavior at each age conformed to a given set of ASBRs.

We will mention one final measure, the child-woman ratio (CWR):

$$\mathrm{CWR} = (\mathrm{P}_{0-4}/\mathrm{F}_{15-44})^{*}1{,}000,$$

where P_{0-4} is the number of children 0–4 and $\mathrm{F}_{15\text{-}44}$ is the number of women 15–44. The CWR is not a true fertility measure. It is simply a ratio of one population subgroup to another. It incorporates the effects of past mortality and migration patterns as well as past fertility behavior. In contrast to the other fertility measures discussed above it does not require any data specifically related to births. This would be a shortcoming for many analytical purposes, but can be very useful for geographic areas lacking vital statistics data.

4.1.2.2 Mortality Rates

Similar to fertility, a number of measures have been developed to reflect the mortality of a population. Discussions of these and other mortality measures can

Table 4.4 Selected Morality Measures, San Diego County, 2010

Age	Population	Deaths	ASDR[a]	Infant Deaths	Births
0 to 4	203,423	261	128.3	230	44,662
5 to 9	194,029	19	9.8		
10 to14	198,716	26	13.1		
15 to 19	225,095	110	48.9		
20 to 24	270,750	165	60.9		
25 to 29	250,737	143	57.0		
30 to 34	220,185	163	74.0		
35 to 39	211,012	242	114.7		
40 to 44	209,551	411	196.1		
45 to 49	219,795	612	278.4		
50 to 54	210,979	802	380.1		
55 to 59	180,305	934	518.0		
60 to 64	149,311	1,082	724.7		
65 to 69	103,241	1,220	1,181.7		
70 to 74	77,313	1,603	2,073.4		
75 to 79	64,347	2,596	4,034.4		
80 to 84	52,564	3,192	6,072.6		
85+	53,960	6,012	11,141.6		
Total	3,095,313	19,593			
Infant Mortality Rate[b]		5.1			
Crude Death Rate[c]		6.3			

[a]Per 100,000 persons
[b]Per 1,000 births
[c]Per 1,000 persons
Sources:
US Census Bureau, 2010 census
California Department of Public Health, Death Records

be found in McGehee (2004) and Smith (1992). Table 4.4 shows the mortality measures discussed below for San Diego County in 2010.

The simplest measure of mortality is the crude death rate (CDR), which is calculated by dividing the number of deaths during a year by the midyear population. It is generally multiplied by 1,000 to reflect the number of deaths per 1,000 persons:

$$\mathrm{CDR} = (\mathrm{D}/\mathrm{P}) * 1{,}000,$$

where D is the number of deaths during the year and P is the midyear population. The CDR provides an indication of the incidence of deaths relative to the overall size of a population. For many purposes, however, the usefulness of the CDR is limited because it does not account for one of the major determinants of mortality,

namely the age structure of the population. A young age structure is the primary reason why the CDR for blacks is lower than the CDR for whites in the United States, and why the CDR for Alaska is lower than the CDR for West Virginia (Smith, Tayman, and Swanson 2001: 51).

The age-specific death rate (ASDR) deals with this problem by focusing on deaths within each age group. It shows the proportion of persons in each age group (x to x+n) that dies during a year:

$$_{n}\mathrm{ASDR}_{x} = {}_{n}D_{x}/{}_{n}\mathrm{Px}^{*}100{,}000,$$

where x is the youngest age in the age interval, n is the number of years in the age interval, ${}_{n}D_{x}$ is the number of deaths of persons between the ages of x and x+n during the year, and ${}_{n}P_{x}$ is the mid-year population of persons between the ages of x and x+n. ASDRs are generally calculated separately for males and females because of their well-known differences in longevity. ASDRs are often expressed in terms of deaths per 100,000 persons because the rates are very small for many age groups. The J-shaped pattern shown in Table 4.4 reflects the relatively high death rates for newborn babies, the considerably lower rates for young children, the slowly increasing rates at the middle ages, and the rapidly increasing rates at the older ages. This general pattern is found for virtually every population and population subgroup throughout the world.

We will mention one final mortality measure, the conventional infant mortality rate, which relates infant deaths (age less than one year) to the number of births:

$$\mathrm{IMR} = (\mathrm{ID}/\mathrm{B})^{*}1{,}000,$$

where ID is infant deaths, and B is births. The IMR usually provides a sufficiently close approximation of the probability of dying between birth and age 1, and is a widely used indicator of the health of a population, especially in less developed areas (McGehee 2004: 283; Reidpath and Allotey 2003).

4.1.2.3 Life Tables and Survival Rates

The life table is a statistical model that summarizes the mortality (and survival) probabilities observed in a particular population during a particular period of time.[3] The empirical foundation of a life table is a complete set of ASDRs for that year. The ASDRs do not provide exact measures of the risk of dying because some people die before midyear. In order to be useful for the construction of life tables, ASBRs must be converted into age-specific probabilities of dying. Techniques for creating these probabilities and constructing the other functions of a life table can be found in Kintner (2004); Namboodiri and Suchindran (1987), and Smith (1992). Life tables are widely used by public health workers, demographers, actuaries, and

many others. For example, they provide information used in setting insurance premiums and annuity payouts; evaluating pension, social security, and retiree health care liabilities, and product life cycles; and determining the effectiveness of public health and criminal justice programs and drug treatments.

Table 4.5 shows a period life table for both sexes in San Diego County in 2010. The functions of a life table are defined below:

1. Proportion dying ($_nq_x$) – The proportion of persons who are alive at exact age x but die before reaching exact age x+n.
2. Number surviving (l_x) – The number of persons who survive to exact age x, out of a beginning cohort of 100,000 live births (called the radix).
3. Number dying ($_nd_x$) – The number of deaths between exact ages x and x+n, out of the number of persons alive at the beginning of that interval.
4. Person-years lived during an age interval ($_nL_x$) – The summed total of person-years lived between exact ages x and x+n, based on each person's record of survival during that age interval.
5. Total person-years yet to be lived (T_x) – The summed total of person-years lived during this and all following age intervals.
6. Life expectancy (e_x) – The average number of years of life remaining to persons alive at exact age x.

All functions in a life table are dependent on other functions in the table, but the $_nq_x$ is independent of the other functions. Once $_nq_x$ is known, the remainder of the table can be derived (Kintner 2004: 317–318).

The life expectancy at birth (e_0) is similar to the TFR in that both measures use hypothetical cohorts and both assume that a given set of age-specific rates will continue indefinitely. One measure shows the average number of children a cohort of women would have if a given set of ASBRs persisted throughout their lifetimes. The other shows the average life span a cohort of newborn babies would have if a given set of ASDRs persisted throughout their lifetimes. Because they have clear intuitive meanings and are unaffected by the age-sex structure of a population, both measures are useful for making comparisons among regions and over time.

Life tables are frequently used to calculate survival rates or the probability of surviving from one age (or age group) to another. In the cohort component model (Chapter 10), survival rates can be used to estimate deaths by age. Survival rates are typically calculated separately for males and females and are often further subdivided by race and ethnicity. The reason for drawing these distinctions is that mortality rates vary from one demographic subgroup to another (Smith, Tayman, and Swanson 2001: 57).

Survival rates are often based on five-year time horizons and five-year age groups and are calculated as[4]:

$$_5S_x = {}_5L_{x+5}/{}_5L_x,$$

where $_5S_x$ is the survival rate, $_5L_{x+5}$ is the number of person-years lived between ages x + 5 and x + 10, and $_5L_x$ is the number of person-years lived between ages x

Table 4.5 Life Table and Survival Rates, Both Sexes, San Diego County, 2010

Age	Proportion Dying During Interval nQx	Number Living at Start of Interval lx	Number Dying During Interval ndx	Stationary population		Avg. Yrs. of Life at Start of Interval ex
				In the Age Interval nLx	In this and All Subsequent Age Intervals Tx	
0 to 1	0.005126	100,000	513	99,539	8,126,982	81.27
1 to 4	0.000781	99,487	78	397,794	8,027,443	80.69
5 to 9	0.000489	99,410	49	496,927	7,629,649	76.75
10 to 14	0.000654	99,361	65	496,643	7,132,722	71.79
15 to 19	0.002440	99,296	242	495,875	6,636,079	66.83
20 to 24	0.003042	99,054	301	494,515	6,140,204	61.99
25 to 29	0.002848	98,752	281	493,059	5,645,689	57.17
30 to 34	0.003695	98,471	364	491,446	5,152,630	52.33
35 to 39	0.005718	98,107	561	489,134	4,661,184	47.51
40 to 44	0.009759	97,546	952	485,352	4,172,049	42.77
45 to 49	0.013826	96,594	1,335	479,634	3,686,697	38.17
50 to 54	0.018828	95,259	1,794	471,811	3,207,063	33.67
55 to 59	0.025569	93,465	2,390	461,353	2,735,252	29.26
60 to 64	0.035588	91,076	3,241	447,275	2,273,900	24.97
65 to 69	0.057390	87,834	5,041	426,570	1,826,625	20.80
70 to 74	0.098561	82,794	8,160	393,568	1,400,055	16.91
75 to 79	0.183238	74,633	13,676	338,978	1,006,487	13.49
80 to 84	0.263610	60,958	16,069	264,616	667,509	10.95
85+	1.000000	44,889	44,889	402,893	402,893	8.98
Total			100,000			

(continued)

Table 4.5 (continued)

Survival Rates		
From Age	To age	Survival Rate
0 to 1	1 to 4	0.994666
1 to 4	5 to 9	0.999365
5 to 9	10 to 14	0.999428
10 to 14	15 to 19	0.998453
15 to 19	20 to 24	0.997259
20 to 24	25 to 29	0.997055
25 to 29	30 to 34	0.996730
30 to 34	35 to 39	0.995296
35 to 39	40 to 44	0.992267
40 to 44	45 to 49	0.988218
45 to 49	50 to 54	0.983691
50 to 54	55 to 59	0.977833
55 to 59	60 to 64	0.969486
60 to 64	65 to 69	0.953708
65 to 69	70 to 74	0.922633
70 to 74	75 to 79	0.861296
75 to 79	80 to 84	0.780629
80 to 84	85+	0.603577

Sources:
California Department of Public Health, Death Records
US Census Bureau 2010 census

and x+5. For San Diego in 2010, the 5-year survival rate for a person aged 65-69 is 0.9226 (see Table 4.5). Survival rates can be calculated for different time horizons and different age groups by changing the subscripts in the equation shown above. For example, a 10-year survival rate for a five-year age group can be calculated as:

$$_5S_x = {}_5L_{x+10}/{}_5L_x.$$

The 10-year survival rate for a person aged 65-69 in San Diego in 2010 is 0.7947 (338,978 / 426,570).

Due to the peculiar nature of mortality patterns in the first year of life, the 0-4 age cohort is often split into two groups: less than 1 and 1-4 and survival rates are then calculated separately for each group:

$$_{1-4}S_{0-1} = (L_{0-1} + L_{1-4})/500,000; \text{ and}$$
$$_{5-9}S_{1-4} = (L_{5-9} * 0.8)/L_{1-4}.$$

For the youngest group, there is no prior L_x for the denominator, so the radix 100,000 is multiplied by five because, hypothetically, 100,000 new born babies are added each year for a five-year period. For ages 1-4, the person years lived in that age group represents 4 years, while for the 5-9 age group it represents 5 years. Therefore, the latter L_x value is adjusted downward by 20% to estimate a 4 year time period.

The procedure for calculating survival rates for the oldest age group is slightly different because it is an open-ended group. For this age group, T-values rather than L-values are used. Suppose that 85+ is the oldest age group to be projected, the five-year survival rate is calculated as:

$$S_{80} = T_{85}/T_{80},$$

where T_{85} and T_{80} are the total person-years lived after ages 85 and 80.

4.1.2.4 Migration Rates

A fundamental methodological problem in the construction of migration rates is choosing the appropriate population base (i.e., the denominator) to use in calculating migration rates. Theoretically, the appropriate base for any rate is the population at risk of the occurrence of the event under consideration. For mortality and fertility, the choice is clear: the population at risk of dying or giving birth is the population of the area under consideration. For migration rates, however, the choice is not so clear. What is the population at risk of migrating?

Most studies simply use the population of the area under consideration as the denominator in the construction of migration rates, regardless of whether those

rates referred to in-migration, out-migration, or net migration (Long 1988; Meuser and White 1989). Yet the population of the area itself is clearly not the population at risk to in-migration; after all, those people are already living in the area. For net migration the issue is even more difficult because net migration is a residual rather than an actual event; consequently, it has no true population at risk. Following Smith, Tayman, and Swanson 2001: 104–109, we suggest the following rules for determining the appropriate at-risk population for computing migration rates:

1. For out-migration rates, it is the population of the area under consideration;
2. For in-migration rates, it is the population of area of origin; that is the rest of the US (or other country) outside the area under consideration;
3. For net migration rates in areas losing population or growing very slowly, it is the population of the area under consideration; and
4. For net migration rates for areas growing rapidly, it is the rest of the US) or other country).

The other major question regarding the construction of migration rates is whether the denominator should reflect the beginning, middle, or end of the migration interval. We suggest using the population at the beginning of the interval because it is unaffected by migration during the interval and corresponds to the base-year population used for making estimates. It is also common to use the base population "survived" to the end of the migration period using the appropriate survival rates (Irwin 1977; Pittenger 1976). This approach is a more complicated to apply but has the advantage of accounting explicitly for deaths of migrants. Both approaches are acceptable and generally yield very similar results.[5]

Table 4.6 illustrates the computation of domestic in-, out-, and net-migration rates for San Diego County based on the 5-year question in the 2000 census. For the net migration rate, we used the San Diego County population in 1995 as the denominator, since the domestic net-migration was slightly negative. In general, San Diego County experienced net domestic in-migration for retirement age groups and net domestic out-migration for other ages, except those aged 10-19 in 1995. The net in-migration in these age groups likely reflects the movement of military personnel into San Diego County.

4.1.3 Indirect Estimates of Net Migration

In Chapter 2 we discussed data sources that provide data on gross migration, or unidirectional population movements into and out of a region. Estimates of net migration can be derived from these gross migration data by subtracting the number of out-migrants from the number of in-migrants. However, there are many circumstances in which gross migration data are not available. Under these circumstances, indirect estimates of net migration can be made by comparing

Table 4.6 Domestic Migration Rates by Age, San Diego County, 1995-2000

Age in		1995–2000 Migrants			1995 Population (in 000 s)			Migration Rate[a]		
1995	2000	In	Out	Net	US	San Diego	Adj. U.S[b]	In	Out	Net[c]
0 to 4	5 to 9	27,378	40,658	−13,280	19,532.0	225.5	19,306.5	1.42	180.30	−58.89
5 to 9	10 to 14	21,284	28,736	−7,452	19,096.0	196.3	18,899.7	1.13	146.39	−37.96
10 to 14	15 to 19	33,864	26,232	7,632	18,853.0	185.6	18,667.4	1.81	141.34	41.12
15 to 19	20 to 24	87,069	48783	38,286	18,203.0	196.0	18,007.0	4.84	248.89	195.34
20 to 24	25 to 29	61,135	74,244	−13,109	17,982.0	197.3	17,784.7	3.44	376.30	−66.44
25 to 29	30 to 34	45,633	54,629	−8,996	18,905.0	228.3	18,676.7	2.44	239.29	−39.40
30 to 34	35 to 39	38,603	45,964	−7,361	21,825.0	237.5	21,587.5	1.79	193.53	−30.99
35 to 39	40 to 44	27,346	32,405	−5,059	22,296.0	225.0	22,071.0	1.24	144.02	−22.48
40 to 44	45 to 49	19,991	20,940	−949	20,259.0	193.2	20,065.8	1.00	108.39	−4.91
45 to 49	50 to 54	15,872	15,585	287	17,458.0	163.0	17,295.0	0.92	95.61	1.76
50 to 54	55 to 59	10,943	10,884	59	13,642.0	115.6	13,526.4	0.81	94.15	0.51
55 to 59	60 to 64	8,651	8,773	−122	11,086.0	91.0	10,995.0	0.79	96.41	−1.34
60 to 64	65 to 69	7,760	6,651	1,109	10,046.0	82.4	9,963.6	0.78	80.72	13.46
65 to 69	70 to 74	6,375	4,930	1,445	9,926.0	80.6	9,845.4	0.65	61.17	17.93
70 to 74	75 to 79	5,210	4,325	885	8,831.0	77.3	8,753.7	0.60	55.95	11.45
75 to 79	80 to 84	3,860	3,447	413	6,700.0	55.6	6,644.4	0.58	62.00	7.43
80+	85+	3,344	3,240	104	8,163.0	65.0	8,098.0	0.41	49.85	1.60
Total		424,318	430,426	−6,108	262,803.0	2,615.2	260,187.8	1.63	164.59	−2.34

[a]Rate per 1,000 persons
[b]US population minus San Diego population
[c]Based on the San Diego County population
Sources:
US Census Bureau, 2000 Census, 1995–2000 County to County Migration Flows Files
US Census Bureau, Resident Population Estimates of the United States by Sex, Race, and Hispanic Origin: April 1, 1990 to July 1, 1999, with Short-Term Projection to November 1, 2000
State of California, Department of Finance, Race/Ethnic Population with Age and Sex Detail, 1990–1999. Sacramento, CA, Revised May 2009

Table 4.7 Estimation of Net Migration by Race and Hispanic Origin, Vital Statistics Method, San Diego County, 2000-2010

Race/ Hispanic Origin	2000	2010	2000 to 2010				
			Change	Births	Deaths	Natural Change	Net Migration
Non-Hispanic	2,062,868	2,103,965	41,097	253,339	137,595	115,744	−74,647
White	1,548,833	1,500,047	−48,786	164,409	149,831	14,578	−63,364
Black of African American	154,487	146,600	−7,887	18,461	9,201	9,260	−17,147
American Indian & Alaskan Native	15,253	14,098	−1,155	702	850	−148	−1,007
Asian	245,297	328,058	82,761	29,704	9,661	20,043	62,718
Native Hawaiian & Other Pac. Is.	12,164	13,504	1,340	1,202	783	419	921
Other races and 2 or more races	86,834	101,658	14,824	38,861	1,118	37,743	−22,919
Hispanic	750,965	991,348	240,383	201,854	23,954	177,900	62,483
Total Population	2,813,833	3,095,313	281,480	455,193	195,398	259,795	21,685

Sources:
US Census Bureau, 2000 and 2010 censuses
E-2 State of California, Department of Finance, California County Population Estimates and Components of Change by Year, July 1, 2000–2010. Sacramento, California, December 2010.
E-3 State of California, Department of Finance, California County Race/Ethnic Estimates and Components of Change by Year July 1, 2000–2008. Sacramento, California, June 2010.

the population at two points in time, measuring the change due to natural increase, and attributing the residual to net migration. Several methods can be used to calculate net migration in this manner.

One is the vital statistics method, in which net migration (NM) is calculated by rearranging the terms of the demographic balancing equation described in Chapter 2:

$$NM = (P_l - P_b) - (B - D),$$

where P_l is the population in a given year, P_b is the population in some earlier year, and B and D are the number of births and deaths occurring between times (b) and (l). The vital statistics method can be used to calculate net migration not only for the entire population, but also for specific subgroups of the population (e.g., race, and ethnicity) as shown in Table 4.7. In San Diego County between 2000 and 2010, Hispanic and non-Hispanic Asians and Native Hawaiians & Other Pac. Is. were the only ethnic/race groups to show positive growth due to migration.

It is very cumbersome to use the vital statistics method to estimate net migration by age and it is rarely used for this purpose (Morrison, Bryan, and Swanson 2004: 505). Instead of explicitly accounting for deaths, the survival rate method uses survival rates to estimate the expected population of each age group at the end of a particular period. Estimates of net migration are then calculated as the difference between the expected population and the observed (census or estimated) population.

Table 4.8 Estimation of Net Migration by Age, Forward Survival Rate Method, San Diego County, 2000-2010

				2010		
Age in 2000	Age in 2010	2000 Pop	Survival Rate[a]	Expected Pop	Census	Net Migration 2000–2010
Births 2005–10	0 to 4	223,754	0.995105	222,659	203,423	−19,236
Births 2000–05	5 to 9	231,439	0.994337	230,128	194,029	−36,099
0 to 4	10 to 14	198,621	0.998613	198,345	198,716	371
5 to 9	15 to 19	212,829	0.997882	212,378	225,095	12,717
10 to 14	20 to 24	199,669	0.995716	198,814	270,750	71,936
15 to 19	25 to 29	199,919	0.994322	198,784	250,737	51,953
20 to 24	30 to 34	230,953	0.993794	229,520	220,185	−9,335
25 to 29	35 to 39	221,273	0.992041	219,512	211,012	−8,500
30 to 34	40 to 44	222,087	0.987599	219,333	209,551	−9,782
35 to 39	45 to 49	235,183	0.980576	230,615	219,795	−10,820
40 to 44	50 to 54	222,080	0.972100	215,884	210,979	−4,905
45 to 49	55 to 59	191,181	0.961886	183,894	180,305	−3,589
50 to 54	60 to 64	161,622	0.947996	153,217	149,311	−3,906
55 to 59	65 to 69	114,391	0.924607	105,767	103,241	−2,526
60 to 64	70 to 74	90,275	0.879923	79,435	77,313	−2,122
65 to 69	75 to 79	81,763	0.794660	64,974	64,347	−627
70 to 74	80 to 84	78,296	0.672353	52,643	52,564	−79
75+	85+	153,691	0.392290	60,292	53,960	−6,332
Total				3,076,194	3,095,313	19,119

[a] For Births 2005–2010 probability of surviving from birth to age 2.5
For Births 2000–2005 probability of surviving from birth to age 7.5
Other ages probability of surviving 10 years
Average of 2000 and 2010 life table survival rates
Sources:
US Census Bureau, 2000 and 2010 censuses
California Department of Public Health, Birth Records

The most common form of this method is called the forward survival rate method, in which net migration is estimated as:

$$NM = {}_nP_{x+y,l} - {}_xS_n({}_nP_{x,b})^6,$$

where ${}_nP_{x,b}$ is the population age x to x+n in year b, ${}_nP_{x+y,l}$ is the population age (x+y) to (x+n+y) in some later year l, y is the number of years between b and l, and ${}_xS_n$ is the y-year survival rate for age group x to x+n. Other approaches to calculating survival rates and deriving net migration estimates can also be applied. Detailed discussions of the survival rate method and other indirect estimates of net migration can be found in Bogue, Hinze, and White (1982) and Morrison, Bryan, and Swanson (2004).

Table 4.8 shows estimates of net migration by age for San Diego County from 2000 to 2010 using the forward survival rate method. When using this method over

a 10-year period, the youngest two age groups in 2010 were not yet born in 2000, so we use births from the first half and second half of the decade and childhood survival rates to generate the expected population in 2010. For the other age groups, the expected population in 2010 reflects 10-year survival rates applied to the 2000 population for each age group.

The net migration for the overall population (19,119) is near the 21,685 estimate produced by the vital events method and shows the survival rates generated slightly fewer deaths than shown in the vital records. If the survival rates are lowered by a factor of 0.9988, around 0.1%, the two estimates of net migration would be almost identical. The large positive net migration for ages 10 to 19 in 2000 may show the impact of the military movements and foreign migration. The dramatic swing in the net migration between the two youngest ages may also indicate the impact of military families having kids after they move in and move out as the kids age and the net undercount of children in the youngest ages in the census.

The major advantage of indirect methods of estimating net migration is that they can be applied when no direct data on in- and out-migration are available (Smith and Swanson 1998). Consequently, they are particularly useful for projections of small areas. However, the accuracy of these estimates depends the accuracy of the underlying population estimates (or counts) and the vital statistics (or survival rate) data. Errors in these data are directly transmitted to the net migration estimate.

4.2 Geographic

4.2.1 Concentration

Population density, discussed in chapter 2, is one way to measure the concentration of population or other activities. Another simple way of depicting concentration is to use a percentage distribution, which measures the relative size of the activity distributed over a set of geographic areas:

$$\begin{aligned} \mathrm{PctD_i} &= \mathrm{P_i}/\mathrm{P}^{*}100, \text{where} \\ \Sigma\mathrm{PctD_i} &= 100, \end{aligned}$$

where P is population; and i is the geographic area. Another common practice is to show the ordinal rank of a given activity across geographic space. Both ranks and percentage distributions facilitate temporal comparisons of geographic concentration.

The Gini concentration ratio (GCR) is a summary measure of the degree of concentration and ranges from 0.0 to 1.0. A GCR of 0.0 indicates the activity is perfectly distributed across geographic areas. A larger GCR indicates a greater the inequality between the population distribution and geographic areas. The GCR

compares the cumulative percentage distributions of the number of areas (Y_i) and Activity in these areas (P_i):

$$GCR = \left(\sum P_i * Y_{i+1}\right) - \left(\sum P_{i+1} * Y_i\right).$$

The index of dissimilarity (IOD) is a popular summary measure used to compare two percentage distributions (e.g., Duncan, Cuzzort, and Duncan 1961: 83-90; Fonseca and Tayman 1989). The IOD measures the percentage that one distribution would have to change to match the other. The IOD ranges from 0 to 100, with 0 meaning no spatial disparity, and 100 being complete disparity between the two groups with no spatial intermingling. The IOD has been criticized because it only measures two groups at a time and is affected by the number and choice of subunits for comparison (Siegel 2002: 26). The IOD is computed by:

$$IOD = 0.5 * \sum |(P_{il}/P_l) - (P_{ib}/P_b)|,$$

where i and P are defined as before; and l and b represent the distributions being compared. Other measures have been proposed to address shortcomings of the IOD and to describe other aspects of segregation and concentration (Massey and Denton 1988, 1998; McKibben and Faust 2004). In the discussion of error measures found in Chapter 14 we discuss IOD again, but use the term "Index of Misallocation"(IOM) to signal that it is being specifically used as a measure of population estimation error.

Table 4.9 shows these various measures of concentration for census divisions in 1990 and 2010. The population distribution across census divisions has changed modestly over the past 20 years. The GCR has declined from 0.067 in 1990 to close to zero in 2010 and the IOD shows that a relatively small change 5% is needed to equate the two distributions. The rapid growth in the Pacific and Mountain Division increased their rankings by 2010, while the ranks of East N. Central, West N. Central, and E. S. Central decreased.

Another widely used measure of concentration is the location quotient (LQ). LQ is a commonly utilized in economic analysis, but has much wider applicability (e.g., Andresen 2007; Barber 1988: 87; Leigh 1970). LQ quantifies how concentrated a particular industry, cluster, occupation, or demographic group is in a geographic area as compared to a larger reference area. For example, the LQ could be used to compare the percent of households with low income in California with the low income percent in the US. An LQ of 1.0 indicates the area has a same concentration as the larger area; a value less than 1.0 indicates the area has a lower concentration; and a value greater than 1.0 indicates the area has a higher concentration of that activity. The LQ is computed by:

$$LQ = \left(A_c / \sum A_c\right) / \left(R_c / \sum R_c\right),$$

where A is the geographic area of interest; R is the reference area; and c is the characteristic.

Table 4.9 Selected Measures of Concentration, Census Divisions, 1990 and 2010

			Percent Distribution		Rank	
Division	1990	2010	1990	2010	1990	2010
New England	13,206,943	14,444,865	5.3%	4.7%	9	9
Middle Atlantic	37,602,286	40,872,375	15.1%	13.2%	4	4
East N. Central	42,008,942	46,421,564	16.9%	15.0%	2	3
West N. Central	17,659,690	20,505,437	7.1%	6.6%	6	7
South Atlantic	43,566,853	59,777,037	17.5%	19.4%	1	1
East S. Central	15,176,284	18,432,505	6.1%	6.0%	7	8
West S. Central	26,702,793	36,346,202	10.7%	11.8%	5	5
Mountain	13,558,776	22,065,451	5.5%	7.1%	8	6
Pacific	39,127,306	49,880,102	15.7%	16.2%	3	2
Total	248,609,873	308,745,538	100.0%	100.0%		
Gini Concent. Ratio 1990[a]	0.067					
Gini Concent. Ratio 2000[a]	0.005					
Index of Dissimilarity	5.0					

[a]Based on the number of counties within each census division.
Sources:
US Census Bureau, 1990 and 2010 censuses

A drawback of the LQ is that a value is calculated for each area being analyzed. For a county with 600 census tracts, the LQ would unwieldy to analyze. The coefficient of localization (CL) complements the LQ, and provides a single number that measures the relative concentration of an activity across all areas. The CL ranges from 0 to 1 and differs from the range of the LQ, which has a minimum value of zero and no upper limit. A CL of zero means the percentage distribution of an activity is evenly spread across the areas in accordance with the percent in the reference area. As CL approaches 1.0, the activity becomes increasingly concentrated in one area. The CL is computed in three steps (Barber 1988: 89):

1. Calculate the share of the reference area activity in each area (A_c / R_c);
2. Calculate the share of the reference area total in each area($\sum A_c$ / $\sum R_c$); and
3. Subtract the value from Step 2 from the value in Step 1; add either all positive or all negative differences.

Table 4.10 presents LQ and CL values for housing structure type in the 18 incorporated cities and unincorporated area in San Diego County, the reference area. Single family units are over-concentrated in 10 of the 19 areas, with the greatest excesses in the Encinitas, Lemon Grove, and Poway. El Cajon has the largest under-representation of single family units with a LQ of 0.75. Multiple family structures containing 2 to 19 units are over-concentrated in seven areas, with the greatest excesses in El Cajon, Imperial Beach, and National City. Poway shows the largest under-representation of these types of units with a LQ of 0.35. Coronado has the largest overconcentration of large multiple family structures with a LQ of 1.75, followed by El Cajon (1.63), La Mesa (1.29), and San Diego (1.27). Santee has the largest under-representation, with a LQ of 0.35. Other units are heavily

Table 4.10 Location Quotients and Coefficient of Localization for Housing Structure Type, Jurisdictions in San Diego County, 2005–2009

Jurisdiction	Single Family[a]	2 to 19	20+	Other[b]
Carlsbad	1.14	0.83	0.72	0.74
Chula Vista	1.05	0.74	1.05	1.44
Coronado	0.93	0.92	1.74	0.02
Del Mar	1.15	0.90	0.77	0.00
El Cajon	0.75	1.25	1.63	1.39
Encinitas	1.25	0.65	0.49	0.65
Escondido	0.94	1.03	0.94	1.92
Imperial Beach	0.79	1.72	0.91	0.59
La Mesa	0.92	1.18	1.29	0.31
Lemon Grove	1.24	0.72	0.53	0.31
National City	0.86	1.30	1.26	0.62
Oceanside	1.08	0.92	0.68	1.29
Poway	1.30	0.35	0.64	1.15
San Diego	0.91	1.21	1.27	0.33
Santee	1.04	0.94	0.35	2.96
Solana Beach	1.09	0.81	1.19	0.09
Vista	0.97	1.10	0.81	1.59
Unincorporated Area	1.19	0.56	0.43	2.35
Coeff. Of Localization	0.058	0.123	0.162	0.367

[a]Single family attached and detached
[b]Mobile homes and other units
Source: US Census Bureau, 2005–2009 ACS

concentrated in Santee and the unincorporated area with LQs of 2.96 and 2.35. Del Mar has no other units and the LQ in Coronado is only 0.02. The CL of 0.058 shows that single family units are the most evenly spread across San Diego County. The distribution of multiple family units is more concentrated than single family units and the concentration is greater for 20+ units (0.162) than for 2 to 19 units (0.123). Other units are the most unevenly distributed across jurisdictions San Diego County with a CL of 0.367, which is over six times larger than the SF CL and over two times larger than the 20+ units CL.

4.2.2 Center of Population and Distance

The measures presented above and in the statistical section of this chapter below do not incorporate a spatial dimension related to fundamental spatial concepts of distance, direction, or relative location. Center of population (CENTP) and distance measures require information on the location of attributes identified by geographic coordinates (or points) such as longitude and latitude that can be represented as X and Y coordinates on a grid system. In this section we present measures that describe the center and variability of a distribution of points.

The mean center is the center of gravity of a point distribution and is a generalization of the arithmetic mean. It is the average of the X and Y coordinates over all points. Perhaps more useful is the weighted mean center that includes the magnitude of activity associated with a point, such as population. The CENTP is an example of a weighted mean center:

$$\text{CENTP} = \sum P_i * X_i / \sum P_i \text{ and } \overline{y} = \sum P_i * Y_i / \sum P_i,$$

where P_i is the population at point i; X_i and Y_i are the horizontal and vertical coordinates. The weighted mean center is affected by extreme values and is influenced by any change of the distribution over the total area (Plane 2004: 98). The weighted Manhattan and Euclidian medians are designed to have equal weights above and below and to the left and right, so they are not influenced by extreme values (Barber 1988: 99-101). Differences between the weighted mean and weighted median centers can be substantial. For example, the location of the 2010 median center of the US population is around 50 miles south of the 1940 mean center of the US population, and about 350 miles east of the 2010 mean center of US population.

Another important characteristic of a spatial distribution is its dispersion or variability. The standard distance (SD) is the spatial equivalent to the standard deviation and is based on deviations between each point and the central point (D_{ic}):

$$D_{ic} = \sqrt{\left(x_i - \bar{x}\right)^2 + \left(Y_i - \bar{Y}\right)^2}$$

and the standard distance by:

$$SD = \sqrt{\sum D^2{}_{ic}/n}.$$

The SD can also be computed for weighted individual points by computing the standard distance around its weighted mean center and for data grouped by area by assuming the activity is concentrated in its geographic center (Barber 1988: 103; Plane 2004: 100). As the size of the areal unit decreases, the computed SD will approach the value from the location of individual points. Standard distances can be mapped as a line segment from the center location origin. The SD is sensitive to extreme observations, and spatial dispersion measures about the median and weighted median centers have been proposed (Barber 1988: 103).

4.2.3 Accessibility and Spatial Interaction

Many practical applications including locating business and public facilities rely on information that measures the accessibility of a particular location or locations with reference to a distribution of population, jobs, houses, or other characteristics.

We discuss commonly used measures and models of accessibility and spatial interaction. More detailed information on these and other measures and models and measures can be found in (e.g., El-Geneidy and Levinson 2006; Guers and van Wee 2004; Handy and Niemeier 1997; Haynes and Fotheringham 1984; Koenig 1980).

One of the most widely used and simple measures of accessibility is the isochronic or cumulative opportunity (IOP), which counts the number of potential opportunities that can be reached within a predetermined travel time or distance[7]:

$$\mathrm{IOP_i} = \sum \mathrm{B_j O_j},$$

where $\mathrm{IOP_i}$ is accessibility measured at point i to potential activity in zone *j*; $\mathrm{O_j}$ are opportunities or characteristics in zone *j*; and $\mathrm{B_j}$ is a binary value equal to 1 if zone *j* falls within the predetermined threshold and 0 otherwise. For example, take a location near Arlington National Cemetery. In 2010, it is estimated that 31,600 and 269,000 people live within a 1-mile and a 3 mile radius of this location (Nielsen Solution Center 2010). These people live in households with median incomes of $84,800 (1-mile) and $82,900 (3-mile).

An alternative to the measures like the IOP is a measure that weights the characteristics by the spatial separation between the characteristics and the location where the accessibility is being measured (Plane 2004: 101). One such aggregate accessibility measure is the population potential (PP) (Hansen 1959). The population or characteristic is adjusted for the distance between its zone and the location zone and receives more weight the closer the distance as follows:

$$\mathrm{PP_i} = \sum \mathrm{P_j}/\mathrm{D_{ji}},$$

where $\mathrm{P_j}$ is the population or characteristic of n areas and $\mathrm{D_{ji}}$ are the distances of these areas from location *i*.

The gravity-based model is a widely used method for measuring accessibility and spatial interaction (El-Geneidy and Levinson 2006; Haynes and Fotheringham 1984: 20-29).[8] Gravity models can be based on several formulations. The gravity model for measuring accessibility (A) is a function of the opportunities (characteristics) in zone *j* and the impedance of traveling from zone *j* to the location point or area *i* ($\mathrm{I_{ji}}$):

$$\mathrm{A_i} = \sum \mathrm{O_j}^{*}(\mathrm{I_{ji}}).$$

The impedance is measured as travel time or cost and can be further specified by travel mode (e.g., auto, transit, non-motorized). Different functional forms (e.g., declining power, negative exponential, and Gamma) can be used to model the impedance function between *j* and *i* (e.g., El-Geneidy and Levinson 2006; Lowry 1964). The basic gravity model shown above has been enhanced to include measures of attractiveness or opportunities in zone *i* and trip end constraints to insure consistency between the regional characteristics and the zone *j* allocations (Putman 1983, 1991).

Table 4.11 Accessibility Based on a Gravity Model

Trip Cost Probabilities from Zone (j) to Point (i)[a]				
	Opportunity Zones (j)			
Location Point (i)	1	2	3	
A	0.05	0.03	0.08	
B	0.15	0.45	0.10	
C	0.20	0.32	0.40	
Population	123	723	424	
	Opportunity Zones (j)			
Location Point (i)	1	2	3	Accessibility
A	6.2	21.7	33.9	61.8
B	18.5	325.4	42.4	386.3
C	24.6	231.4	169.6	425.6

[a]Trip cost probabilities are based on a modified gamma function ($Cost_{ji}$-1.5 * exp(-2$Cost_{ji}$))

Table 4.11 considers a hypothetical area containing three opportunity zones and three point locations and calculates the accessibility of these points to the population based on a gravity model. The top panel of the table shows trip cost probabilities from each opportunity zone to each point location. For example, the accessibility of location point A to all opportunity zones reflects its high travel costs relative to the other points, as indicated by the low probabilities. The lowest travel cost is found between opportunity zone 2 and location point B. This pair has the highest probability at 0.45. As a result, location point A has by far the lowest accessibility measure of 61.8. The accessibility measures are much closer in values for location points B and C. For location point B, most of its value (84.2%) comes from its relatively low cost of travel to opportunity zone 2 and the relatively large population residing in this zone.

4.3 Statistical

4.3.1 Descriptive

Descriptive statistics quantify three general characteristics of a data distribution: location, variability, and shape. Location refers to various points of the data distribution, including central tendency. Variability describes how spread out or closely clustered a set of data is. The shape of a distribution is usually characterized by its symmetry or lack thereof and the extent of its peakedness or flatness.

One way to measure location is to divide the data into equal parts and determine the end points of these parts. Common divisions include quartiles (4 parts), deciles (5 parts), and percentiles (100 parts). A percentile is the value of a variable below which a certain percent of observations fall. Commonly used percentiles are the 25^{th} and 75^{th}, which represent the middle half of the data set, and the 50^{th} percentile that corresponds to the median.

The median is one of several measures of central tendency that describe the location of the middle of the distribution. The arithmetic mean or average is the most commonly used measure of central tendency and is computed by summing the observations and dividing by size of the population (N) or sample (n)[9]:

$$\mu = \sum X/N; \qquad \text{Population}$$
$$\bar{x} = \sum x/n. \qquad \text{Sample}$$

As discussed in Chapter 2, the mean may not accurately reflect the middle of the distribution when the distribution is skewed. A trimmed mean is less susceptible to the effects of extreme scores. It is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores. For data collected over time, the arithmetic mean gives the wrong answer (Levin and Rubin 1998: 78). In this situation, the geometric mean (GM) is the appropriate measure of central tendency for the average rate of change:

$$GM = \sqrt[n]{(X_1 * X_2 * X_3 \ldots * X_n)}.$$

The simplest measure of variability is the range or the difference between the minimum and maximum values of the data set. The range ignores most of the data and can be heavily influenced by outlying observations. The interquartile range minimizes the effect of outliers by taking the difference between the 25th and 75th percentiles; the interquartile deviation divides the interquartile range in half. The interquartile range does not take advantage of all of the information, does not measure the variability within the middle of the distribution, and by design ignores the variability at the extremes.

The most widely used measures of variation are the variance and its square root, the standard deviation. These measures describe how far each observation lies from the mean of the data set[10]:

$$\sigma^2 = \Sigma(X_i - \mu)^2/N; \qquad \text{Population Variance}$$
$$\sigma = \sqrt{\sigma^2}; \qquad \text{Population Standard Deviation}$$
$$s^2 = \Sigma(x_i - \bar{x})^2/(n-1); \qquad \text{Sample Variance}$$
$$s = \sqrt{s^2}. \qquad \text{Sample Standard Deviation}$$

The standard deviation is easier to interpret than the variance because it is expressed in the same units as the observations. The sample variance uses (n-1) and not (n) in the denominator because it provides a better estimate of the population variance (Levin and Rubin 1998: 102).

The standard deviation is an absolute measure of dispersion that is not comparable across data sets that have widely different means or different units of measurement. For example, a data set has a mean of 50 and standard deviation of 10. If every observation is multiplied by 3, the mean and standard deviation become 150

and 30. In this case the variation of the data has not been altered, but just comparing the two standard deviations would provide a misleading picture. What is needed is a measure of relative variability or the coefficient of variation (CV). The CV is useful because the standard deviation should always be interpreted in the context of its mean. The coefficient of variation is a dimensionless number that relates the mean to the standard deviation of a distribution:

$$\begin{aligned} CV &= \sigma/\mu^{*}100; && \text{Population} \\ cv &= s/\bar{x}^{*}100. && \text{Sample} \end{aligned}$$

Skewness and kurtosis are characteristics of the shape of the distribution. Skewness refers to the distribution's symmetry, while kurtosis its peakedness. A negative skewness indicates that the tail on the left side of the distribution is longer than the right side and the bulk of the values (including the median) lie to the right of the mean. A positive skewness indicates that the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean. A zero value indicates that the values are relatively evenly distributed on both sides of the mean, typically, but not necessarily, implying a symmetric distribution. A normal distribution is mesokurtic and has as kurtosis value of 3.0. Values below 3.0 indicate a flatter distribution and above 3.0, a more peaked distribution. Skewness and kurtosis are computed by:

$$\begin{aligned} SK &= \Sigma(X_i - \mu)^3/(N^{*}\sigma^3); && \text{Population Skewness} \\ sk &= \Sigma(x_i - \bar{x})^3/((n-1)^{*}s^3); && \text{Sample skewness} \\ KT &= \Sigma(X_i - \mu)^4/(N^{*}\sigma^4); && \text{Population kurtosis}[11] \\ kt &= \Sigma(x_i - \bar{z})^4/((n-1)^{*}s^4). && \text{Sample kurtosis} \end{aligned}$$

Table 4.12 displays the descriptive statistics for household size for all census tracts in Salt Lake County, Utah and Cumberland County, Maine. These two areas were chosen because Utah typically has the highest household sizes in the US and Maine the lowest. The household size in Salt Lake County was among the highest of any Utah County and that in Cumberland County was among the lowest in Maine. The mean household size is around 30% higher in Salt Lake County (3.00 vs. 2.32). There is not much difference is the three measures of central tendency in either county, indicating the census tract distributions are not greatly impacted by outliers. The minimum census tract value is similar in both counties, but the maximum value is over a person larger in Salt Lake County. As a result, the range for census tracts in Salt Lake County is larger, as are the other measures of variability that are approximately double those for Cumberland County. The relative variation is also higher for Salt Lake County as seen by the CV of 21.3, which is 44% higher than the CV in Cumberland County (14.8). The census tract distributions for both counties have a slight left-skewness, but do not indicate great departures from symmetry. Census tracts in Salt Lake County have a flatter distribution than those in Cumberland County.

Table 4.12 Descriptive Statistics, Persons per Household by Census Tract, Salt Lake County and Cumberland County, 2005–2009

Statistical Measure	Salt Lake County	Cumberland County
Location		
Minimum	1.31	1.36
25th Percentile	2.54	2.12
Mean	3.00	2.32
Median	2.99	2.37
Trimmed Mean (5%)	3.01	2.32
75th Percentile	3.47	2.54
Maximum	4.19	3.10
Variability		
Range	2.88	1.74
Interquartile Range	0.93	0.42
Semi-Quartile Range	0.47	0.21
Variance	0.409	0.119
Standard Deviation	0.639	0.345
Coefficient of Variation	21.3	14.8
Distribution Shape		
Skewness	−0.251	−0.382
Kurtosis	2.406	3.332
Number of Observations	193	61

Source: US Census Bureau, 2005–2009 ACS

4.3.2 Inferential

Let's now assume that the data in Table 4.12 represent a random or probability sample of census tracts in the two counties. The various statistics shown represent point estimates of their respective population parameters. To develop interval estimates that account the fact that we are analyzing a sample and not a population, two pieces of information are required: 1) a measure of the variability of the sampling distribution; and 2) confidence level. The sampling distribution variability is measured by the standard error, which relates the sample standard deviation and the sample size. Each sample statistic has a standard error (SE). Some examples of standard errors are shown below:

$$SE_{\bar{x}} = s/\sqrt{n} \qquad \text{Mean}$$

$$SE_{\bar{p}} = \sqrt{(\bar{p}^{*}\bar{q})/n} \qquad \text{Proportion}^{12}$$

$$SE_{median} = 1.25 * SE_{\bar{x}}. \qquad \text{Median}$$

The SE is influenced by the variability in the sample and sample size. The variability in sampling distribution declines with increases in the sample size. But since the formula uses the square root of the sample size, there is a diminishing effect of sample size increases. The SE for the median is 25% larger than the mean, in large samples with a normal distribution, and therefore the median varies more from sample to sample than the mean.

The confidence level determines the area under the curve of the probability function that that underlies the sampling distribution (e.g., normal curve or t-distribution).[13] For example, a 95% confidence interval would have 2.5% of the area in both tails of the curve. The corresponding standard or Z score is then used to compute confidence intervals. For example, standard scores under a normal curve for 90%, 95%, and 99% confidence intervals are 1.64, 1.96, and 2.58. The general formula for the confidence interval adds and subtracts the standard error times the appropriate standard score for the desired level of confidence to the sample statistic. For example, a confidence interval of the mean under the normal distribution is:

$$\bar{x} \pm SE_{\bar{x}} {}^{*}Z.$$

Hypothesis testing begins with an assumption, called the null hypothesis, made about a population parameter. Sample data are used to determine the difference between the hypothesized value and sample statistic. Smaller differences increase the likelihood that the hypothesized value is correct. Larger differences decrease the likelihood. Using a test statistic and the appropriate sampling distribution (e.g., normal, t, F or Chi-square), one computes the likelihood or probability of getting this result assuming the null hypothesis is true, or the p-value.[14] In general a test statistic is computed by taking the difference between the hypothesized value and sample statistic and dividing that difference by the standard error. One accepts the null hypothesis if the p value is $\geq \alpha$ and rejects it if the p value is $< \alpha$. Remember, α is a Type 1 error or the probability of rejecting a null hypothesis when it is true.

Table 14.3 presents selected confidence intervals and hypothesis tests treating the census tract data as representing two independently drawn samples. Confidence intervals for the mean are given in the top panel. Even though the sample standard deviation is larger for Salt Lake County (see Table 4.12), the standard errors are similar for both sets of census tracts due to the larger sample size in Salt Lake County (193 vs. 61). The intervals indicate a narrow margin of error around the mean for the census tracts in both counties; the difference between the upper and lower limits is around 3%.

The middle panel shows the results of a hypothesis test for a normal distribution (D'Agostino, Belanger, and D'Agostino 1990). Our test criterion is $\alpha = 0.05$. This procedure tests three null hypotheses: 1) the skewness is zero, 2) the kurtosis is that of a normal distribution; and 2) both the symmetry and kurtosis are from a normal distribution. All three null hypotheses are accepted for Cumberland County census tracts, suggesting a normal distribution. For Salt Lake County, the assumption of symmetry is accepted, but the hypothesis that its distribution has the kurtosis associated with a normal distribution is rejected.

The final panel shows the difference of mean household size test. We expect that the mean household size for census tracts in Salt Lake County would be greater than the mean for Cumberland County. The null hypothesis is set up so not accepting it would support our expectation, which is what the p-value indicates. One assumption of a difference of means test is the variance of both samples is equal. We ran a hypothesis test and rejected the null hypothesis of equal variances, so we show the difference of means test with and without that assumption.

Table 4.13 Inferential Statistics, Persons per Household by Census Tract, Salt Lake County and Cumberland County, 2005–2009

	95% Confidence Interval Around the Mean		
	Standard Error	Lower Limit	Upper Limit
Salt Lake County	0.046	2.95	3.04
Cumberland County	0.044	2.28	2.36
	Test of Normality (p-value)		
	Skewness	Kurtosis	Joint
Salt Lake County	0.149	0.028	0.037
Cumberland County	0.200	0.355	0.272
	Differences of Means Test (Ho: Cumberland $\geq$ Salt Lake)		
	T-Statistic	p-value	
Equal Variances	7.9	0.000	
Unequal Variances	10.6	0.000	

Source: US Census Bureau, 2005–2009 ACS

4.3.3 Regression

To illustrate the results from a regression analysis, we use the data and ratio correlation model for the 39 counties in Washington State discussed in Chapter 8. The dependent variable is the county's share of the state population in 2000 divided by its share in 1990. The three independent (explanatory) variables, measured in the same way, are: registered voters, registered automobiles, and school enrollment grades 1–8. Table 4.14 contains a variety of statistics from this multiple regression.

The top panel shows the general output for a regression analysis. The explanatory variables explain 79.4% of the variation in the population ratio. Taking into account the number of parameters in the equation reduces the r-square to 77.6%. The standard error of 0.038 measures the overall fit of the equation, but is not a standardized measure like the r-square and is sensitive to the units of measurement of the dependent variable. The analysis of variance table shows the overall explanatory value of the regression is statistically significant from zero. The unstandardized regression coefficients (b) are all positive as expected, and represent the effect of the variable on the population ratio taking into account or controlling for the other independent variables. The standardized regression coefficient (BETA) is used to gauge their relative strengths in explaining movements in the population ratio. School enrollment is the strongest predictor, followed by automobile registration, and finally by voter registration.

The standard error, T-stat, and p-value are used to test the null hypothesis the slope (b) equal to zero. The coefficients for school enrollment and automobile registration are significantly different from zero at $\alpha = 0.05$, but not voter registration. The zero-order correlation between voter registration and population is 0.563 and is significantly different from zero, but its effect on population is diminished once the other variables are taken into account. Finally, 95% confidence intervals are presented for the coefficients of each independent variable. When a coefficient is significantly different from zero, the endpoints of the interval will have the same sign; otherwise the signs will differ, as is the case for the voter registration variable.

Table 4.14 Regression Statistics for Ratio-Correlation Model, Washington State Counties, 2000/1990

Multiple R	0.891
R Square	0.794
Adj. R Square	0.776
Std. Error	0.038
Observations	39

Analysis of Variance Table

	DF	SS	MS	F	Sig. F
Regression	3	0.198	0.066	44.855	4.40E-12
Residual	35	0.052	0.001		
Total	38	0.250			

Coefficients, Standard Errors, T- and p-values, and Confidence Intervals

	b	Beta	Std. Error	T-stat	p-value	Lower 95%	Upper 95%
Intercept	0.195		0.071	2.734	0.010	0.050	0.341
Voters	0.093	0.141	0.062	1.505	0.141	−0.033	0.219
Autos	0.336	0.410	0.079	4.246	0.000	0.175	0.497
Enrollment	0.398	0.545	0.063	6.344	0.000	0.271	0.525

Influence and Leverage Values

	Minimum	Maximum	Mean	Standard Deviation	Cut-off Point	Exceed[1] Cut-off
Cook's Distance	0.000	0.156	0.029	0.042	0.103	7.7%
DFBeta-Voters	−0.031	0.330	0.000	0.011	0.320	0.0%
DFBeta-Auto	−0.027	0.046	0.000	0.014	0.320	0.0%
DFBeta-Enrollment	−0.026	0.033	0.000	0.110	0.320	0.0%

Multicollinearity Assessment

Correlation Matrix

	Voters	Autos	Enrollment
Voters	1.000	0.554	0.358
Autos	0.554	1.000	0.425
Enrollment	0.358	0.425	1.000

Variance Inflation Factors

Voters	Autos	Enrollment
20.7	24.5	31.1

[1]Percent of the sample larger than cut-off value

We calculated Cook's D and DFBETA to evaluate the impact of influential observations on the regression results (Chatterjee and Hadi 1998; Chapter 4). Cook's D assesses the influence (i.e., scaled distance) of an observation on the estimated set of coefficients. Values exceeding the conventional cut-off point (4/n) indicate an observation that may excessively influence the regression results. The DFBETA diagnostic assesses the effect of an individual observation on each estimated parameter in the model; for each parameter estimate, DFBETA calculates for each observation the standardized difference in the parameter estimate due to deleting the observation. Absolute values exceeding the conventional cut-off point ($2/\sqrt{n}$) indicate that a particular observation may be excessively influential. These results are shown in the middle panel of Table 14.4. Influential observations do not appear to be a significant problem in this regression model. Three counties exceed the Cook's D threshold, while the DFBETA shows no influential points for any variable. We reran the regression removing the counties that exceeded the Cook's D threshold and the results were not materially different from the original model.

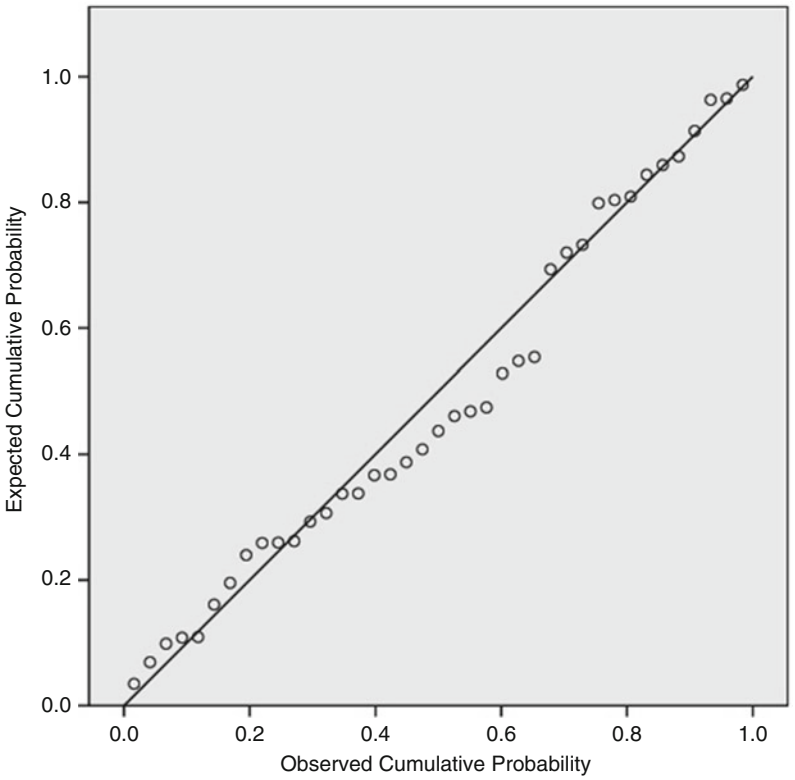


Fig. 4.1 Normal Probability Plot of Regression Residuals, Ratio-Correlation Model

The last panel of the table shows statistics to evaluate multicollinearity. The usual approach is to examine a correlation matrix of the independent variables. These correlations are moderate in value and range from 0.358 to 0.554. While a high correlation between independent variables can point to a collinearity problem, the absence of a high correlation cannot be interpreted as evidence of no problem (Belsley, Kuh, and Welsch 1980:92). It is possible for three variables to be collinear, while no two pairs alone are highly correlated. Another way to examine multicollinearity is through the variance inflation factor (VIF), which measures the proportion of the variance in an independent variable associated with the other independent variables. The common rule of thumb is a VIF greater than 10 indicates excessive or severe multicollinearity (e.g., Kennedy 1992: 183; Marquardt 1970; Neter, Wasserman, and Kutner 1989: 409), but in some circumstances the VIF threshold could be considerably larger than 10 (O'Brien 2007). Based on the criteria of 10, it appears that multicollinearity a potential issue in this regression.

We also plotted residuals against predicted y-values and each independent variable to check for heteroscedasticity and model misspecification, and created normal probability plots to check the normality assumption (Draper and Smith 1981: Chapter 3). Figure 4.1 shows the normal probability plot. Under the normality

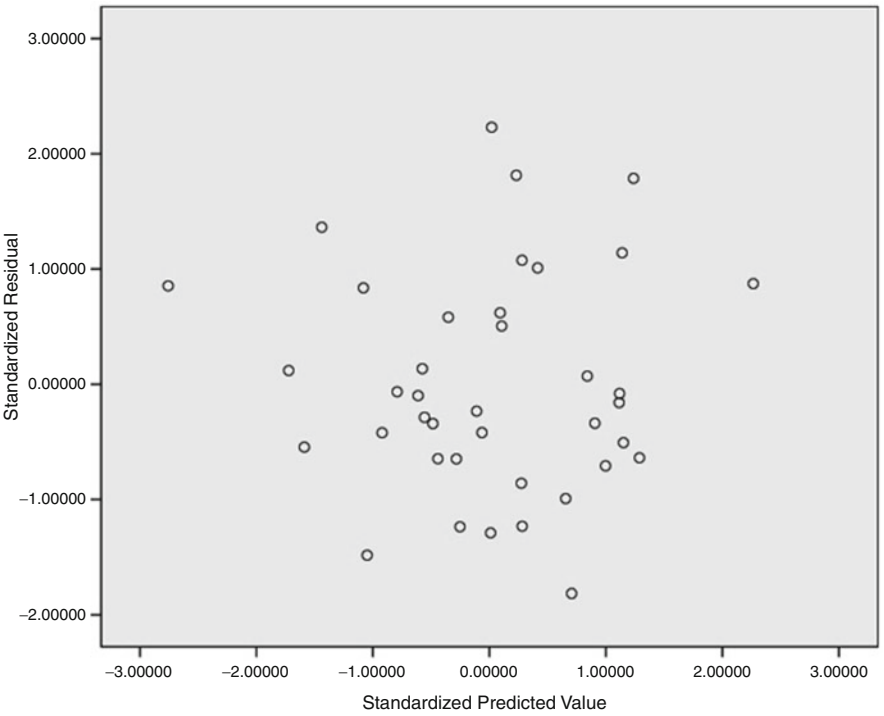


Fig. 4.2 Scatter Plot of Regression Residuals versus Predicted Values, Ratio-Correlation Model

assumption, all of the residual points would fall directly on the 45 degree line. There is some departure from normality is the middle of the curve and may indicate excessive skewness in the residuals with too many errors in one direction. The plot of the residuals against the predicted y-values are random and do not indicate the presence of heteroscedasticity or model misspecification (see Figure 4.2). Plots of the residuals against each independent variable did not reveal any abnormalities (data not presented). If heteroscedasticity is present, the simplest solution is to use heteroscedasticity-robust standard errors (Stock and Watson 2003: 126-129). In our example, there was little difference between the standard errors for the coefficients between the heteroscedasticity-robust standard errors and standard errors not corrected for heteroscedasticity.

4.4 Data Display

This final section of this chapter discusses commonly used statistical graphical and mapping techniques for analyzing and displaying information. Statistical graphics are used to achieve four broad objectives (Jacoby 1997: 2-4): 1) exploring the

contents of a dataset; 2) finding structure in the data; 3) checking assumptions of statistical models, as just described; and 4) communicating the results of the analysis. Statistical graphics provide useful summaries of large, complicated datasets and emphasize the important features of the data. Graphical tools are not as reliant on underlying assumptions of the data (e.g., interpreting the mean in a skewed distribution), are less subject to misrepresentation, and facilitate a greater interaction between the researcher and data by highlighting interesting and unusual aspects of quantitative data. Further details on the design principles and practices of constructing, using, and interpreting statistical graphics can be found in Jacoby (1997, 1998), Schmid (1983), and Tufte (1990, 1997, 2001).

Many of the above comments apply to mapping techniques for spatial data analysis. Maps provide a visual representation of change over space and are becoming indispensible for creating, evaluating, and disseminating population estimates, especially given the increasingly detailed spatial resolution of these estimates. The information age has introduced a new cartographic product that is changing the face of mapping: digital data for computerized mapping and analysis. Even more significant, mapping databases and GIS tools are accessible that do not require extensive expertise in geographic information systems and cartographic methods. For better or worse, the ability to make maps has been democratized. Additional information on the principles of map design and techniques for creating maps of geospatial data can be found in Krygier and Wood (2011) and Tyner (2010).[15]

4.4.1 Statistical Graphics

4.4.1.1 Univariate Data

The histogram is one of the most common ways of displaying univariate data. It is a two dimensional graph that shows the frequency on the vertical axis and the variable categories on the horizontal axis. Despite their widespread use, histograms have several disadvantages when used for continuous data (Jacoby 1997: 14): 1) minor changes in the bin definitions can greatly impact the visual display; 2) narrow bins may produce erratic looking graphs, while wide bins may distort/mask important distribution features; and 3) variability within a bin is masked. So the choice of bins can greatly impact substantive conclusions from a histogram. The smooth histogram overcomes some of these problems, by showing local densities within the distribution as a smooth continuous function (e.g., Silverman 1986; Tarter and Kronmal 1976). Line graphs are useful for presenting data over time. Joining up the points on a line graph gives an instant picture of past trends.

Figure 4.3 shows a histogram and line graph of household income classes in the highest and lowest income census tracts in San Diego County. The histogram more clearly identifies the income classes, but the advantage of using the line graph for showing trends is also evident. Lines will show changes in trends more clearly

Fig. 4.3 Household Income Distribution Histogram and Line Chart, Low and High Income Census Tracts, San Diego, County, 2005–2009
Source: US Census Bureau, 2005–2009 ACS

than bars, because the area of the bars detracts from the trend. Both figures show the much larger concentration of households with incomes of $200,000 or more in the high income census tract and conversely the much larger concentration of households with incomes under $30,000 in the low income census tract.

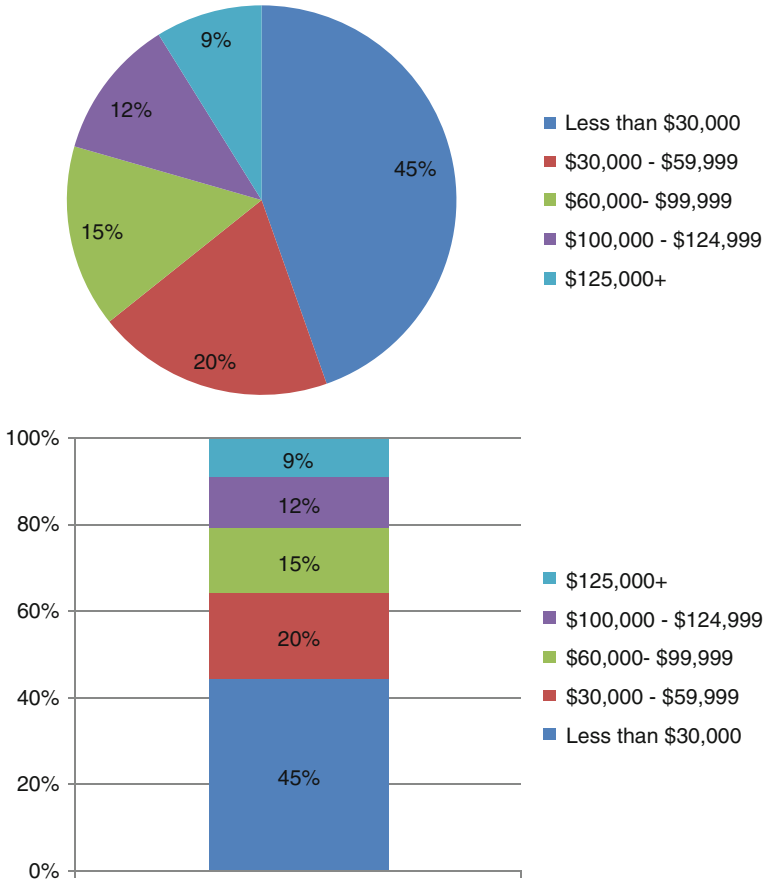


Fig. 4.4 Household Income Distribution Pie and Stacked-Bar Chart, Low Income Census Tract, San Diego, County, 2005–2009
Source: US Census Bureau, 2005–2009 ACS

Pie charts are used to show the relative sizes of subgroups that make up a whole. Pie charts are ubiquitous in business, journalism, and the public sector. However, this type of graph has been criticized because of the difficulties in comparing sections within a pie chart or data across different pie charts (e.g., Schmid 1983:65; Tufte 2001: 178). Pie charts can be an effective way of displaying information in some cases, in particular if the intent is to compare the size of a slice with the whole pie, rather than comparing the slices themselves (Spence 2005; Spence and Lewandowsky 1990). An alternative to the pie chart is the stacked bar chart, which shows the parts of the whole in a vertical representation. Figure 4.4 compares the pie and stacked-bar chart for the low income census tract, where the categories have been collapsed to improve readability. A less widely used but informative graphic is the ogive, which is a line chart of a cumulative

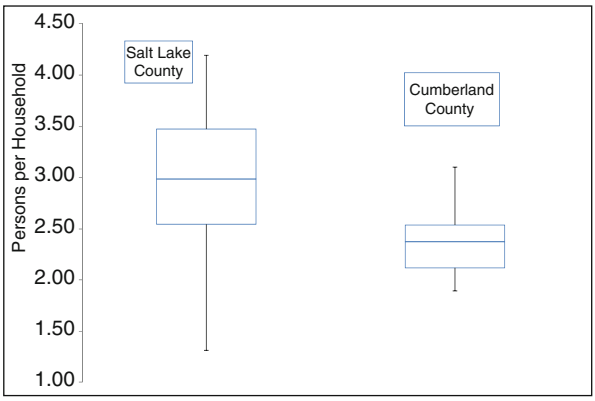
probability distribution. Such a distribution provides information on the percent of the observations above (or below) a certain value. For example, 80% of the households in the low income census tract have incomes less than $100,000, compared to 30% in the high income tract (ogive not shown).

Several graphical options exist for displaying a univariate data distribution. Univariate scatterplots show each observation plotted along a scale line representing the range of values. This graph displays all of the information without the potential loss or distortion of a histogram, but it limited to relatively small datasets (Jacoby 1997: 30). A quantile plot is a two dimensional graph that shows the data value on the vertical axis plotted against its quantile value. A quantile is the probability of an observation being less than or equal to certain position in a cumulative distribution. The shape of a quantile plot helps identify the shape of a distribution. For example, a symmetrical distribution will have an s-shaped quantile plot. Quantile plots show all of the data and can be used for datasets of virtually any size, since they represent the shape of a monotonic array rather than locations of individual plotting symbols (Jacoby 1997: 36). There are many variants of the univariate scatterplot theme (e.g., dot plots, symmetry plots, quantile-quantile plot) (e.g., Friendly 1991: Chapter 3; Mitchell 2008: Chapters 3 and 7).

Similar to a histogram, the stem-and-leaf plot is useful for visualizing the shape of the data, but unlike a histogram it retains the original data to two significant digits (Emerson and Hoaglin 1983). This display consists of a leaf, which is the last digit, and the rest of the number is the stem. It is sorted in ascending order by stem and ascending order within each leaf. With very small data sets a stem-and-leaf plot can be of little use, as a reasonable number of data points are required to establish definitive distribution properties. With very large data sets (greater than 300), the stem-and-leaf plot will become cluttered, since each data point must be represented numerically.

The box plot shows a five-number summary of a dataset and provides a visual impression of several important aspects of a data distribution: location, spread, skewness, tail length, and outliers (Emerson and Strenio 1983). The main component of a box plot is a box whose endpoints represent the middle of the distribution bounded by the 25^{th} and 75^{th} percentiles. A crossbar in the box shows the median, and the tails are represented by a line drawn from each end of the box to a remote point not considered an outlier. The distance to the remote point from the end of the box is 1.5 times the interquartile range. Outliers beyond the remote point are represented by asterisks. The relative position of the median in the box and the length and direction of the tails depict the distribution shape. For example, a median closer to the lower end of the box with a long upper tail indicates a right-skewed distribution.

Figure 4.5 shows box and stem-and-leaf plots for household size for census tracts for Salt Lake and Cumberland Counties. The differences between the shapes of the two distributions are evident from these graphs. Both distributions show some departures from a perfectly symmetrical distribution. The long lower tail for Salt Lake County census tracts is prominent.



Salt Lake County

Stem	Leafs
1	335678899999
2	000111122222222333444444455555555555566666666666777777777777888888888999999999999
3	000000011222222222222333333333333344444444445555555566666666667777777777788888888899999
4	00011111222

Cumberland County

Stem	Leafs
1	4469999
2	000000001111112222333344444444455555555556666777778 89
3	01

Fig. 4.5 Persons per Household By Census Tract Box and Stem-and-Leaf Plots, Salt Lake County and Cumberland County, 2005–2009
Source: US Census Bureau, 2005–2009 ACS

4.4.1.2 Bivariate and Multivariate Data

The basic visualization tool for bivariate data is the scatterplot that illustrates the joint distribution of the two variables in a rectangular coordinate system. The value of one variable is shown on the X-axis and the corresponding value of the other variable is shown on the Y-axis. Scatterplots help identify strength and direction of a bivariate relationship and outlying observations. One of the most powerful features of a scatterplot is its ability to show non-linear relationships. The basic scatterplot can be enhanced to provide more information about the underlying pattern of the data, the univariate distributions of the variables, and the individual observations in the data set (Jacoby 1997: 54-63). Enhancements include: 1) adding marginal box plots; 2) labeling points; and 3) slicing the data to examine the conditional relationship within each slice. Another strategy for illuminating the bivariate relationship is to fit a smooth curve to points on the scatterplot to summarize changes in the central tendency of one variable at different locations of the other variable. Nonparametric smoothing techniques are used to fit a curve without having to know in advance the functional specification between the two variables (Goodall 1990). The scatterplot in Figure 4.6 shows a moderately strong relationship between the automobile registration and population variables used in the ratio-correlation model discussed above.

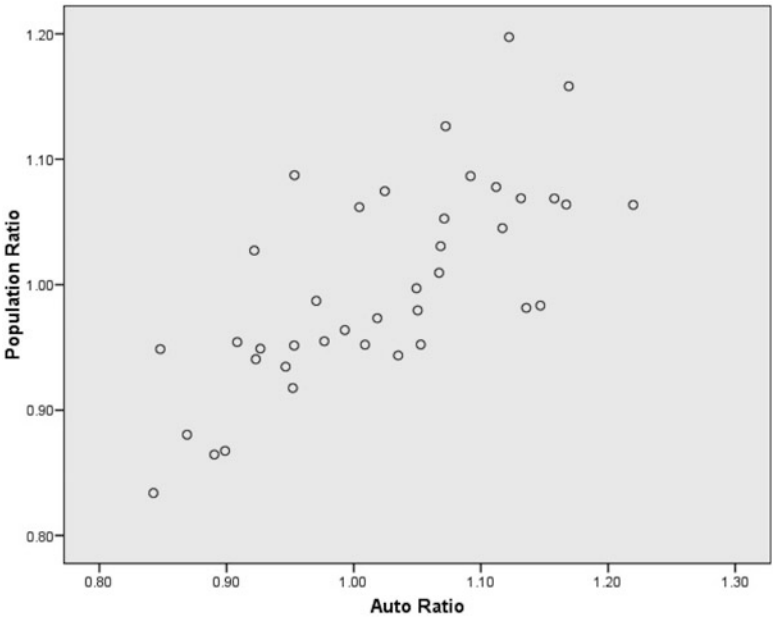


Fig. 4.6 Scatter Diagram of the Relationship between Automobile Registration and Population, Ratio-Correlation Model

Multivariate data pose special challenges for statistical graphics. Information can vary along several dimensions (variables) in a display medium that is inherently two-dimensional in nature. All multivariate graphics require changing or expanding the familiar visual metaphors we use for two variables, and a wide variety of methods have been developed (e.g., Chambers, Cleveland, Kleiner, and Tukey 1983; Friendly 1991: Chapters 8 and 9); Gnanadesikan 1997; Jacoby 1998; Monette 1990). We discuss some of the more popular approaches.

Multiple code plotting uses alternative symbols for representing the different dimensions. For qualitative variables, the standard bivariate plot will show the relationship for each category using different symbols. For multiple quantitative variables, a glyph plot can show all variables simultaneously. Three dimensional (3-D) scatterplots can be used to examine the relationship between a dependent variable and two independent variables simultaneously. There are problems in discerning any more than general patterns and accurate visual estimates from 3-D scatterplots, but incorporating interactive, dynamic technology display increases their utility (Sung, Shirley, and Baer 2008).

Multivariate data can be shown as a series of bivariate scatterplots (e.g., Jacoby 1998: 39). The scatterplot matrix facilitates a comprehensive understanding of multiple bivariate scatterplots by portraying a square, symmetrical table with k rows and columns. Each intersection of row i and column j contains a scatterplot showing variable X_i as the horizontal axis and variable X_j as the vertical axis.

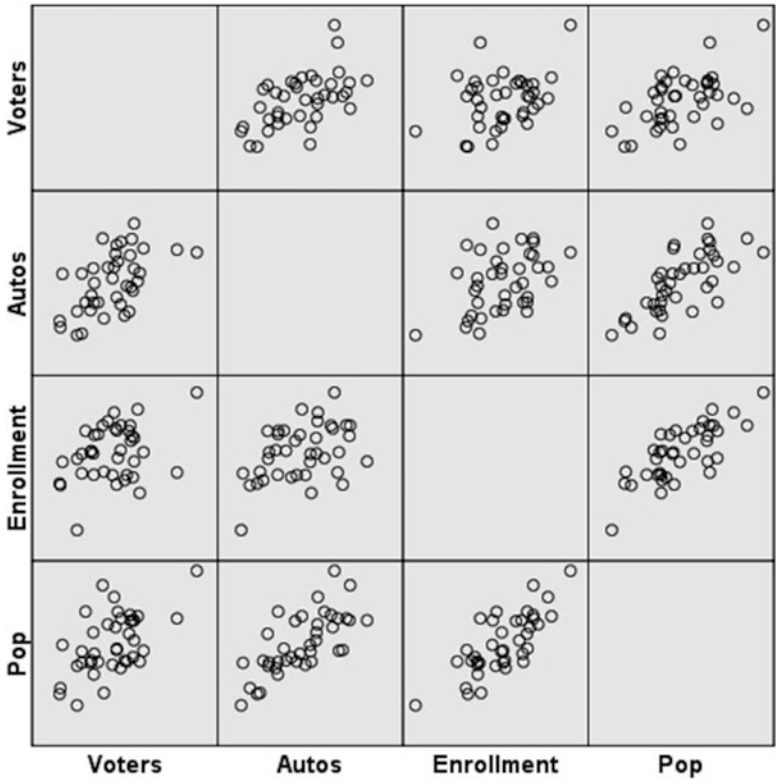


Fig. 4.7 Scatterplot Matrix of the Variables, Ratio-Correlation Model

Because the matrix is symmetric, the same variables appear in panel *ji* with the horizontal and vertical positions of the variables reversed. One drawback of the scatterplot matrix is that it does not show multivariate structure because each scatterplot within a panel is constructed independent of information in other panels. Figure 4.7 shows the scatterplot matrix of all variables used in the ratio-correlation model. That the population is more closely associated with school enrollment and automobile registration is clearly apparent in this matrix. Potential outliers are also apparent, most notably in the relationship between automobile and voter registrations.

Multivariate structure is the understanding of non-random patterns of several variables simultaneously. Conditioning plots show how a variable is affected by several other variables, which addresses the problem with the scatterplot matrix. A conditioning plot shows the bivariate relationship holding constant or conditioned on the values of other variables. It is a multi-panel display that shows the bivariate relationship within slices of the conditioning variable. These slices follow the principle of small multiples (Tufte 2001: 170–175).

Finally, the bi-plot allows the joint relationships between the observations and variables of a data matrix to be displayed graphically. Observations are displayed as

points while variables are displayed either as vectors, linear axes, or nonlinear trajectories in a common space (Greenacre 2010). Bi-plots are easy to interpret. Correlations among the variables are represented by the angles of the vectors. The observations are represented by points and their distances in the two-dimensional space determine the similarity between their profiles. A profile for observation *i* is its array of scores on all k variables.

4.4.2 Maps

Combining science, aesthetics, and technique cartography builds on the premise that reality can be modeled in ways that communicate spatial information effectively. In the 21st century it is possible to find a map of virtually anything. However, finding or making a meaningful map is the objective, and a well-designed map "is convincing because it implies authenticity (MacEachren 1994: 9). A good map provides a compromise between portraying the items of interest in the right place for the map scale used, against the need to annotate that item with text or a symbol, which takes up space on the map medium and may cause some other item of interest to be displaced. In this last section of the chapter, we provide examples of cartographic mapping techniques. These examples are not meant to be comprehensive or even representative of the huge variety of different styles and types of maps, but are intended to provide a glimpse into the efficacy of this display medium.

Figure 4.8 presents a map of the inventory of real property within the city of Middletown, Connecticut compiled from tax maps, recorded deeds, and plats. Maps showing geographic boundaries and their attributes are useful for validating model inputs and evaluating the estimates themselves. Maps like this provide a convenient and easy to understand platform for obtaining outside review and greatly enhance the transparency of modeling data. They are even more productive when made available digitally.

Figure 4.9 shows the widely used thematic map with shading patterns. The dramatic change in population pre- and post-Katrina can easily be understood. The population estimates, built from the ZIP + 4 level and mapped at the census tract level, reveal where Louisiana's population has shifted as a result of the hurricane's impact. Seventy percent of the census tracts in the state saw an increase in population totaling 11,746 people, but 30 percent of the census tracts experienced a population decrease of 460,190, resulting in Louisiana's state population decreasing by an estimated 448,444.

Figure 4.10 is an example of corridor mapping that also includes pie charts to indicate the share of daily trips by trip purpose (home to work, school, or college; home to other places; and trips not originating at home). It shows daily trip-end data for five selected highway corridors in the San Diego region in the year 2030. Each selected corridor is surrounded by a one-mile buffer area. This map helps transportation planners evaluate future travel demands and travel model results and allows decision makers to see the travel impacts of land use and transportation policies.

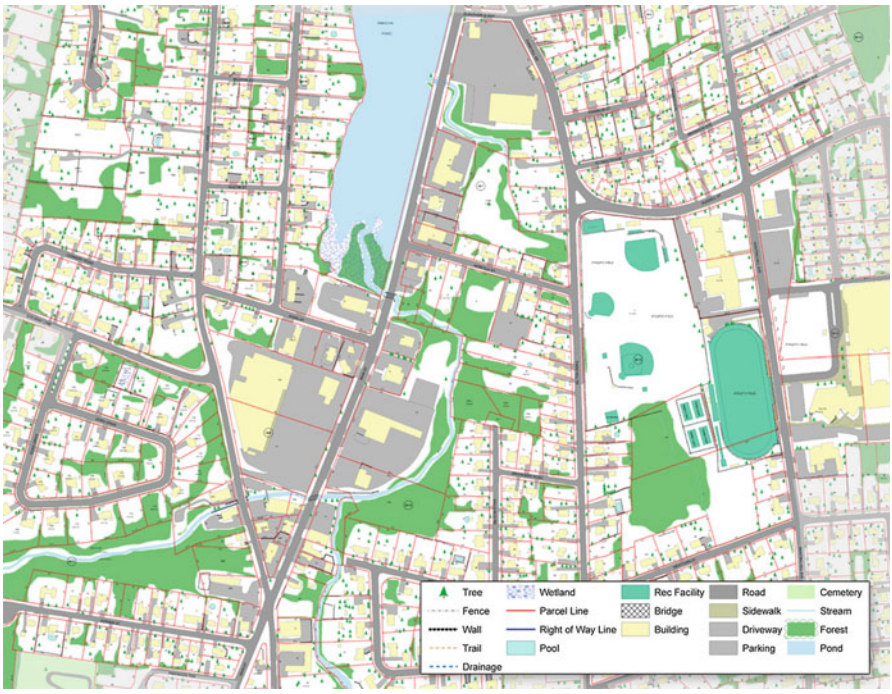


Fig. 4.8 City of Middletown, Connecticut, Tax Map 28
Source: Heidi Krueger, Applied Geographics, Inc., and Frank Marchese, City of Middletown, ESRI Map Book No. 21 http://www.esri.com/mapmuseum/mapbook_gallery/volume21/planning7.html

Figure 4.11 illustrates the use of circle symbology to depict the future number of dwelling units located throughout Guam. The US will relocate thousands of military personnel from Okinawa, Japan to the island of Guam. This relocation will have a major impact on the island's infrastructure. Housing, schools, and services will have to be built. New roads and new utilities will be needed, and existing roads will have to be upgraded. This map is part of a series created to show the distribution of different demographic variables depending on various development scenarios on the island of Guam. This type of map gives planners a good way to visualize the development alternatives and will help them manage the growth of Guam upon the relocation of several US military bases.

Figure 4.12 shows the distribution of economic activity across the continental US using gross domestic product (GDP) per day as the measure. The scale of the economic activity is represented by the height from a 3D surface model. GDP is measured using employee data from Dun & Bradstreet and combining it with GDP by industry data from the Bureau of Economic Analysis. This map is used to respond quickly to requests from federal and state agencies for economic impact analyses related to hazardous events. It also provides a unique perspective on economic activity that moves beyond tabular representations of economic data.

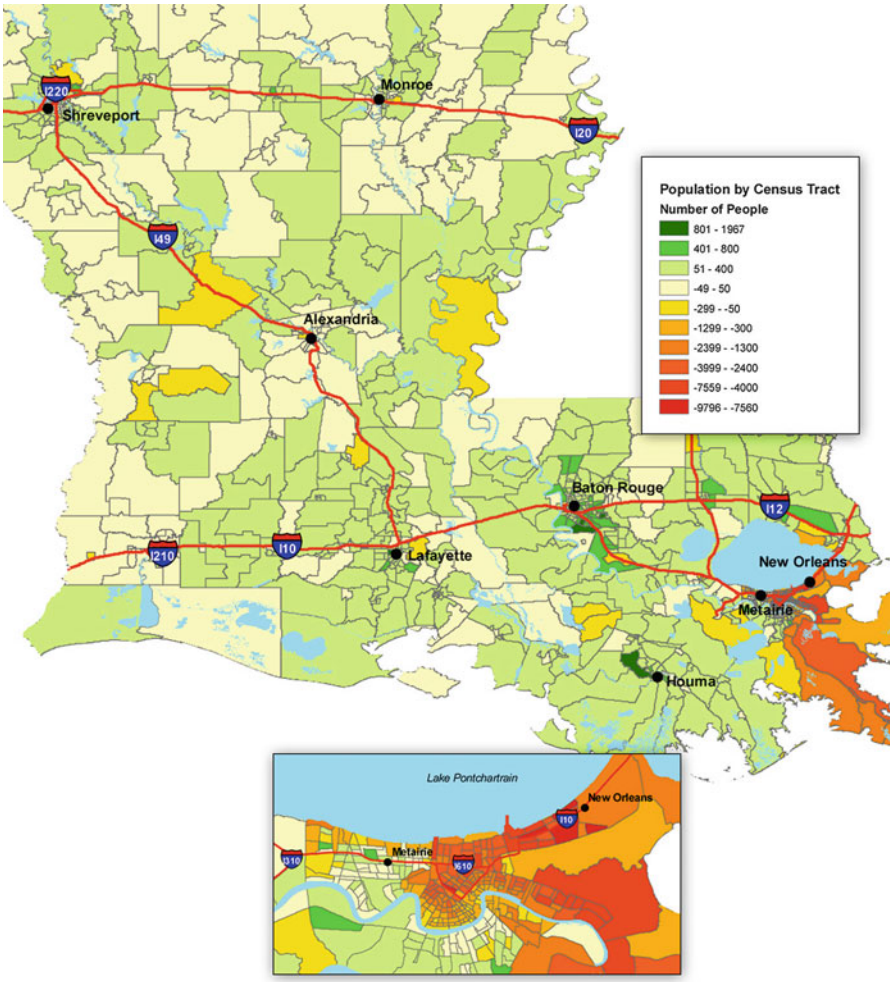


Fig. 4.9 Population Displacement—The Impact of Hurricane Katrina on the State of Louisiana Source: Mapping Analytics, LLC, Gene Rinas, ESRI Map Book No. 21 http://www.esri.com/mapmuseum/mapbook_gallery/volume21/business2.html

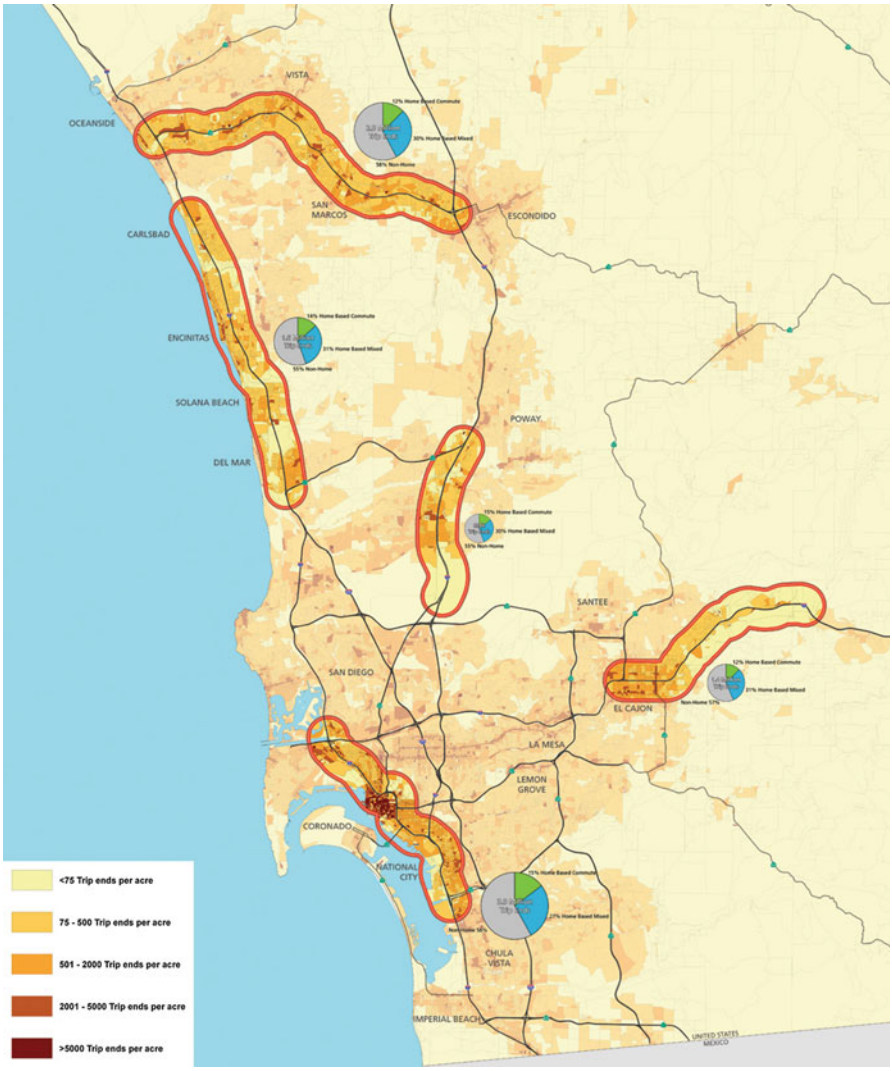


Fig. 4.10 San Diego Region—Trip Proximity Analysis along Selected Highway Corridors Source: Joaquin S. Ortega, San Diego Association of Governments, ESRI Map Book No. 23 http://www.esri.com/mapmuseum/mapbook_gallery/volume23/transportation13.html

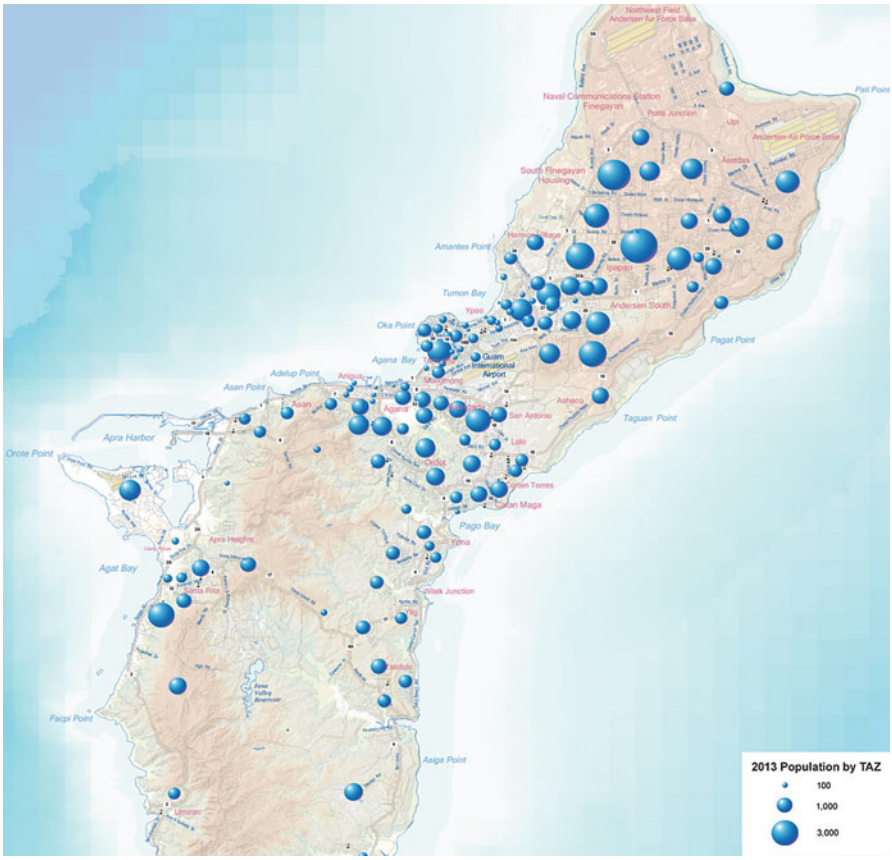


Fig. 4.11 Island of Guam Demographics
Source: Parsons Corporation, ESRI Map Book No. 24 http://www.esri.com/mapmuseum/mapbook_gallery/volume24/sustainable3.html

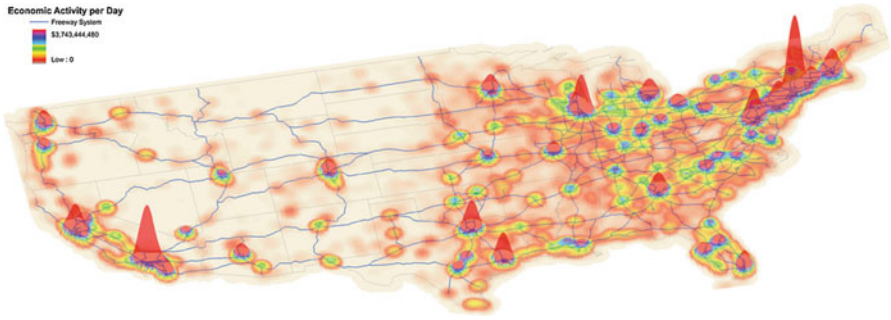


Fig. 4.12 US Centers of Economic Activity (Continental United States)
Source: Los Alamos National Laboratory, ESRI, Map Book No. 24 http://www.esri.com/mapmuseum/mapbook_gallery/volume24/business2.html

Endnotes

1. If r is expressed as a percentage average annual growth rate the 100 in this formula is not needed.
2. Demographic measures typically refer to a calendar year. Sometimes a three-year average of demographic events is used to smooth out the effects of annual fluctuations.
3. Life tables can be classified as complete or abridged. Complete life tables provide data by single year of age; abridged life tables provide data by age group (usually five-year groups, with the youngest group subdivided at age one). There are two types of life tables. A period life table is based on the ASDRs calculated for a particular period of time (usually one, two, or three years). A cohort life table is based on the mortality patterns actually experienced by members of a particular birth cohort (e.g., all persons born in 1910) over their lifetimes.
4. Calculating survival rates for one-year age groups requires an unabridged life table, but the approach is the same. For example, a five-year survival rate for a one-year age group can be calculated as: $S_x = L_{x+5}/L_x$
5. The migration rates must be applied in a manner consistent with the way they were computed; for example, if rates were based on the unadjusted population at the beginning of the migration interval, they must be applied to the unadjusted population at the beginning of the estimation interval.
6. The reverse survival rate method can also be used to estimate net migration by age. Here the survival rate is divided into the age group at the end of the period and compared to the appropriate beginning-period age group. The two methods yield identical results and in practice the forward survival rate is used (Siegel 2002: 22).
7. While travel time and distances are usually used in the IOP, one could set the predetermined limits based on the travel cost of reaching an area.
8. Migration flows between places i and j, have also been modeled using a gravity function. In most gravity models, migration is directly proportional to size of the origin and destination areas and inversely proportional to the intervening distance between them (Rodrigue, Comtois, and Brian 2009: 216; Tarver and McLeod 1973; Zipf 1946).
9. A variant of the arithmetic mean is the weighted average. In an arithmetic mean each of the data points contributes equally to the final average, while in a weighted average some data points contribute more than others.
10. A property of the mean is that the sum of the deviations from each observation and the mean will equal zero. One way to handle this is to take the absolute value of the differences before summing, which is what the mean deviation does. The variance on the other hand squares the differences. The mean deviation is adequate for purely descriptive purposes, but it is not useful for statistical inference and is rarely used (Blalock 1972: 80).
11. Some computer programs (e.g., excel and SPSS) subtract 3 from the kurtosis formula. In this case, 0.0 indicates a mesokurtic distribution; a negative value a flatter distribution; and a positive value a more peaked distribution. Stata uses the formula in the text.
12. $\bar{p}$ is the estimated proportion and $\bar{q}$ is its compliment.
13. Intervals can also be computed that do not require any specific assumption about the shape of the probability distribution (e.g., Hahn and Meeker 1991: Chapter 5).
14. Associated with every statistical test are model and measurement requirements, and the test is valid only if these requirements hold. Parametric tests, as described in the text, make the most stringent model and measurement assumptions. If those stringent assumptions are correct, parametric methods have the most statistical power. However, if those assumptions are incorrect, parametric methods can be very misleading. For that reason they are not considered robust. Most parametric hypothesis tests have non-parametric equivalents that require fewer and less restrictive assumptions (e.g., Gibbons and Chakraborti 2011; Siegel 1956). Nonparametric tests are often referred to as distribution-free tests.
15. ESRI publishes map books that contain a wide range of maps for many different uses in the private and public sector. These books are available online at http://www.esri.com/mapmuseum/index.html.

References

Andresen, M. A. (2007). Location quotients, ambient populations, and the spatial analysis of crime in Vancouver, Canada. *Environment and Planning A, 39*, 2423–2444.

Barber, G. M. (1988). *Elementary statistics for geographers*. New York: The Guilford Press.

Belsley, D., A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Blalock, H. M. (1972). *Social statistics, Second Edition*. New York: McGraw Hill.

Bogue, D. J., Hinze, K. E., & White, M. J. (1982). *Techniques for estimating net migration*. Chicago: University of Chicago Press.

Nielsen Solution Center. (2010). Pop-Facts: Demographic snapshot report. (http://www.claritas.com/samples/sitereports/demographic_snapshot_10.pdf).

Chambers, J. M., Cleveland, W. S., Tukey, P. A., & Kleiner, B. (1983). *Graphical methods for data analysis*. Pacific Grove: Wadsworth and Brooks/Cole.

Chatterjee, S., & Hadi, A. S. (1998). *Sensitivity analysis in linear regression*. New York: John Wiley.

D'Agostino, R. B., Belanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(3), 316–321.

Dharmalingam, A. (2004). Reproductivity. In J. S. Siegel, & D. A. Swanson (Eds.), *The methods and materials of demography* (pp. 429–454). New York: Elsevier Academic Press.

Draper, N., & Smith, H. (1981). *Applied regression analysis, second edition*. New York: John Wiley & Sons.

Duncan, O. D., Cuzzort, R. P., & Duncan, B. (1961). *Statistical geography: Problems in analyzing areal data*. Glencoe: Free Press.

El-Geneidy, A. M., & Levinson, D. M. (2006). Access to destinations: Development of accessibility measures. St. Paul: Minnesota Department of Transportation.

Emerson, J. D., & Hoaglin, D. C. (1983). Stem-and-Leaf Displays. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 1–32). New York: John Wiley & Sons.

Emerson, J. D., & Strenio, J. (1983). Boxplots and Batch Comparisons. In D. C. Hoaglin, F. Mosteller & J. Tukey, W. (Eds.), *Understanding robust and exploratory data analysis* (pp. 58–96). New York: John Wiley & Sons.

Estee, S. (2004). Natality-Measures based on vital statistics. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 371–406). New York: Elsevier.

Fonseca, L., & Tayman, J. (1989). Post-censal estimates of household income distributions. *Demography, 26*(1), 149–159.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics 4th edition*. New York: W. W. Norton and Company.

Friendly, M. (1991). *SAS systems for statistical graphics*. Cary: SAS Institute.

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference, Fifth Edition*. Boca Raton: Chapman & Hall/CRC.

Gnanadesikan, R. (1997). *Methods for statistical analysis of multivariate observations, 2nd Edition*. New York: John Wiley & Sons.

Goodall, C. (1990). A survey of smoothing techniques. In J. Fox & J. S. Long (Eds.), *Modern Methods in Data Analysis* (pp. 126–176). Newbury Park: Sage Publications.

Greenacre, M. (2010). *Bi-plots in practice*. Bilbao: BBVA Foundation.

Guers, K. T., & van Wee, B. (2004). Accessibility evaluation of land use and transport strategies: Review and research directions. *Journal of Transport Geography, 12*(2), 127–140.

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners*. New York: John Wiley & Sons.

Handy, S. L., & Niemeier, D. A. (1997). Measuring accessibility: An exploration of issues and alternatives. *Environment and Planning A, 29*(7), 1175–1194.

Hansen, W. G. (1959). How accessibility affects land use. *Journal of the American Institute of planners, 25*, 72–77.
Haynes, K. E., & Fotheringham, A. S. (1984). *Gravity and spatial interaction models*. Beverly Hills: Sage Publications.
Irwin, R. B. (1977). Guide for local area population projections. Washington, DC: US Bureau of the Census.
Jacoby, W. G. (1997). *Statistical graphics for bivariate and univariate data*. Thousand Oaks: Sage Publications.
Jacoby, W. G. (1998). *Statistical graphics for visualization of multivariate data*. Thousand Oaks: Sage Publications.
Kennedy, P. (1992). *A guide to econometrics*. Oxford: Blackwell.
Keyfitz, N. (1977). *Applied mathematical demography*. New York: John Wiley & Sons.
Kintner, H. J. (2004). The life table. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 301–340). New York: Elsevier Academic Press.
Koenig, J. G. (1980). Indicators of urban accessibility: Theory and application. *Transportation, 9*, 145–172.
Krygier, J., & Wood, D. (2011). *Making maps: A visual guide to map design for GIS, Second Edition*. New York: The Guilford Press.
Leigh, R. (1970). The use of location quotients in urban base studies. *Land Economics, 46*(2), 202–205.
Levin, R. I., & Rubin, D. S. (1998). *Statistics for management, Seventh Edition*. Englewood Cliffs; Prentice Hall.
Long, L. H. (1988). *Migration and residential mobility in the United States*. New York: Russell Sage Foundation.
Lowry, I. S. (1964). *A model of metropolis*. Santa Monica: The Rand Corporation.
MacEachren, A. M. (1994). *How maps work*. New York: The Guilford Press.
Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics, 12*, 591–612.
Massey, D. S., & Denton, N. A. (1998). The elusive quest for the perfect index of concentration: Reply to Egan, Anderton, and Weber. *Social Forces, 76*(3), 1123.
Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces, 67*, 281–315.
McGehee, M. (2004). Mortality. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 265–300). New York: Elsevier Academic Press.
McKibben, J. N., & Faust, K. A. (2004). Population distribution: Classification of residence In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 105–123). New York: Elsevier Academic Press.
Meuser, P. R., & White, M. J. (1989). Explaining the association between rates of in-migration and out-migration. *Papers of the Regional Science Association, 67*, 121–134.
Mitchell, M. N. (2008). *A visual guide to Stata graphics, 2nd Edition*. College Station: Stata Press.
Monette, G. (1990). Geometry of multiple regression and interactive 3-D graphics. In J. Fox & J. S. Long (Eds.), *Modern Methods in Data Analysis* (pp. 209–256). Newbury Park: Sage.
Morrison, P. A., Bryan, T. M., & Swanson, D. A. (2004). Internal migration and short-distance mobility. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 493–522). New York: Elsevier Academic Press.
Namboodiri, K., & Suchindran, C. M. (1987). *Life table techniques and their applications*. Orlando: Academic Press.
Neter, J., Wasserman, W., & Kutner, M. H. (1989). *Applied linear regression models*. Homewood: Irwin.
O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* (41), 673–690.

Pittenger, D. B. (1976). *Projecting state and local populations*. Cambridge, MA: Ballinger Publishing Company.
Plane, D. A. (2004).Population distribution–Geographic In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 105–120). New York: Elsevier Academic Press.
Pullum, T. J. (2004). Natality- Measures based on censuses and surveys. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 407–428). New York: Elsevier Academic Press.
Putman, S. H. (1983). *Integrated urban models*. London: Piton Limited.
Putman, S. H. (1991). *Integrated urban models: II*. London: Pion Limited.
Reidpath, D. D., & Allotey, P. A. (2003). Infant mortality rate as an indicator of population health. *Journal of Epidemiology and Community Health, 57*(5), 344–346.
Rodrigue, J. P., Comtois, C., & Brian, S. (2009). *The geography of transport systems*. New York: Routledge.
Schmid, C. F. (1983). *Statistical graphics: Design principles and practices*. New York: John Wiley & Sons.
Siegel, J. S. (2002). *Applied demography: Applications in business, government, law, and public policy*. San Diego: Academic Press.
Siegel, J. S., & Swanson, D. A. (Eds.). (2004). *The Methods and Materials of Demography*, *Second Edition*. New York: Elsevier Academic Press.
Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw Hill.
Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.
Smith, D. P. (1992). *Formal Demography*. New York: Plenum Press.
Smith, S. K., & Swanson, D. A. (1998). In defense of the net migrant. *Journal of Economic and Social Measurement, 24*(249–264).
Smith, S. K., Tayman, J. & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic / Plenum Publishers.
Spence, I. (2005). No humble pie: The origins and uses of a statistical chart. *Journal of Educational and Behavioral Statistics, 30*, 353–368.
Spence, I., & Lewandowsky, S. (1990). Graphical perception. In J. Fox & J. S. Long (Eds.), *Modern Methods of Data Analysis* (pp. 13–57) Newbury Park: Sage.
Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics*. Boston: Addison Wesley.
Sung, K., Shirley, P., & Baer, S. (2008). *Essentials of interactive computer graphics*. Wellesley: A.K. Peters, Ltd.
Tarter, M. E., & Kronmal, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician, 30*, 105–112.
Tarver, J. D., & McLeod, R. D. (1973). A test and modification of Zipf's hypothesis for predicting interstate migration. *Demography, 10*(2), 259–275.
Tufte, E. R. (1990). *Envisioning information*. Cheshire: Graphics Press.
Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire: Graphics Press.
Tufte, E. R. (2001). *The visual display of quantitative information, Second Edition*. Cheshire: Graphics Press.
Tyner, J., A. (2010). *Principles of map design*. New York: The Guilford Press.
Zipf, G. K. (1946). The P1P2/D hypothesis: On the intercity movement of persons. *American Sociological Review, II*, 677–686.

Chapter 5
Overview of Estimation Methods

Our purpose in this Chapter is to provide a general roadmap of subnational population estimation methods, which are covered in detail in subsequent Chapters. We set the stage for this overview by differentiating between pre-censal, inter-censal, and post-censal estimates and classifying population estimation methods. Finally we describe the variety of methods currently used to estimate population. This focus of this Chapter is on population estimates based on usual residence or de jure. We cover de facto (physically present) population estimates in Chapter 16.

5.1 Classification of Estimates and Methods

5.1.1 Pre-censal, Inter-censal, and Post-censal Estimates

Estimates are commonly divided into two types on the basis of their time reference and derivation (Bryan 2004: 523; Raymondo 1992: 99,123). These two types, which employ different methodologies, are: (1) inter-censal estimates, which relate to a date between two censuses and take the results of these censuses into account; and (2) post-censal estimates, which relate to a date following the latest census, but prior to a subsequent census. Post-censal estimates take the last census and possibly earlier censuses into account. Post-censal estimates can be generally viewed as extrapolations, and inter-censal estimates as interpolations. Though extrapolation techniques may be used in post-censal estimates, post-censal estimates are commonly made with symptomatic data, or data related to changes in population. There are also pre-censal estimates for dates prior to the advent of census taking (e.g., before 1790 in the US, before 1871 in Canada, and before 1801 in the United Kingdom). Pre-censal estimates are the province of historical demography presented in Chapter 17.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_5,

5.1.2 *Classification Schemes*

Murdock and Ellis (1991: 181) identified four broad categories of techniques used to estimate population: (1) Extrapolative (e.g. linear trend; shift-share); (2) Symptomatic (e.g., censal ratio; housing unit method); (3) Regression-based (e.g., ratio correlation); and 4) Component (e.g., cohort survival, component methods). The Symptomatic and Regression-based procedures both use symptomatic data (as defined in Chapter 3), but the form of the variables and statistical procedures are quite distinct between them. While this scheme covers the major techniques for estimating population, it excludes sample-based techniques (See Chapter 11) and other techniques such as dual system estimators (Chapter 12).

A more general schema categorizes estimation methods into two types: (1) "flow" and (2) "stock" (Long 1993). In general, stocks typically have a certain value at a point in time (e.g., the population size in 2011), while a flow (or "rate") changes a stock over time; stock and flows are the basic building blocks for systems dynamics models (e.g., Forrester 1958). The component technique, for example, is a flow method because it estimates each component of population change since the last census. The censal ratio method, for example, is a stock method that estimates population based on its relationship to other variables such as school enrollment, employment, automobile registrations, total number of deaths (and births), and tax returns.

Another schema places estimation methods into three categories: (1) analytical and statistical models that use data symptomatic of population and its changes; (2) mathematical models that use historical census data; and (3) sample based (Judson and Swanson 2011: 13-14). Methods falling into the first category have generally been developed by and for applied demographers, most of who work for national, state, and local governments. Methods falling into the second category have generally developed by and for academic demographers, most of who work at universities and research institutes. The methods falling into the third category have generally been developed by and for statisticians and survey research scientists, but they also are widely used by demographers. There also are techniques that combine methods from two or even all three categories.

Table 5.1 shows a classification scheme for population estimation methods that combines elements from the three approaches discussed above. It adds sample based and other methods to the Murdock and Ellis scheme, along with a stock and flow dimension. Some broad categories are reported in finer detail to reflect where the methods differ by stock and flow. The Hamilton-Perry method is not a strictly component method, but we place it under this rubric because it is a shorthand or simplified version of the cohort-component model. This scheme identifies a method as flow if it is predominately based on the change formulation, and the identification is based on the most frequent application of a method. For example, the censal ratio method is almost always based on the relationship of a symptomatic indicator to the population at the last census, but this method could be based on the change in population relative to the change in a symptomatic indicator between the last two censuses.

Table 5.1 Classification of Estimation Methods

Estimation Method	Stock	Flow
Extrapolation		
Simple	x	
Complex	x	
Ratio		
Constant Share	x	
Shift Share	x	
Share-of-Growth		x
Symptomatic		
Housing Unit	x	x
Censal ratio	x	
Regression		
Ratio Correlation	x	
Difference Correlation		x
Rate Correlation	x	
Lagged Correlation	x	
Component		
Component Method II		x
Cohort-Component		x
Hamilton-Perry	x	
Composite	x	x
Sample Based	x	
Other	x	

Most estimation methods are stock-based or have a stock-based component whose formulation has a point in time focus (13 of the 17 methods). The component method II, cohort-component, share-of-growth, and difference-correlation methods are flow based. The housing unit and composite methods use both stocks and flows. In the housing unit method the flow component is related to estimating changes in housing units and the stock component is related to determining the occupancy and/or household size characteristics at a point in time. In the composite method the flow approach is often applied to estimate the population under 65 years of age and the stock approach is often applied to estimate the population 65 years and older.

5.2 Estimation Methods

5.2.1 Extrapolation

Extrapolation techniques rely solely on the pattern of past population changes to estimate the post-censal population. The method assumes that trends in the post-censal period will be similar to past trends. Extrapolation methods generally are simple to implement and require limited data; for example, the constant share method requires data only for a single point in time. These techniques are most likely to be used for post-censal periods relatively close to the last census, for completing estimates when resources are limited, or for estimating small areas and demographic subgroups (e.g., race).

Following Smith, Tayman, and Swanson (2001: 161-162), we examine three general categories of extrapolation methods: simple, complex, and ratio. Simple methods have simple mathematical structures and require data for only two points in time. We cover methods that assume linear, exponential, and geometric patterns of population change. Complex methods have more complex mathematical structures, require data from a number of points in time, and require statistical methods to estimate model parameters. We cover five complex methods: linear trend models, exponential trend models, polynomial curve fitting, logistic curve fitting, and ARIMA models. Ratio methods are those where a population of a subgroup is expressed as a proportion of a larger population (e.g., city population as a share of the county population; Asians as a share of the total population). Three ratio methods are discussed: constant share; shift share, and share-of-growth.

5.2.2 Housing Unit

The Housing Unit method is one of the most widely used techniques for subnational population estimates (Bryan 2004b: 550; Jarosz 2008; Smith and Cody 2004). This method can be applied at virtually any level of geography, can accommodate a variety of data sources and application techniques, and can produce estimates that are at least as accurate as other post-censal estimation techniques (Lowe, Myers, and Weisser 1984; Smith 1986; Smith & Cody 2004). Since 1996 the US Census Bureau has relied exclusively on the housing unit method for subcounty population estimates (US Census Bureau 1998).

The housing unit method is based on the fact that almost everyone lives in some type of housing structure, whether a single family unit, an apartment, a mobile home, a college dormitory, or a state prison. The population can therefore be estimated as the number of occupied housing units (households) times the average number of persons per household (PPH), plus the number of persons living in group quarters. Occupied units can be estimated directly or derived from an estimate of housing units (total and occupied) by applying a vacancy rate factor. The efficacy of the housing unit method depends on accurate data reflecting the change in housing and accurate information on the vacancy rate, persons per household, and group quarters population at the post-censal estimation date. A variety of data sources and estimation techniques are used to estimate these factors (Smith and Cody 2004).

5.2.3 Regression

Regression techniques employ the statistical procedures of simple or multiple regression where symptomatic data are the independent variables and the population is the dependent variable. This method assumes that the statistical relationship between symptomatic data and the corresponding population remains unchanged

over time (e.g., Mandell and Tayman 1982; Tayman and Schafer 1985). The types of symptomatic data that have been used in regression models are births, deaths, school enrollment, tax returns, motor vehicle registrations, employment, voter registration, and sales taxes. Regression methods are most often applied to estimate the population of counties within a state, but if symptomatic data are available, they could be used in any nested geographic system.

The most common regression-based approach to estimating population is the ratio-correlation method (Schmitt and Crosetti 1954). A multiple regression equation is derived to express the relationship between (1) the ratio at two census points of an area's share of the total for the larger area for several symptomatic series and (2) the ratio at two census points of an area's share of the population of the larger area. Modifications to the ratio variable construction include taking the difference between the shares over the inter-censal period (Schmitt and Grier 1966; O'Hare 1976; Swanson 1978), and using the ratio of the natural logarithm of the shares (Swanson and Tedrow 1984). Modifications to the basic regression model have included creating separate equations for different subgroups of counties (stratification) and using dummy variables to represent demographic and socio-economic features of counties (Martin and Serow 1978; Purcell 1970; Rosenberg 1968). Another alternative that has been proposed is the use of the simple, unweighted average of the estimates from simple regressions instead of estimates from a single multiple regression equation (Namboodiri and Lalu 1971).

5.2.4 Censal Ratio

The censal ratio method is among the earliest approaches for estimating post-censal estimates and its beginning is tied to the publication of Bogue's (1950) vital rates method. The vital rates method derives local population estimates from estimates of post-censal local birth and death rates and post-censal values for births and deaths. In the early 1970s the more encompassing censal ratio term came into common use as the method was expanded to include symptomatic indicators other than births and deaths (Voss, Palit, Kale, and Krebs 1995). The basic approach of a censal ratio method is to establish the ratio between a symptomatic indicator and the population at the time of the last census; update the ratio to the post-censal time point; and derive the estimate from the updated ratio and the value of the post-censal symptomatic indicator. Censal ratios are often updated using a synthetic approach that ties changes in the local ratio to changes in the ratio for a larger area (see Chapter 11).

5.2.5 Component

Component methods generally use estimates of births, deaths, and migration for the post-censal period to derive a population estimate. This approach to estimating population is attractive because it can provide a more complete explanation of

the reasons behind the population change than other estimation techniques. Component methods that estimate births, deaths, and migration are most often applied at the county, state, and national levels, but can be applied at any level providing the appropriate data is available (Murdock, Hwang, and Hamm 1995). Births and deaths are available for the post-censal period through the national vital statistics system; although there may be a lag between the latest information and the post-censal time point. The information on post-censal migration is meager at best and various approaches have been designed to estimate this component of population change (e.g., Murdock, Hwang, and Hamm 1995; Bryan 2004b: 540-544).

The component method can be applied to total population and non-age related demographic characteristics by simply adding births, subtracting deaths, and either adding or subtracting the change due to migration from the latest census population. The cohort-component method is used when estimates are needed by age and for the components of change (Smith, Tayman, and Swanson 2001). The simplest cohort-component framework breaks the population into age and sex groups, but this method can handle further disaggregation by other demographic characteristics. Estimates of births, deaths, and migration by age are based on age-sex specific survival rates from a life table and migration rates, which can be net or gross in- and out-migration rates. Age-specific fertility rates are used to generate births. For post-censal estimates, the age pattern of births, deaths, and migration is often adjusted to independently derived control totals.

Although not strictly a component technique, the Hamilton-Perry method offers an alternative approach for estimating population by age that requires only population data by age from two time points (Hamilton and Perry 1962). As such, it can be applied quickly and easily and is particularly suited from subcounty areas that usually lack the necessary data for the cohort-component method (Smith and Tayman 2003; Swanson, Schlottmann, and Schmidt 2010: Chapter 3). The Hamilton-Perry method is based on cohort-change ratios that combine the effects of mortality and migration and uses a child-woman ratio to estimate the youngest age group.

The composite method, developed by Bogue and Duncan (1959), is a portfolio of separate methods each tailored to particular segments of the population. The results of this portfolio are put together to estimate the total population. While not strictly a component method, the composite method can incorporate procedures for estimating the components of change. Many alternative portfolios of methods are possible, such as using the component method for estimating the population under 65 years of age and censal ratio method using Medicare data to estimate the population 65 years and older.

5.2.6 Sample Based

Under the sample based rubric we cover sample based, synthetic, and SPREE methods. These methods are interconnected to each other, but also connect to

other estimation methods. Sample based methods often rely on estimates of the population and their characteristics and their predominant use is designing, analyzing, and adjusting samples. Chapters 2 and 4 provide an overview of the design and implementation of sample surveys and the statistical tools for analyzing their results.

Synthetic methods are used to estimate the population or demographic characteristics of a smaller area based on trends in a larger area, such as shown for the censal ratio method. The synthetic method assumes the local trend changes at the same rate as the larger area trend. The rate of change in a local area can differ greatly and even be in the opposite direction of trends in the larger area, biasing the results of synthetic estimation (Voss, Palit, Kale, and Krebs 1995).

Chambers and Feeney (1977) and Purcell and Kish (1980) proposed structure preserving estimation (SPREE) as a generalization of synthetic estimation that makes fuller use of reliable direct estimates. Within a log-linear model framework, SPREE uses the method of iterative proportionate fitting that adjusts a multiple dimensional matrix to independently derived marginal totals for each dimension (e.g., Deming 1943: Chapter VII).

5.2.7 Other Methods

Methods that are sufficiently different from those previously discussed include structural models, administrative records, dual system estimators, social network analysis, and imputation and related methods. While different, the methods covered here have pieces in common with other post-censal population estimation methods.

Demographers and others often face questions that cannot be answered using estimation (and projection) methods based solely on demographic factors - the demographic consequences of the closing of a large manufacturing plant, for example. Structural models can produce population estimates that can account for factors such as the economy, environment, land use, housing, and the transportation system (Smith, Tayman, and Swanson 2001: Chapters 9 and 10).

The administrative records method is most associated with using tax return information to estimate migration rather school enrollment as part of the component method (Starsinic, Lee, Goldsmith, and Spar 1995). Administrative records may have a role beyond their use for estimating migration. Swanson and Walashek (2011) propose a re-vamped US census based neither on the current system, self-enumeration, nor its predecessor door-to-door canvassing. Instead, they propose the Census-Enhanced Master Address File (CEMAF) system built on a combination of four elements: (1) administrative records; (2) the continuously updated Master Address File; (3) survey data; and (4) modeling and imputation techniques. CEMAF could also deliver population estimates that are timely, comprehensive, and internally consistent and also estimates of housing and demographic and socio-economic characteristics for the US and subnational areas. The dual system estimates method represents a specific application of the general theme of record

matching underlying the CEMAF. The dual system method matches files from the Current Population Survey and IRS to obtain population estimates by age (15 to 64), sex, race, region, etc. (Causey 1984).

Social network analysis is a method for estimating the hard to count populations. It represents an interdisciplinary approach developed from the interplay of social theory; application; and formal mathematical, statistical, and computing methods (Wasserman and Faust 1994:10). The general approach involves asking people how many people they know in various populations whose size is well known. From this information, the network scale up method is used to derive estimates of the population group of interest (Killworth, Johnsen, Bernard, Shelley, and McCarty 1990).

Imputation is the general term to describe the assignment of information to cases with missing values due to non-response in a survey or census. The problems of missing data are well known and include less efficient estimators, the inability to use standard complete-data analysis methods, and possible biases because respondents are often systematically different from non-respondents (Rubin 1987:1). Four common imputation methods are: (1) deductive, which is based on other information available from the case in question; (2) hot-deck, which is based on information from "closest-matching" cases; (3) mean-value, which uses the average as the source of assignment; and (4) regression-based, which missing values are estimated from independent factors in cases with no missing values. Imputation can be based on a single value or derived from a distribution of missing values (multiple imputation).

5.2.8 Inter-censal

After a census is conducted, the post-censal estimates created for the prior decade by definition become inter-censal estimates. The post-censal estimate corresponding to the census data inevitably turns out to be different from the census count. This difference is known as the error of closure, which represents the cumulative decade error in the estimation procedure as well as any error in the two censuses. To create a set of inter-censal estimates consistent with the two censuses, the error of closure is allocated to each inter-censal year. Different approaches have been developed to handle this allocation for population totals and errors of closure in demographic composition (Bryan 2004b: 535-538 and 551-552). For example, the error of closure can be distributed using the share of the decade change in total population in each inter-censal year.

In some applications, inter-censal estimates may be needed where no post-censal estimates are available or they do not contain the demographic detail required. The common way to handle such a situation is using methods of interpolation (Raymondo 1992: 102-110). The chief difference among interpolation methods is their assumption about the nature of growth over the inter-censal period. The most common assumption is that change occurs linearly, but a non-linear (geometric or exponential growth) assumption may be appropriate in areas of rapid change.

For more complicated growth patterns, polynomial or osculatory interpolation methods are available (e.g., Judson and Popoff 2004: 685-692). For example, say only total population estimates exist for the inter-censal years, but information is needed by age. One approach is to interpolate the percentage distributions by age between the two censuses for each inter-censal year and then apply the interpolated distributions to the inter-censal total population (e.g., Espenshade and Tayman 1982).

References

Bogue, D. J. (1950). A technique for making extensive population estimates. *Journal of the American Statistical Association, 45*, 149–163.

Bogue, D. J., & Duncan, B. B. (1959). A composite method for estimating post-censal population for small areas by age, sex, and color (Vol. Vital Statistics Special Report, 47). Washington, DC: National Office of Vital Statistics.

Bryan, T. (2004b). Population estimates. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 523–560). New York: Elsevier Academic Press.

Causey, B. D. (1984). Dual system estimation based on iterative proportional fitting. *SRD* Research Report Number*: CENSUS/SRD/RR-84/03*. Washington, DC: Bureau of the Census.

Chambers, R., & Feeney, G. (1977). *Log linear models for small area estimation*. Belconnen, ACT: Australian Bureau of Statistics.

Deming, W. E. (1943). *Statistical adjustment of data*. New York: Dover Publications.

Espenshade, T. J., & Tayman, J. (1982). Confidence intervals for post-censal population estimates. *Demography, 19*(2), 191–210.

Forrester, J. W. (1958). Industrial dynamics: A major breakthrough for decision makers. *Harvard Business Review, 36*(4), 37–66.

Hamilton, C. H, & Perry, J. (1962). A short method for projecting population by age from one decennial census to another. *Social Forces, 41*, 163–170.

Jarosz, B. (2008). Using assessor parcel data to maintain housing unit counts for small area population estimates. In S. H. Murdock, & D. A. Swanson (Eds.), *Applied Demography in the 21st Century* (pp. 89–101). Dordrecht, Heidelberg, London, and New York: Springer.

Judson, D. H., & Popoff, C. L. (2004). Selected general methods. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 677–732). New York: Elsevier Academic Press.

Judson, D. H., & Swanson, D. A. (2011). *Estimating characteristics of the foreign-born by legal status: An evaluation of data and method*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Killworth, P. D., Johnsen, E. C., Bernard, R. H., Shelley, G. A., & McCarty, C. (1990). Estimating the size of personal networks. *Social Networks, 12*(4), 289–312.

Long, J. F. (1993). Post-censal population estimates: States, counties, and places. Technical Working Paper No. 3. Washington, DC: US Bureau of the Census.

Lowe, T. J., Myers, W. R., & Weisser, L. M. (1984). A special consideration in improving housing unit estimates: The interaction effect. Paper presented at the annual meeting of the Population Association of America, Minneapolis, MN.

Mandell, M., & Tayman, J. (1982). Measuring temporal stability in regression models of population estimation. *Demography, 19*(1), 135–146.

Martin, J. M., & Serow, W. J. (1978). Estimating demographic characteristics using the ratio-correlation method. *Demography, 15*(2), 223–234.

Murdock, S. H., & Ellis, D. R. (1991). *Applied demography: An introduction to basic concepts, methods, and data*. Boulder: Westview Press.

Murdock, S. H., Hwang, S., & Hamm, R. R. (1995). Component methods. In N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, & P. R. Voss (Eds.), *Basic methods for preparing small-area estimates* (pp. 10–53). Madison: Applied Population Laboratory, University of Wisconsin.

Namboodiri, K., & Lalu, N. (1971). The average of several regression estimates as an alternative to the multiple regression estimate in post-censal and inter-censal estimates: A case study. *Rural Sociology, 36*(187–194).

O'Hare, W. (1976). Report on a multiple regression method for making population estimates. *Demography, 13*(3), 369–380.

Purcell, D. E. (1970). Improving population estimates with the use of dummy variables. *Demography, 7*(1), 87–92.

Purcell, N. J., & Kish, L. J. (1980). Post-censal estimates for local areas (or domains). *International Statistical Review, 48*, 3–18.

Raymondo, J. C. (1992). *Population estimation and projections*. New York: Quorum Books.

Rosenberg. H. M. (1968). Improving current population estimates through stratification. *Land Economics, 44*, 331–338.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Schmitt, R. C., & Crosetti, A. H. (1954). Accuracy of ratio-correlation method for estimating post-censal population. *Land Economics, 30*(3), 279–280.

Schmitt, R. C., & Grier, J. M. (1966). A method for estimating the population of minor civil divisions. *Rural Sociology, 31*, 355–361.

Smith, S. K. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association, 81*, 287–296.

Smith, S. K., & Cody, S. (2004). An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review, 23*, 1–24.

Smith, S. K., & Tayman, J. (2003). An evaluation of population projections by age. *Demography, 40*(4), 741–757.

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

Starsinic, D. E., Lee, A. S., Goldsmith, H. F., & Spar, M. A. (1995). The Census Bureau's administrative records method. In N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, & P. R. Voss (Eds.), *Basic methods for preparing small-area estimates* (pp. 54–70). Madison: Applied Population Laboratory, University of Wisconsin.

Swanson, D. A. (1978). An evaluation of the ratio and difference regression methods for estimating small, highly concentrated populations. *Review of Public Data Use, 6*, 18–27.

Swanson, D. A., Schlottmann, A., & Schmidt, B. (2010). Forecasting the population of census tracts by age and sex: An example of the Hamilton-Perry method in action. *Population Research and Policy Review, 29*(1), 47–63.

Swanson, D. A., & Tedrow, L. M. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography, 21*(3), 373–382.

Swanson, D. A., & Walashek, P. J. (2011). *CEMAF as a census method: A proposal for a redesigned census and independent US Census Bureau*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Tayman, J., & Schafer, E. (1985). The impact of coefficient drift and measurement error on the accuracy of ratio correlation population estimates. *The Review of Regional Studies, 15*(2), 3–10.

US Census Bureau. (1998). Subcounty population estimates methodology. (http://www.census.gov/population/methods/e98scdoc.txt).

Voss, P. R., Palit, C. D., Kale, B. D., & Krebs, H. J. (1995). Censal ratio methods. In N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, & P. R. Voss (Eds.), *Basic methods for preparing small-area estimates* (pp.70–89). Madison: Applied Population Laboratory, University of Wisconsin.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Chapter 6
Extrapolation Methods[1]

Extrapolation techniques rely solely on the pattern of past population changes to estimate the post-censal population, and they assume trends in the post-censal period will be similar to historical trends. This method involves fitting mathematical models to historical data and using these models to estimate population. Relatively low costs and small data requirements make extrapolation methods useful, not only in demography, but in other fields as well (e.g., Armstrong 2001: 217; Granger 1989: Chapters 2, 3, and 4; Mahmoud 1984; Makridakis, Wheelwright, and Hyndman 1989: Chapters 4 and 7; Schnaars 1986. Although trend extrapolation methods are associated more frequently with population projections, they are useful for post-censal estimates relatively close to the last census, for completing estimates when resources are limited, or for estimating small areas and demographic subgroups (e.g. Murdock and Ellis 1991: 184; Baker et al. 2008).

Although there are many different methods by which historical values can be extrapolated, it is convenient to organize them into three categories. Simple extrapolation methods have simple mathematical structures and require data for only two dates. We discuss three simple methods: linear change, geometric change, and exponential change. Complex extrapolation methods require data for additional time points, have more complicated mathematical structures, and require statistical estimation of model parameters. We cover five complex methods: linear trend, polynomial curve, exponential curve, logistic curve, and ARIMA time series models. The final category, ratio extrapolation methods, involves the two populations: the population of a subgroup or "child" (e.g., county, Hispanic origin); and the population of its larger "parent" (e.g., state, total population). We cover three methods: Constant-Share, Shift-Share, and Share-of-Growth.

We illustrate extrapolation methods using annual total population data from 1980 to 2000 for two counties in Washington State: Island and Walla Walla (Forecasting Division 2010). We use the 20 year base period for all 11 methods.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_6,

Table 6.1 Population of Washington State and Island and Walla Walla Counties, 1980-2000

Year	Washington State	Island	Walla Walla
1980	4,132,353	44,048	47,435
1981	4,229,278	45,443	47,134
1982	4,276,549	46,559	47,712
1983	4,307,247	47,551	48,248
1984	4,354,067	48,225	48,345
1985	4,415,785	49,661	48,287
1986	4,462,212	51,024	48,163
1987	4,527,098	52,436	48,170
1988	4,616,886	54,370	48,085
1989	4,728,077	56,523	48,277
1990	4,866,692	60,195	48,439
1991	5,021,335	62,107	50,220
1992	5,141,177	63,947	51,119
1993	5,265,688	64,193	52,812
1994	5,364,338	66,239	53,836
1995	5,470,104	66,462	53,269
1996	5,567,764	67,856	55,047
1997	5,663,763	68,967	55,238
1998	5,750,033	69,609	55,521
1999	5,830,835	70,512	55,108
2000	5,894,121	71,558	55,180
Av. Annual Change	88,088.4	1,375.5	387.3
Percent Change	42.6%	62.5%	16.3%
Av. Annual Rate of Growth[a]	1.78	2.43	0.76

[a]Exponential rate of growth

For methods requiring only two time points, we use the population in 1980 and 2000. For the complex methods requiring more data we use all 21observations. For each method, we produced annual estimates over a simulated post-censal period from 2000 to 2010. To illustrate the methods, we calculate a post-censal estimate for 2010, and then discuss estimates for selected post-censal years (2002, 2005, and 2010) later in the Chapter.

Table 6.1 and Figure 6.1 show the base data for Island and Walla Walla Counties. The table also includes data for Washington State, which are needed to apply the ratio methods. During the 1980 to 2000 period, Island County grew faster than the state (62.5% vs. 42.6%), while Walla Walla County was among the slowest growing counties in the state, increasing by 16.3%. In 1980, Island County had around 3,000 fewer people than Walla Walla County. By 2000, the population of Island County exceeded that of Walla Walla County by more than 16,000 people. We return to this fact in our summary comments on trend extrapolation methods.

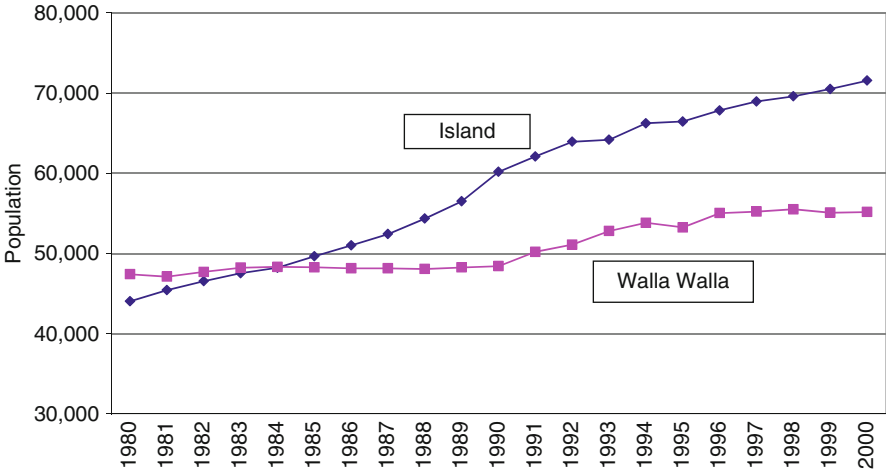


Fig. 6.1 Population of Island and Walla Walla Counties, 1980-2000

6.1 Simple Extrapolation

6.1.1 Linear Change

This method assumes that the post-censal population will change by the same amount over a given period, usually a year, as occurred during the base period. Average annual absolute change (aaac) during the base period is:

$$\text{aaac} = (\text{P}_\text{l} - \text{P}_\text{b})/\text{y},$$

where aaac is the average annual absolute change in the base period; P_l is the population in the launch year (usually the latest census); P_b is the population in the base year (earliest year of the data); and y is the number of years in the base period (i.e., the number of years between the base and launch years). An estimate using the linear change method is:

$$\text{P}_\text{t} = \text{P}_\text{l} + \text{z}(\text{aaac}),$$

where P_t is the population in the post-censal year and z represents the years between the post-censal estimate date and the last census.

The average annual absolute change between 1980 and 2000 and the 2010 population estimate for Island County are:

$$\text{aaac}[(71,558 - 44,048)/20] = 1{,}375.5; \text{ and}$$
$$\text{P}_{2010}[71,558 + (10\,^{*}1375.5)] = 85{,}313.$$

The corresponding calculations for Walla Walla County are:

$$\text{aaac}[(55{,}180 - 47{,}435)/20] = 387.3; \text{ and}$$
$$P_{2010}[55{,}180 + (10 * 387.3)] = 59{,}053.$$

6.1.2 Geometric Change

This method assumes that the population will change by the same percentage rate in the post-censal period as during the base period. The average geometric rate of population change during the base period is:

$$r = [(P_1/P_b)^{(1/y)}] - 1,$$

where r is the average geometric rate of change; P_l is the population in the launch year; P_b is the population in the base year; and y is the number of years in the base period. An estimate using the geometric change method is:

$$P_t = (P_1)[(1 + r)^z],$$

where P_t is the population in the post-censal year; and z is the number of years in the post-censal period.

The annual rate of geometric change between 1980 and 2000 and the 2010 population estimate for Island County are:

$$r([(71{,}558/44{,}048)^{(1/20)}] - 1) = 0.02456; \text{ and}$$
$$P_{2010}[(71{,}558)[(1 + 0.02456)^{(10)}] = 91{,}208.$$

The corresponding calculations for Walla Walla County are:

$$r([(55{,}180/47{,}435)^{(1/20)}] - 1) = 0.00759; \text{ and}$$
$$P_{2010}[(55{,}180)(1 + 0.00759)^{10}] = 59,514.$$

6.1.3 Exponential Change

The exponential change approach is closely related to the geometric, but it views change as occurring continuously rather than at discrete intervals. The exponential rate of population change during the base period is computed as:

$$r = [\ln(P_1/P_b)]/y,$$

where r is the average annual exponential rate of change; ln represents the natural logarithm; P_l is the population in the launch year; P_b is the population in the base

year; and y is the number of years in the base period. A population estimate using the exponential change method is:

$$P_t = P_1 e^{rz},$$

where P_t is the population in the post-censal year, e is the base of the system of natural logarithms (approximately 2.71828), and z is the number of years in the post-censal period.

The annual rate of exponential change from 1980 to 2000 and the 2010 population estimate for Island County are:

$$r([\ln(71{,}558/44{,}048)]/20) = 0.02426; \text{ and}$$
$$P_{2010}[71{,}558 * (e^{0.02426 * 10})] = 91{,}205.$$

The corresponding calculations for Walla Walla County are:

$$r([\ln(55{,}180/47{,}435)]/20) = 00756; \text{ and}$$
$$P_{2010}[55{,}180 * (e^{.00759 * 10})] = 59{,}513.$$

6.2 Complex Extrapolation

Complex extrapolation methods differ from simple extrapolation methods in several respects. Complex methods require additional time points over the base period and thus can provide a more complete picture of the historical pattern of population change. Their more complex mathematical structures provide a wider range of assumptions regarding post-censal trends. Finally, the statistical algorithms for estimating complex model parameters provide a basis for constructing probabilistic intervals around post-censal population estimates (Espenshade and Tayman 1982; Swanson and Beck 1994). We discuss the uncertainty in population estimates in Chapter 14. These features do not guarantee that complex extrapolation methods provide more accurate estimates than either simple or ratio extrapolation methods, and complex extrapolation methods are considerably more difficult to implement.

Three basic steps are typically followed when applying complex extrapolation methods. The first is to assemble historical population data for different dates during the base period; typically annual data for population. The data must be based on consistently defined geographic boundaries for each time point; adjustments will be required in areas that have experienced shifts in boundaries, which is typically the case for subcounty areas. The second step is to estimate the parameters of the model selected to generate the estimate; a process known as curve fitting (Alinghaus 1994). Typically, graphs and statistical measures are used to determine how well a given model is fits the data for base period, but the choice of a particular model also reflects judgment about the nature of change during the post-censal period. The final step is to generate post-censal estimates using the model(s) selected.

We develop a variety of complex extrapolation models for Island and Walla Walla Counties. In these types of models, population is the dependent variable and time is the independent variable. Time can be measured using the original units or as recoded values. For ease of interpretation, we express time as integers ranging from 1 to 31 (e.g., 1 = 1980, 2 = 1981, . . ., 21 = 2000, 22 = 2001, . . ., 31 = 2010). The decision on the measurement of time is not substantively important, as long as a consistent coding scheme is used in both the base and post-censal periods. The only impact of the coding scheme will be on the equation intercept; none of model statistics, slope parameters, or post-censal estimates are affected.

6.2.1 Linear Model

A linear model is based on the equation for a straight line. It assumes that a population will change by a constant numerical amount. This assumption is identical to that underlying the simple linear method discussed earlier, but the model is different:

$$P_i = a + [(b)(T_i)] + e_i + cb$$

where i is the time point; P is the population; T is the time variable; a is the constant or intercept term, and b is the slope; e is the error term of the equation (see Chapter 2); and cb is a calibration factor. The slope represents the annual change in population; a positive slope reflects an increasing population and a negative slope reflects a decreasing population in the post-censal period. The calibration factor requires explanation. In any curve fitting procedure, it will be unusual for the estimated and observed values in the launch year to be identical. An additive calibration factor, based on the launch year residual, adjusts the post-censal estimates so they are consistent with the launch year population. The calibration factor is computed by subtracting the predicted population from the observed population. If the residual is negative/positive (predicted value is too high/low), the estimates will be adjusted downward/upward by a constant amount.

The Ordinary Least Squares (OLS) regression results, calibration factor, and the 2010 population estimate for Island County, using the value of 31 for time are:

$$\text{cb } 71{,}558 - 73{,}487 = -1{,}929; \text{ and}$$
$$P_{2010}\, 41{,}913.2 + (1{,}503.5 * 31) - 1{,}929 = 86{,}593; \text{ adjusted } r^2 = 0.979.$$

The corresponding results for Walla Walla County are:

$$\text{cb } 55{,}180 - 55{,}555 = -375; \text{ and}$$
$$P_{2010}\, 45{,}454.0 + (481.0 * 31) - 375 = 59{,}990; \text{ adjusted } r^2 = 0.867.$$

The linear model fits the data for Island County better than Walla Walla County as evidenced by its higher adjusted r^2. The slopes indicate that Island and Walla Walla Counties will increase by 1,503 and 481 annually during the post-censal period. These increases are somewhat, but not dramatically, higher than the average change over the base period.

6.2.2 *Polynomial Model*

Polynomial models can be used for basing post-censal estimates on non-linear patterns (i.e., the annual change is a constant percentage value). The general formula for a polynomial curve is:

$$Y_i = a + (b_1)(X_i) + (b_2)(X_i^2) + (b_3)(X_i^3) + \cdots + (b_n)(X_i^n).$$

Unlike the linear model, a polynomial model has more than one term for the time variable, represented by raising its value to different powers. The coefficients for a polynomial model can be estimated using OLS procedures and include a measure of the linear trend (b_1) and measures of the non-linear pattern (b_2, b_3, ..., b_n). The highest exponent in the equation is called the degree of the polynomial. The linear model previously discussed is a first degree polynomial; a second degree or quadratic polynomial contains X and X^2; a third degree polynomial contains X, X^2; and X^3; and so forth.

To illustrate the use of a polynomial curve, we use the quadratic form:

$$P_i = a + [(b_1)(T_i)] + [(b_2)(T_i^2)] + e_i + cb.$$

A quadratic curve can produce a variety of growth scenarios, depending on the signs and magnitudes of the two slope coefficients (see Figure 6.2). The curves in this figure are derived from slopes of 1.0 and 0.1 on the linear and squared term and time values from 1 to 11.

A population growing/declining at an increasing rate will occur when both coefficients are positive/negative. A positive linear and negative squared term will cause a population to grow at a decreasing rate, with the possibility of change turning negative. A negative linear and positive squared term will cause a population to decline at a decreasing rate, with the possibility of change turning positive. Although any degree can be used, polynomials higher than a "second" or at most a "third" degree are seldom used for population estimation.

The quadratic regression results, calibration factor, and the 2010 population estimate for Island County are:

$$\text{cb } 71,558 - 72,609 = -1,051; \text{ and}$$

$$P_{2010}\, 40,744.3 + (1808.4 * 31) + (-13.86 * 31^2) - 1{,}051 = 82{,}434,$$

$$\text{adjusted } r^2 = 0.981.$$

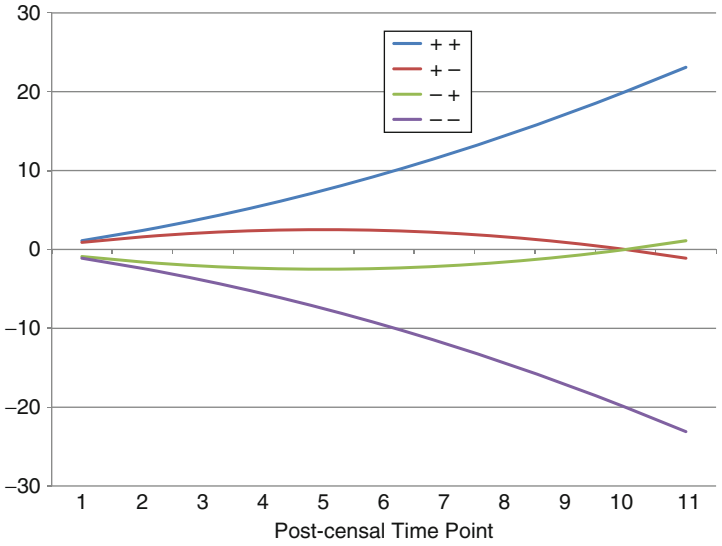


Fig. 6.2 Alternative Patterns of Change from a Quadratic Curve: Signs of the Linear and Squared Terms

The corresponding results for Walla Walla County are:

$$\text{cb } 55{,}180 - 56{,}809 = -1{,}629; \text{ and}$$

$$P_{2010}\, 47{,}124.4 + (45.24 * 31) + (19.81 * 31^2) - 1{,}629 = 65{,}935,$$

$$\text{adjusted } r^2 = 0.907.$$

For Island County, the squared term is not significant at $\alpha = 0.05$ and the adjusted r^2 is only 0.002 higher than the linear model. However, the negative on the squared term causes the 2010 estimate in this model to be lower than the estimate from the linear model (82,434 vs. 86,593). For Walla Walla County, the squared term is significant at $\alpha = 0.05$ and the adjusted r^2 is 0.04 points higher than the linear model. The addition of the squared term causes the linear term to lose its statistical significance. The slope of the linear term in the quadratic model (45.24) is considerably smaller than the slope in the linear model (481.0). The prominence of the squared term in the Walla Walla County quadratic model causes the post-censal estimate to be 9.9% higher than the estimate based on the linear model (65,935 vs. 59,990).

6.2.3 Exponential Model

Non-linear trends in the historical data also can be projected using curves based on logarithmic or other transformations of the base data (e.g., Draper and Smith 1981: Chapter 5; Isserman 1977; Stock and Watson 2003: Chapter 6). Transformations

include using the inverse of time to model areas where the population change is asymptotic; the power function that uses the natural logarithm of time and population; the logarithmic function that uses the natural logarithm of time; and the exponential function that uses the natural logarithm of population.

We illustrate the use of other non-linear functions using the exponential model:

$$\ln(P_i) = a + [(b)(T_i)] + e_i + cb; \text{ and}$$
$$P_i = e^{\ln(Pi)}.$$

When the population has been transformed by taking the natural logarithm, the equation yields the post-censal estimate transformed value of the population. The value of the population itself is obtained by the rules of natural logarithms (i.e., e raised to the power of the equation result). The slope of an exponential model estimates the average annual rate of growth.

The exponential model regression results, calibration factor, and the 2010 population estimate for Island County are:

$$\text{cb } 11.17826 - 11.22556 = -0.04730;$$
$$\ln(P_{2010})10.6746 + (0.02624 * 31) - 0.04730 = 11.44074; \text{ adjusted } r^2 = 0.972;$$
$$\text{and } P_{2010}\, e^{11.44062} = 93,036.$$

The corresponding results for Walla Walla County are:

$$\text{cb } 10.91836 - 10.92663 = -0.00827;$$
$$\ln(P_{2010})\ 10.7294 + (0.00939 * 31) - 0.00827 = 11.01222; \text{ adjusted } r^2 = 0.871;$$
$$\text{and } P_{2010}\, e^{11.01222} = 60,610.$$

The exponential model fits the data for Island County better than Walla Walla County as evidenced by its higher adjusted r^2. The slopes indicate that Island and Walla Walla Counties will increase by 2.6% and 0.9% annually during the post-censal period. These rates are somewhat higher than the rates observed over the base period.

6.2.4 Logistic Model

Unlike the extrapolation methods considered so far, the logistic approach explicitly allows an upper limit on the ultimate size of the population. It is designed to yield an S-shaped pattern representing an initial period of slow growth rates, followed by a period of increasing growth rates, and finally a period of declining growth rates that approach zero as a population approaches its upper limit. The logistic model is consistent with Malthusian and other theories of constrained population growth.

Keyfitz (1968: 215) provides the following formula for a 3-parameter logistic curve:

$$P_i = a/[1 + (b(e^{-cTi}))],$$

where a reflects the upper asymptote; b and c are parameters that define the shape of the logistic curve; and e is the base of the natural logarithm. Some software packages (e.g. SPSS) require the magnitude of the upper asymptote prior to the estimation of other model parameters, while other packages (e.g., NCSS) estimate all parameters within the context of the model. However, like parameters in an ordinary regression model (e.g., the intercept term), the estimated parameters may not be consistent with a substantive interpretation (e.g., *a* represents an actual upper population limit). Other specifications are available for the logistic curve, some including more than three parameters (Pielou 1969: 19-32; Sieber and Wild 1989: 331). Other functions that contain asymptotic ceilings or floors on population include the modified exponential, Gompertz, and hyperbolic (Davis 1995; Pittenger 1976: 57-67).

The 3-parameter logistic regression model regression results, calibration factor, and the 2010 population estimate for Island County are:

$$\text{cb } 71{,}558 - 72{,}443 = -885; \text{ and}$$
$$P_{2010}\ (92{,}078.29/1 + (1.2472 * e^{(-0.072686(31))})) - 885 = 80{,}526;$$
$$\text{adjusted } r^2 = 0.984.$$

The corresponding results for Walla Walla County are:

$$\text{cb } 55{,}180 - 55{,}752 = -572; \text{ and}$$
$$P_{2010}\ (19{,}774{,}450/(1 + (432.731 * e^{(-0.009605(31))})) - 572 = 60{,}783;$$
$$\text{adjusted } r^2 = 0.883.$$

The estimated a parameter is more logical for Island County (92,078) than the same parameter for Walla Walla County (19,774,450), but the 2010 post-censal estimates for both counties do not appear unreasonable. Estimates based on the logistic curve are sensitive to the value of the upper asymptote. Increasing the a parameter to 120,000 in Island County raises the 2010 population estimate to 84,990; 5.5% greater than the estimate based on the software generated parameter. Reducing the a parameter to 80,000 lowers the 2010 population estimate to 76,860.

6.2.5 Arima Model

The final complex extrapolation method we illustrate is the Autoregressive Integrated Moving Average (ARIMA) model popularized by Box and Jenkins (1976). Some feel ARIMA models are preferable to regression-based complex

extrapolation methods because they produce more accurate coefficient estimates and smaller errors over the post-censal period (e.g., Granger and Newbold 1986: 205-215; Jenkins 1979: 88-94; McDonald 1979). The dynamic and stochastic framework of ARIMA models also provide a statistical basis for developing probabilistic intervals around post-censal estimates (e.g., Box and Jenkins 1976: Chapter 5; Nelson 1973: Chapter 6). However, the methods used in ARIMA modeling are considerably more complex than the other extrapolation methods, and are more difficult to implement and explain to users. We provide a general overview of ARIMA modeling and suggest consulting standard texts for more details on implementing and using this modeling framework (e.g., Box and Jenkins 1976; Brockwell and Davis 2002; Chatfield 2000; Jenkins 1979; Montgomery, Jennings, and Kulahci 2008; Yaffee and McGee 2000).

ARIMA models attempt to uncover the stochastic mechanisms that generate the historical population series and then use this information as a basis for developing post-censal estimates. Three processes describe the stochastic mechanism and specify the structure of an ARIMA model: (1) autoregressive; (2) differencing; and (3) moving average.

The autoregressive process has a memory, which is based on the correlation of each observation with all preceding observations. The impact of earlier observations is assumed to diminish exponentially over time. The number of preceding observations incorporated into the model determines its order. For example, in a first-order autoregressive process, the current observation is explicitly a function only of the immediately preceding observation. However, the immediately preceding observation is a function of the one before it, which is a function of the one before it, and so forth. Consequently, all preceding observations influence current observations, albeit with a declining impact. In a second-order autoregressive process, the current observation is explicitly a function of the two immediately preceding observations; again, all preceding observations have an indirect impact.

A stationary time series (i.e., one with constant differences over time) is needed to properly construct an ARIMA model. The differencing process is used to achieve such a series. First differences (i.e., observation minus its preceding value) are usually sufficient, but second differences (i.e., differences between differences) have been found to be applicable to human populations (McNown and Rogers 1989; Saboia 1974; Tayman, Smith, and Lin 2007). Logarithmic and square root transformations may also be useful for stabilizing the variance of a time series.

The moving average represents an event that has a substantial but short-lived impact on a time series pattern. The order of the moving average process defines the number of time periods affected by a given event.

The general ARIMA model is expressed as ARIMA(p,d,q), where p is the order of the autoregressive term, d is the degree of differencing, and q is the order of the moving average term. ARIMA models based on time intervals of less than one year may also require seasonal terms for p, d, and q; seasonal terms are not usually relevant when modeling population. The first step in developing an ARIMA model is to identify the best values for p, d, and q, which typically range from 0 to 2. The d value is determined first because a stationary series is required to properly identify the autoregressive and moving average processes (Box and Jenkins 1976: 174;

Granger 1989: 72). As a rule, the time series should contain enough observations for model identification and parameter estimation. Convention suggests a minimum of 50 observations for ARIMA modeling (e.g., McCleary and Hay 1980: 20; Meyler, Kenny and Quinn 1998; Saboia 1974), but there is no hard and fast rule; for example, some say the minimum should be 60 and others say 30 (Yaffee and McGee 2000: 4).

The traditional approach for identifying the best values for p, d, and q focuses on assessing the patterns of the autocorrelation function (ACF) and partial autocorrelation function (PACF) (Box and Jenkins 1976: Chapter 6). This quasi formal approach to identification is subjective and highly dependent on the skill and interpretation of the analyst (Granger and Newbold 1986: 77-78; Meyler, Kenny, and Quinn 1998). To help with this problem more objective methods have been developed, such as statistical tests for stationarity (Dickey, Bell, and Miller 1986; Elliot, Rothenberg, and Stock 1996; Phillips and Perron 1988) and statistics such as the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) for selecting the best values for p and q while avoiding a model with too many parameters (Brockwell and Davis 2002: 187-193). The usual practice is to evaluate several tentative models before the best model is selected. An acceptable ARIMA model will have random residuals, no significant values in the ACF and PACF, and the smallest possible values for p, d, and q. The Portmanteau test is used to evaluate the null hypothesis of randomness in model residuals (Ljung and Box 1978).

We identified the "best" parameters for p, d, and q using the ACF, PACF, statistical tests, and the AIC and BIC for Island and Walla Walla Counties. We used annual observations from 1980 to 2000 to be consistent with the base period of the other extrapolation methods.[2] Calculation of post-censal estimates from the other complex methods requires only the proper value for the time variable and the model parameters; they are easy to recreate. It is more difficult to recreate post-censal estimates for ARIMA models. Beyond the launch year post-censal values, or a combination of the observed and post-censal values are used to generate the post-censal estimate at a subsequent time point. The temporal sequence of values required will depend on the degree of differencing and the order of the autoregressive and moving average parameters. Box and Jenkins (1976: 135-138) and Nelson (1973: 144-147) provide details on computing values past the launch year for a wide variety of ARIMA models

For Island County, an ARIMA(1,2,0) model is selected, which has a first order-auto regressive process, second-degree differencing, and no moving average. The coefficients and 2010 post-censal estimate, adjusted with a calibration factor, from this model are: constant (-16.444); ar1 (-0.593); and 2010 population estimate (80,828). ARIMA models with second-order differences and a constant follow a quadratic trend (Tayman, Smith, and Lin 2007), but the negative coefficients on the constant and ar1 term decelerates the non-linear growth potential of the Island County ARIMA model.

For Walla Walla County, an ARIMA(0,1,0) model is selected, which has neither an autoregressive nor moving average process and a first-degree difference. This model has only a constant term (387.2), and yields a 2010 population estimate of 58,738. An ARIMA(0,1,0) model is also known as a random walk process

(Pearson 1905). Under a random walk model, a post-censal estimate is computed by adding the constant term to the prior year population. If the constant term is not statistically significant, it can be ignored and the post-censal estimate is simply the launch year value (Granger and Newbold 1986: 40).

6.3 Ratio Extrapolation

Unlike other trend extrapolation methods, ratio extrapolation methods involve the relationship between two entities: a larger unit or parent and a subunit or child. The entities can be geographically-based such as census tracts within a city or reflect demographic subgroups, such as race groups that comprise the total population. Ratio extrapolation methods are often used where there is a perfect hierarchical structure; that is, where the subunits are mutually exclusive and exhaustive and can be aggregated to the parent level culminating in one all-inclusive unit. For example, census blocks in the US can be aggregated successively into block groups, census tracts, counties, states, and finally the entire United States. Ratio methods also can be used where there is not a perfect hierarchy. For example, the urbanized area of cities that nests within the urbanized area of a county, and not the entire county area.

Similar to simple extrapolation methods, ratio methods have small data requirements and are easy to apply. These methods are also self-normalizing in that the sum of subunit estimates will equal the estimate for the parent. We discuss three commonly used ratio methods: 1) Constant-Share; 2) Shift-Share: and 3) Share-of-Growth. Ratio methods require an independent post-censal estimate for the parent, which for most examples in the Chapter is the State of Washington; we also use the Shift-Share method to produce 2010 post-censal estimates by Hispanic Origin for Island County. The post-censal estimates for Washington State are: 6,041,710, 6,256,400, and 6,733,250, for the years 2002, 2005, and 2010 (Forecasting Division 2010).

6.3.1 Constant-Share

In the Constant-Share method, the child's share of the parent population is held constant at a level observed during the base period. Typically, it is the share observed in the launch year. The post-censal estimate is made by applying the child's share to independent estimate for the parent. The Constant-Share method is:

$$P_{it} = (P_{il}/P_{jl})(P_{jt}),$$

where P_{it} is the estimate for child (i) in the post-censal year (t); P_{il} is the population of the child in the launch year; P_{jl} is the population of the parent (j) in the launch year; and P_{jt} is the post-censal estimate of the parent.

Table 6.2 Population Estimates: Constant-Share Method Island County, Walla Walla County, and Balance of State, 2010

	2000		2010[a]	2000–2010 Change	
	Population	Share	Estimate	Number	Percent
Island County	71,558	0.01214	81,742	10,184	14.2
Walla Walla County	55,180	0.00936	63,023	7,843	14.2
Balance of State	5,767,383	0.97850	6,588,485	821,102	14.2
State Total	5,894,121	1.00000	**6,733,250**	839,129	14.2

[a]2010 pop = 2000 share * 2010 State estimate

The computation of the 2010 post-censal estimates for Island County, Walla Walla County, and the remainder of the State using the Constant-Share method is shown in Table 6.2. This example held the shares constant at the 2000 launch year value, but shares for any point in time or an average of shares from prior time points could also be used. The constant method requires data from only one point in time, making it useful for areas where changing geographic boundaries or lack of information and/or time make it impossible to construct a geographically consistent historical series. The main drawback of this method is that it assumes that all the subunits will change at the same rate as the parent. In many instances, this will not be a reasonable assumption.

6.3.2 Shift-Share

Unlike the Constant-Share method, the Shift-Share method is designed to deal with historical changes in population shares. Different mathematical functions can be used for extrapolating the historical trend in the shares (Gabbour 1993), and we describe a method that assumes a linear trend in the shares over the post-censal period. The Shift-Share method is:

$$P_{it} = (P_{jt})[(P_{il}/P_{jl}) + ((z/y)((P_{il}/P_{jl}) - (P_{ib}/P_{jb})))],$$

where the child is denoted by i; the parent by j; z is the number of years in the post-censal period; y is the number of years in the base period; and b, l, and t refer to the base, launch, and post-censal years. The (z/y) term implements the linear trend and relates the length of the base and post-censal periods. For example with a 20 year base period length, a post-censal estimate one year past the launch year would apply (1/20) of the historical share change to the launch year share; two years past would apply 2/20 of the historical change, and so forth.

The computation of the 2010 post-censal estimates for Island County, Walla Walla County, and the remainder of the State using the Shift-Share method is shown in Table 6.3. As mentioned earlier, ratio methods can also be used to estimate demographic subgroups of the total population. Table 6.4 illustrates the

Table 6.3 Population Estimates: Shift-Share Method Island County, Walla Walla County, and Balance of State, 2010

	1980		2000		Change in Share		2010		2000–2010 Change	
	Population	Share	Population	Share	1980–2000	2000–2010[a]	Share[b]	Estimate[c]	Number	Percent
Island County	44,048	0.01066	71,558	0.01214	0.00148	0.00074	0.01288	86,724	15,166	34.4
Walla Walla County	47,435	0.01148	55,180	0.00936	−0.00212	−0.00106	0.00830	55,886	706	1.5
Balance of State	4,040,870	0.97786	5,767,383	0.97850	0.00064	0.00032	0.97882	6,590,640	823,257	20.4
State Total	4,132,353	1.00000	5,894,121	1.00000	0.00000	0.00000	1.00000	**6,733,250**	839,129	20.3

[a]2000–2010 change in share = 0.5 * 1980–2000 change in share
[b]2010 share = 2000 share + 2000–2010 change in share
[c]2010 pop = 2010 share * 2010 State estimate

Table 6.4 Population Estimates by Hispanic Origin: Shift-Share Method, Island County, 2010

	Shares			2010		2000–2010 Change	
	1990	2000	Change	Share[a]	Population[b]	Number	Percent
Non-Hispanic	0.96667	0.96027	−0.00640	0.95387	82,805	14,090	20.5
Hispanic- Mexican	0.01964	0.02376	0.00412	0.02788	2,420	720	42.4
Hispanic Other	0.01369	0.01597	0.00228	0.01825	1,584	441	38.6
	1.00000	1.00000	0.00000	1.00000	**86,809**[c]	15,251	21.3

[a]2010 share = 2000 share + 1990–2000 change in share
[b]2010 pop = 2010 share * 2010 Island County total population estimate
[c]2010 total population is the average of all 11 extrapolation methods for Island County

use of the Shift-Share method to estimate the 2010 population by Hispanic Origin for Island County. This example uses a 10-year base period from 1990 to 2000. Because the length of the base and post-censal periods are the same the (z/y) term equals 1.0 and the entire 1990 to 2000 change in shares is applied to the 2000 shares to estimate the 2010 shares. The Hispanic Origin estimates require an independent 2010 total population estimate for Island County. For this estimate, we use the average of the 11 estimates for Island County discussed in this Chapter.

The Shift-Share method can lead to substantial population losses in areas that grew very slowly (or declined) during the base period or unreasonably high estimates for places that have grown very rapidly. This inherent problem with the Shift-Share method is more acute when applying it to projections covering long-range horizons (e.g., 20 or 30 years), but this method could be problematic for post-censal estimates in areas with extreme changes in population during the base period.

6.3.3 Share-of-Growth

The Share-of-Growth ratio method deals with shares of population change rather than population size. This method, also known as the apportionment method, assumes the child's share of population change in the parent area will be the same over the post-censal period as it was during the base period. This Share-of-Growth method is:

$$P_{it} = P_{il} + [((P_{il} - P_{ib})/(P_{jl} - P_{jb}))(P_{jt} - P_{jl})],$$

where the components are defined as those in the Shift-Share method.

The computation of the 2010 post-censal estimate for Island County, Walla Walla County, and the remainder of the State using the Share-of-Growth method is shown in Table 6.5. The Share-of-Growth method may provide more reasonable post-censal estimates than either the Constant-Share or Shift-Share methods. However, it runs into problems when a child's population change has the opposite sign of the parent's change. For example, say during the base period the parent grew by

Table 6.5 Population Estimates: Share-of-Growth Method Island County, Walla Walla County, and Balance of State, 2010

	Change 1980–2000		Population		Change 2000–2010	
	Number	Share	2000	2010[a]	Number[b]	Percent
Island County	27,510	0.01561	71,558	84,657	13,099	18.3
Walla Walla County	7,745	0.00440	55,180	58,872	3,692	6.7
Balance of State	1,726,513	0.97999	5,767,383	6,589,721	822,338	14.3
State Total	1,761,768	1.00000	5,894,121	6,733,250	**839,129**	14.2

[a]2010 pop = population 2000 + population change 2000–2010
[b]pop change 2000–2010 = share of pop change 1980–2000 * State pop change 2000–2010

2,500 and a child declined by 500; the child's share of the parent's change would be computed as -500 / 2,500 or -0.2. If the parent grows by 5,000 during the post-censal period, the decline in the child would be estimated at -1,000. This is not likely a reasonable result; if anything the child is more likely to decline by a smaller amount or could even increase. In this situation, the child's share should be adjusted and the remaining shares modified so they sum to 1.0. Some applications assume zero change where change in the child is in the opposite direction of the parent's change (Pittenger 1976: 101). Adjusting a distribution with negative and positive values can be done using the plus-minus method discussed in Chapter 13.

6.4 Analyzing Estimation Results

Using the 11 extrapolation methods and the 20-year base period from 1980 to 2000, we estimated the total population for the years 2002, 2005, and 2010 in the simulated post-censal period (see Table 6.6). What can these data tell us about the behavior of post-censal estimates from ratio extrapolation methods?

Except for the complex polynomial, Constant-Share and Shift-Share, post-censal estimates are close in value for Walla Walla County; the range of estimates for 2010 is only 2,045 if these three methods are excluded. The reasons for the exceptions are clear. As previously noted, estimates based on the polynomial equation are driven mainly by the squared term, which translates into relatively large increases in Walla Walla population. The Constant-Share-method also yields relatively high estimates because it assumes that Walla Walla County grows at the same rate as the State during the post-censal period; it grew much more slowly during the base period. The Shift-Share method assumes that Walla Walla's share of the State population will continue to decline. Despite the post-censal growth in the State's population, the Shift-Share method indicates Walla Walla's population remains relatively constant. Because the Share-of-Growth method relies on shares of the State's population change, its estimates are more in line with the other extrapolation methods for Walla Walla County. We believe these results would occur in most areas characterized by slow to moderate population changes.

Table 6.6 Population Estimates Based on Alternative Extrapolation Methods, Island and Walla Walla Counties, 2002, 2005, and 2010

	Island			Walla Walla		
Extrapolation Method	2002	2005	2010	2002	2005	2010
Simple						
Linear Change	74,309	78,436	85,313	55,955	57,116	59,053
Geometric Change	75,116	80,788	91,208	56,021	57,306	59,514
Exponential Change	75,116	80,786	91,205	56,021	57,306	59,513
Complex						
Linear	74,565	79,075	86,593	56,142	57,585	59,990
Quadratic	73,955	77,343	82,434	57,014	60,061	65,935
Exponential	75,413	81,588	93,036	56,227	57,833	60,610
Logistic	73,710	76,592	80,526	56,258	57,914	60,783
ARIMA	73,792	76,547	80,828	55,640	56,801	58,738
Ratio						
Constant-Share	73,346	75,953	81,742	56,550	58,560	63,023
Shift-Share	74,241	78,268	86,742	55,270	55,244	55,886
Share-of-Growth	73,862	77,213	84,657	55,829	56,744	58,872
Projection Range						
Numeric Difference	2,067	5,635	12,510	1,744	4,817	10,049
Percentage Difference	3%	7%	16%	3%	9%	18%

The pattern of the estimates shows more variability for the faster growing Island County. We believe the greater variability of estimates developed from different extrapolation methods would occur in rapidly growing areas. With the exception of the quadratic model, the assumption of linear versus non-linear growth makes a larger difference in the Island County estimates. The polynomial model has a negative sign on the squared term, which lowers the estimate relative to other non-linear extrapolation methods. Estimates based on the logistic and ARIMA models are the lowest of any method, perhaps unreasonably too low. The modest difference between the calibrated asymptote in the logistic model (92,078) and the launch year population (71,558) and the negative coefficient on the AR1 term act to slow down the pace of growth during the post-censal period. The ratio methods also show considerable variability, but now Shift-Share is the highest and the Constant-Share is the lowest, reversing the pattern seen for Walla Walla County. This result is expected because Island County grew faster than the State during the base period.

6.5 Conclusions

Extrapolation methods have a number of useful characteristics. Simple and ratio methods have modest data requirements and most require population data from two points in time; the Constant-Share method requires data from only a single time point. These methods are easy to apply and explain to users, can be applied in a timely manner, and have low resource requirements. Complex extrapolation methods require

data from a number of time points and require much greater modeling and statistical skills than simple or ratio extrapolation methods. The lack of sufficient data often precludes their use in many small areas. However, compared to other estimation techniques that use symptomatic data or produce more detailed demographic characteristics, even complex extrapolation methods are characterized by timeliness, low resources, and small data requirements. In addition, complex extrapolation methods can be used to develop probabilistic intervals around post-censal estimates.

Extrapolation methods have some notable short-comings for post-censal estimation. They do not account for differences in some demographic characteristics, notably age, or for differences in the components of growth. They provide limited information on the demographic characteristics of the population, and can lead to unreasonable results even over the relatively short post-censal time period. Perhaps most importantly, they ignore relevant and known information that tracks changes in the population over the post-censal period. Finally, how does one decide on which extrapolation method or methods to use? Should an average of estimates from several methods be calculated? If so, what methods should be included and should the average be weighted or not? If a weighted average is desired, how are the weights determined? Despite the mechanical/objective nature of extrapolation methods, their proper use requires judgment and subjective choices about time periods, functional forms, and so forth.

Endnotes

1. Adapted from Chapter 8, "Trend Extrapolation Methods", in S. Smith, J. Tayman, and D. "Swanson. Projecting State and Local Populations: Methodology and Analysis. New York, NY: Kluwer Academic/Plenum Press. 2001.
2. The 21 observations are fewer than recommended for ARIMA modeling, so we also examined a larger dataset based on 41 annual observations from 1960 to 2000. The orders of the p, d, and q for the ARIMA models did not change and the post-censal estimates were close to those based on shorter data series. The 2010 post-censal estimates for Island and Walla Walla Counties were 1.8% lower and 1.0% higher using the longer data series.

References

Alinghaus, S. L. (1994). *Practical handbook of curve fitting*. New York: CRC Press.

Armstrong, J. S. (2001). Extrapolation of time series and cross-sectional data. In J. S.

Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 217–244). Norwell: Kluwer Academic Publishers.

Baker, J., Runa, X., Alcantara, A., Jones, T., Watkins, K., McDaniel, M., et al. (2008). Density-dependence in urban housing unit growth: An evaluation of the Pearl-Reed model for predicting housing unit stock at the census tract level. *Journal of Economic and Social Measurement*, 33, 155–163.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting, Second Edition*. Dordrecht, Heidelberg, London, and New York: Springer.
Chatfield, C. (2000). *Time series forecasting*. Boca Raton: Chapman & Hall/CRC.
Davis, H. C. (1995). *Demographic projection techniques for regions and smaller areas*. Vancouver: UBC Press.
Dickey, D. A., Bell, W. R., & Miller, R. B. (1986). Unit roots in time series models: Tests and implications. *American Statistician*, 74, 427–431.
Draper, N. R., & Smith, H. (1981). *Applied regression analysis, Second Edition*. New York: John Wiley & Sons.
Elliot, G., Rothenberg, T. J., & Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, 64(4), 813–836.
Espenshade, T. J., & Tayman, J. (1982). Confidence intervals for post-censal population state estimates. *Demography*, 19(2), 191–210.
Forecasting Division. (2010). April 1 inter-censal and post-censal estimates for the state and counties: 1960-2010. Olympia, WA: Office of Financial Management, State of Washington.
Gabbour, I. (1993). SPOP: Small area population projection. In R. E. Klosterman, R. K. Brail, & E. G. Bossard(Eds.), *Spreadsheet models for urban and regional analy*sis (pp. 69–84). New Brunswick: Rutgers University, Center for Urban Policy Research.
Granger, C. W. (1989). *Forecasting in business and economics*, Second Edition. San Diego: Academic Press.
Granger, C. W., & Newbold, P. (1986). *Forecasting economic time series, Second Edition*. San Diego: Academic Press.
Isserman, A. M. (1977). The accuracy of population projections for subcounty areas. *Journal of the American Institute of Planners*, 43, 247–259.
Jenkins, G. M. (1979). *Practical experiences with modeling and forecasting time series*. Jersey, Channel Islands: Gwilym Jenkins & Partners (Overseas) Ltd.
Keyfitz, N. (1968). *An introduction to the mathematics of population*. Reading: Addison Wesley.
Ljung, G. M., & Box, G. E. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65, 297–303.
Mahmoud, E. (1984). Accuracy in forecasting: A survey. *Journal of Forecasting*, 3, 139–159.
Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1989). F*orecasting: Methods and applications, 3 rd edition*. New York: John Wily & Sons.
McCleary, R., & Hay, R. A. (1980). *Applied time series analysis for the social sciences*. Beverly Hills: Sage Publications.
McDonald, J. (1979). A time series approach to forecasting Australian total live-births. *Demography*, 16(4), 575–601.
McNown, R., & Rogers, A. (1989). Forecasting mortality: A parameterized time series approach. *Demography*, 26, 645–660.
Meyler, A., Kenny, G., & Quinn, T. (1998). Forecasting Irish inflation using ARIMA models Technical Paper Series 3/RT/98. Dublin, Ireland: Central Bank and Financial Services Authority of Ireland.
Montgomery, D. C., Jennings, C. J., & Kulahci, M. (2008). *Introduction to time series analysis and forecasting*. Hoboken: John Wiley & Sons.
Murdock, S. H., & Ellis, D. R. (1991). *Applied demography: An introduction to basic concepts, methods, and data*. Boulder: Westview Press.
Nelson, C. R. (1973). *Applied time series analysis for managerial forecasting*. San Francisco: Holden-Day.
Pearson, K. (1905). The problem of the random walk. *Nature*, 72, 294.
Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, 75, 335–346.
Pielou, E. C. (1969). *An introduction to mathematical ecology*. New York: John Wiley & Sons.
Pittenger, D. B. (1976). *Projecting state and local populations*. Cambridge: Ballinger Publishing Company.

Saboia, J. L. (1974). Modeling and forecasting populations by time series: The Swedish case. *Demography*, 11, 483–492.

Schnaars, S. P. (1986). A comparison of extrapolation models on yearly sales forecasts. *International Journal of Forecasting*, 2, 71–85.

Sieber, G. A., & Wild, C. J. (1989). *Nonlinear regression*. New York: John Wily & Sons.

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

Stock, J. H., & Watson, M. W. (2003). *Introduction to econometrics*. Boston: Addison Wesley.

Swanson, D. A., & Beck, D. (1994). A new short-term county population projection method. *Journal of Economic and Social Measurement*, 20, 25–50.

Tayman, J. Smith, S. K., & Lin, J. (2007). Precision, bias, and uncertainty for state population forecasts: An exploratory analysis of time series models. *Population Research and Policy Review*, 26, 347–369.

Yaffee, R. A., & McGee, M. (2000). *An introduction to time series analysis and forecasting: With applications of SAS and SPSS*. San Diego: Academic Press.

Chapter 7
Housing Unit Method

The housing unit method is based on the fact that almost everyone lives in some type of housing structure, whether a single family unit, an apartment, a mobile home, a college dormitory, or a state prison. Recall that the demographic balancing equation is an exact identity of population change (see Chapter 3). In a similar vein, the housing unit method provides an exact determination of the total population; any error is due to inaccuracies in estimates of its elements, not an inherent flaw in the method itself (Lowe, Pittenger, and Walker 1977; Swanson, Baker, and Van Patten 1983). It is one of the most widely used techniques for subnational population estimates (Bryan 2004b: 550; Byerly 1990). One reason for the wide spread use of the housing unit method is it can be applied at virtually any level of geography, especially at detailed spatial resolutions (Jarosz 2008; Tayman 1994). Second it can accommodate a variety of data sources and application techniques (Lowe, Myers, and Weisser 1984; Smith and Cody 2004). Finally, the housing unit method can produce estimates that are at least as accurate as other post-censal estimation techniques (Hoque 2010; Smith 1986; Smith and Mandell 1984; Starsinic and Zitter 1968).

This chapter begins by illustrating the general framework of the housing unit method for estimating post-censal population. We show two methods. One requires a post-censal estimate of the vacancy rate and the other provides a direct estimate of households. The four subsequent sections discuss data and procedures for estimating the individual elements of the housing unit method; namely, housing units, vacancy rates, persons per household, and group quarters population. The chapter ends with some concluding remarks about the housing unit method.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_7,

7.1 Components of the Housing Unit Method

7.1.1 Population

The housing unit method relies on the straightforward assumption that nearly all people sleep under some kind of shelter. Within this general framework the post-censal population estimate of a given place at time (t) is:

$$P_t = (HU_t * OCCR_t) * PPH_t + GQ_t,$$

where P is population, HU is housing units (vacant and occupied units); OCCR is occupancy rate (compliment of the vacancy rate); PPH is average number of persons per household; and GQ is group quarters population. A common way to estimate housing unit change is by permit data on housing units built, demolished, or annexed since the last census (c):

$$\Delta HS_{c\ to\ t} = NHS_{c\ to\ t} - DHS_{c\ to\ t} \pm AHS_{c\ to\ t},$$

where NHS is new housing units; DHS is demolished housing units; AHS is housing units lost or gained due to annexations; and c to t is the time between the last census and the post-censal estimate date (t)[1]. The post-censal estimate of housing units is then given by:

$$HS_t = HS_c + \Delta HS_{c\ to\ t}.$$

The housing unit method contains both stocks and flows using the formulation shown in the above three equations. The flow component is used to estimate the change in housing units over the post-censal period, and stock components are used to estimate households and household population at the post-censal time point.

These equations represent the basic formation of the housing unit method. The most used refinement is to disaggregate the method by housing structure type (s):

$$P_t = \sum_s ((HS_{c,s} + \Delta HS_{c\ to\ t,s}) * OCCR_{t,s} * PPH_{t,s}) + GQ_t.$$

As discussed in the next section, using building permits rather than certificates of occupancy requires assumptions about the lag between permit issuance and when the units are ready for occupancy. Using a single lag time for all building types may produce inaccuracies since the lag time for multiple family structures is usually longer than for single family units (Smith 1986). Occupancy rates and PPH values differ by structure type and using overall rates can lead to errors if the mix of housing units has changed significantly in the post-censal period. Aside from these methodological advantages, information on the dynamics of housing markets is often more useful when disaggregated by housing type (Myers and Doyle 1990).

Table 7.1 Housing Unit Method: Building Permits Olympia, Washington, April 1, 2010

	Housing Units					
	Single-Family	2 Units	3–4 Units	5+ Units	Other[a]	Total
Apr. 1, 2000	11,089	758	1,132	5,907	852	19,738
2000 to 2010[b]						
Completions[c]	1,330	40	124	154	n/a	1,648
Demolitions	117	8	0	5	n/a	130
Annexations	172	18	36	11	n/a	237
Net Change	1,385	50	160	160	−33	1,722
Apr. 1, 2010	12,474	808	1,292	6,067	819	21,460
Occupancy Rate	0.95027	0.92625	0.87523	0.88988	0.86813	
Households	11,854	748	1,131	5,399	711	19,843
Persons per HH	2.5454	1.7172	2.1568	1.6660	1.6948	
Household Pop	30,174	1,284	2,439	8,995	1,205	44,097
Group Quarters	Nursing[d] Homes	Dorms	Mental[e] Health	Military	Other	Total
	571	0	410	0	419	1,400
Total Population						**45,497**

[a]Mobile homes, trailers, and other units
[b]Apr. 1, 2000 to March 31, 2010
[c]Certificates of occupancy
[d]Includes convalesence facilities
[e]Includes corrections facilities
Source: State of Washington, Office of Financial Management, Forecasting Division, March 2, 2011

Table 7.1 shows April 1, 2010 population estimates for the City of Olympia, Washington made with the disaggregated housing unit method using building permit data. The estimated vacancy rates and PPH values vary substantially by structure type. Single family units have the highest occupancy rates and PPH. Occupancy rates range from 0.868 for other units to 0.950 for single family units and PPH values range from 1.67 for structures with 5 or more units to 2.54 for single family units. Total housing units, households, and household population are determined using a bottom up approach by summing over the structure type, preserving the dynamics reflected in the occupancy rates and PPH values.

Studies have found that better estimates of households can be made from electric customer (EC) data than from building permit information (Starsinic and Zitter 1968; Smith and Cody 2004; Smith and Lewis 1980, 1983). EC data are often of better quality than building permit data and households can be directly estimated, eliminating the intermediate steps of estimating housing unit change and occupancy rates. Starsinic and Zitter (1968) proposed the change in electric customers as an indicator of the net change in households:

$$HH_t = HH_c + (EC_t - EC_c),$$

where HH is the number of households; EC is the number of electric customers; c is the last census; and t is the post-censal time point. This technique assumes a one-to-one correspondence between households and electric customers.

Table 7.2 Household Estimates using Residential Electric Customers, Sarasota County Jurisdictions, April 1, 2010

	2000		Electric Customers			2010 Households	
Jurisdiction	Households	Ratio[a]	2000	2010	2000–2010	Difference[b]	Ratio[c]
Longboat Key	4,280	0.71836	5,958	5,970	12	4,292	4,289
NorthPort	9,111	0.88914	10,247	25,606	15,359	24,470	22,767
Sarasota	23,427	0.89086	26,297	26,644	347	23,774	23,736
Venice	9,680	0.72537	13,345	16,402	3,057	12,737	11,898
Unincorporated Area	103,439	0.80857	127,929	144,232	16,303	119,742	116,623
Sarasota County	149,937	0.81587	183,776	218,854	35,078	185,015	178,557

[a]2000 Households divided by 2000 electric customers
[b]2000 households + 2000-2010 change in electric customers
[c]2000 Ratio * 2010 electric customers
Sources: Bureau of Economic and Business Reseach, University of Florida
US Census Bureau, 2000 Decennial Census

A major issue with the difference approach is there is often not a one-to-one correspondence between households and electric customers. Seasonal residents may occupy housing units; master meters can serve more than one household; and separate meters may be installed non-housing uses (Smith and Cody 2004). To overcome these issues, households can be estimated using a censal ratio approach:

$$HH_t = HH_c/(EC_c)^* EC_t.$$

The ratio of households to electric customers is computed at the time of the latest census. It is often held constant over the post-censal period, as it appears to remain stable in many places (Smith 1986). If, for example, the seasonal and permanent populations are growing at different rates, it may be useful to change the ratio over the post-censal period. Extrapolation of the ratio from previous censuses and professional judgment can be used to adjust the ratio up or down. A study in Florida found household estimates from the censal ratio method were more accurate than estimates based on the difference in electric customers (Smith and Lewis 1983).

Table 7.2 shows April 1, 2010 household estimates for jurisdictions in Sarasota County, Florida using the difference and censal ratio methods applied to EC data. The difference method yields higher estimates for all places. The estimates are similar between the methods for Longboat Key and Sarasota, areas with little change in electric customers over the decade, but diverge more for Venice, North-Port and the unincorporated area. In NorthPort, the gain in electric customers is 68.5% larger than the number of households in 2000. Its censal ratio estimate is 7% lower than the estimate based on the change, similar to the difference seen in Venice. For the unincorporated area, the censal ratio estimate is 2.5% lower than the estimate based on the difference approach. The finding of a greater upward bias in the difference approach is consistent with the results from other studies (Smith and Lewis 1980, 1983).

Examining the 2000 censal ratio illustrates the lack of a one-to-one relationship between households and EC. All ratios are below one, ranging from 0.718 in Longboat Key to 0.891 in Sarasota. The effect of seasonal units is most clearly

seen for Longboat Key. In 2000, 92% of its vacant units were classified as vacant for seasonal, recreational, or occasional use. The corresponding figures for NorthPort, Sarasota, and the unincorporated area were 66%, 34%, and 61%.

7.1.2 Housing Units

Building permits, certificates of occupancy, and electric customers are most often used as symptomatic indicators of post-censal housing unit change (Smith 1986). Administrative records such as property tax files, voter registration, postal address lists, and aerial photographs can also be used to track housing unit change. All these indicators have strengths and weaknesses that center on their ability to accurately measure housing unit change and the resources required to collect, analyze, and integrate them into the housing unit method framework.

7.1.2.1 Building Permit and Completions

The most comprehensive source of building permit data is Building Permits Survey conducted by the Census Bureau (see Chapter 3). These data cover permits for new construction by type of unit for every county and jurisdiction in the US, are available monthly and in an annual aggregation, and are easily accessible via the Internet.[2] The Census Bureau's residential permit data no longer include demolitions, demolitions were last collected in 1995, and these data do not cover mobile homes and other units.

Building permit information can also be directly obtained from the local agency granting the permits. Local agencies are likely to be the best source for subcity-level permit information, and for most agencies issuing permits the quality is generally good. Permits obtained local agencies are subject to inconsistencies including variations in definitions, data format, and geographic accuracy (Jarosz 1998). Many jurisdictions only provide hard-copy information which can introduce additional error and require additional resources for data entry and verification. Errors in geocoding permits to subcity areas are not uncommon and identification of structure type can be ambiguous, especially the classification of 1-unit attached structures. Some local agencies provide a permit record for multi-unit buildings, but do not indicate the number of units. Some agencies provide only the permits authorized, while others provide both authorized and completed housing units.

While most areas of the county require building permits, some small towns and sparsely populated rural areas do not issue them. Building permits for mobile homes are often of questionable quality (Smith 1986). These problems include issuing permits for spaces in parks rather than for the mobile home unit, double counting when ownership changes, and not tracking mobile homes. Building permits are frequently issued for remodels and garages and other external structures and can be included with permits for new units. Perhaps the most difficult problem with building permit data is they represent the intention to build. Some units get built quickly, others after a long delay, and still others never get built. To address this

Table 7.3 Comparison of Building Permits and Completions Olympia, Washington, Apr. 1, 2000 to March 31, 2010

	Single-Family	2 Units	3–4 Units	5+ Units	Total
Permits					
Local	1,324	66	183	168	1,741
Census Bureau	1,459	52	241	307	2,059
Completions[a]	1,330	40	124	154	1,648
Ratio					
Permits[b]	0.907	1.269	0.759	0.547	0.846
Permits and Compl.[c]	1.005	0.606	0.678	0.917	0.947

[a] Certificates of occupancy
[b] Local permits / Census Bureau Permits
[c] Completions / Local Permits
Sources: State of Washington, Office of Financial Management, Forecasting Division, March 2, 2011
US Census Bureau, Censtats

issue, lagged times are used to represent an assumed time frame for completion (Smith and Cody 2004). Using lagged times still assumes that every permit will eventually get built. The likelihood for eventual construction is influenced by housing type, housing market conditions, and geographic area. To our knowledge, this information, if available at all, would have to come from a specific local agency. Certificates of occupancy or completions solve these problems because they are issued only after the housing unit is built and ready for occupancy. Completion data may provide a more accurate assessment of housing unit change than building permits, but completions are not as widely available.

Table 7.3 provides a comparison of permits and certificates of occupancy for new housing units from a local agency and permits from the Census Bureau issued between 2000 and 2010 for Olympia, Washington. There is considerable variation between the permits data collected in the Census Bureau survey and those from the local agency. The difference between the local agency and Census Bureau data is the smallest for single family units; the local agency reported 10 percent fewer permits. More pronounced differences are seen for multiple family units, especially for structures with 5+ units.

The comparison of permits and completions from the local agency show a close correspondence with single family units, but proportionately fewer multiple family units actually got built. During the 2000s, Olympia added 2,348 housing units according to the decennial censuses. All symptomatic indicators understate this change, with undercounts ranging from 14% to 30%. These results, while illustrative in nature, do indicate the potential uncertainty and variability in using building permit data to estimate housing change.

7.1.2.2 Electric Customers

While the quality of EC data can vary from company to company, large companies generally have very accurate information, sometimes separately by housing structure type (Smith 1986). EC data can often be obtained for spatially-detailed areas

(e.g., census tract) and even by address, facilitating subcity-level estimation (Rynerson and Tayman 1998; Tayman 1994). Address-level information is especially valuable. In conjunction with a GIS procedure known as Admatch, an address can be assigned to any spatial location (see Chapter 2). EC data also offer one-stop shopping by covering multiple agencies, saving the considerable time and effort required to collect data from these agencies individually. For example, Florida has 53 power companies that cover its roughly 400 cities and towns and 67 counties.

EC data do have some potential drawbacks. Companies may not be willing or make it very difficult to obtain their customer information. EC data may not always distinguish between active and inactive meters or between residential and non-residential customers (Smith 1986). Replacement meters can lead to a double count, if the record for the original record is maintained. Master-metered apartments have more than one dwelling unit serviced by a meter, and the dwelling unit counts are not always provided. EC data may not accurately account for demolitions or unit conversions and may not distinguish seasonal and non-seasonal households.

7.1.2.3 Parcel Files

Administrative records such as assessor's property tax (parcel) file may provide very useful information for estimating housing units. The availability of GIS software, robust desktop and server computing systems, data management capabilities, and digitization of administrative records make using parcel files feasible. Parcel files are increasing being incorporated into small-area data systems and can form a foundation for estimating housing units in the post-censal period (e.g., Brown 1999; Fairfax County 2010: Appendix A; Jarosz 2008; Puget Sound Regional Council 2009). Parcel files often contain the most consistent and up-to-date information on housing units. San Diego County, for example, provides parcel file updates on a quarterly basis. In addition to the number of units, parcel files contain other housing attributes such as structure type, age of structure, and assessed value. Like EC data, a parcel file offers one-stop shopping and provides very detailed spatial resolution. Parcels are the atomic unit of development and this fine level of geographic granularity makes it easier to review, find, and correct errors compared to aggregate data (Jarosz 2008). A study of San Diego County cities and census tracts found that housing unit estimates based on the parcel file had less bias and lower absolute percent errors than estimates based on EC data (Rynerson and Tayman 1998).

The quality of parcel files can vary a great deal from place to place. Some agencies do not have electronic records or may not be willing to provide data. Non-taxable housing units such as those on federal, state, and tribal lands will not be in a parcel file. The quality of housing unit counts on parcels containing multiple family structures can vary considerably and determining units lost may be problematic. Reconciling parcel records and housing unit census is labor intensive, requires a significant commitment of time and resources, and requires ancillary data sources (Jarosz 2008). Definitional differences between the census and assessor rules can be

particularly problematic. For example, the census counts time shares as residential housing units while the assessor in San Diego counts them as non-residential structures.

Little is known about the accuracy of housing unit estimates based on parcel files. Table 7.4 compares 2010 parcel file-based housing unit estimates with 2010 census housing unit counts for jurisdictions in San Diego County. The estimates for 2010 have a modest bias and relatively low absolute percent errors. The MALPE and MAPE (based on the last column and its absolute values) are -1.6% and 1.6%. Average absolute percent errors across jurisdictions range from 0.0% to 5.8%, and the estimates are at or above the census in only Oceanside and Lemon Grove. A less favorable picture emerges when examining the error of the housing unit change. The parcel file missed 12.3% of the housing unit change in San Diego County. Even after discounting jurisdictions with relatively little housing change, the parcel file systematically underestimates housing unit change, and sometimes by large amounts. These results, while illustrative in nature, do indicate the potential for parcel file information to deviate from housing unit trends.

7.1.2.4 Other Data Sources

The preceding sources are the most commonly used data to estimate post-censal housing units, but other data also can track housing unit change including telephone customer data and water and gas utility information. The correspondence between these data sources and housing unit change is generally not as close the other data sources described above (Smith 1986). Many housing units have no phones, especially since the advent of cell phones, or unlisted numbers; do not require gas or use bottled gas; or obtain water from wells.

Aerial photography can ascertain changes in housing units, but it tends to more effective for very specific geographic areas rather than for broad-based estimation. Aside from its expense, estimating housing units through aerial photography is very time consuming and labor intensive. Unit counts derived from this source can be ambiguous. It is relatively straightforward to identify single family detached houses, but it is much more difficult to determine unit counts in multiple family structures. With aerial photography it is not always evident whether a building represents a housing unit or some other type of non-housing use and it is impossible to identify housing units in mixed used projects (e.g., retail on the first floor and housing lofts on the remaining floors).

Despite these issues, aerial photography is useful visual validation and external checking of housing unit estimates from other sources, especially for very small geographic areas, such as block groups and parcels (Jarosz 2008). Figures 7.1 and 7.2 contain examples of aerial photography, taken close to the 2000 census date, used for ground truthing in San Diego County. In Figure 7.1, block 1074 clearly shows no units, but the census counts 10 units. Census counts show 332 housing units for blocks (1003-1012) in the bottom left of Figure 7.2, but the parcel file has no units. This area, formerly a navy base, was converted to a private-sector mixed use project. Apparently the ownership change and attributes were never made to the

Table 7.4 Accuracy of Parcel-Based Housing Unit Estimates, San Diego County Jurisdictions, 2010

Jurisdiction	2010			Change 2000–2010		Error 2000–10 Change		Pct.Error
	2000	Census[a]	Estimate[b]	Census	Estimate	Number	Percent	2010 Est.
Carlsbad	33,798	44,673	43,844	10,875	10,046	−829	−7.6	−1.9
Chula Vista	59,495	79,416	78,244	19,921	18,749	−1,172	−5.9	−1.5
Coronado	9,494	9,634	9,562	140	68	−72	−51.4	−0.7
Del Mar	2,557	2,596	2,542	39	−15	−54	−138.5	−2.1
El Cajon	35,190	35,850	35,644	660	454	−206	−31.2	−0.6
Encinitas	23,843	25,740	24,877	1,897	1,034	−863	−45.5	−3.4
Escondido	45,050	48,044	47,682	2,994	2,632	−362	−12.1	−0.8
Imperial Beach	9,739	9,882	9,860	143	121	−22	−15.4	−0.2
La Mesa	24,943	26,167	25,614	1,224	671	−553	−45.2	−2.1
Lemon Grove	8,722	8,868	8,868	146	146	0	0.0	0.0
National City	15,422	16,762	15,787	1,340	365	−975	−72.8	−5.8
Oceanside	59,581	64,435	64,758	4,854	5,177	323	6.7	0.5
Poway	15,714	16,715	16,364	1,001	650	−351	−35.1	−2.1
San Diego	469,689	516,033	511,820	46,344	42,131	−4,213	−9.1	−0.8
San Marcos	18,862	28,641	27,744	9,779	8,882	−897	−9.2	−3.1
Santee	18,833	20,048	19,837	1,215	1,004	−211	−17.4	−1.1
Solana Beach	6,456	6,540	6,521	84	65	−19	−22.6	−0.3
Vista	29,814	30,986	30,716	1,172	902	−270	−23.0	−0.9
Unincorporated Area	152,947	173,756	169,142	20,809	16,195	−4,614	−22.2	−2.7
San Diego County	1,040,149	1,164,786	1,149,426	124,637	109,277	−15,360	−12.3	−1.3
							MAPE[c]	1.6
							MALPE[d]	−1.6

[a]April 1, 2010
[b]January 1, 2010
[c]Mean absolute percent error
[d]Mean algebraic percent error
Sources:
US Census Bureau, 2000 and 2010 Decennial Censuses
San Diego Association of Governments, 2010 Population Estimates

Fig. 7.1 Ground Truthing Using Aerial Photography: Census Problem
Source: Jarosz (2008)

Fig. 7.2 Ground Truthing Using Aerial Photography: Parcel File Problem
Source: Jarosz (2008)

parcel file. To-date aerial photography has played a limited role in the estimation of housing units; but perhaps its efficacy may increase going forward (e.g. Lo 1995; Wicks, Swanson, Vincent, and De Almeida 1999; Wang and Wu 2010). We discuss this possibility in Chapter 18.

7.1.3 Occupancy Rates

Occupancy rates are needed to convert the post-censal housing stock estimate into an estimate of households, or occupied units. The occupancy rate is the compliment of the more widely presented vacancy rate. The occupancy rate can be just for all units. If the post-censal housing stock is estimated for separate structure types, estimates of occupancy rates specific to each structure type should be used. Occupancy rates are typically highest for single family units and tend to decrease with increases in the number of units in a structure; mobile homes often have the lowest occupancy rate of any housing unit type (see Table 7.1).

A common practice is to hold the vacancy rate constant at the last census (e.g., US Census Bureau 2009, 1983). This is a reasonable assumption if the economic and housing market conditions at the time of the census do not change and is likely more valid when the estimate date is relative close to the last census. Even if market conditions are stable, using the occupancy rate from the last census can be problematic in areas with substantial housing unit growth around the time of the census or that have experienced significant boundary changes. Census occupancy rates in these areas tend to be low because of the time required to sell these units and also because the census may count a housing unit before it is ready for occupancy. In this situation, it is likely more accurate to raise increase the occupancy rate in the first few years of the post-censal period, before holding it constant.

Models of housing supply and demand have been developed primarily in the field of housing economics (e.g., Edelstein and Tang 2007; Gabriel and Nothaft 2001; Hendershott, MacGregor and Tse 2002). These models attempt to explain fluctuations in occupancy rates using statistical models and factors such as housing supply and demand, home prices, rents, and population and employment change. Their primarily aim is to provide a better understanding of housing market dynamics rather than providing specific estimates of occupancy rates for use in the housing unit method.

Information on post-censal occupancy rates can also be estimated from direct survey methods such as the "windshield survey" or direct canvassing of an area (e.g., Alaska Department of Community and Economic Development 2004; Lowe, Pittenger, and Walker 1977; Swanson, Baker, and Van Patten 1983). These methods are very labor intensive, time consuming and are most practical for areas characterized by single family units and relatively small population size.

7.1.3.1 American Community Survey (ACS)

The American Community Survey (ACS) offers the most comprehensive data platform for estimating post-censal occupancy rates. Since there is no longer a census long form, occupancy rates (and persons per household) by structure type will come from ACS data. An important issue in considering the ACS as a source of occupancy rate data is it does not have the same residency definition as the decennial census; the ACS uses essentially a de facto residency rule and the decennial census uses a de jure residency rule (See Chapter 3). While the difference in rules may not matter nationally or for states, it could be substantial at sub-state levels, especially for areas with large numbers of seasonal units.

The ACS controls housing units and household characteristics to post-censal estimates of housing units and controls population-related variables to population estimates stratified by age, sex, Hispanic origin, and race. Both the housing unit and population controls are subject to random error and errors that are systematically biased either upward or downward based on characteristics of the controlling unit such as population size and growth rate. Housing unit controls have been found to have less error than the population controls stratified by age, sex, and Hispanic Origin (Citron and Kalton 2007: Chapter 5), which suggests that point estimates for ACS occupancy rates may be more accurate than point estimates for ACS persons per household variables.

To illustrate ACS estimates of occupancy rates, we compare ACS rates with the 2010 census rates for incorporated cities in San Diego County, California. Three ACS rates are compared based on the: 1) 2009 annual estimates; 2) estimates based on the 3-year period 2007 to 2009; and 3) estimates based on the 5-year period from 2005 to 2009.[3] We also show annual ACS data for Maricopa County, Arizona from 2000 to 2009, along with the 2000 and 2010 census occupancy rates.

The ACS occupancy rate estimates are generally in line with the 2010 census for jurisdictions in San Diego County (see Table 7.5). Average absolute percent differences are between 1.5% and 2.0%, and the ACS shows a small downward bias that ranges from an average of -0.5% to -1.4%. Also, the estimates generally do not vary greatly between the three ACS samples. The greatest variability occurs in Carlsbad and El Cajon. In Carlsbad, both multi-year accumulations are very close to the census and the 2009 annual estimate is 2.4% above. The opposite pattern occurs in El Cajon; its 2009 estimate is very close to the census and both multi-year accumulations are lower than the census by around three percent.

The occupancy rate trend for Maricopa County appears reasonable; although the estimate for 2003 seems out of line with the other data points (see Figure 7.3). The ACS seems to have picked up the impact of the housing meltdown; the occupancy rate shows a substantial drop between 2006 and 2008. While the 2010 census rate is outside of the 90% confidence interval for the 2009 estimate, the 2009 point estimate and census value are close.

Obviously, we cannot draw any firm conclusions or generalizations about the efficacy of ACS occupancy rates from these comparisons. For these areas, the ACS

Table 7.5 ACS and Decennial Census Occupancy Rates, San Diego County Incorporated Cities

	2010	American Community Survey			% Difference from the Census		
	Census	2009	2007–09	2005–09	2009	2007–09	2005–09
Carlsbad	0.926	0.948	0.925	0.920	2.4	−0.1	−0.6
Chula Vista	0.951	0.886	0.889	0.905	−6.8	−6.5	−4.8
Coronado	0.769	–	0.807	0.804	–	4.9	4.6
Del Mar	0.795	–	–	0.780	–	–	−1.9
El Cajon	0.952	0.953	0.922	0.919	0.1	−3.2	−3.5
Encinitas	0.936	–	0.930	0.926	–	−0.6	−1.1
Escondido	0.947	0.952	0.934	0.941	0.5	−1.4	−0.6
Imperial Beach	0.922	–	0.875	0.889	–	−5.1	−3.6
La Mesa	0.937	–	0.931	0.937	–	−0.6	0.0
Lemon Grove	0.951	–	0.945	0.937	–	−0.6	−1.5
National City	0.925	–	0.910	0.919	–	−1.6	−0.6
Oceanside	0.919	0.918	0.911	0.912	−0.1	−0.9	−0.8
Poway	0.965	–	0.952	0.965	–	−1.3	0.0
San Diego	0.936	0.921	0.923	0.927	−1.6	−1.4	−1.0
San Marcos	0.950	0.951	0.954	0.953	0.1	0.4	0.3
Santee	0.963	–	0.946	0.963	–	−1.8	0.0
Solana Beach	0.864	–	–	0.856	–	–	−0.9
Vista	0.946	0.956	0.927	0.941	1.1	−2.0	−0.5
San Diego County	0.933	0.918	0.917	0.924	−1.6	−1.7	−1.0
				MAPD[a]	1.6	2.0	1.5
				MALPD[b]	−0.5	−1.4	−0.9

[a]Mean absolute percent difference
[b]Mean algebraic percent difference
Sources: US Census Bureau, 2010 Census; American Communty Survey 2009, 2007–2009, and 2005–2009

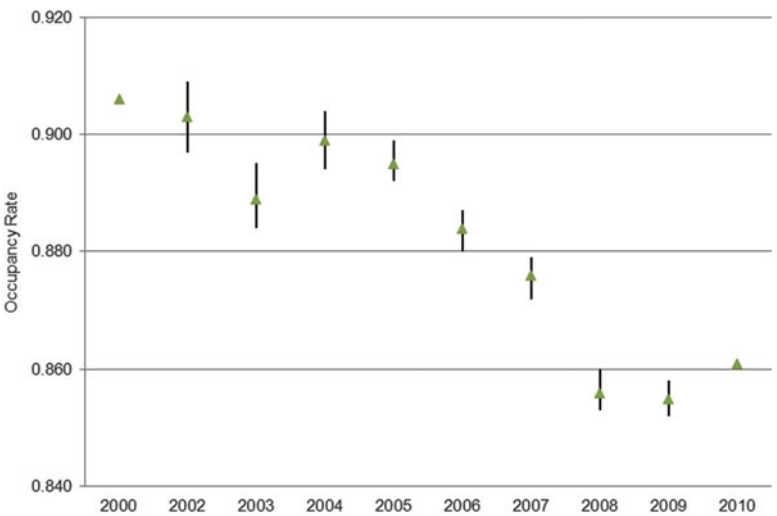


Fig. 7.3 Occupancy Rates, Maricopa County, 2000–2010
Note: Line endpoints represent the limits of a 90% confidence interval
Sources: US Census Bureau, 2000 and 2010 Decennial Censuses; 1-yr ACS, 2002–2009

estimates of occupancy rates are quite good and do not appear to be impacted by the large number of seasonal units in Maricopa County; according to the 2009 ACS 22.7% (52,300) of the vacant units (230,145) in Maricopa County are classified for seasonal, recreational, or occasional use. Much work is needed to determine the efficacy of ACS occupancy rate estimates and how they best can be used in the housing unit method. Some suggestions for future analysis include: 1) testing a broad-based a representative sample at various levels of geography; 2) examining occupancy rates by structure type; 3) testing the ACS against other sources of data and methods for estimating post-censal occupancy rates; and 4) providing guidelines for using the ACS occupancy rate estimates.

7.1.3.2 Postal Service Deliveries and Real Estate Vacancy Surveys

Postal service deliveries and real estate vacancy surveys may be useful for updating occupancy rates during the post-censal period. Two important considerations in using these data are the definitions of a housing unit and a vacant unit, which differ from the definitions used in the decennial census. Postal statistics do not have a distinct definition of vacant; they classify deliveries as Possible and Active. Possible deliveries include Active deliveries and their difference is Inactive deliveries, which are roughly defined as vacant. Active deliveries loosely represent housing units; although they represent any address where mail is delivered (i.e., post-office boxes). Lowe and Mohrman (2003a) contain additional details on other differences between the census and postal delivery data.

Real estate vacancy surveys use non-random sampling procedures; rely on data provided by apartment managers; and cover only apartments on the rental market. Thus, they represent a potentially biased subset of multiple family unit types covered in the census. Additionally, units rented for temporary use and under renovation are counted as occupied. Lowe (2000a) contains additional details on other differences between the census and apartment vacancy survey data.

Lowe and Mohrman (2003a) and Lowe (2000a) compared postal delivery data and real estate vacancy surveys to census data for counties and cities in Washington State. These studies found that:

1. Real estate vacancy rates were on average five percent points lower than 1990 census rates for a sample of 15 cities in the Seattle area, ranging from -8.3% to 0.6%;
2. Postal statistics for all deliveries were the best match to the housing counts in the 2000 census. Postal counts excluding post office boxes fell notably short in nearly all counties. The MAPE across counties was 12.6% for all deliveries and 24.5% when post office boxes were excluded;
3. Postal data substantially understated 2000 census vacancy rates by an average of 11 percentage points across counties. Postal and census vacancy rates were close

in metropolitan counties with an average difference of two percentage points versus an average of 12 percentage points for non-metropolitan counties. The largest differences between census and postal vacancy rates were found in counties with substantial seasonal housing; and

4. Postal vacancy rate trends tracked what might be expected based on changes in the economy and population. Postal rates showed the largest increases in counties with struggling economies and net out migration of population.

Because of differences in coverage and definitions with the census, postal data and real estate surveys should be used to adjust census rates rather than as direct estimates of the post-censal occupancy rates (Lowe and Mohrman 2003b). How can these adjustments be made? The procedures discussed above for converting electric customer data into post-censal household estimates can also be used with postal delivery and real estate survey data. Given the lack of a one-to-one correspondence with census occupancy rates, the ratio procedure might be preferable. Another option is to model the census occupancy rate as a function of either the postal delivery rate or the real estate survey rate.

Tayman and Rynerson (1997) demonstrated the viability of this modeling approach for estimating post-censal income distributions for census tracks in San Diego County. The household income distribution is characterized by three parameters, which determine the probabilities of households locating in each income group. These parameters were estimated by fitting a modified lognormal curve to census household income distributions (Fonseca and Tayman 1989). Federal and state tax returns from the California Franchise Tax Board (FTB) provided personal income distributions for census tracks at a census time point, which were used to estimate the same three parameters. Three census-tract level regression models were estimated using the census parameter as the dependent variables and the corresponding FTB parameter as the independent variable. Post-censal estimates for the three parameters were derived using the estimated regression model from the census time point and census-tract level post-censal values for the parameters obtained from the calibration of post-censal FTB data.

7.1.4 Persons Per Houshold

All applications of the housing unit method require an estimate of the average number of persons per household (PPH) to convert a post-censal household estimate into a household population estimate. The PPH can be just for all units. If post-censal households are estimated for separate structure types, estimates of PPH specific to each structure type should be used. PPH is typically highest for single family units and lower for multiple family housing and mobile homes (see Table 7.1).

Small shifts in PPH in moderate to large cities can have a dramatic impact on population estimates based on the housing unit method (Lowe 2000b).

Information on post-censal PPH can be based on data from special censuses conducted by the Census Bureau (US Census Bureau 2011). Post-censal estimates of PPH are computed for places taking a special census and then are used in estimate PPH is places with similar characteristics (Lowe, Pittenger and Walker 1977). A large scale, continuous special census program proved successful in Washington State, but special censuses are costly. Today they are primarily used with a goal of obtaining a higher population count for revenue sharing and other benefits. Sample surveys can also be used to estimate PPH, but to provide accurate estimates, especially for subcounty areas, samples must be accurately drawn and quite large. The costs of such surveys generally preclude their use for estimating PPH. Trends in PPH for large states, regions, and the US from the Current Population Survey have been used to estimate local overall PPH and PPH by structure type (Smith 1986; Smith and Lewis 1980).

7.1.4.1 American Community Survey (ACS)

Swanson and Hough (2007) evaluated PPH values from the ACS as to their suitability for use in the housing unit method. ACS PPH values were compared to estimated PPH values based on geometric trend extrapolation for 18 of the 36 counties that were 1999 ACS test sites. These 18 sites had ACS data online annually from 2001 to 2006 and for five three-year periods between these dates. Single-year ACS PPH values exhibited the least systematic change over time and considerable directional volatility from year to year; three counties had directional changes three or more times, two changes in nine counties, and one change in six counties. Three-year ACS PPH values were more directionally stable over time; two counties changed direction twice, one change in seven counties, and no directional change in nine counties.

This volatility in the ACS PPH estimates is not desirable for those making population estimates. There is an expectation of demographers and their stakeholders that PPH estimates exhibit systematic changes unless there is compelling, substantive evidence to justify their temporally instability. ACS PPH values are subject to sample and non-sample errors and can vary year to year because of statistical fluctuations. Swanson and Hough (2007) offer a preliminary conclusion that ACS PPH values exhibit too little systematic change over time to be usable by demographers preparing post-censal population estimates. They point out that additional research, along the lines discussed previously for occupancy rates, is required to confirm this finding and to determine how best to use ACS PPH values for post-censal estimation.

To illustrate ACS estimates of PPH, we compare ACS rates with the 2010 census rates for incorporated cities in San Diego County and over time for Maricopa County, as was done above for occupancy rates. ACS PPH values line up fairly well with the census for jurisdictions in San Diego County (see Table 7.6). Average absolute percent

Table 7.6 ACS and Decennial Census Persons per Household, San Diego County Incorporated Cities

	2010	American Community Survey			% Difference from the Census		
	Census	2009	2007–09	2005–09	2009	2007–09	2005–09
Carlsbad	2.525	2.522	2.573	2.536	−0.1	1.9	0.4
Chula Vista	3.207	3.299	3.267	3.174	2.9	1.9	−1.0
Coronado	2.307	–	2.304	2.270	–	−0.1	−1.6
Del Mar	2.016	–	–	2.028	–	–	0.6
El Cajon	2.842	2.961	2.925	2.883	4.2	2.9	1.5
Encinitas	2.450	2.523	2.532	–	3.0	3.4	
Escondido	3.117	3.190	3.121	3.095	2.3	0.1	−0.7
Imperial Beach	2.821	–	2.806	2.678	–	−0.5	−5.1
La Mesa	2.301	–	2.378	2.290	–	3.3	−0.5
Lemon Grove	2.961	–	2.921	2.766	–	−1.4	−6.6
National City	3.408	–	3.458	3.396	–	1.5	−0.3
Oceanside	2.805	2.729	2.818	2.801	−2.7	0.5	−0.1
Poway	2.930	3.001	2.997	–	2.4	2.3	
San Diego	2.599	2.654	2.623	2.612	2.1	0.9	0.5
San Marcos	3.049	3.183	3.105	3.105	4.4	1.8	1.9
Santee	2.717	–	2.796	2.822	–	2.9	3.9
Solana Beach	2.277	–	–	2.279	–	–	0.1
Vista	3.131	3.129	3.213	3.130	−0.1	2.6	0.0
San Diego County	2.754	2.815	2.790	2.760	2.2	1.3	0.2
				MAPD[a]	2.4	1.7	1.7
				MALPD[b]	1.6	1.5	−0.1

[a]Mean absolute percent difference
[b]Mean algebraic percent difference
Sources: US Census Bureau, 2010 Census; American Communty Survey 2009, 2007–2009, and 2005–2009

differences are between 1.7% and 2.4%, and the ACS shows an upward bias of around 1.5% for the 1-year and 3-year estimates. The 5-year ACS PPH estimates are unbiased.

ACS PPH estimates, on average, have larger absolute percent errors than the occupancy rate estimates for the 1-year and 5-year samples (2.4% vs. 1.6% and 1.5% vs. 1.7%), but are more accurate for the 3 year sample (1.7% vs. 2.0%). ACS PPH estimates, on average, show a greater bias than the occupancy rate estimates in the 1-year sample (1.6% vs. -0.5%); a similar level of bias (opposite in sign in the 3-year sample (-1.5% vs. -1.4%), and less bias in the 5-year sample (-0.1% vs. -0.9%). To examine the variability of the ACS estimates across the ACS samples, we examined the range of the absolute percent errors in jurisdictions with two or more ACS samples. ACS PPH estimates show greater variability than the ACS occupancy rate estimates based on the range. The average of the range for the ACS PPH estimates is 2.6 with a low of 0.1 and a high of 5.2. For the ACS occupancy rate estimates, the average is 1.2 with a low of 0.3 and a high of 3.4.

The ACS PPH trend is Maricopa County is not reasonable (see Figure 7.4). PPH values remain relatively stable at around the 2000 census value from 2002 to 2005; the statistical fluctuations during this time are clearly evident. The PPH estimate

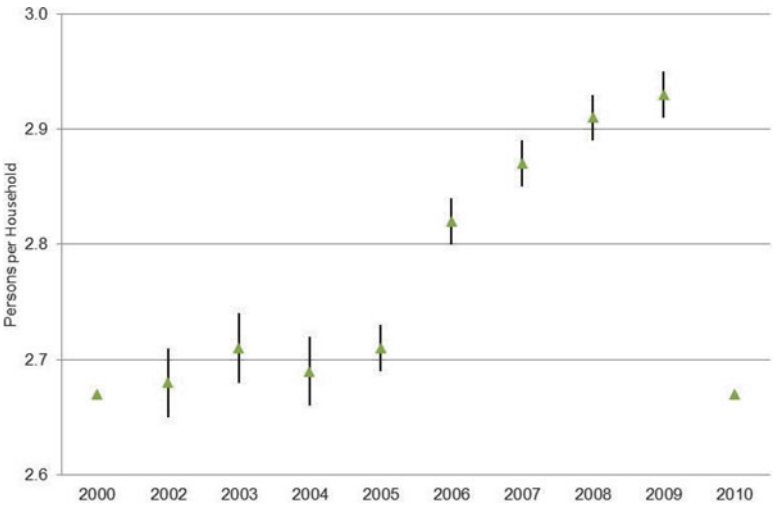


Fig. 7.4 Persons per Household, Maricopa County, 2000–2010
Note: Line endpoints represent the limits of a 90% confidence interval
Sources: US Census Bureau, 2000 and 2010 Decennial Censuses; 1-yr ACS, 2002–2009

show a sharp increase after 2005; rising from 2.71 to 2.93 in 2009. This dramatic increase was incorrect, the 2010 census PPH showed a stable pattern since 2000, and illustrates the errors that can occur when controlling the ACS to post-censal population estimates that miss the mark. The 2009 ACS household population estimate for Maricopa County (3,977,398) is 213,458 persons higher than the 2010 census (3,763,940).

7.1.4.2 Trend Methods

Most applications of the housing unit method hold PPH values constant at a previous value or have extrapolated historical trends (Cai and Spar 2008; Smith, Nogle, and Cody 2002; Swanson and Hough 2007; Swanson, Baker and Van Patten 1983; US Census Bureau 2009). The simplest method is to use the PPH value from the most recent census:

$$\mathrm{PPH_t} = \mathrm{PPH_c}.$$

This approach can provide relatively accurate estimates when the post-censal date is relatively close to the census or the PPH has shown stability over time. Other techniques that use historical trends have been found to produce more accurate post-censal estimates of PPH rather than using the last census value (Smith and Lewis 1980, 1983; Starsinic and Zitter 1968; Swanson and Hough 2007). Two extrapolation techniques that have been used to estimate PPH were discussed in Chapter 6.

One technique assumes a linear trend between the last two censuses (c and c-1) will extend into the post-censal period, and the other technique assumes a geometric trend. PPH estimates for the post-censal time period (t) from these trend models are:

$$PPH_t = PPH_c + (x/y)(PPH_c - PPH_{c-1}), \text{ Linear trend}$$

$$PPH_t = (PPH)_c[(1 + r)^{[x]}], \text{ and Geometric trend}$$

$$r = [(PPH_c/PPH_{c-1})^{1/y}] - 1,$$

where x is the number of years between t and c and y is the number of years between the last two censuses (e.g., 10 in the US and 5 in Canada).

A third technique makes three modifications to the linear trend method (Smith and Lewis 1980). The first modification uses national percentage change since the last census.[4] The second modification adjusts the US percentage change upward or downward depending on the level of the local PPH at the time of the last census. Studies have shown that PPH declines/increases tend to be greater/lesser in places with a large PPH values than in places with small PPH values (Smith and Lewis 1980; Serow, Eberstein, Mayberry, and Rives 1984). The final modification adjusts for the change in local housing mix, which accounts for the difference in household sizes between single family and other housing unit types. The first two modifications are quantified in the following relationship:

$$D_l = [(PPH_{l,c} - L)/(PPH_{us,c} - L)D_{us}, \text{ and}$$

$$D_{us} = (PPH_{us,t}/PPH_{us,c}) - 1,$$

where D is the proportional change; L is the lower bound for PPH, l is the local area, and us is the nation. The lower bound (L) represents the level below which PPH will no longer decline. As PPH approaches its lower bound the percentage must become smaller. The conceptual lower limit for PPH is 1.0, but it is not likely his threshold would ever be reached. Smith and Lewis (1980) used 1.5 as L, but found that the exact value of L is not critical to the formulation. The effect of changes in the local housing mix on PPH is estimated as follows:

$$PPHW_{l,c} = \Sigma w_{l,s,c} * PPH_{l,s,c}; \text{ and}$$

$$w_{l,s,c} = H_{l,s,c}/H_{l,c},$$

where PPHW is the PPH for all households; H is households; s is structure type. Overall PPH comes from the decennial census and PPH and households by structure type came from the long form prior to 2010 and now from the ACS. Therefore, the PPHW created from a weighted average will not necessary match the overall

PPH at the census time point. The next step is to estimate the post-censal PPHW at time (t) due to changes in the structure type mix:

$$\mathrm{PPHW}_{l,t} = \Sigma w_{l,s,t} * \mathrm{PPH}_{l,s,c};\ \text{and}$$

$$w_{l,s,t} = \mathrm{H}_{l,s,t}/\mathrm{H}_{l,t}.$$

The proportionate change in PPH due to changes in structure type mix is:

$$\mathrm{DM}_l = (\mathrm{PPHW}_{l,t} - \mathrm{PPHW}_{l,c})/\mathrm{PPHW}_{l,c}.$$

Combing the proportionate change in local PPH due to national trends and the proportionate change by to structure type shifts, the estimated local area PPH is:

$$\mathrm{DT}_{l,t} = (\mathrm{D}_l + \mathrm{DM}_l);\ \text{and}$$

$$\mathrm{PPH}_{l,t} = (1 + \mathrm{DT}_{l,t})\mathrm{PPH}_{l,c}.$$

Table 7.7 uses the three trend methods to produce April 1, 2010 estimates of PPH for jurisdictions in Sarasota County, Florida. The linear and geometric trend methods yield virtually identical results, which is expected since the change in PPH during the 1990s was relatively small. During the 1990s, the PPH declined in all jurisdictions, except NorthPort. These trends continued during the 2000s; albeit at a slower pace, which is evident when comparing the 2010 estimates and 2010 census values. Changes in US PPH and in local household structure type mix are required to develop estimates using the adjusted trend method (AdjTrend). Change in the US PPH was derived from the 2000 and 2010 censuses and post-censal estimates of households by structure type were taken from the 5-year ACS (2005–2009). The D component shows the direct relationship between the size of the adjustment and PPH level in 2000 built into this method. The DM component shows, except for Longboat Key, the change in household structure type mix causes an upward adjustment to the PPH. According to the ACS between 2000 and 2010, the single family share of households declined from 0.417 to 0.386 in Longboat key; was stable in the City of Sarasota, Venice, and the unincorporated area; and increased by 0.014 in NorthPort and by 0.018 in Sarasota County.

7.1.4.3 Regression Methods

As previously discussed, post-censal estimates of PPH are traditionally based on the latest census, historical trends, and or post-censal values for larger areas. These approaches are easy to implement and produce good estimates when PPH values remain constant or follow stable trends, but are inaccurate when PPH values and trends change rapidly (Smith, Nogle, and Cody 2002). PPH estimates may be

Table 7.7 Estimates of Average Household Size: Alternate Extrapolation Methods, Sarasota County Jurisdictions, April 1, 2010

Place	2000 pph	Change 1990–2000		Trend	Adjustment	Components	2010 Estimates			2010
		Linear	Geo. Rate[a]	D[b]	DM[c]	DT[d]	Linear[e]	Geo.[f]	AdjTrend[g]	Census
Longboat Key	1.78	−0.05	−0.00277	0.00383	−0.00359	0.00024	1.73	1.73	1.78	1.77
NorthPort	2.48	0.15	0.00626	0.01339	0.00659	0.01998	2.63	2.64	2.53	2.55
Sarasota	2.12	−0.02	−0.00094	0.00847	0.00057	0.00904	2.10	2.10	2.14	2.09
Venice	1.76	−0.04	−0.00224	0.00355	0.00047	0.00402	1.72	1.72	1.77	1.74
Unincorporated Area	2.15	−0.04	−0.00184	0.00888	0.00150	0.01038	2.11	2.11	2.17	2.10
Sarasota County	2.13	−0.10	−0.00458	0.00861	0.00563	0.01424	2.03	2.03	2.16	2.13

[a]Geometric rate of change
[b]US trend adjusted for PPH value and lower limit
[c]Adjustment for changes in structure type mix
[d]Total adjustment (D + DM)
[e]2000 pph + Linear
[f]2000 pph * (1 + Geo. Rate)^{10}
[g]2000 pph * (1 + DT)
Sources: US Census Bureau, 1990, 2000, and 2010 Decennial Censuses, and 2005–2009 ACS

improved by developing regression models that relate changes in PPH to changes in symptomatic indicators of household size.

This idea is not new. Regression models of PPH were suggested during the 1970s (Comprehensive Planning Organization of the San Diego Region 1974; Voss and Krebs 1979). Lowe (2000b) developed a regression model for counties in Washington State using school enrollment in grades K-8, sum of births 4 years prior to the estimation date, and population over 65+. This model did not perform as well as expected when PPH estimates for 2000 were compared to the census; it did not capture PPH increases and tended to overstate the magnitudes of PPH decline. Kimpel and Lowe (2007) developed an alternative specification of the prior model using population over 65+, sum of births 14 years prior to the estimation date, Hispanic population, and the PPH from the prior census to control for the base level of PPH. The longer lag on the birth variable eliminated the need for the school enrollment variable, which they believed was less reliable than the birth data. This model has not yet been tested against the 2010 census.

The most comprehensive investigation to date of regression models for PPH was conducted by Smith, Nogle, and Cody (2002). Their analysis was based on a 462 counties in Florida, Illinois, Texas, and Washington and analyzed PPH estimates for 1980, 1990, and 2000. Regression-based PPH estimates were compared to the census data for these years and were also compared against commonly used methods to estimate PPH. For independent variables were evaluated: 1) births per household; 2) school enrollment in grades K-12 per household; 3) Medicare enrollees age 65 and older per household; and 4) average number of exemptions per federal income tax return (IRS). The IRS variable was added after models with the first three independent variables were evaluated. Four regression models, which used different forms of the variables, were tested: 1) original variable at a single point in time; 2) ratio of county to state values at a single point in time; 3) arithmetic change in the variable; and 4) arithmetic change in the ratio. Model 4 is a variant of the difference correlation method of population estimate discussed in Chapter 8. The also analyzed the average of the estimates from the four models.[5]

The main findings and conclusions from this study were:

1. MAPEs based on the four regression models generally fell within a relatively small range;
2. The PPH estimates from the four models had little systematic bias;
3. The average method generally produced better results than did the individual regression models;
4. The regression models generally improved upon traditional estimation methods in both levels of bias and precision;
5. The regression models produced substantially fewer large errors (5% or more) than did the traditional methods;
6. Inclusion of the IRS variable raised the explained variance; improved the precision of the PPH estimates; and had no consistent impact on bias;
7. IRS data are excellent indicators of changes in PPH and are available for counties and subcounty areas (i.e., cities and zip codes);

Table 7.8 Group Quarters Share of Total Population, San Diego County Incorporated Cities and Census Tracts, 2000

	Percent of Total Population						
	All Group Quarters	Correction Facilities	Nursing Homes	Other Institutional	College Dorms	Military	OtherNon-institutional
County	3.4%	0.4%	0.3%	0.1%	0.5%	1.5%	0.7%
	Maximum Percent of Total Population						
Incorporated Cities[a]	27.1%	1.3%	1.4%	0.1%	1.2%	25.1%	2.5%
Census Tracts[b]	100.0%	81.7%	12.8%	11.8%	73.8%	99.7%	96.2%

[a]18 incorporated cities
[b]437 census tracts with non-zero group quarters population
Source: US Census Bureau, 2000 Decennial Census

8. PPH estimates from state-specific regression models did not perform better compared to regression models based on the entire sample of all states; and
9. The greatest advantage of regression models over traditional methods of estimating PPH may be their ability to perform well when PPH is changing rapidly or following unusual patterns and to reduce the number of large errors.

7.1.5 Group Quarters Population

The number of people living in group quarters facilities is the last piece of information required for the housing unit method. For large group quarters facilities (i.e., college dorms, correctional facilities, and military shipboard and barracks) information is usually available from the administrators of these facilities. Many state agencies that produce population estimates also maintain a list of group quarters facilities (e.g., Mohrman 2007). For the group quarters population not in large facilities such as nursing homes; half-way houses; and monasteries, it may be reasonable to assume no change since the last census or to assume they are growing at the same rate as the population in households (Smith 1986). The group quarters population usually accounts for a small proportion of the total population, but it can have a major impact on the population estimate in places with colleges, military installations, and correctional facilities.

Table 7.8 shows the group quarters population represents 3.4% of the total population in San Diego County, with military group quarters population comprising 1.5%. San Diego County has a large presence of naval and marine installations. The impact of the group quarters population becomes greater at finer spatial resolutions. In one city (Coronado), 25.1% of its population is comprised of military group quarters and the maximum percentages that occur in a city for the other categories are higher than seen for the County, with the exception of the Other Institutional. The group quarters population in several categories comprises over

70% of the total population in certain census tracts and the maximum percentages are all substantially higher than the corresponding values for the cities and the County.

7.2 Conclusions

The housing unit method is a comprehensive method for estimating post-censal population. It has a long and successful track record, is conceptually simple, easy to explain, can be understood by non-experts in demographic methods, and can be applied at virtually any geographic level (from states, to counties, census tracks, blocks, and parcels). The geographic scalability of the housing unit method is a tremendous advantage over most other estimation methods that are applicable for counties and higher level geographies. Another advantage of the housing unit method is along with population it also provides current information about trends in the housing market. The housing unit method is also very flexible and can use different techniques and data sources.

The housing unit method is a general approach to post-censal population estimation rather than a specific set of techniques. The housing unit method can be adapted to use data sources and techniques that work the best for any geographic area. When applied properly and carefully, the housing unit method can yield accurate estimates. The method will produce inaccurate estimates when applied carelessly using poor data and assumptions. Judgments about the efficacy of a particular set of estimates from the housing unit method must be based on the validity of the data and assumptions for the particular application, not on an assessment of the validity and performance of the method in general (Smith 1986).

The housing unit method, while simple in concept, requires a major commitment of time and resources to produce good results. There are a number of different techniques and data sources that must be carefully evaluated and accessed to determine which combination is best suited for a particular application. Considerable effort is required to ensure that data series are consistent over time (e.g., reporting procedures have not changed), have been adjusted for boundary changes, and reflect unique local conditions. The techniques for estimating occupancy rates and PPH must be scrutinized to determine which one or combination should be used. The housing unit method will not perform well unless sufficient resources are devoted to its implementation.

Endnotes

1. Housing units can change location if they are physically moved and they can be gained or lost due to conversions (e.g., single family house subdivided into apartments or retail space converted to a loft in a mixed use development. These events usually represent a very small

part of the overall change in housing units, but may be more significant in certain areas such as those undergoing revitalization. If data are available, housing unit moves and conversions should be taken into account.
2. The Censtats building permit website is at http://censtats.census.gov/bldg/bldgprmt.shtml.
3. Ideally, the comparisons should use estimates for 2010, 2009 to 2011, and 2008 to 2013, where the 2010 census is the center year of the accumulated samples in the ACS.
4. One could use the proportionate change from any larger area such as a state trend to estimate county PPH or a county trend to estimate census track PPH. Synthetic estimation discussed in Chapter 11 is an approach that uses trends from a large area to estimate local PPH values.
5. Averages reduce the chances of making large errors and have often have been found to produce more accurate results than the results for an individual method (e.g., Granger 1989: Chapter 8; Smith and Mandell 1984).

References

Alaska Department of Community and Economic Development. (2004). Housing unit method manual: Population estimate instructions and reporting forms. Juneau, AK: Alaska Department of Community and Economic Development.

Brown, W. (1999). Use of property tax records and household composition matrices to improve the household units method for small area population estimates. Paper presented at the US Census Bureau, Population Estimation Conference, Suitland, MD.

Bryan, T. (2004b). Population estimates. In J. S. Siegel, & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 523–560). New York: Elsevier Academic Press.

Byerly, E. (1990). State and local agencies preparing population and housing estimates Current Population Reports, Series P-25 (Vol. No. 1063). Washington, DC: US Bureau of the Census.

Cai, Q., & Spar, M. (2008). An evaluation of housing unit-based estimates in Virginia. Charlottesville, VA: Weldon Cooper Center for Public Service, University of Virginia.

Citron, C. F., & Kalton, G. (Eds.). (2007). *Using the American Community Survey: Benefits and challenges*. Washington, DC: The National Academies Press.

Comprehensive Planning Organization of the San Diego Region. (1974). A model for estimating household size in the San Diego region. San Diego, CA: Comprehensive Planning Organization of the San Diego Region.

Edelstein, R., & Tang, D. (2007). Dynamic residential housing cycle analysis. T*he Journal of Real Estimate Finance and Economics*, 35(3), 295–313.

Fairfax County. (2010). Demographic reports of 2010: County of Fairfax, Virginia. Fairfax, VA: Fairfax County, Department of Neighborhood and Community Services.

Fonseca, L., & Tayman, J. (1989). Post-censal estimates of household income distributions. *Demography*, 26, 149–160.

Gabriel, S. A., & Nothaft, F. E. (2001). Rental housing markets: The incidence and duration of vacancy and natural vacancy rate. *Journal of Urban Economics*, 49(1), 121–149.

Granger, C. W. (1989). *Forecasting in business and economics*, Second Edition. San Diego: Academic Press.

Hendershott, P. H., MacGregor, B. D., & Tse, R. Y. (2002). Estimation of the rental adjustment process. *Real Estate Economics*, 30(2), 165–183.

Hoque, N. M. (2010). An evaluation of small area population estimates produced by component method II, ratio-correlation, and housing unit methods for 1990. *The Open Demography Journal*, 3, 18–30.

Jarosz, B. (2008). Using assessor parcel data to maintain housing unit counts for small area population estimates. In S. H. Murdock, & D. A. Swanson (Eds.), *Applied Demography in the 21st Century* (pp. 89–101). Dordrecht, Heidelberg, London, and New York: Springer.

Kimpel, T., & Lowe, T. J. (2007). Estimating household size for use in population estimates. Research Brief No. 47. Olympia, WA: Washington State Office of Financial Management.

Lo, C. P. (1995). Automated population dwelling unit estimation from high-resolution satellite images: A GIS approach. *International Journal of Remote Sensing*, 16, 17–34.

Lowe, T. J. (2000a). Occupied versus rented: Census & real estate occupancy rates and their use in population estimates. Research Brief No. 9. Olympia, WA: Washington State Office of Financial Management.

Lowe, T. J. (2000b). Developing trends in household size for use in population estimates. Research Brief No. 10. Olympia, WA: Washington State Office of Financial Management.

Lowe, T. J. & Mohrman, M. (2003a). Developing postal delivery data for use in population estimates. Research Brief No. 17. Olympia, WA: Washington State Office of Financial Management.

Lowe, T., J. Mohrman, M. (2003b). Use of postal delivery data in the population estimation process Research Brief No. 18. Olympia, WA: Washington State Office of Financial Management.

Lowe, T. J., Myers, W. R., & Weisser, L. M. (1984). A special consideration in improving housing unit estimates: The interaction effect. Paper presented at the annual meeting of the Population Association of America, Minneapolis, MN.

Lowe, T. J., Pittenger, D. B., & Walker, J. R. (1977). Making the housing unit method work: A progress report. Paper presented at the annual meeting of the Population Association of America, St. Louis, MO.

Myers, D. and Doyle, A. (1990. Age-specific population-per-household ratios: Linking population age structure and housing characteristics. In D. Myers (Ed.), *Housing Demography: Linking Demographic Structure and Housing Markets* (pp. 109–132). Madison: University of Wisconsin Press.

Mohrman, M. (2007). Tracking group quarter facility data: The Washington State experience. Paper presented at the Applied Demography Conference, San Antonio, TX.

Rynerson, C., & Tayman, J. (1998). An Evaluation of Address-Level Administrative Records Used to Prepare Small Area Population Estimates. Paper presented at the annual meeting of the Population Association of America, Chicago, IL.

Puget Sound Regional Council. (2009). Analysis and forecasting at PSRC. Seattle, WA: Puget Sound Regional Council.

Serow, W. J., Eberstein, I. W., Mayberry, L. A., & Rives, N. W. (1984). Can simple techniques produce useful population estimates. Paper presented at the North American Meeting of the Regional Science Association, Washington, DC

Smith, S., K. (1986). A review and evaluation of the housing unit method of population estimation. *Journal of the American Statistical Association*, 81, 287–296.

Smith, S. K., & Cody, S. (2004). An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review*, 23, 1–24.

Smith, S. K., & Lewis, B. B. (1980). Some new techniques for applying the housing unit method of local population estimation. *Demography*, 17(3), 323–339.

Smith, S. K., & Lewis, B. B. (1983). Some new techniques for applying the housing unit method of local population estimation: Some further evidence. *Demography*, 20(3), 407–413.

Smith, S. K., & Mandell, M. (1984). A comparison of population estimation methods: Housing unit versus component II, ratio-Correlation, and administrative records. *Journal of the American Statistical Association*, 79, 282–289.

Smith, S. K., Nogle, J., & Cody, S. (2002). A regression approach to estimating the average number of persons per household. *Demography*, 39(4), 697–712.

Starsinic, D. E., & Zitter, M. (1968). The accuracy of the housing unit method is preparing population estimates for cities. *Demography*, 5, 475–484.

Swanson, D. A., Baker, B., & Van Patten, J. (1983). Municipal population estimation: Practical and conceptual features of the housing unit method. Paper presented at the annual meeting of the Population Association of America, Pittsburgh, PA.

Swanson, D. A., & Hough, G. C. (2007). An evaluation of persons per household (PPH) data generated by the American Community Survey: A demographic perspective. Paper presented at the annual meeting of the Southern Demographic Association, Birmingham, AL.

Tayman, J. (1994). Estimating population, housing, and employment for micro-geographic areas. In K. V. Rao & J. W. Wicks (Eds.), *Studies in Applied Demography*. Bowling Green: Population and Society Research Center, Bowling Green State University.

Tayman, J., & Rynerson, C. (1997). An integrated system for estimating subcounty household income distributions. Paper presented at the annual meeting of the Population Association of America, Washington, DC

US Census Bureau. (1983). A survey of agencies using the housing unit method. US Department of Commerce. Washington, DC

US Census Bureau. (2009). Methodology for the subcounty total resident population estimates (Vintage 2009): April 1, 2000 to July 1, 2009. (http://www.census.gov/popest/topics/methodology/2009-su-meth.pdf).

US Census Bureau. (2011). Special census program. (http://www.census.gov/regions/specialcensus/).

Voss, P. R., & Krebs, H. C. (1979). The use of federal revenue sharing data for improving estimates of average household size for minor civil divisions. Technical Series (70–5). Madison, WI: University of Wisconsin, Applied Population Laboratory.

Wang, L., & Wu, C. (2010). Population estimation using remote sensing and GIS technologies International. *Journal of Remote Sensing* 31(21), 5569 – 5570

Wicks, J. W., Swanson, D, A., Vincent, R. K., & De Almeida, J. P. (1999). Population estimates from remotely sensed data: A discussion of recent technological developments and future research plans. Paper presented at the Canadian Population Society Meetings, Lennoxville, Quebec.

Chapter 8
Regression Methods

8.1 Introduction

Regression-based methods for estimating population date back to E. C. Snow (1911), who published "The application of the method of multiple correlation to the estimation of post-censal populations" in the *Journal of the Royal Statistical Society*. Snow's paper represents the first published description of the use of multiple regression in the estimation of population. It also discusses other methods, pointing out their strengths and weaknesses, then describes the model framework and the data used in the regression application, and applies it to districts in the U. K. In addition to being the first published report in English of the use of regression for population estimates, it sets the stage for subsequent papers by discussing it relative to other methods. A discussion is published with the paper that contains many important insights that are today commonplace in the use of multiple regression not only for making population estimates, but for general use.

One of the insights (Snow 1911: 625) is given by David Heron, who suggests that one of the shortcomings acknowledged by Snow was to "control" the sum of the estimates for individual districts to an estimate for the who country ("Estimate for the whole country/sum of estimates for individual districts). Another is provided by G. Udny Yule, who contributed substantially to the development of multiple regression as a modern analytic technique (Stigler 1986: 345-361). Yule (Snow 1911: 621) noted that Snow demonstrated that a multiple regression model built using data over one decade had coefficients that could be used for the subsequent decade with the insertion of the new set of values for the independent variables. Yule also agreed with Snow that the ex post facto tests performed by Snow suggested that using variables constructed on relative (percent) change would perform better than variables constructed on the basis of absolute change (Snow 1911: 622). Finally, among many comments useful still today for those interested in

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_8,

regression based methods for estimating population, are the following: Greenwood' remarks on the impact of skewed distributions (Snow 1911: 626); Baines' (Snow 1911: 626) comments on using ratios, and the importance of data quality by virtually all of the discussants (Snow 1911: 621-629).

Snow's (1911) seminal paper is based on the premise that the relationship between symptomatic indicators and the corresponding population remains unchanged over time. His work and the insights provided by the discussants of his paper have led to three related but distinct approaches: ratio-correlation; difference-correlation; and average ratio methods.

8.2 Ratio-Correlation and Its Variants

The most common regression-based approach data to estimating the total population of a given area is the ratio-correlation method. Introduced and tested by Schmitt and Crosetti (1954) and again tested by Crosetti and Schmitt (1956), this multiple regression method involves relating between changes in several variables known as symptomatic indicators on the one hand to population changes on the other hand. The symptomatic indicators that are used reflect the variables related to population change that are available such that those that yield an optimal model are chosen. Examples of symptomatic variables that have been used for this purpose are births, deaths, school enrollment, tax returns, motor vehicle registrations, employment data, and registered voters. The ratio-correlation method is used where a set of areas (e.g., counties) are structured into a geographical hierarchy (e.g. the populations of counties within a given state sum to the total state population). It proceeds in two steps. The first is the construction of the model and the second is its implementation – actually using it to create estimates for given years.

Because the method looks at change, population data from two successive censuses are needed to construct the model along with data for the same years representing the symptomatic indicators. During its implementation step the ratio-correlation method requires symptomatic data representing the year for which an estimate is desired and an estimate of the population for the highest level of geography (e.g., the state as a whole) that is independent of the ratio-correlation model.

The ratio-correlation method expresses the relationship between (1) the change over the previous inter-censal period (e.g., 1990 to 2000) in an area's share (e.g., a given county) of the total for the parent area (e.g., the state as a whole) for several symptomatic series and (2) the change in an area's share of the population of the parent area. The method can be employed to make estimates for either the primary or secondary political, administrative and statistical divisions of a country (Bryan 2004). In the US, the variables selected usually vary from state to state and because of the small number of counties in some states, certain states were combined and estimated in one regression equation.

In general terms, the ratio-correlation model is formally described as follows (Swanson and Beck 1994):

$$P_{i,t} = a_0 + \sum (b_j)^* S_{i,j,t} + \varepsilon_i \tag{8.1a}$$

where

a_0 = the intercept term to be estimated
b_j = the regression coefficient to be estimated
ε_i = the error term
j = symptomatic indicator ($1 \leq j \leq k$)
i = subarea ($1 \leq j \leq n$)
t = year of the most recent census

and

$$P_{i,t} = (P_{i,t}/\sum P_{i,t})/(P_{i,t-z}/\sum P_{i,t-z}) \tag{8.1b}$$

$$S_{i,j,t} = (S_{i,t}/\sum S_{i,t})_j/(S_{i,t-z}/\sum S_{i,t-z})_j \tag{8.1c}$$

where

z = number of years between each census for which data are used to construct the model
p = population
s = symptomatic indicator

Once a ratio-correlation model is constructed, a set of population estimates for time t + k is developed in a series of six steps. First, $(S_{i,t+k}/\sum S_{i,t+k})_j$ is substituted into the numerator of the right side of equation 8.1c for each symptomatic indicator j and $(S_{i,t}/\sum S_{i,t})_j$ into the denominator of the right side of equation 8.1c for each symptomatic indicator j, which yields $S_{i,j,t+k}$. Second, the updated model with the preceding substitution of symptomatic data for time t + k is used to estimate $P_{i,t+k}$. Third, $(P_{i,t}/\sum P_{i,t})$ is substituted into the denominator of $P_{i,t+k}$, which yields $P_{i,t+k} = (P_{i,t+k}/\sum \mathbf{P_{i,t+k}})/(P_{i,t}/\sum P_{i,t})$, where $\sum \mathbf{P_{i,t+k}}$ represents the independently estimated population of the "parent" area of the i subareas for time t + k (Note that this estimate is given in boldface and is done by a method exogenous to the ratio-correlation model (e.g., a component method)). Fifth, since $P_{i,t+k}$, $(P_{i,t}/\sum P_{i,t})$ and $\sum \mathbf{P_{i,t+k}}$ are all known values, the equation $P_{i,t+k} = (P_{i,t+k}/\sum \mathbf{P_{i,t+k}})/(P_{i,t}/\sum P_{i,t})$ is manipulated to yield an estimate of the population of area i at time t + k:

$$(P_{i,t+k})^*(P_{i,t}/\sum P_{i,t})^*(\sum \mathbf{P_{i,t+k}}) = \hat{P}_{i,t+k} \tag{8.1d}$$

As equation 8.1d shows, it is important to remember that an independent estimate of the population for the "parent" geography ($\sum \mathbf{P_{i,t+k}}$) of the i subareas

is required when using the ratio-correlation model to generate population estimates. The sixth and final step is to effect a final "control" so that the sum of the i subarea population estimates is equal to the independently estimated population for the parent of these i subareas: $\sum P_{i,t+k} = \sum \mathbf{P_{i,t+k}}$, which is accomplished as follows:

$$P_{i,t+k} = (P_{i,t+k}/\Sigma P_{i,t+k})^{*}(\Sigma \mathbf{P_{i,t+k}}). \tag{8.1e}$$

It is obvious from the preceding definitions that we are focusing on the ratio-correlation method as a means of developing post-censal estimates. However, it can be used to develop inter-censal estimates, a topic we cover at some length in Chapter 17.

As an empirical example of ratio-correlation model, we use data for the 39 counties of Washington state. We used excel to construct a ratio-correlation model using 1990 and 2000 census data in conjunction with three symptomatic indicators: (1) registered voters; (2) registered automobiles, and (3) public school enrollment in grades 1-8. The raw 1990 and 2000 input data for this model are provided in an appendix at the end of this chapter as Tables 8.2.a through 8.2.d. We then use 2005 symptomatic indicators to construct a set of county estimates for 2005. The input data for 2000 and 2005, along with the results of the calculations leading to the estimates are shown as Tables 8.2.e through 8.2.h at the end of this chapter.

A summary of the model and its characteristics is provided in Exhibit 8.1.

Exhibit 8.1 Example Ratio-Correlation Model

$$P_{i,t} = 0.195 + (0.0933^{*}\text{Voters}) + (0.3362^{*}\text{Autos}) + (0.3980^{*}\text{Enroll})$$
$$[p<.001]\ [p = 0.14] \qquad [p<.001] \qquad [p<.001]$$

where

$$P_{i,t} = (P_{i,2000}/\sum P_{i,2000})/(P_{i,1990}/\sum P_{i,1990})$$
$$S_{i,1,t} = (\text{Voters}_{i,2000}/\sum \text{Voters}_{i,2000})/(\text{Voters}_{i,1990}/\sum \text{Voters}_{i,1990})$$
$$S_{i,2,t} = (\text{Autos}_{i,2000}/\sum \text{Autos}_{i,2000})/(\text{Autos}_{i,1990}/\sum \text{Autos}_{i,1990})$$
$$S_{i,3,t} = (\text{Enroll}_{i,2000}/\sum \text{Enroll}_{i,2000})/(\text{Enroll}_{i,1990}/\sum \text{Enroll}_{i,1990})$$
$$R^2 = 0.794$$
$$\text{adj } R^2 = 0.776$$

Although the coefficient for Voters is not statistically significant, we elected to retain this symptomatic indicator in the model so that we would have a model with three independent variables, a feature that as explained later, can assist in dealing with "model invariance."

The amount of "explained variance" ($R^2 = 0.794$) is typical for a ratio-correlation model. Do not be alarmed that this level is not sufficient to have a "good

model." That is, neither believe that a good ratio-correlation model should have a very high level of explained variance (e.g., $R^2 > 0.9$) nor expect one. This is the case because the structure of the ratio-correlation model reflects the "stationarity" achieved by taking ratios over time (Swanson 2004). Note that the coefficients approximately sum to 1.00. This also is a universal feature of the ratio-correlation model, one which can be exploited in a model with three symptomatic indicators, as is discussed shortly.

In using this model to construct a set of county population estimates for 2005, we follow the six steps just described. First, $(S_{i,2005}/\sum S_{i,2005})_j$ is substituted into the numerator of the right side of the model for each symptomatic indicator j and $(S_{i,2000}/\sum S_{i,2000})_j$ into the denominator of the right side of the model for each symptomatic indicator j, which yields $S_{i,j,2005}$. Second, the updated model with the preceding substitution of symptomatic data for 2005 is used to estimate $P_{i,2005}$. Third, $(P_{i,2000}/\sum P_{i,2000})$ is substituted into the denominator of $P_{i,2005}$, which yields $P_{i,2005} = (P_{i,2005}/\sum \mathbf{P_{i,2005}})/(P_{i,2000}/\sum P_{i,2000})$, where $\sum \mathbf{P_{i,2005}}$ represents the independently estimated population of the state as a whole, which is the parent area of the 39 counties for 2005. Fifth, since $P_{i,2005}$, $(P_{i,2000}/\sum P_{i,2000})$ and $\sum \mathbf{P_{i,2005}}$ are all known values, the equation $P_{i,2005} = (P_{i,2005}/\sum \mathbf{P_{i,2005}}) / (P_{i,2000}/\sum P_{i,2000})$ is manipulated to yield an estimate of the population of county i in the year 2005:

$$(P_{i,2005})^*(P_{i,2000}/\sum P_{i,2000})^*(\sum \mathbf{P_{i,2005}}) = \hat{P}_{i,2005}$$

The sixth and final step is to control the 2005 population estimates of the 39 counties so that they sum to the independently estimated 2005 population for the state of Washington as a whole:

$$\hat{P}_{i,2005} = (P_{i,2005}/\sum P_{i,2005})^*(\sum \mathbf{P_{i,2005}})$$

The final "controlled" population estimates are shown in Table 8.1. The appendix shows the results of these steps in detail.

An acute observer may notice that except when $k = z$, the use of the model for estimating population corresponds to a shorter length of time than that used to calibrate the model. For example, if one constructs a model using 1990 and 2000 data for the 39 counties in the state of Washington it corresponds to a ten year period of change in both population shares and shares of symptomatic variables. However, in using this same model to estimate the populations of the 39 counties in 2003, the time period now corresponds to a three year period of change in both population shares and shares of symptomatic variables. Swanson and Tedrow (1984) addressed this temporal inconsistency by using a logarithmic transformation. They called the resulting model the "rate-correlation" model. This is one of several variants of the basic ratio-correlation regression technique. We discuss this variation in Chapter 17 and provide an empirical example based on the same Washington state data used in the example for ratio-correlation found in this chapter.

Table 8.1 2005 County Population Estimates for the state of Washington

County	Estimated 2005 Population
Adams	18,125
Asotin	20,706
Benton	155,792
Chelan	66,727
Clallam	66,870
Clark	393,823
Columbia	4,284
Cowlitz	95,522
Douglas	40,065
Ferry	7,295
Franklin	59,650
Garfield	2,266
Grant	79,475
GHarbor	68,680
Island	74,802
Jefferson	26,994
King	1,793,565
Kitsap	239,943
Kittitas	36,560
Klickitat	18,979
Lewis	69,010
Lincoln	9,982
Mason	53,729
Okanogan	38,740
Pacific	21,099
Pend Oreille	12,093
Pierce	758,454
SanJuan	15,363
Skagit	110,607
Skamania	10,104
Snohomish	652,045
Spokane	442,581
Stevens	41,795
Thurston	230,361
Wahkaikum	4,043
WallaWalla	58,906
Whatcom	180,956
Whitman	40,906
Yakima	235,504
State of Washington	6,256,400

Another variant is known as the "difference-correlation" method. Similar in principle to the ratio-correlation method, the difference-correlation method differs in its construction of a variable that is used to reflect change over time. Rather than making ratios out of the two proportions at two points in time, the difference correlation method employs the *differences* between proportions (Schmitt and Grier 1966; O'Hare 1980; Swanson 1978a). One advantage of this method is that

in taking differences one never has to worry about dividing by zero, which, of course, is undefined in terms of simple algebra, such as is found in the ratio-correlation and difference-correlation operations. This is not a hypothetical problem. In cases involving small populations one could expect that some of the ratios involving the symptomatic indicators would be zero (Swanson 1978a). For example, if one used the ratio-correlation model to estimate the non-white population of Garfield County, Washington, one would encounter division by zero. This is where the difference-correlation form may be preferred.

Another variant was proposed by Namboodiri and Lalu (1971). Known as the "average regression" technique, Namboodiri and Lalu (1971) examined the use of the simple, unweighted average of the estimates provided by a number of simple regression equations, each of which relates the population ratio to *one* symptomatic indicator ratio (As discussed in Chapter 9, this turns out to be very similar to using an average of several censal ratio estimates). Using the insights provide by Namboodiri and Lalu (1971), Swanson and Prevost (1985) demonstrated that the ratio-correlation model can be interpreted as a demographic form of "synthetic estimation" that is composed of a set of weighted censal-ratio estimates, with the regression coefficients serving as the weights (See Chapter 11).

Another variation on the ratio-correlation method is to use administrative data to refine the definition of the population being estimated. In the United States, it has been possible to use Medicare and related data to obtain an estimate of the population aged 65 years and over down to the county level (Bryan 2004; Murdock et al. 1995; US Census Bureau 2010). With these data, the model is then used to estimate the population under the age of 65. Variations on this theme include obtaining separate estimates of the population not living in households (e.g., students living in dormitories, military personnel living in barracks and on ships, patients in long-term care, and prisoners) and then using the model to estimate the household population (Feeney et al. 1995). Combining this approach with the one using Medicare data, one could develop a model to estimate the household population under the age of 65.

Bryan (2004) observes that one of the shortcomings of the ratio-correlation method and related techniques is that substantial time lags can occur in obtaining the symptomatic indicators needed for producing a current population estimate. That is, suppose that it is the year 2014 and a current (2014) estimate is desired, but the most current symptomatic indicators are for 2012. What can one do? One answer to this question is "lagged ratio-correlation," which was introduced by Swanson and Beck (1994). In this variant of ratio-correlation, the ratios of proportional symptomatic indicators precede the ratios of population proportions by "m" years in model construction so that:

$$S_{i,jt-m} = \left(S_{i,t-m} / \sum S_{i,t-m}\right)_j / \left(S_{i,(t-m)-z} / \sum S_{i,(t-m)-z}\right)_j \qquad (8.1f)$$

where

m = number of years that symptomatic indicators precede the population proportions

When the lagged ratio-correlation is used to estimate a population, the only change to the six steps described earlier for the basic form of ratio-correlation is that $(S_{i,t+k}/\sum S_{i,t+k})_j$ is substituted into the numerator of the right side of equation 8.1c for each symptomatic indicator j in place of $(S_{i,(t-m)+k}/\sum S_{i,(t-m)+k})_j$ and $/(S_{i,(t-m)}/\sum S_{i,(t-m)})_j$ into the denominator of the right side of equation 8.1c for each symptomatic indicator j in place of $(S_{i,t}/\sum S_{i,t})_j$.

Because ratio-correlation and its variants are grounded in regression, they are connected to the inferential and other statistical tools that come with it (Swanson 1989; Swanson and Beck 1994). This is a theme to which we return in Chapter 14. In using these tools, it is important to keep in mind an important point, which is that within this framework, "uncertainty" is generally based on the "frequentist" view of sample error. Thus, as discussed by Swanson and Beck (1994), the construction of confidence intervals around estimated values means, for example, that one perceives (whether implicitly or explicitly) the following: the data used in model construction are a random sample drawn from a universe; the model would fit perfectly were it not for random error; and, any subsequent observations of independent variables placed into the model and used to generate dependent variables are drawn from the same universe. Since a given model is constructed from data using observations from all known cases (e.g., all 39 counties in Washington), the "universe" represented by the county data is a "superpopulation". This means, as we discussed in Chapter 4 and as noted by D'Allesandro and Tayman (1980), the observed values are a random manifestation of all the possible observations that could have occurred.

Technically speaking, this makes it difficult to interpret confidence intervals in an actual estimation or projection application or an ex post facto test because we can never observe the regression surface for this superpopulation (specifically, the set of county populations forming the expected values of this regression surface). What we do observe is a census count. This census count has two distinct uses. First, it must be viewed as an estimator during the model construction phase (as are all of the symptomatic indicators). However, when we use a given model to estimate or project the number of persons in a given county, we must view the number that is (or could be) generated by a complete enumeration as a parameter. Thus, in using the term "confidence intervals" one (implicitly or explicitly) assumes that a census count is used to generate an estimate or projection. Consequently, when a confidence band is placed around estimated or projected figures, the band is an interval estimator for a parameter (Swanson and Beck 1994).

Given these qualification, Swanson and Beck (1994) conducted ex post facto examinations on estimates produced by the lagged ratio-correlation model and their "forecast intervals" for total populations of the 39 counties in Washington State in 1970, 1980, and 1990. For the 1970 set of county population projections, they found that the 2/3 forecast intervals contained the 1970 census figure in more than two-thirds (30 of the 39 counties) as did the 1990 results (31 of 39 counties). For the 1980 set, the 2/3 forecast interval contained the 1980 census figure in just less than two-thirds (24 of the 39 counties). Swanson and Beck (1994) argued that these findings are of interest from an application standpoint because if the

2/3 forecast intervals contained substantially less than two-thirds of the actual county populations, one would have a misplaced sense of accuracy in the ability of the given models to accurately estimate and project county populations. Since the intervals did contain more than two-thirds of the actual county population figures in both 1970 and 1990 and nearly two-thirds in 1980, they argued that the results of this case study revealed an intuitively appealing view of the accuracy of these particular models (Swanson and Beck 1994).

The findings by Swanson and Beck (1994) suggest that, among other useful features, one can construct confidence and "forecast" intervals around the estimates produced by ratio-correlation and its variants that are both statistically and substantively meaningful.

Given that the input data are of good quality, the accuracy of the regression-based techniques largely depends upon the validity of the central underlying assumption: that the observed statistical relationship between the independent and dependent variables in the past inter-censal period will persist in the current post-censal period. The adequacy of this assumption (that the model is invariant) is dependent on several conditions (Swanson 1980; Mandell and Tayman 1982; McKibben and Swanson 1997; Tayman and Schafer 1985).

In an attempt to deal with model invariance, Ericksen (1973, 1974) introduced a method of post-censal estimation in which the symptomatic information is combined with sample data by means of a regression format. He considered combining symptomatic information on births, deaths, and school enrollment with sample data from the Current Population Survey. Swanson (1980) took a different approach to the issue of model invariance and presented a mildly restricted procedure for using a theoretical causal ordering and principles from path analysis to provide a basis for modifying regression coefficients in order to improve the estimation accuracy of the ratio-correlation method of population estimation.

Ridge Regression also represents a method for dealing with model invariance. Swanson (1978b) and D'Allesandro and Tayman (1980) examined this approach to multiple regression and found that it offered some benefits. Ridge Regression also represents a way to deal with another possible problem with the regression approach, which is multi-collinearity, a condition whereby the independent variables are all highly correlated. This condition can result in type II errors (finding that given coefficients are not shown to be statistically significant when in fact they are) when one evaluates the statistical significance of the coefficients associated with the symptomatic indicators used in a given model. One also can use the standard diagnostic tools associated with regression to evaluate and this issue and overcome it without resorting to ridge regression, if an evaluation suggests it is present (Fox 1991). Swanson (1989) demonstrated another way to deal with model invariance by using the statistical properties of the ratio-correlation method in conjunction with the Wilcoxon matched-pairs signed rank test and the "rank-order" procedure he introduced (Swanson 1980).

Judgment is also important in the application of ratio-correlation, as the analyst must take into account the reliability and consistency of coverage of each variable (Tayman and Schafer 1985). The increasing availability of administrative data

allows many possible combinations of variables. High correlation coefficients for two past inter-censal periods would *suggest* that the degree of association of the variables is not changing very rapidly. In such a case, the regression based on the last inter-censal period should be applicable to the current post-censal period. Furthermore, it is assumed that deficiencies in coverage in the basic data series will remain constant, or change very little, in the present period (Tayman and Schafer 1985).

In addition to the issue of time lags in the availability of symptomatic indicators, Bryan (2004) notes two other shortcomings of regression-based techniques: (1) the use of multiple and differing variables (oftentimes depending on the place being estimated) and in some instances averaging the results of multiple estimates, which makes it difficult to decompose error; and (2) this process may compromise the comparability of estimates between different subnational areas. In regard to decomposing error, this is a feature of all of the estimation methods that do not deal directly with the components of population change. In regard to comparability, we note that this is an issue when different regression models are used (e.g., the ratio-correlation model used to estimate the populations of the 75 counties of Arkansas is different from the ratio-correlation model used to estimate the populations of the 39 counties of Washington state.

In regard to the issue of decomposing error, McKibben and Swanson (1997) argue that at least some of the shortcomings in accuracy of population estimates would be better understood by linking these methods with the substantive socio-economic and demographic dynamics that clearly must be underlying the changes in population that the methods are designed to measure. They provide a case study of Indiana over two periods, 1970-1980 and 1980-1990, which was selected because a common population estimation method exhibits a common problem over the two periods: its coefficients change. The authors link these changes to Indiana's transition to a post-industrial economy and describe how this transition operated through demographic dynamics that ultimately affected the estimation model.

8.3 Summary

Regression-based methods have very limited application in the preparation of estimates of population composition, such as age-sex groups for small geographic areas. It is possible, of course, to apply the age distribution at the last census date to a pre-assigned current total for the area, or to extrapolate the last two census age distributions to the current date and apply the extrapolated distribution to the current total. Spar and Martin (1979) found, for example, that the ratio-correlation method is more accurate than others in estimating the populations of Virginia counties by race and age.

While the regression approach has its limitations, as suggested by this overview, it is clear it has strong advantages, given the availability of good quality data to implement and test it. This is especially the case for the ratio-correlation method. Among its many advantages is the fact that regression has

a firm foundation in statistical inference, which leads to the construction of meaningful measures of uncertainty around the estimates it produces, as demonstrated by Swanson and Beck (1994). As discussed in Chapters 14 and 15, this gives the ratio-correlation method an important advantage in terms of error and uncertainty assessment over the methods that are not linked to statistical inference. Further, as suggested by Snow (1911) and those who discussed his groundbreaking use of multiple regression for population estimation, it is important to use variables that represent some measure of relative change over time, which the ratio-correlation method does. Although ratio-correlation is inherently a cross-sectional model rather than a time series, Swanson (2004) suggests that one of the reasons for its consistently good performance, may be due to the fact that the formation of the change in ratios provides some of the benefits associated with "stationarity," which as we discussed in Chapter 6 is an important characteristic in the development of a good ARIMA model.

The basic assumption underlying the regression methods discussed here is the same as those underlying the trend extrapolation methods discussed in Chapter 6 – in terms of the change in a variable of interest specified by a particular method—the future will be just like the past. This is the source of model invariance and one must always ask in using a regression-based method what sort of changes are expected to occur over time and how can they be accommodated? These questions can be set in terms of spatial and temporal heterogeneity, spatial autocorrelation, spatial dependence and spatial interaction. Some progress toward answering these questions appears to have been made (D'Allesandro and Tayman 1980; Ericksen 1973; Mandell and Tayman 1982; McKibben and Swanson 1997; Swanson 1978b; Swanson 1980; Tayman and Schafer 1985), but more work is needed. A factor favoring the success of these endeavors is that these regression models are firmly embedded in the theory, substance, and issues of spatial demography, which is discussed in Chapter 12.

Appendix

Table 8.2a Registered Voters, 1990 and 2000 Data

COUNTY	Number Year = 2000	Number Year = 1990	Proportion Year = 2000	Proportion Year =1990	Ratio of 2000 Prop/1990 Prop
Adams	6,098	5,553	0.00196738	0.002499767	0.787025521
Asotin	12,987	8,597	0.004189959	0.00387007	1.082657236
Benton	75,315	53,452	0.024298665	0.024062227	1.009826097
Chelan	32,803	24,043	0.010583139	0.010823321	0.977808879
Clallam	39,068	28,085	0.012604398	0.012642888	0.996955607
Clark	167,584	88,903	0.054067151	0.040021032	1.350968445
Columbia	2,671	2,256	0.000861737	0.001015573	0.848523475
Cowlitz	49,643	34,503	0.01601618	0.015532048	1.031169905
Douglas	16,855	11,320	0.005437881	0.005095869	1.067115429
Ferry	3,856	2,486	0.00124405	0.001119111	1.111642059
Franklin	16,321	13,228	0.005265598	0.005954785	0.884263396
Garfield	1,670	1,537	0.000538787	0.000691904	0.778702686
Grant	29,970	21,391	0.009669136	0.009629483	1.004117935
GHarbor	32,038	29,613	0.010336329	0.01333074	0.775375474
Island	38,265	24,325	0.012345329	0.010950267	1.12739976
Jefferson	17,330	11,413	0.005591129	0.005137735	1.088247842
King	1,001,339	765,692	0.323059164	0.344687849	0.937251385
Kitsap	125,219	82,518	0.040399051	0.037146727	1.087553441
Kittitas	16,417	12,836	0.00529657	0.00577832	0.916628084
Klickitat	11,717	7,943	0.003780223	0.003575662	1.057209207
Lewis	40,913	27,990	0.013199645	0.012600122	1.047580719
Lincoln	6,656	5,495	0.002147406	0.002473657	0.868109854
Mason	27,238	18,108	0.008787719	0.00815159	1.078037328
Okanogan	18,159	14,987	0.005858587	0.006746625	0.868372958
Pacific	12,697	9,906	0.004096397	0.004459336	0.918611473
PendOreille	6,903	4,851	0.002227095	0.002183751	1.019848515
Pierce	325,079	229,449	0.104879316	0.103289942	1.015387506
SanJuan	9,228	6,919	0.002977203	0.003114693	0.955857879
Skagit	55,780	38,696	0.017996143	0.01741959	1.033097962
Skamania	5,586	3,946	0.001802195	0.001776352	1.014548749
Snohomish	303,110	196,968	0.09779152	0.088668128	1.102893707
Spokane	209,404	165,189	0.067559419	0.07436233	0.908516708
Stevens	25,481	14,406	0.008220863	0.006485079	1.267658073
Thurston	119,016	79,381	0.038397795	0.035734559	1.074528289
Wahkaikum	2,455	1,944	0.00079205	0.000875121	0.90507445
WallaWalla	24,411	20,614	0.007875652	0.009279704	0.848696416
Whatcom	90,987	60,874	0.029354878	0.027403353	1.071214827
Whitman	25,273	18,842	0.008153756	0.008482012	0.961299834
Yakima	94,011	73,148	0.030330502	0.03292868	0.921096825
check sum	3,099,553	2,221,407	1.0000	1.0000	
STATE	3,099,553	2,221,407			

Table 8.2b Registered Autos, 1990 and 2000 Data

COUNTY	Number Year = 2000	Number Year = 1990	Proportion Year = 2000	Proportion Year = 1990	Ratio of 2000 Prop/1990 Prop
Adams	9,144	7,476	0.002950103	0.003365435	0.876588954
Asotin	10,375	8,964	0.003347257	0.00403528	0.829497968
Benton	80,977	62,203	0.02612538	0.028001622	0.932995226
Chelan	39,153	31,360	0.012631821	0.014117179	0.894783691
Clallam	35,697	29,592	0.011516822	0.013321287	0.864542744
Clark	183,053	139,958	0.059057871	0.063004213	0.937363832
Columbia	2,186	2,226	0.000705263	0.001002068	0.703807786
Cowlitz	52,461	47,555	0.016925344	0.021407603	0.790623007
Douglas	13,008	12,107	0.004196734	0.005450149	0.770021861
Ferry	2,384	1,943	0.000769143	0.000874671	0.879351522
Franklin	27,518	24,762	0.008878054	0.011146989	0.796453117
Garfield	1,263	1,247	0.000407478	0.000561356	0.725881898
Grant	35,188	28,154	0.011352605	0.012673949	0.895743254
GHarbor	33,310	32,097	0.010746711	0.014448951	0.743771032
Island	37,675	28,462	0.012154978	0.0128126	0.94867382
Jefferson	14,459	10,170	0.004664866	0.00457818	1.018934751
King	1,083,380	975,138	0.349527819	0.438973137	0.796239654
Kitsap	125,716	101,075	0.040559397	0.045500442	0.891406658
Kittitas	16,405	13,174	0.005292699	0.005930476	0.892457708
Klickitat	9,820	8,351	0.003168199	0.003759329	0.842756427
Lewis	36,164	34,157	0.011667489	0.015376291	0.758797358
Lincoln	5,566	5,632	0.001795743	0.00253533	0.708287578
Mason	25,701	18,893	0.008291841	0.00850497	0.974940622
Okanogan	18,420	15,046	0.005942792	0.006773185	0.877400015
Pacific	10,214	9,204	0.003295314	0.00414332	0.795331737
PendOreille	5,709	4,486	0.001841878	0.002019441	0.912073511
Pierce	349,476	308,937	0.112750451	0.139072669	0.810730479
SanJuan	8,063	5,917	0.002601343	0.002663627	0.97661673
Skagit	66,322	49,147	0.021397279	0.022124266	0.967140723
Skamania	4,149	3,104	0.00133858	0.001397313	0.957967535
Snohomish	332,324	278,326	0.10721675	0.125292664	0.855730473
Spokane	231,030	202,904	0.074536554	0.091340308	0.816031341
Stevens	16,866	12,789	0.00544143	0.005757162	0.945158355
Thurston	121,894	104,118	0.039326316	0.046870294	0.839045632
Wahkaikum	1,634	1,513	0.000527173	0.0006811	0.774002197
WallaWalla	24,258	22,549	0.00782629	0.010150774	0.771004254
Whatcom	90,938	70,164	0.029339069	0.031585387	0.928881103
Whitman	17,061	16,285	0.005504342	0.007330939	0.750837213
Yakima	117,751	99,187	0.037989671	0.04465053	0.850822406
check sum	3,296,712	2,828,372	1.0636	1.2732	
STATE	3,296,712	2,828,372			

Table 8.2c Enrollment in Grades 1-8, 1990 and 2000 Data

COUNTY	Number Year = 2000	Number Year = 1990	Proportion Year = 2000	Proportion Year = 1990	Ratio of 2000 Prop/1990 Prop
Adams	2,417	2,277	0.000779745	0.001025026	0.76070721
Asotin	2,183	2,212	0.00070436	0.000995765	0.707355068
Benton	18,719	15,296	0.006039116	0.006885726	0.87704854
Chelan	8,268	6,567	0.002667485	0.002956234	0.902325116
Clallam	6,424	6,439	0.002072702	0.002898613	0.715066772
Clark	42,803	30,613	0.013809333	0.013780906	1.002062827
Columbia	381	521	0.000122885	0.000234536	0.523951293
Cowlitz	11,789	10,538	0.003803339	0.00474384	0.801742579
Douglas	3,979	3,285	0.001283695	0.001478792	0.868069579
Ferry	816	896	0.000263264	0.000403348	0.652696401
Franklin	6,980	5,760	0.002252063	0.002592951	0.868532899
Garfield	295	311	9.5175E-05	0.000140001	0.679814927
Grant	10,776	8,281	0.003476627	0.003727818	0.932617293
GHarbor	7,778	8,129	0.002509452	0.003659392	0.685756503
Island	6,433	5,803	0.002075538	0.002612308	0.794522595
Jefferson	2,282	2,145	0.00073618	0.000965604	0.762403811
King	173,328	145,005	0.055920321	0.065276197	0.856672483
Kitsap	27,470	23,320	0.008862526	0.010497851	0.844222898
Kittitas	2,907	2,637	0.000937955	0.001187085	0.790132316
Klickitat	2,365	2,370	0.000762987	0.001066891	0.715150057
Lewis	7,901	8,124	0.002549003	0.003657142	0.696993252
Lincoln	1,475	1,466	0.000475943	0.000659942	0.721188755
Mason	5,281	4,448	0.001703768	0.002002335	0.8508909
Okanogan	4,895	4,449	0.001579241	0.002002785	0.788522402
Pacific	2,068	2,069	0.000667125	0.000931392	0.71626711
PendOreille	1,242	1,150	0.000400677	0.00051769	0.773971288
Pierce	85,065	70,118	0.027444386	0.03156468	0.869465072
SanJuan	1,175	949	0.000379132	0.000427207	0.887467517
Skagit	12,035	9,713	0.003882792	0.004372454	0.88801211
Skamania	835	877	0.000269339	0.000394795	0.682224832
Snohomish	73,759	56,030	0.023796657	0.025222753	0.943459945
Spokane	48,216	43,219	0.015555879	0.019455687	0.799554304
Stevens	3,938	3,898	0.001270386	0.001754744	0.723972616
Thurston	23,806	20,459	0.007680617	0.009209929	0.833949692
Wahkaikum	318	287	0.000102595	0.000129197	0.794098348
WallaWalla	6,082	5,650	0.001962199	0.002543433	0.771476591
Whatcom	17,695	14,297	0.005708817	0.006436011	0.887011641
Whitman	3,120	3,079	0.001006639	0.001386058	0.726259907
Yakima	31,436	26,359	0.010142062	0.011865903	0.854723186
check sum	668,735	559,046	0.2158	0.2517	
STATE	668,735	559,046			

Table 8.2d Total Population, 1990 and 2000 Data

COUNTY	Number Year = 2000	Number Year = 1990	Proportion Year = 2000	Proportion Year = 1990	Ratio of 2000 Prop/1990 Prop
Adams	16,428	13,603	0.005300119	0.006123596	0.865523901
Asotin	20,551	17,605	0.006630311	0.007925157	0.836615678
Benton	142,475	112,560	0.045966305	0.050670589	0.907159495
Chelan	66,616	52,250	0.021492131	0.023521129	0.913737242
Clallam	64,525	56,464	0.020817518	0.025418125	0.819002903
Clark	345,238	238,053	0.111383158	0.107163163	1.039379154
Columbia	4,064	4,024	0.001311157	0.001811465	0.723810362
Cowlitz	92,948	82,119	0.02998755	0.036967111	0.811195376
Douglas	32,603	26,205	0.010518613	0.011796578	0.891666538
Ferry	7,260	6,295	0.002342273	0.00283379	0.826551571
Franklin	49,347	37,473	0.015920683	0.016869038	0.943781286
Garfield	2,397	2,248	0.000773337	0.001011971	0.764189025
Grant	74,698	54,758	0.024099604	0.024650143	0.977665896
GHarbor	67,194	64,175	0.02167861	0.028889348	0.750401489
Island	71,558	60,195	0.023086555	0.027097691	0.851974986
Jefferson	25,953	20,146	0.008373143	0.009069027	0.923268049
King	1,737,034	1,507,319	0.560414357	0.678542473	0.825909031
Kitsap	231,969	189,731	0.074839501	0.085410283	0.876235257
Kittitas	33,362	26,725	0.010763488	0.012030663	0.894671151
Klickitat	19,161	16,616	0.006181859	0.007479944	0.82645794
Lewis	68,600	59,358	0.022132224	0.026720903	0.828273803
Lincoln	10,184	8,864	0.003285635	0.003990264	0.823412987
Mason	49,405	38,341	0.015939395	0.017259782	0.923499229
Okanogan	39,564	33,350	0.012764421	0.015013008	0.850224126
Pacific	20,984	18,882	0.006770008	0.008500018	0.796469874
PendOreille	11,732	8,915	0.003785062	0.004013222	0.943147843
Pierce	700,820	586,203	0.22610357	0.263888157	0.856815905
SanJuan	14,077	10,035	0.004541623	0.004517407	1.005360465
Skagit	102,979	79,555	0.033223823	0.035812888	0.927705773
Skamania	9,872	8,289	0.003184975	0.003731419	0.853556112
Snohomish	606,024	465,642	0.195519806	0.209615798	0.932753198
Spokane	417,939	361,364	0.134838475	0.162673477	0.82889035
Stevens	40,066	30,948	0.01292638	0.013931711	0.927838668
Thurston	207,355	161,238	0.066898356	0.072583727	0.921671543
Wahkaikum	3,824	3,327	0.001233726	0.001497699	0.82374758
WallaWalla	55,180	48,439	0.017802567	0.021805549	0.816423687
Whatcom	166,814	127,780	0.053818728	0.057522102	0.935618244
Whitman	40,740	38,775	0.013143831	0.017455153	0.753005741
Yakima	222,581	188,823	0.071810677	0.085001533	0.844816262
check sum	5,894,121	4,866,692	1.9016	2.1908	
STATE	5,894,121	4,866,692			

Table 8.2e Registered Voters, 2000 and 2005 Data

COUNTY	Number Year = 2005	Number Year = 2000	Proportion Year = 2005	Proportion Year = 2000	Ratio of 2005 Prop/2000 Prop
Adams	6,477	6,098	0.001846242	0.00196738	0.938426384
Asotin	11,805	12,987	0.003364966	0.004189959	0.803102325
Benton	85,586	75,315	0.024395931	0.024298665	1.004002932
Chelan	37,395	32,803	0.010659288	0.010583139	1.007195336
Clallam	43,520	39,068	0.012405194	0.012604398	0.984195647
Clark	207,611	167,584	0.059178646	0.054067151	1.094539755
Columbia	2,542	2,671	0.000724586	0.000861737	0.840843924
Cowlitz	53,914	49,643	0.01536796	0.01601618	0.95952715
Douglas	16,994	16,855	0.004844069	0.005437881	0.890800781
Ferry	4,088	3,856	0.001165267	0.00124405	0.936672121
Franklin	21,235	16,321	0.006052948	0.005265598	1.149527149
Garfield	1,524	1,670	0.00043441	0.000538787	0.806273207
Grant	32,760	29,970	0.009338101	0.009669136	0.965763711
GHarbor	36,647	32,038	0.010446074	0.010336329	1.010617382
Island	43,688	38,265	0.012453081	0.012345329	1.008728237
Jefferson	21,165	17,330	0.006032995	0.005591129	1.079029809
King	1,082,406	1,001,339	0.308535298	0.323059164	0.955042706
Kitsap	138,956	125,219	0.039608826	0.040399051	0.980439512
Kittitas	19,817	16,417	0.005648753	0.00529657	1.066492593
Klickitat	12,163	11,717	0.003467012	0.003780223	0.917145013
Lewis	38,007	40,913	0.010833736	0.013199645	0.820759649
Lincoln	6,642	6,656	0.001893274	0.002147406	0.881656249
Mason	31,083	27,238	0.008860079	0.008787719	1.008234247
Okanogan	20,066	18,159	0.005719729	0.005858587	0.976298476
Pacific	13,195	12,697	0.003761179	0.004096397	0.918167693
PendOreille	7,486	6,903	0.002133853	0.002227095	0.958132743
Pierce	405,023	325,079	0.11545011	0.104879316	1.10079007
SanJuan	11,246	9,228	0.003205625	0.002977203	1.076723584
Skagit	63,185	55,780	0.01801062	0.017996143	1.000804414
Skamania	6,305	5,586	0.001797214	0.001802195	0.997235871
Snohomish	352,238	303,110	0.100403967	0.09779152	1.02671445
Spokane	251,184	209,404	0.071598947	0.067559419	1.05979223
Stevens	28,414	25,481	0.008099292	0.008220863	0.985211881
Thurston	137,742	119,016	0.03926278	0.038397795	1.022526959
Wahkaikum	2,592	2,455	0.000738839	0.00079205	0.932818677
WallaWalla	29,279	24,411	0.008345856	0.007875652	1.059703579
Whatcom	106,094	90,987	0.03024165	0.029354878	1.030208693
Whitman	21,082	25,273	0.006009336	0.008153756	0.737002132
Yakima	97,052	94,011	0.027664266	0.030330502	0.912093896
STATE	3,508,208	3,099,553	1.0000	1.0000	

Table 8.2f Registered Autos, 2000 and 2005 Data

COUNTY	Number Year = 2005	Number Year = 2000	Proportion Year = 2005	Proportion Year = 2000	Ratio of 2005 Prop/2000 Prop
Adams	12,064	9,144	0.003438793	0.002950103	1.165651813
Asotin	11,853	10,375	0.003378648	0.003347257	1.009378178
Benton	103,288	80,977	0.029441812	0.02612538	1.126942914
Chelan	40,826	39,153	0.01163728	0.012631821	0.921267009
Clallam	43,880	35,697	0.01250781	0.011516822	1.086047029
Clark	238,323	183,053	0.067932973	0.059057871	1.150278066
Columbia	2,602	2,186	0.000741689	0.000705263	1.05164913
Cowlitz	59,836	52,461	0.017056001	0.016925344	1.007719636
Douglas	23,100	13,008	0.006584558	0.004196734	1.568971966
Ferry	2,767	2,384	0.000788722	0.000769143	1.025455079
Franklin	35,678	27,518	0.010169865	0.008878054	1.145505997
Garfield	1,413	1,263	0.00040277	0.000407478	0.988445079
Grant	42,352	35,188	0.01207226	0.011352605	1.063391227
GHarbor	38,934	33,310	0.011097974	0.010746711	1.032685607
Island	47,153	37,675	0.013440765	0.012154978	1.105782723
Jefferson	18,982	14,459	0.00541074	0.004664866	1.159891708
King	1,227,244	1,083,380	0.349820763	0.349527819	1.000838114
Kitsap	152,831	125,716	0.043563837	0.040559397	1.074075061
Kittitas	20,690	16,405	0.005897598	0.005292699	1.114289372
Klickitat	11,859	9,820	0.003380358	0.003168199	1.066965344
Lewis	39,820	36,164	0.011350524	0.011667489	0.972833523
Lincoln	6,025	5,566	0.001717401	0.001795743	0.956373605
Mason	34,352	25,701	0.009791894	0.008291841	1.180907111
Okanogan	21,622	18,420	0.006163261	0.005942792	1.037098412
Pacific	12,270	10,214	0.003497512	0.003295314	1.061359329
PendOreille	7,157	5,709	0.002040073	0.001841878	1.107604487
Pierce	436,245	349,476	0.124349811	0.112750451	1.102876387
SanJuan	10,736	8,063	0.003060252	0.002601343	1.176412351
Skagit	81,691	66,322	0.023285677	0.021397279	1.088254146
Skamania	5,032	4,149	0.001434351	0.00133858	1.071546273
Snohomish	412,919	332,324	0.117700832	0.10721675	1.09778399
Spokane	277,551	231,030	0.07911475	0.074536554	1.06142216
Stevens	20,268	16,866	0.005777309	0.00544143	1.061726194
Thurston	163,196	121,894	0.046518336	0.039326316	1.182880611
Wahkaikum	2,080	1,634	0.000592895	0.000527173	1.124669752
WallaWalla	29,277	24,258	0.008345286	0.00782629	1.066314496
Whatcom	115,773	90,938	0.033000609	0.029339069	1.124800811
Whitman	20,277	17,061	0.005779874	0.005504342	1.050057184
Yakima	141,179	117,751	0.040242483	0.037989671	1.059300628
STATE	3,973,145	3,296,712	1.1325	1.0636	

Table 8.2g Enrollment in Grades 1-8, 2000 and 2005 Data

COUNTY	Number Year = 2005	Number Year = 2000	Proportion Year = 2005	Proportion Year = 2000	Ratio of 2005 Prop/2000 Prop
Adams	2,482	2,417	0.000707381	0.000779745	0.907195775
Asotin	2,077	2,183	0.00059204	0.00070436	0.840536749
Benton	19,064	18,719	0.005434222	0.006039116	0.899837281
Chelan	7,930	8,268	0.002260533	0.002667485	0.847439938
Clallam	5,899	6,424	0.001681528	0.002072702	0.811273366
Clark	46,759	42,803	0.013328426	0.013809333	0.965175193
Columbia	389	381	0.000110871	0.000122885	0.902233821
Cowlitz	11,373	11,789	0.003241755	0.003803339	0.852344476
Douglas	4,067	3,979	0.001159361	0.001283695	0.903143919
Ferry	736	816	0.000209651	0.000263264	0.796354155
Franklin	8,701	6,980	0.002480283	0.002252063	1.101338148
Garfield	241	295	6.87473E-05	9.5175E-05	0.7223256
Grant	10,846	10,776	0.003091595	0.003476627	0.889251387
GHarbor	7,155	7,778	0.00203952	0.002509452	0.812735113
Island	5,909	6,433	0.00168447	0.002075538	0.811582196
Jefferson	1,933	2,282	0.000551099	0.00073618	0.748592414
King	170,347	173,328	0.048556614	0.055920321	0.868317855
Kitsap	25,376	27,470	0.007233434	0.008862526	0.816181917
Kittitas	2,964	2,907	0.000844947	0.000937955	0.900840028
Klickitat	1,984	2,365	0.000565508	0.000762987	0.741176146
Lewis	7,682	7,901	0.002189579	0.002549003	0.85899443
Lincoln	1,341	1,475	0.000382349	0.000475943	0.80335081
Mason	5,074	5,281	0.001446394	0.001703768	0.848938059
Okanogan	4,021	4,895	0.001146141	0.001579241	0.725754324
Pacific	1,817	2,068	0.000518037	0.000667125	0.776520715
PendOreille	1,110	1,242	0.000316458	0.000400677	0.789807647
Pierce	84,043	85,065	0.023956174	0.027444386	0.872898863
SanJuan	1,126	1,175	0.000320819	0.000379132	0.846193378
Skagit	12,072	12,035	0.003441122	0.003882792	0.886249222
Skamania	748	835	0.000213169	0.000269339	0.791451626
Snohomish	73,322	73,759	0.020900101	0.023796657	0.878278846
Spokane	46,975	48,216	0.013389944	0.015555879	0.860764266
Stevens	3,754	3,938	0.00107015	0.001270386	0.842381765
Thurston	24,096	23,806	0.006868415	0.007680617	0.894253064
Wahkaikum	302	318	8.60838E-05	0.000102595	0.839061039
WallaWalla	6,027	6,082	0.001717988	0.001962199	0.875542265
Whatcom	17,575	17,695	0.005009683	0.005708817	0.877534391
Whitman	2,891	3,120	0.000824028	0.001006639	0.818593144
Yakima	31,688	31,436	0.009032589	0.010142062	0.890606697
STATE	661,898	668,735	0.1887	0.2158	

Table 8.2h Estimated Population 2005

COUNTY	Number Year = 2000	Proportion Year = 2000	Estimated Ratio of 2005 Prop / 2000 Prop	Estimated Proportion Year = 2005	Estimated Population 2005 Not Controlled	Estimated Population 2005 Controlled
Adams	16,428	0.002787184	1.063487897	0.002964136	18,545	18,125
Asotin	20,551	0.003486695	0.971181915	0.003386215	21,186	20,706
Benton	142,475	0.024172391	1.054035167	0.025478551	159,404	155,792
Chelan	66,616	0.011302109	0.965545336	0.010912699	68,274	66,727
Clallam	64,525	0.010947349	0.998966852	0.010936039	68,420	66,870
Clark	345,238	0.05857328	1.099587137	0.064406425	402,952	393,823
Columbia	4,064	0.000689501	1.016129849	0.000700622	4,383	4,284
Cowlitz	92,948	0.015769612	0.990626693	0.015621798	97,736	95,522
Douglas	32,603	0.005531444	1.184544909	0.006552244	40,993	40,065
Ferry	7,260	0.001231736	0.968611432	0.001193073	7,464	7,295
Franklin	49,347	0.008372241	1.165182116	0.009755185	61,032	59,650
Garfield	2,397	0.000406676	0.91106728	0.00037051	2,318	2,266
Grant	74,698	0.012673306	1.025583671	0.012997536	81,318	79,475
GHarbor	67,194	0.011400173	0.985248907	0.011232008	70,272	68,680
Island	71,558	0.012140572	1.007627662	0.012233176	76,536	74,802
Jefferson	25,953	0.004403201	1.002602877	0.004414662	27,620	26,994
King	1,737,034	0.2947062	0.995305428	0.29332268	1,835,144	1,793,565
Kitsap	231,969	0.039355996	0.99707038	0.039240697	245,505	239,943
Kittitas	33,362	0.005660216	1.056326591	0.005979037	37,407	36,560
Klickitat	19,161	0.003250866	0.954783049	0.003103872	19,419	18,979
Lewis	68,600	0.011638716	0.969691291	0.011285961	70,609	69,010
Lincoln	10,184	0.001727823	0.944850982	0.001632536	10,214	9,982
Mason	49,405	0.008382081	1.048303049	0.008786961	54,975	53,729
Okanogan	39,564	0.006712451	0.943852979	0.006335567	39,638	38,740
Pacific	20,984	0.003560158	0.969194674	0.003450486	21,588	21,099
PendOreille	11,732	0.001990458	0.993572129	0.001977664	12,373	12,093
Pierce	700,820	0.118901529	1.043206233	0.124038816	776,036	758,454
SanJuan	14,077	0.002388312	1.052024748	0.002512563	15,720	15,363
Skagit	102,979	0.017471477	1.035336839	0.018088864	113,171	110,607
Skamania	9,872	0.001674889	0.986583624	0.001652418	10,338	10,104
Snohomish	606,024	0.102818385	1.037135674	0.106636615	667,161	652,045
Spokane	417,939	0.070907774	1.020769569	0.072380498	452,841	442,581
Stevens	40,066	0.006797621	1.005540852	0.006835285	42,764	41,795
Thurston	207,355	0.03517997	1.070883741	0.037673658	235,701	230,361
Wahkaikum	3,824	0.000648782	1.019015058	0.000661119	4,136	4,043
WallaWalla	55,180	0.009361871	1.029031869	0.009633664	60,272	58,906
Whatcom	166,814	0.02830176	1.045653176	0.029593826	185,151	180,956
Whitman	40,740	0.006911972	0.967871665	0.006689902	41,855	40,906
Yakima	222,581	0.037763222	1.019901547	0.038514769	240,964	235,504
STATE					6,401,438	6,256,400
	5,894,121	1.0000		1.0232	6,256,400	

References

Bryan, T. (2004). "Population Estimates." pp. 523–560 in J. Siegel and D. A. Swanson (Eds.) *The Methods and Materials of Demography, 2nd Edition*. Amsterdam, The Netherlands: Elsevier Academic Press

Crosetti, A., and R. Schmitt. (1956). "A Method of Estimating the Inter-censal Population of Counties." *Journal of the American Statistical Association* 51: 587–590.

D'Allesandro, F. and J. Tayman. (1980). "Ridge Regression for Population Estimation: Some Insights and Clarifications." *Staff Document No. 56*. Olympia, Washington, Office of Financial Management

Ericksen, E. (1974). "A Regression Method for Estimating Population Changes of Local Areas." *Journal of the American Statistical Association* 69: 867–875.

Ericksen, E. (1973). "A Method for Combining Sample Survey Data and Symptomatic Indicators to obtain Population Estimates for Local Areas." *Demography* 10: 137–160.

Feeney, D., J. Hibbs, and T. Gillaspy. (1995). "Ratio-Correlation Method." pp. 118–136 in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (Eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI" Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Fox, J. (1991). *Regression Diagnostics*. Quantitative Applications in the Social Sciences Series, no. 79. London, England: Sage Publications.

Mandell, M., and J. Tayman. (1982). "Measuring Temporal Stability in Regression Models of Population Estimation." *Demography* 19 (1): 135–146.

McKibben, J., and D. Swanson. (1997). "Linking Substance and Practice: A Case Study of the Relationship between Socio-economic Structure and Population Estimation." *Journal of Economic and Social Measurement* 24 (2): 135–147.

Murdock, S. S. Hwang, and R. Hamm. (1995). "Component Methods" pp. 10–53 in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (Eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI" Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Namboodiri, N. K., and N. Lalu. (1971). "The Average of Several Simple Regression Estimates as an Alternative to the Multiple Regression Estimate in Post-censal and Inter-censal Population Estimation: A Case Study." *Rural Sociology* 36: 187–194.

O'Hare, W. (1980). "A Note on the use of Regression Methods in Population Estimates." *Demography* 17 (3): 341–343.

Schmitt, R., and A. Crosetti. (1954). "Accuracy of the Ratio-correlation Method for Estimating Post-censal Population." *Land Economics* 30: 279–281.

Schmitt, R., and J. Grier. (1966). "A Method of Estimating the Population of Minor Civil Divisions." *Rural Sociology* 31: 355–361

Snow, E.C. (1911). "The application of the method of multiple correlation to the estimation of post-censal populations." *Journal of the Royal Statistical Society* 74 (part 6): 575–629 (pp. 621–629 contain the discussion).

Spar, M. and J. Martin. (1979). "Refinements to Regression-based Estimates of Post-censal Population Characteristics." *Review of Public Data Use* 7: 16–22.

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University.

Swanson, D. (2004). "Advancing Methodological Knowledge within State and Local Demography: A Case Study." *Population Research and Policy Review* 23 (4): 379–398

Swanson, D. (1980). "Improving Accuracy in Multiple Regression Estimates of County Populations Using Principles from Causal Modeling. *Demography* 17 (November):413–427.

Swanson, D. (1978a). "An Evaluation of Ratio and Difference Regression Methods for Estimating Small, Highly Concentrated Populations: The Case of Ethnic Groups." *Review of Public Data Use* 6 (July):18–27.

Swanson, D. (1978b). "Preliminary Results of an Evaluation of the Utility of Ridge Regression for Making County Population Estimates." Presented at the Annual Meeting of the Pacific Sociological Association, Spokane, WA.

Swanson, D. (1989). "Confidence Intervals for Post-censal Population Estimates: A Case Study for Local Areas." *Survey Methodology* 15 (2): 271–280.

Swanson, D. and D. Beck. (1994). "A New Short-term County Population Projection Method." *Journal of Economic and Social Measurement* 21:25–50.

Swanson, D. and R. Prevost. (1985). "A New Technique for Assessing Error in Ratio-Correlation Estimates of Population: A Preliminary Note." *Applied Demography* 1 (November): 1–4.

Swanson, D. and L. M. Tedrow. (1984). "Improving the measurement of temporal change in regression models used for county population estimates" *Demography* 21 (3): 373–381.

Tayman, J., and E. Schafer. (1985). "The Impact of Coefficient Drift and Measurement Error on the Accuracy of Ratio-Correlation Population Estimates." *The Review of Regional Studies.* 15 (2): 3–1.

US Census Bureau. (2010). Appendix A, Source Notes and Explanations, pp A1–A78 in *State and Metropolitan Area Data Book: 2010*. Washington, DC: US Census Bureau (http://www.census.gov/compendia/databooks/pdf_version.htm).

Chapter 9
Censal-Ratio Methods

9.1 Introduction

In the late 17th century, John Graunt estimated the population of London and then of the whole of England and Wales using what today is known as a censal-ratio method (Devlin 2008: 93-94). Not long afterward, in the 18th century, the French mathematician, Laplace, also used a censal-ratio method in combination with recorded births and a population sample to estimate the population of France (Stigler 1986:163-164). However, methodological development really only took off in the late 1930s and early 1940s, fueled in large part by the need for low-cost and timely information generated by the great depression of the 1930s and World War II (Bryan 2004; Eldridge 1947; Shryock 1938; Shryock and Lawrence 1949). In modern times, the censal-ratio method is usually traced to Bogue (1950) who introduced the "vital rates method."

9.2 Approaches

The censal-ratio method can be implemented in several different ways. The most basic approach is to use relationships between symptomatic indicators and population counts in census years to estimate populations in non-census years and applying these relationships to symptomatic indictors available in the years for which estimates are desired. Borrowing from the notation in the preceding chapter on regression methods, the general form of this approach is as follows.

$$R_{i,j,t} = S_{i,j,t}/P_{i,t} \tag{9.1a}$$

where

R = Censal-ratio
P = population

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_9,

S = symptomatic indicator
j = indicator ($1 \leq j \leq k$)
i = subarea ($1 \leq j \leq n$)
t = year of the most recent census

Once a censal-ratio is constructed, a population estimate for time t + k is developed by dividing the t + k value of the symptomatic indicator ($S_{i,j,t+k}$) by the ratio ($R_{i,j,t}$) to yield an estimate of $P_{i,t+k}$:

$$\hat{P}_{i,t+k} = S_{i,j,t+k}/R_{i,j,t} \tag{9.1b}$$

If area i has a parent area for which an independently-derived population estimate is available, then, as was the case for the ratio-correlation model discussed in Chapter 8, it is common is to effect a final "control" so that the sum of the i subarea population estimates is equal to the independently estimated population for the parent of these i subareas, $\sum \mathbf{P_{i,t+k}}$, which is accomplished as follows:

$$\hat{P}_{i,t+k} = (\hat{P}_{i,t+k}/\Sigma\hat{P}_{i,t+k})^{*}(\Sigma\hat{P}_{i,t+k}). \tag{9.1c}$$

It should be noted that as long as the algebra yields an estimate of P_i at time t + k, it is immaterial if $R_{i,j,t} = P_{i,t}/ S_{i,j,t}$ or if $R_{i,j,t} = S_{i,j,t}/ P_{i,t}$. In the case of the latter version, Equation [9.1a] and [9.1b] become, respectively, [9.1d] and [9.1e]:

$$R_{i,j,t} = P_{i,jt}/S_{ij,,t} \tag{9.1d}$$

$$\hat{P}_{i,t+k} = (R_{i,j,t})/(S_{i,j,t+k}) \tag{9.1e}$$

One advantage of using Equations [9.1a] and [9.1b] over [9.1d] and 9.1c] is that the resulting ratio of interest is easier to interpret. As the following example shows, if one uses deaths as the symptomatic indicator, then the ratio is the crude death rate. Similarly, if one uses births, the resulting ratio is the crude birth rate.

Here, we provide two examples of the censal-ratio method, one for an area with a large population and the other for an area with a small population. The large population example is for a 2006 estimate of the population of the state of Washington and the other one for a 2006 estimate of the population of Garfield County, which is one of the smallest counties in the state of Washington.

In 2000, the count of the state population was 5,894,143 and the number of reported deaths in 2000 and 2006, was 43,904 and 45,878, respectively (State of Washington 2009a). Using Equation [9.1a] we estimate the ratio of deaths to population at time = t (2000) as 0.0074 = 43,904/ 5,894,143. We can interpret the ratio, 0.0074, as the crude death rate for the state of Washington in 2000. Using Equation [9.1b] our 2006 estimate is 6,199,730 = (45,878/0.0074). This estimate compares favorably with the state's official 2006 population estimate of 6,376,600 (State of Washington 2009a:3), with a numerical difference of -176,870 and a relative difference of -2.77% (-2.77 = (6,199,730 - 6,376,600/6,376,600)*100).

Turning now to Garfield County, its 2000 census population was 2,976 (State of Washington 2010). There were 20 deaths reported for Garfield County in 2000 (State of Washington 2002) and 28 for 2006 (State of Washington 2009b). Using Equation [9.1a] we estimate the ratio of deaths to population at time = t (2000) as 0.0067 = 20/ 2,976. We can interpret the ratio, 0.0067, as the crude death rate for Garfield County in 2000. Using Equation [9.1b] our 2006 estimate is 4,166 = (28/0.0067). This estimate does not compare favorably with the state's official 2006 population estimate of Garfield County, which is 2,400 (State of Washington 2010). This is an absolute difference of 1,766 and relative difference of 73.6% (7.36 = (4,166-2,400/2,400)*100).

What accounts for the small difference between the 2006 estimates for the state as a whole and the large difference between the 2006 estimates for Garfield County? Clearly, the accuracy of the Censal-ratio method depends on the stability of the relationship between P and S over time, whether the first approach (equations [9.1a] and [9.1b]) is used or the latter (equations [9.1d] and [9.1e]). Thus, "invariance" is an underlying assumption of any censal-ratio method, as it is for the ratio-correlation method discussed in Chapter 8.

How can we examine the stability of the relationship between P and S over time? The answer to this question is found in the work of Voss et al. (1995). They argue that a symptomatic indicator can be viewed as the outcome of a random variable, which leads to using the statistical properties in the symptomatic indicator that can be used for purposes of estimation. This is a very insightful contribution that Voss and his colleagues use to illustrate how censal-ratio estimators can be examined and improved (Voss et al. 1995: 73-79).

It is useful to consider deaths as an example because everybody is at risk of dying. Using deaths, Voss and his colleagues (1995) looked at the crude death rate of a given area i as the marginal probability of death for the area's inhabitants. This leads to looking at the distribution of deaths in a given area i as (approximately) binomial or Poisson with parameter d, where d is defined as follows.

$$d_{i,t} = D_{i,t}/P_{i,t} \tag{9.2a}$$

where

i = area (i = 1 to n)
t = time
D = deaths
P = population

Still following Voss et al. (1995), equation [9.2] can be re-written so that the Expected number of deaths at time = k in area i is:

$$E[D_{i,t}] = d_{i,t}{}^{*}P_{i,t} \tag{9.2b}$$

The preceding leads to defining the variance of $D_{i,t}$:

$$V[D_{i,t}] = P_{i,t}{}^{*}(d_{i,t}(1 - d_{i,t})) \tag{9.2c}$$

Which for a given area i with a very small population (and hence, a very small number of deaths leads to a variance that is approximately

$$P_{i,t}{}^{*}d_{i,t} \tag{9.2d}$$

So, if we define the variance of $D_{i,t}$ as is done in Equation [9.2c], we have a binomial distribution and if we define it as in Equation [9.2d], we have a Poisson distribution. If d is assumed to be known, then the recorded number of deaths at time t in area i, $D_{i,t}$, leads to an estimate of $P_{i,t}$ that comes with the following statistical properties (Voss et al. 1995: 74):

(1) $\hat{P}_{i,t}$ is an unbiased estimator for $P_{i,t}$ since

$$E[\hat{P}_{i,t}] = E[D_{i,t}/d_{i,t}] = P_{i,t} \tag{9.2e}$$

(2) The variance of $\hat{P}_{i,t}$ is

$$V[\hat{P}_{i,t}] = (P_{i,t}{}^{*}(1 - d_{i,t})/d_{i,t} \tag{9.2f}$$

and
(3) the coefficient of variation for $\hat{P}_{i,t}$ is

$$CV[\hat{P}_{i,t}] = [(1 - d_{i,t})/(d_{i,t}{}^{*}P_{i,t})]^{1/2} \tag{9.2g}$$

As can be seen in Equation [9.2 g], the coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean. It is most useful for variables that are always positive, which is the case for population estimates made using censal-ratio methods. In general terms, as the "sample" size decreases, the size of the CV increases and a large CV indicates that the sampling error is large relative to the estimate, and thus the user is less confident that the estimate is close to the population value (see, e.g., US Census Bureau 2008: A-13). In terms of its use with a censal-ratio estimator, as the number of events measured by the symptomatic indicator (e.g., deaths) decreases, the size of the CV increases. Thus, as observed by Voss et al. (1995: 75-76) for symptomatic data with a small count, the natural variation induced in the estimate by the binomial process would tend to be the dominant source of error.

Using again our example data for the state of Washington in conjunction with equations [9.2f] and [9.2 g], we find that the variance of our 2000 estimator is 790,612 = (5,894,143*(1-0.0074)/0.0074) and the CV is 0.0048 = [(1-0.0074)/(0.0074*5,894,143)]^{½}. The low CV reflects the fact that we have a relatively large count for our symptomatic indicator, deaths, which suggests that the natural variation in deaths is not a dominant source of error in our censal-ratio estimate of 6,199,730 for the state's population in 2006. A very difference picture emerges when we construct the CV for Garfield County. Here we find that the variance of

our 2000 estimator is 441,203 = (2,976*(1-0.0067)/0.0067) and the CV is .050 = $[(1\text{-}0.0067)/(0.0067*2{,}976)]^{½}$. Garfield County's CV is about ten times larger than the CV for the state of Washington. It suggests that natural variation in deaths is a dominant source in our censal-ratio estimate of 4,166. The CV for Garfield County illustrates three major points. First, it represents the 'instability' inherent to such a small population and its relatively small number of deaths, which in changing from 20 deaths in 2000 to 28 in 2006 produced a population estimate that is 73 percent larger than the state's official estimate. Second, it illustrates the need to use a range of methods in dealing with small populations and the importance of embedding their estimates within a larger context. Third, the first and second points suggest that the symptomatic data themselves should be embedded within a larger context.

In developing such a context, the US National Center for Health Statistics, for example, uses three year averages centered on the year of interest as a way to improve the accuracy of death rate estimates made using small death counts (NCHS 1994: 30). This provides a large context, but it must be done carefully. As illustrated by Voss et al. (1995: 76-79), this strategy leads to biased estimates of the death rate except under very limited conditions. Thus, while a death rate for that is built using a centered three-year average may have a smaller variance than its one-year counterpart, it is likely to be biased. Voss and his colleagues (Voss et al. 1995) experimented with alternatives to three-year centered averages that would not have the bias likely in this approach but would have a smaller variance than a death rate constructed from a single year's worth of mortality data. They found that an autoregressive approach (see, e.g., the discussion on ARIMA in Chapter 6) yielded satisfactory results. This extended the 'larger data context' substantially. This suggests that it will work well in those areas with a long history of annual counts of symptomatic data, but not in areas that have only a limited number of annual counts available.

Another variation on fundamental form of the censal-ratio method is to use symptomatic indicators for a "parent area" if they are only available for the areas in which estimates are desired in census years (and not in the years for which estimates are desired). This is another way of increasing the context for symptomatic indicators representing small populations. This approach, described by Voss et al. (1995), is similar to the "synthetic method" of estimation, which is described in detail in Chapter 11 and, therefore, not discussed here.

Yet another variation on the fundamental form of the censal-ratio method is the Composite Method (Bryan 2004: 550-551). This method generally involves computing age-specific death rates in a census year as illustrated in Equation [9.1a] and then using the reported deaths by age in the estimation year in conjunction with equation [9.1b] to develop estimates of the population by age, which are then adjusted using other information and summed to obtain an estimate of the total population.

As an example of one of the many variations on the composite method, we use school enrollment data to estimate the 2010 population of Inyo County, California and compare our estimate to the 2010 Census number. Because we use school enrollment data, the example here will have similarities to the example of

Component Method II we provide for Inyo County in Chapter 10, where we generate a 2010 estimate that like this one can be compared to the 2010 census number.

Our example "jumps off" from the 2000 census population base. We use enrollment in grades K-4 and 5-9 in our example because they correspond very closely with age groups 5-9 and 10-14. In terms of our "census ratios" we use the following ratios for 2000: (1) the ratio of the population aged 5-9 enumerated in the 2000 census to the enrollment reported for 2000-1 (fall 2000 enrollment) in grades K-4; (2) the ratio of the population aged 10-14 enumerated in the 2000 census to the enrollment reported for 2000-1 (Fall 2000 enrollment) in grades 5-9; (3) the ratio of the population aged 0-4 enumerated in the 2000 census to the population aged 5-14 enumerated in the 2000 census; (4) the ratio of the population aged 15 years and over enumerated in the 2000 census to the population aged 5-14 enumerated in the 2000 census. With these ratios and the 2010-11 enrollment in grades K-4 and 5-9 (Fall 2010 enrollment) we can estimate the 2010 total population of Inyo County as follows.

For 2000-1 (Fall 2000 enrollment), the California Department of Education (2011a) shows 1,170 students enrolled in grades K through 4 and 1,382 in grades 5-9, respectively, for Inyo County. Our census 2000 population aged 5-9 is 1,184, which yields a ratio of 1.01197 (= 1,184/1,170) relative to enrollment in grades K-4. Our census 2000 population aged 10-14 is 1,360, which yields a ratio of 0.98408 (=1,360/1,382) relative to enrollment in grades 5-9. Our census 2000 population aged 0-4 in 2000 is 961, which yields a ratio of 0.37775 (=961/2,544) relative to our census 2000 population aged 5-14. Our census 2000 population aged 15 years and over is 14,440, which yields a ratio of 5.67610 (=14,440/2,544) relative to our census 2000 population aged 5-14.

For Fall 2010-11 (Fall 2010 enrollment), the California Department of Education (2011b) shows 999 students enrolled in grades K through 4 and 1,317 in grades 5-9, respectively, for Inyo County. Multiplying the 999 students in grades K-4 by 1.01197 provides an estimated population aged 5-9 of 1,011. Multiplying the 1,317 students in grades 5-9 by 0.98408 provides an estimated population aged 5-9 of 1,338. Multiplying the estimated 2010 population of 2,349 for age group 5-14 by 0.37775 yields an estimate of 887 for the 2010 population of Inyo County aged 0-4. Multiplying the estimated 2010 population of 2,349 for age group 5-14 by 5.67610 yields an estimate of 13,335. Adding together our estimates for age groups 0-4, 5-14, and 15+ gives an estimated total 2010 population of 16,572. relative to the 2010 census number of 18,546 (US Census Bureau 2010), our estimate has an absolute difference of -1,974 and a relative difference of -10.65%.

9.3 Summary

As evidenced by its ubiquity today and John Graunt's use of it over 300 years ago, the Censal-ratio method has staying power. The examples, the wide-spread use and the staying power of censal ratio methods suggest that there are useful symptomatic indicators readily available. Bryan (2004) advises that to be able to consistently use

a given censal ratio method, accurate and comparable data must be available at frequent intervals, including the census date. Bryan (2004) also advises that the number of annual cases should be high in relation to population size, which is something that the "statistical interpretation" we just examined can assist in evaluating. Bryan (2004) also notes that if a small area has a large proportion of military personnel or group quarters that these should be withdrawn before the calculations are made, then added again at the estimate date. Following this suggestion results in what amounts to a composite method, which has features in common with Component Method II, as we will show in Chapter 10. Bryan (2004) also suggests that averages can be taken of the estimates resulting from different symptomatic indicators since the averaging process may partly offset opposite biases characteristic of the birth-rate estimate and the death-rate estimate. For example, if a population estimate is too low as a result of an overestimate of the birth rate, the other population estimate is likely to be too high as a result of an underestimate of the death rate, because an age distribution that favors a high birth rate also generally favors a low death rate.

As discussed in Chapter 8 in regard to regression-based techniques, a wide range of variables have been considered as symptomatic indicators. The list includes school enrollment or school census data, number of electric, gas, or water meter accounts, bank receipts, building permits issued, voter registration rolls, welfare rolls, motor vehicle registrations, birth statistics, death statistics, tax returns, and covered employment.

It is safe to say that some variation of the censal-ratio method is widely used in the development of population estimates. In point of fact, most estimation methods are, at their core, censal-ratio methods. For example, the Housing Unit Method can be viewed as a censal-ratio method where housing units serve as symptomatic indicators. However, in the case of the Housing Unit Method, we believe that it is sufficiently important and its use so widespread that it warranted a separate chapter. The ratio-correlation regression method discussed in Chapter 8 also uses censal ratios in the form of its symptomatic indicators.

The symptomatic indictors chosen to use in a given application of the censal-ratio method are typically those that represent data related to population that are collected for administrative or legal purposes. The "timing" between the collection of population and that of a given symptomatic indicator (or set of symptomatic indicators) generally determines if the ratio is to be used for estimation or projection. If the timing between past correspondence of population and its symptomatic indicators is synchronous, then the ratio is generally used to estimate the population for both a post-censal and an inter-censal estimate. If the timing is lagged, such that values of population are collected at points in time beyond the time for which values of symptomatic indicators are collected, then the ratio method use current values of S to project future values of P to obtain a current post-censal estimate, such as the case both with the "lagged ratio-correlation model" discussed in Chapter 8 and "Component Method II," which is discussed in Chapter 10.

References

Bogue, Donald J. (1950). "A Technique for Making Extensive Population Estimates." Journal of the American Statistical Association. Vol. 45 (June): 149–163.

Bryan, T. (2004). "Population Estimates." pp. 523–560 in J. Siegel and D. A. Swanson (eds.) *The Methods and Materials of Demography, 2nd Edition*. Amsterdam, The Netherlands: Elsevier Academic Press

California Department of Education. (2011a). "Enrollment by Grade for 2001-2001, Inyo County." California Department of Education Online Query System (http://dq.cde.ca.gov/dataquest/).

California Department of Education. (2011b). "Enrollment by Grade for 2010–2011, Inyo County." California Department of Education Online Query System (http://dq.cde.ca.gov/dataquest/).

Devlin, K. (2008). *The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern*. New York, NY. Basic Books.

Eldridge, H. (1947). "Problems and Methods of Estimating Post-censal Population." *Social Forces* 24: 41–46.

Shryock, H. (1938). "Methods of Estimating Post-censal Population." *American Journal of Public Health* 28: 1042–1047

Shryock, H., and N. Lawrence. (1949). "The Current Status of State and Local Population Estimates in the Census Bureau." *Journal of the American Statistical Association* 44 (246): 157–173.

State of Washington. (2010). *April 1st Official Population Estimates*. Olympia, WA: Washington State Office of Financial Management. (http://www.ofm.wa.gov/pop/april1/default.asp).

State of Washington. (2009a). *Washington State Data Book 2009*. Olympia, WA: Washington State Office of Financial Management. (http://www.ofm.wa.gov/databook/default.asp).

State of Washington, (2009b). *Washington State Vital Statistics and Induced Terminations of Pregnancy 2006*. Olympia, WA: Center for Health Statistics, Washington State Department of Health (http://www.doh.wa.gov/ehsphl/chs/chs-data/public/AnnSum_2006.pdf).

State of Washington, (2002). *Washington State Vital Statistics 2000*. Olympia, WA: Center for Health Statistics, Washington State Department of Health. (http://www.doh.wa.gov/ehsphl/chs/chs-data/public/AnnSum_2000.pdf).

Stigler, S. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University.

US Census Bureau. (2008). *A Compass for Understanding and Using American Community Survey Data: What General Data Users need to Know*. Washington, DC: US Census Bureau.

US Census Bureau. (2010). "GCT-PL2, Population and Housing Occupancy Status: 2010 - State – County / County Equivalent, Inyo County, California (http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?fpt=table).

US National Center for Health Statistics (NCHS). (1994). Technical Appendix From Vital Statistics of the United States, 1994, Mortality. Washington DC: US Department of Health and Human Services, Center for Disease Control and Prevention, National Center for Health Statistics (http://www.cdc.gov/nchs/data/statab/techap94.pdf).

Voss, P. R., C. Palit, B. Kale, and H. Krebbs. (1995). "Censal Ratio Methods" pp. 70–88 in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (Eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI" Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Chapter 10
Component Methods

10.1 Introduction

These methods have a long history, but not always under the names we now know them by (Shryock and Lawrence 1949). There are several methods of population estimation that belong to the "component" family. All of the component methods are based on the fundamental demographic equation:

$$P_{i,t+k} = P_{i,t} + B_i - D_i + I_i - O_i \tag{10.1a}$$

where

$P_{i,t}$ = Population of area i at time t (the launch date)
$P_{i,t+k}$ = Population of area i at time t + k (the estimate date)
B_i = Births in area i between time t and t + k
D_i = Deaths in area i between time t and t + k
I_i = In-migrants in area i between time t and t + k
O_i = Out-migrants in area i between time t and t + k

This deceptively simple equation can be displayed in a number of forms (Hoque 2010; Murdock et al. 1995; Zitter and Shryock 1964). For example, it is common to combine in-migrants and out-migrants into net number of migrants and use the fundamental equation to estimate net migration between two censuses one taken at time t and the other at time t + k:

$$N_i = P_{i,t+k} - P_{i,t} - B_i + \tag{10.1b}$$

where

$P_{i,t}$ = Population of area i at time t
$P_{i,t+k}$ = Population of area i at time t + k
B_i = Births in area i between time t and t + k

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_10,

D_i = Deaths in area i between time t and t + k
Ni = I_i - O_i =
(In-migrants to area i between time t and t + k) -
(Out-migrants from area i between time t and t + k)

To be exactly true, the fundamental equation must apply to a defined population (e.g., the resident population) of a fixed area i and there must be no measurement errors. For example, if we are using it to estimate the resident population of area i at time t + k, then all births and deaths used must be to the resident population of area i between time t and time t + k while all in- and out-migrants during the same period also apply to this same resident population and $P_{i,t}$ is measured without error.

The fundamental equation can be applied to age, sex, race, and ethnic segments of the population. In the case of an age group the age specification of the group changes over the period. For example, if t is 10 years, then one should compare age x at time 0 with x + 10 at time t. Put another way, this age group is a "cohort" that is followed over time. In conjunction with age groups, and the use of future fertility, mortality and net migration rates, the expanded version of the fundamental equation can be used to make both estimates and projections. This is known as the "cohort-component method" (Smith et al. 2001), where "cohort" is defined as before and "component" is used to refer to the three components of population change, fertility, mortality, and migration.

All of the component methods generally employ counts of births and deaths because they are generally available every year from vital statistics records while migration data are only available in countries with well-maintained population registers (e.g., Finland). They tend to vary in how the migration component is estimated. Adapting a classification system provided by Murdock et al. (1995: 14), we identify five component methods:

1. Component Method I
2. Component Method II
3. Administrative Records Method
4. Cohort-component Method
5. Hamilton-Perry Method

Murdock et al. (1995: 13) identify data on four key components of the population of a given area i for which an estimate is desired:

(1) The size of the population for time = t (the launch date);
(2) The number of births between the base date, time = t, and the estimate date, time = t + k;
(3) the number of deaths between the base date, time = t, and the estimate date, time = t + k; and
(4) the magnitude and direction of (net) migration between base date, time = t, and the estimate date, time = t + k.

Each of the component methods deals with these four data requirements in related, but different ways. We start with Component Method I.

10.2 Component Method I

Component Methods I and II (CM I and CM II, respectively) employ school enrollment data as a way to estimate net migration (Bryan 2004). However, they vary in how they account for fertility and mortality. CM I assumes that the migration rate of school-age children in a given area i can may be estimated as the difference between the percent change in the population of school age in area i and the corresponding difference for the a "parent area" that is subject to zero in-migration and zero out-migration (e.g., in the US, area i is a given county within a state and the parent area is the United States as a whole). That is, the relative change in school age children in the parent area is assumed to be due solely to births and deaths. This relative change for the parent area is then applied to the population of area i at time t to get an estimate of the net change in the population of area i due to births and deaths between t and t + k. The net migration rate of the total population of area i is then assumed to be the same as the migration rate of the school-age population. This migration rate is multiplied by the total population of area i at time t and this product along with the estimated net population change due to births and deaths are algebraically added to P at time t obtain an estimate of P at time t + k.

The assumption in CM I that the relative change in school are children for the parent area is solely due to births and deaths during the period is a strong assumption as is the corresponding one that this change fits the change due solely to births and deaths in area i. Moreover, it is not needed when vital records data on births and deaths are available for area i. Thus, CM II was developed.

10.3 Component Method II

CM II is based on an estimate of net migration that finds the difference between a current estimate of school-age children (e.g., time = t + k) in area i with the expected number "survived" from the last census (e.g., time = t) of area i and then converting the difference to a migration rate that is applied to the entire population of area i at time t. The net migration component is estimated in six steps: (1) Enrollment in selected grades (e.g. grades 2 to 8 or in grades Kindergarten to 9) at time = t + k is adjusted to approximate the population of corresponding elementary school age on the basis of the relative size of these two groups at the last census (relating local school enrollment data to a census count at time t); (2) next, the "expected" population (assuming no net migration) of elementary school age for area i for time t + k is found by "surviving" the population in the same cohort from time t (including, if necessary, births subsequent to time t) to t + k (This is usually done using survivorship probabilities found a life table that is assumed to apply to area for the period t to t + k; (3) the net migration of children of school age is estimated as the difference between the "actual" population of school age and the "expected" population of school age; (4) the estimated net migration of school-age

children is converted into the estimated net migration of the remainder of the population by dividing these other population groups by the number of school age children at the time of the last census; (5) the estimated net number of migrants in each age group is then summed to obtain an estimate of the net number of migrants for the total population; and (6) in the final step, the total population is obtained by using the fundamental demographic equation: adding to the population in the last census, the net number of migrants and the number of births during the period and subtracting the number of deaths.

Where administrative records data are available on the population aged 65 years and over (e.g., in the U.S, Medicare data), it is not uncommon to use CM II to develop an estimate of the population less than 65 years with appropriate adjustments to the six steps just described and then use the administrative records data to estimate the population age 65 years and over (Murdock et al. 1995). The two groups are added together to get an estimate of the total population in what could be termed a composite method (Bogue 1950; Bogue and Duncan 1959). There are more variations on the basic idea (Bryan 2004; US Census Bureau 2010; Zitter and Shryock 1964), but these six steps essentially describe CM II.

CM II assumes: (1) there has been no change since the last census in the ratio of the population of elementary school age to the number enrolled in the elementary grades; and (2) that the ratio of the net migration rate of the total population to the migration rate of the school-age population of area i for the period t to t + k corresponds to that for the net migration of adults for this are over the same period.

As an example of CM II, we develop an estimate of the 2010 population of Inyo County, California from a "jump off" from the 2000 census and then compare it to the 2010 census count for Inyo County. We use grades K-9 in this example because they correspond very closely to ages 5-14. We proceed in accordance with the six steps described earlier, which contain, as appropriate, information on the data sources.

Step 1. Enrollment in grades K-4 and 5-9 in the year 2000-2001 (Fall, 2000) for Inyo County, California is reported by the California Department of Education (no date) as 1,170 and 1,382, respectively. The US Census (2001a) reports 1,184 persons aged 5-9 and 1,360 persons aged 10-14 in the 2000 census. The ratio of the population aged 5-9 to the K-4 enrollment is 1,184/1,170 = 1.01197 and the ratio of the population aged 10-14 to the 5-9 enrollment is 1,360/1,382 = 0.98408 (We note that the Fall 2000 enrollment is higher than the corresponding population as of the Spring of 2000).

The enrollment in grades K-4 and 5-9 in the year 2010-2011 (Fall, 2010) for Inyo County is reported by the California Department of Education (no date) as 999 and 1,317, respectively. Using our adjustment factors from 2000, we estimate the population aged 5-9 in 2010 as 1.01197*999 = 1,011 and the population aged 10-14 in 2010 as 0.98408*1,317 =1,296. Thus, our total 2010 population aged 5-14 based on school enrollment is 1,011 + 1,296 = 2,307.

Step 2. The "expected" population aged 5-9 in 2010 is found by "surviving" those born during the period 2000-2004; and the expected population aged 10-14 in 2010 is found by surviving those aged 0-4 counted in the 2000 census. In regard

to the former, there were 932 births reported for Inyo County residents from 2000 to 2004 (California Department of Public Health, no date). In regard to the latter, there were 961 persons aged 0-4 reported for Inyo County in the 2000 Census (US Census Bureau 2001a). In terms of surviving the births from 2000-2004 and the population aged 0-4 in 2000, we apply survivorship values from the 1995 California Life table (California Department of Public Health 1999), which is the life table that is closest to the time period for which we want survivorship values. For the former, a 10 year survivorship value of 0.988932 was used to generate 930 expected survivors aged 5-9 in 2010; for the latter, a 10 year survivorship value of 0.99826 was used to generate 960 survivors aged 10-14 in 2010. To find the estimated number of survivors aged 0-4 in 2010, we multiplied a five year survivorship value of 0.99991 taken from the closest life table (the 1995-97 California Life table, California Department of Public Health 1999) by the reported 1,118 births to Inyo County residents from 1995 to 1999 (California Department of Public Health, No Date) and found 1,117 survivors.

Step 3. As shown in Step 1, our total population aged 5-9 in 2010 based on school enrollment is 1,011 and our estimated population aged 10-14 based on school enrollment is 1,296. Subtracting our expected population aged 5-9 from the estimated population aged 5-9 we estimate the net number of migrants aged 5-9 as of 2010 to be 1,011 - 930 ≈ 81. Doing the same for the population aged 10-14 we estimate the net number of migrants to be 1,296 - 960 ≈ 336. Thus, the total number of net migrants aged 5-14 in 2010 is 81 + 336 ≈ 417

Step 4. The ratio of the population aged 15 years and over to the population aged 5-14 in 2000 (US Census Bureau 2001a) is 14,440/1,360 ≈ 5.67610. Applying this ratio to the estimated 2010 net number of migrants aged 5-14, we estimate the net number of migrants aged 15 years and over in 2010 to be 5.67610*417 ≈ 2,368.

To estimate the net number of migrants aged 0-4 in 2010 we use the ratio of the population aged 0-4 to the population aged 5-14 in 2000 (US Census Bureau 2001a, which is 961/2,544 ≈ 0.37775. Applying this ratio to the estimated 2010 population aged 5-14 of 2,307, we estimate the number aged 0-4 in 2010 to be 0.37775*2,307 ≈ 871. From this we subtract the survivors from the 1,118 births reported in 2005-2009: 871 – 1,118 ≈ -247

Step 5. Adding together our estimates of net migrants aged 0-4, 5-14 and 15+, we find that the estimated net migration for the entire population of Inyo County from 2000 to 2010 to be -247 + 417 + 2,368 ≈ 2,538

Step 6. Applying the fundamental demographic equation, we add to the 2000 population (US Census Bureau 2001), the estimated net number of migrants found in step 5 and the reported number of births (California Department of Public Health, No Date), and subtract the reported number of deaths (California Department of Public Health, No Date): 17,945 + 2,538 + 2,050 – 2,062 ≈ 20,471. Comparing our CM II estimate for Inyo County, California of 20,471 to the 2010 census number of 18,546 (US Census Bureau 2011a), we find an absolute error of -1,943 and a relative error of -10.5%.

Before we move on, it is worthwhile to note that other variations in the use of school data to estimate net migration in a component model are possible. One is the

"grade-progression method," which determines the annual net migration of school-age children by comparing the number of children enrolled in, for example, grades 2 to 7 in one year with the number enrolled in grades 3 to 8 in the following year. The remaining steps in a school-progression approach are those described earlier for CM II.

10.4 Administrative Records Method

An administrative records data source upon which to base estimates of migration are tax-return records. The US Census Bureau uses them for making state and county population estimates (Bryan 2004; Long 1993; Starsinic et al. 1995). This method obtains and uses births and deaths in the same manner as CM II. However, where CM II uses school enrollment data to estimate net migration, this method uses the annual tax return record to estimate domestic net migration (Bryan 2004).

A highly useful feature of this method for the United States is that the US Internal revenue Service makes available online at no cost the annual gross in and out flows of tax return filers and their dependents (Swanson and McKibben 2010).

10.5 Cohort-Component Method

The cohort-component method was introduced by Cannan (1895), subsequently used by Bowley (1924), and later re-discovered independently by Whelpton (1928). It is the most widely used method for producing population projections. Since it is used for projections it also can be used for estimates. Whether used for projections or estimates, the basic framework is the same as shown in Equations [10.1a] and [10.1b], but with age and sex details. We only provide an overview of the cohort-component method here. Full implementation details are found in Smith et al. (2001).

The cohort-component method divides the population at time = t (the launch date) population into age-sex groups (i.e., cohorts) and accounts separately for the fertility, mortality, and migration behavior of each cohort as it passes from he launch date at time = t to the estimate data at time = t + k. The division of the population into age groups was an important methodological advance (de Gans 1999). Not only does this account for the differences in mortality, fertility, and migration rates among different age groups at a particular time, but it also allows for changes in these rates for individual cohorts as they cycle through time.

Age cohorts can be defined in a number of ways, but cohort-component models typically use either single years or 5-year groups. The oldest age group is virtually always "open-ended," usually 75+ , 85+, or 90+. Age groups are typically divided by sex and are sometimes further subdivided by race, ethnicity, and other ascribed characteristics.

The cohorts are cycled through time in "intervals," where the components of change are applied to the cohorts in each interval as appropriate to bring them forward in time from the launch date. It is customary that the width of the number of years used to define the cohorts corresponds to the number of years in the temporal interval (i.e., 5-year age cohorts when the cohort-component method uses 5-year intervals).

The first step in the process is to establish the launch year (time = t) population and calculate the number of persons in it who survive to the estimation date (time = t + k) . This is done by applying age-sex-specific survival rates to each age-sex group in the launch year population. These can be "controlled" so that the numbers they generate match reported deaths for each interval (e.g., year) up to the estimate date.

The second step is to calculate migration for each age-sex group in each interval from time = t to time = t + k. The third step is to calculate the number of births in each interval. This is usually done by applying age-specific birth rates to the female population in each age group. As was the case with the age-sex specific survival rates, these can be "controlled" so that the numbers they generate match reported deaths for each year up to the estimate date.

The fourth and final step in the process is to add the number of births (distinguishing between males and females) to the rest of the population. These calculations provide an estimate of the population by age and sex at the end of each interval. This population then serves as the starting point for the following interval. The process is repeated until the estimate date is reached.

As can be gleaned from the preceding discussion. The cohort-component method is both data-intensive and computationally-intensive (Bryan 2004; George et al. 2004; Smith et al. 2001; Swanson et al. 2010). This is especially the case when it is used for estimation purposes, which typically means that a one year interval is used to cycle cohorts defined by single years of age through time. Unfortunately, this also means that data problems tend to increase as the level of demographic detail increases and as population size declines. Fortunately, there is a variation of the cohort-component method that can be used to overcome these problems. It is to this variation, the Hamilton-Perry Method, that we now turn.

10.6 Hamilton-Perry Method

The Hamilton-Perry Method is a variant of the cohort-component method that has far less intensive input data requirements than the full-blown version does (Hamilton and Perry 1962; Swanson et al. 2010). Instead of mortality, fertility, migration, and total population data, which are required by the full-blown cohort-component method, the Hamilton-Perry method requires data only from the two most recent censuses (Smith et al. 2001: 153-158; Swanson et al. 2010).

The Hamilton-Perry method moves a population by age (and sex) from time t to time t + k using cohort-change ratios (CCR) computed from data in the two most recent censuses. As shown by Swanson et al. (2010), the formula for a CCR is:

$$ {}_nCCR_{i,x} = {}_nP_{i,x,t} / {}_nP_{i,x-k,t-k} \tag{10.2a} $$

where

${}_nP_{i,x,t}$ is the population aged x to x + n in area i at the most recent census (t),
${}_nP_{i,x-k,t-k}$ is the population aged x-k to x-k + n in area i at the second most recent census (t-k), and k is the number of years between the most recent census at time t for area i and the one preceding it for area i at time t-k.

The basic formula for moving a population into the future to do an estimate (or a projection) is:

$$ {}_nP_{i,x+z,t+k} = ({}_nCCR_{i,x})^{*}({}_nP_{i,x,t}) \tag{10.2b} $$

where

${}_nP_{i,x+k,t+k}$ is the population aged x + k to x + k + n in area i at time t + k
${}_nCCR_{i,x} = {}_nP_{i,x,t} / {}_nP_{i,x-k,t-k}$

and

${}_nP_{i,\ x,t}$ is the population aged x to x + n in area i at the most recent census (t),

Given the nature of the CCRs, 10-14 is the youngest age group for which projections can be made if there are 10 years between censuses. To project the population aged 0-4 and 5-9 one can use the Child Woman Ratio (CWR). It does not require any data beyond what is available in the decennial census. For projecting the population aged 0-4, the CWR is defined as the population aged 0-4 divided by the population aged 15-44. For projecting the population aged 5-9, the CWR is defined as the population aged 5-9 divided by the population aged 20-49. Here are the CWR equations for males and females aged 0-4 and 5-9, respectively.

$$ \textit{Females } 0{-}4\text{: } {}_5F_{0,t+k} = ({}_5F_{0,t} / {}_{30}F_{15,t})^{*}({}_{30}F_{15,t+k}) \tag{10.3a} $$

$$ \textit{Males } 0{-}4\text{: } {}_5M_{0,t+k} = ({}_5M_{0,t} / {}_{30}F_{15,t})^{*}({}_{30}F_{15,t+k}) \tag{10.3b} $$

$$ \textit{Females } 5{-}9\text{: } {}_5F_{5,t+k} = ({}_5F_{5,t} / {}_{30}F_{20,t})^{*}({}_{30}F_{20,t+k}) \tag{10.3c} $$

$$ \textit{Males } 5{-}9\text{: } {}_5M_{5,t+k} = ({}_5M_{5,t} / {}_{30}F_{20,t})^{*}({}_{30}F_{20,t+k}) \tag{10.3d} $$

where

F is the female population,
M is the male population,
t is the year of the most recent census
and t + k is the estimation year

Projections of the oldest age group differ slightly from projections for the age groups from 10-14 to the last closed age group (e. g., age group 80-84). For example, if the final closed age group is 80-84, with 85+ as the terminal open-ended age group, then calculations for the $CCR_{i,x+}$ require the summation of the three oldest age groups to get the population age 75+ at time t-k:

$$CCR_{i,75+} = P_{i,85+,t}/P_{i,75+,t-k} \tag{10.4a}$$

The formula for estimating the population 85+ of area i for the year t + k is:

$$P_{i,85+,t+k} = P_{i,85+,t}/_{i,75+,t-k})^{*}{}_{i,75+,t} \tag{10.4b}$$

Table 10.1 provides an example of the Hamilton-Perry Method. It uses 1990 census data by age and sex (US Census Bureau 1992) and 2000 census data by age and sex (US Census Bureau 2001a) to generate a 2010 population estimate of 7,314 persons for census tract 1.01 in Clark County, Nevada. The US Census (2011b) reported 6, 851 persons in Census tract 1.01 for the 2010 Census. Thus, our estimate has an absolute error of 463 persons, and a relative error of 6.76%.

You may notice that there is a 2005 estimate by age group for Census Tract 1.01 in Table 10.1. This was included to illustrates how one can use the Hamilton-Perry Method in conjunction with an interpolation technique to obtain an estimate for a year that is not equal to the length of time between the last census and the one preceding it (e.g., in the US this would be ten years while in Canada it would be five years). To do this, we first set the Hamilton-Perry Method up as a projection with the length of the horizon being equal to the time between the most recent census and the one preceding it and then use interpolation to get to the year for which a post-censal estimate is desired. For example, as is shown in Table 10.1, if the last census was 2000 and we wanted an estimated for Census Tract 1.01 in Clark County for the year 2005 we could project the population by age group to 2010 and then interpolate between the 2000 census counts by age and the 2010 "projection" by age shown in Table 10.1 and then sum up the interpolated age group numbers. Using the geometric ratio of change, this yields the estimate of 6,857 for the total population of Census Tract 1.01 2005.

A disadvantage of the Hamilton-Perry method, is that it can lead to unreasonably high estimates in rapidly growing places and unreasonably low projections in places experiencing population losses (Smith et al. 2001: 159; Swanson et al. 2010). Geographic boundary changes are an issue, even with census tracts. Since the Hamilton-Perry and other extrapolation methods are based on population changes within a given area, it is essential to develop geographic boundaries that remain constant over time. For some sub-county areas, this presents a major challenge,

Table 10.1 Example Tract 101, Clark County, Nevada, 1990 to 2010

TRACT 101 CLARK COUNTY*					
YEAR and TRACT Identifier	1990 32003000101	2000 32003000101	CCR	2005 Estimate	2010 Estimate
Total Population: 0 to 4 years	345	441	0.37436	508	576
Total Population: 5 to 9 years	389	560	0.32580	521	482
Total Population: 10 to 14 years	372	566	1.64058	645	723
Total Population: 15 to 19 years	373	458	1.17738	559	659
Total Population: 20 to 24 years	299	358	0.96237	451	545
Total Population: 25 to 29 years	389	372	0.99732	414	457
Total Population: 30 to 34 years	384	448	1.49833	492	536
Total Population: 35 to 39 years	421	510	1.31105	499	488
Total Population: 40 to 44 years	461	442	1.15104	479	516
Total Population: 45 to 49 years	392	428	1.01663	473	518
Total Population: 50 to 54 years	334	374	0.81128	366	359
Total Population: 55 to 59 years	346	341	0.86990	357	372
Total Population: 60 to 64 years	356	268	0.80240	284	300
Total Population: 65 to 69 years	348	221	0.63873	219	218
Total Population: 70 to 74 years	208	233	0.65449	204	175
Total Population: 75 to 79 years	104	199	0.57184	163	126
Total Population: 80 to 84 years	50	115	0.55288	122	129
Total Population: 85 years and over	37	67	0.35079	100	134
Total Population	5608	6401		6857	7314

Data from 1990 and 2000 census counts for Clark County, NV (Las Vegas)

however. Fortunately, as discussed by Swanson et al. (2010), there are ways of overcoming these limitations of the Hamilton-Perry Method. They include:

1. Control Hamilton-Perry projections to independent projections produced by
2. some other method;
3. Calibrate Hamilton-Perry projections to post-censal population estimates

4. Set limits on population change (i.e., establish "ceilings" and "floors"); and
5. Account for all boundary changes;

10.7 General Comments on Component Methods

A major strength of the component methods is that they account for the three components of population change, births, deaths, and migration. This makes them straightforward both in terms of understanding them and explaining them. Importantly, they also take advantage of the fact that vital statistics data are widely available. As such they are very useful for developing post-censal estimates and inter-censal estimates. However, as we discuss in Chapter 17, by reversing them (running them backwards), they also can useful for developing pre-censal as well as inter-censal estimates.

Another disadvantage is that the availability of vital statistics data may lag behind the year for which an estimate is desired (e.g., an estimate as of July 1st, 2010 and only calendar year birth and death data for 2009 are available. This means that these data must be "projected." This can be done by using one of the trend extrapolation methods discussed in Chapter 6.

References

Bowley, A. (1924). "Births and Population in Great Britain." *The Economic Journal* 34: 188–192.

Bogue, D. (1950). "A Technique for making Extensive Population Estimates." *Journal of the American Statistical Association* 45: 149–163.

Bogue, D., and B. Duncan. (1959). "A Composite Method for Estimating Post-censal Population of Small Areas by Age, Sex, and Color." *Vital Statistics Special Reports* 47 (6).

Bryan, T. (2004). Population Estimates." pp. 523–560 in J. Siegel and D. A. Swanson (eds.) *The Methods and Materials of Demography, 2nd Edition*. Amsterdam, The Netherlands: Elsevier Academic Press

California Department of Health Services. (1999). *California Life Expectancy Abridged Life tables by Race/Ethnicity for California 1995-97*. Sacramento, CA: California Department of Health Service, Center for Health Statistics, "Data Matters," Report Register No. DM99-10000.

California Department of Education. (No Date). California Department of Education online query system (http://dq.cde.ca.gov/dataquest).

California Department of Public Health. (No Date). California Department of Public Health Online Query System (http://www.apps.cdph.ca.gov/vsq).

Cannan, E. (1895). "The Probability of Cessation of the Growth of Population in England and Wales during the next Century." *The Economic Journal* 5: 506–515.

De Gans, H. (1999). *Population Forecasting 1895–1945: The Transition to Modernity*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

George, M.V., S. Smith, D. A. Swanson, and J. Tayman. (2004). Population Projections." pp. 561–601 in J. Siegel and D. Swanson (eds.) *The Methods and Materials of Demography, 2nd Edition, Revised*. Amsterdam, The Netherlands: Academic/Elsevier Press.

Hamilton, C. H. and J. Perry. (1962). "A Short Method for Projecting Population by age from one Decennial Census to Another." *Social Forces*. 41: 163–170.

Hoque, M. N. (2010). "An Evaluation of Small Area Population Estimates Produced by Component Method II, Ratio-correlation, and Housing Unit Methods." *The Open Demography Journal* 3: 18–30.

Long, J. (1993). "Post-censal Population Estimates: States, Counties, and Places." Population Division Working Paper No. 3. Washington, DC: US Census Bureau (http://www.census.gov/population/www/documentation/twps0003.html).

Murdock, S. S. Hwang, and R. Hamm. (1995). "Component Methods" pp. 10–53 in N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (Eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, WI" Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Shryock, H. and N. Lawrence. (1949). "The Current Status of State and Local Population Estimates in the Census Bureau." *Journal of the American Statistical Association* 44: 157– 173.

Smith, S., J. Tayman, and D. A. Swanson. (2001). *State and Local Population Projections: Methodology and Analysis*. New York, NY: Kluwer Academic Press.

Starsinic, D., A. Lee, H. F. Goldsmith, and M. Spar. (1995). "The Census Bureau's Administrative Records Method." Pp. 54–69 in N. Rives, W. Serow, A. Lee, H. F. Goldsmith, and P. Voss (eds.) *Basic Methods for Preparing Small-Area Population Estimates*. Madison, Wisconsin: Applied Population Laboratory, Department of Rural Sociology, University of Wisconsin.

Swanson, D.A. and J. McKibben. (2010). "Urban-Suburban Migration Patterns in the United States, 2004-2008: The Beginning of the End for Suburbanization?" Presented at the 2010 Conference of the European Population Association, Vienna, Austria, September 1–4.

Swanson, D. A., A. Schlottmann, and R. Schmidt. (2010). "Forecasting the Population of Census Tracts by Age and Sex: An Example of the Hamilton-Perry Method in Action." *Population Research and Policy Review* 29 (1): 47–63.

US Census Bureau. (1992). Table QT-P1A. Age and Sex for the Total Population: 1990." Summary Tape File 1, Tract 1.01, Clark County, Nevada. (http://factfinder.census.gov).

US Census Bureau. (2001a). "Table P12, Sex by Age." Census 2000 Summary File 1, Inyo County California (http://factfinder.census.gov).

US Census Bureau. (2001b). "Table P12, Sex by Age." Census 2000 Summary File 1, Census tract 1.01, Clark County, Nevada (http://factfinder.census.gov).

US Census Bureau. (2010). Appendix A, Source Notes and Explanations, pp A1–A78 in *State and Metropolitan Area Data Book: 2010*. Washington, DC: US Census Bureau (http://www.census.gov/compendia/databooks/pdf_version.html).

US Census Bureau. (2011a). "Table GCP-PL2, Population and Housing Occupancy Status: 2010 – State- County- County Equivalent." (http://factfinder2.census.gov).

US Census Bureau. (2011b). "Table QT-PL, Race, Hispanic or Latino, Age, and Housing Occupancy: 2010 Census Redistricting Data (Public Law 94–171)." (http://factfinder2.census.gov).

Whelpton, P. (1928). "The Population of the United States, 1925 to 1975." *American Journal of Sociology* 34: 253–270.

Zitter, M. and H. S. Shryock. (1964). "Accuracy of Methods of Preparing Post-censal Population Estimates for States and Local Areas." *Demography* 1: 227–241.

Chapter 11
Sample Based Methods

The methods discussed in this chapter are based on concepts discussed in chapters 2 and 4. They also are interconnected and connected to methods discussed earlier. Specifically, we noted in Chapter 8 that the ratio-correlation method can both be informed by sample data (Ericksen 1973, 1974) and viewed as a form of synthetic estimation (Swanson and Prevost 1985), a subject we take up in this chapter. Moreover, it is possible to use methods discussed in chapters 7, 8, 9, 10 with sample data. Conversely, it is the case that some of the methods discussed here, particularly synthetic estimation, do not require sample data for their use. In this regard, the placement of synthetic estimation in this chapter reflects its origins in sample methods and the needs of survey statisticians to leverage the resources they had available (Steinberg 1979; US NCHS 1968). As will be seen in this chapter, demographers use a form of synthetic estimation that is not dependent on sample information.

In this chapter we start with a very brief discussion of *sample based* methods, move on to *synthetic methods* and then to *SPREE*, which is followed by a brief description of the *RSS* method. We then touch on *Bayesian* statistical methods. We conclude the chapter with observations on all four approaches, their interconnections and their connections to other methods.

Of the methods discussed in this chapter, the synthetic methods are the ones most likely to be used to make population estimates. Consequently, we spend most of our time on this approach. Before we start with sample based methods, it is useful to mention that Swanson and Pol (2008) observe that there are two distinct traditions in regard to population estimates (1) demographic; and (2) statistical. As Swanson and Pol (2008) explain,

> "Demographic methods are used to develop estimates of a total population as well as the ascribed characteristics – age, race, and sex - of a given population. Statistical methods are largely used to estimate the achieved characteristics of a population – educational attainment, employment status, income, and martial status, for example Among survey statisticians, the demographer's definition of an estimate is generally termed an "indirect estimate" because unlike a sample survey, the data used to construct a demographic estimate are symptomatic indicators of population change (e.g., K-12 enrollment data, births, deaths,) and do not directly represent the phenomenon of interest. Among demographers, the term "indirect estimate" has a different meaning."

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_11, © Springer Science+Business Media B.V. 2012

So, in the field of demography a direct estimate refers to the measurement of demographic phenomena using data that directly represent the phenomena of interest, while among statisticians, it is used to describe estimates obtained by survey sampling. In terms of an indirect estimate, demographers, usually use this term in referring to the measurement of demographic phenomena using data that do not directly represent the phenomena of interest (e.g., a child woman ratio instead of a crude birth rate). Among survey statisticians, this term refers to an estimate not based on a sample survey, for example, a model based estimate (Schaible 1993). Unless specifically stated, we will use the demographic definitions of direct estimate and indirect estimate here, respectively.

11.1 Sample Based Methods

As alluded to in the introduction to this chapter and in chapters 2 and 4, sample based methods are rarely to estimate the ascribed characteristics of a human population, such as age, race, and sex, much less the size of the population itself. Instead, sample based methods often rely on estimates made by demographers of total populations and their achieved characteristics (e.g., age, race, and sex) in designing, analyzing, and adjusting samples (Kish 1965; US Census Bureau 2009). Thus, we only present a very brief discussion of sample based methods here. However, we note that sample surveys are used in conjunction with the "Housing Unit Method" (chapters 7) of estimating population (Lowe and Mohrman 2003; Swanson et al. 1983).

In a sample based approach, one is typically interested in estimating some "parameter" of the distribution of a random variable, such as an arithmetic mean (Rao 2003: xxi). However, survey statisticians also are interested in estimated counts of characteristics of interest, such as the number of unemployed persons in labor market areas (Feeney 1987). In addition to being interested in estimating means and counts, survey statisticians also are interested in errors associated with these estimates (Pfeffermann 2002: 125).

Given that a probability sample was designed and selected and that measurement error as well as non-response and other forms of bias were minimal, one could have a good (unbiased) estimate of the parameter of interest along with information about its precision (e.g., a confidence interval). Moreover, it is relatively straightforward to explain to a non-technical audience how a sample was obtained, as well as its validity, and level of precision. The major disadvantage of this approach is that while it is less resource needy than a full-blown census, it still requires a lot of resources in the form of money, time, and technical expertise (Levy 1979; Swanson 1981; Swanson et al. 1983; US Census Bureau 2009). In addition, especially when dealing with small areas, the statistical precision may not be sufficient to determine differences between current and past values of a parameter of interest or between parameters measured at the same point in time in different

areas (US Census Bureau 2008). Yet another disadvantage of the sample based approach is that it is not uncommon to run into poorly planned and poorly executed surveys, often using something other than probability based sampling. The results from such sample surveys run a high risk of being inaccurate and there is virtually no way to assess their validity, as can be done with well executed surveys based on probability sampling (Trochim 2006).

Because there are many excellent books and manuals for conducting sample surveys, we do not describe the steps needed to design and implement sample surveys as well as analyze their results (Babbie 2009; Cochran 1977; Dillman et al. 2008; Groves et al. 2009; Kish 1965; Salant and Dillman 1994). Rather, we move directly to describing sample based methods that are likely to be of interest to those producing and using population estimates and then touch upon related methods so that demographers and others not familiar with them some idea of what they are.

11.1.1 Synthetic Methods

Ford (1981) observed that the problem of constructing county or other small area estimates from survey data has been an important topic and large-scale surveys and even complete census counts were often used to solve the problem. Because of the resource needs of this approach, attention turned to possible alternatives for obtaining small area information in the 1970s (US NCHS 1968; Ford 1981). One of the alternatives that gained a lot of attention was synthetic estimation, which according to Ford (1981) emerged because of a 1978 workshop on Synthetic Estimates for Small Area Estimates co-sponsored by the National Institute on Drug Abuse (NIDA) and the National Center for Health Statistics (NCHS). This same workshop resulted in a monograph edited by Steinberg (1979).

In the "Introduction" to the NIDA/NCHS monograph, Steinberg (1979) cites "The Radio Listening Survey," discussed in Hansen et al. (1953) as an early example of the employment of the synthetic method. In this survey, questionnaires were mailed to about 1,000 families in each of 500 county areas and personal interviews were conducted with a sub-sample of the families in 85 of these county areas who were mailed questionnaires (Hansen et al. 1953: 483-484). Knowing in advance that the mail-out portion would yield a low level of responses (about 20 percent of those mailed questionnaires responded), the data collected in the personal interviews were used to obtain estimates not affected by non-response. The relationships between the data in the 85 county areas that were collected from the personal interviews and the mailed questionnaires were then applied to the county areas for which only mail-out/mail-back was done to improve the estimates for these areas (Hansen et al. 1953: 483). While the radio listening study did not use the hallmark of synthetic estimation, which is taking information from a "parent" area and applying it to its subareas, the idea behind it is similar.

As was discussed in regard to sample based estimates, in most cases, synthetic estimation is used to estimate "achieved characteristics" and often relies on estimates made by demographers of total populations and their achieved characteristics (e.g., age, race, and sex) in developing the estimates (Causey 1988; Cohen and Zhang 1988; Gonzalez and Hoza 1978; Levy 1979). However, it need not be confined to this use. Before we turn to a demographic interpretation of synthetic estimation, it is useful to spend some time on its statistical interpretation.

Cohen and Zhang (1988) provide an informal statistical definition of a synthetic estimator that we adapt as follows. First, assume that one is interested in obtaining estimates of an unknown characteristic, x_i over a set of i sub-regions ($i = 1,\ldots, n$). Second, suppose one has census counts p_i, ($i = 1,\ldots,n$), for each of the sub-regions and both a census count, P, and a "known" value of X, for the parent region, where $\sum p_i. = P$ and $\sum x_i. = X$, respectively. Third, suppose that the estimated values of x_i for the subareas must sum to the known value X for the parent area. In this case, Cohen and Zhang (1988: 2) define the statistical synthetic estimate as:

$$\hat{X}_i = (X/P)^*(p_i). \tag{11.1}$$

Basically, equation [11.1] shows that the estimated characteristic (x_i) for a given subarea i is found by multiplying the known value of population for sub-area i, p_i, by the "known" ratio of the characteristic (X) to population (P) for the parent area. It is inevitably the case that the "known" value of X for the parent area is taken from a sample survey (Steinberg 1979). Cohen and Zhang (1988) go on to show how the basic idea given in Equation [11.1] can be extended to include demographic subgroups (e.g., by age, race, and sex). Similar examples are provided by Levy (1979).

As a simple example that shows how Equation [11.1] would be applied, suppose we have 50,000 people in a parent area (P = 50,000) and 1,000 have a characteristic (X = 1,000) that we are interested in estimating for its three subareas, which have, respectively 30,000, 15,000, and 5,000 people, respectively.

From a statistical perspective, synthetic estimates are generally held to be "biased." That is, there is a difference between the estimator's expected value and the true value of the parameter being estimated (see, e.g., Weisstein 2011). The bias basically comes from the fact that the ratio of x_i to p_i in a given subarea i is not the same as the ratio for the parent area. That is, $X/P \neq x_i/p_i$.

With this simple introduction to systematic estimation, we now turn to how synthetic estimation works from the standpoint of demographers. The key difference for demographers is that unlike statisticians, it is the population of area i (pi) that is

Table 11.1 Example of Synthetic Estimation

Sub-area	Population	Parent Area Ratio (X/P)	Estimated number with Characteristic x
1	30,000	(1000/50000)	6,000
2	15,000	(1000/50000)	3,000
3	5,000	(1000/50000)	1,000

"unknown" rather than some characteristic (xi) of this population. To implement synthetic estimation, demographers find "characteristics" that are available for both the parent area and its subareas. These characteristics are known to demographers as "symptomatic indicators," a term discussed in many of the chapters preceding this one, especially in chapters 8 and 9. So, for demographers, Equation [11.1] becomes

$$\hat{p}_i = (S_{j,i})/(\mathbf{S_j}/\mathbf{P}) \tag{11.2}$$

where

$\mathbf{P}$ = population of the parent area
$\mathbf{Sj}$ = value of symptomatic indicator j for the parent area
$S_{j,i}$ = value of symptomatic indicator j for subarea i ($1 \leq i \leq n$)
p_i = estimated population for subarea i ($1 \leq i \leq n$) and so, we can identify the ratio S_j/P as
$\mathbf{R_j} = \mathbf{(S_j/P)}$

As is the case for the synthetic estimators used by statisticians (Equation [11.1]), the basic form of the synthetic estimator used by demographers (as shown in Equation [11.2]) can be expanded. One expansion is to put the synthetic estimation process in motion using a regression framework. This can be done as follows.

$$p_{i,t} = a_0^*(\mathbf{P}_t)^*(p_{i,t-z}/\mathbf{P}_{t-z}) + bj^*[(S_{j,i,t})/((S_{j,i,t-z}/p_{i,t-z})^*(\mathbf{S}_{j,t}/(\mathbf{S}_{j,t-z}/\mathbf{P}_{t-z})))] + \varepsilon_i \tag{1.3}$$

where

a_0 = the intercept term to be estimated
bj = the regression coefficient to be estimated using symptomatic indicator j
ε_i = the error term
$s_{j,i}$ = symptomatic indicator ($1 \leq j \leq k$) in subarea i ($1 \leq i \leq n$)
t = year of the most recent census
z = number years to the census preceding the most recent census

and

$\mathbf{P}$ = population of the parent area
$\mathbf{Sj}$ = value of symptomatic indicator j for the parent area
p_i = estimated population for subarea i ($1 \leq i \leq n$)

Once the preceding regression model is constructed, it can be used to estimate the population of each area i for a year k years subsequent to the last census (time = t) as follows:

$$\hat{p}_{i,t+k} = [a_0^*(\mathbf{P}_{t+k})^*(p_{i,t}/\mathbf{P}_t)] + [bj^*((S_{j,i,t+k})/((S_{j,i,t}/p_{i,t})^*(S_{j,t+k}/(S_{j,t}/P_t))))] \tag{11.4}$$

Equations [11.3] and [11.4] should be familiar. They can be algebraically manipulated to become a bi-variate form (i.e., a regression model with only one independent variable) of the ratio-correlation model discussed in Chapter 8, which we show here. First, borrowing from Equation [8.1a] in Chapter 8 we show here the simple bi-variate ratio-correlation regression model that is algebraically equivalent to Equation [11.3] :

$$P_{i,t} = a_0 + (b_j)^* S_{i,jt} + \varepsilon_i \tag{11.5}$$

where

a_0 = the intercept term to be estimated
b_j = the regression coefficient to be estimated
ε_i = the error term
j = symptomatic indicator ($1 \leq j \leq k$)
i = subarea ($1 \leq j \leq n$)
t = year of the most recent census

and

$$P_{i,t} = (P_{i,t}/\Sigma P_{i,t})/(P_{i,t-z}/\Sigma P_{i,t-z}) \tag{11.6}$$

$$S_{i,jt} = (S_{i,t}/\Sigma S_{i,t})_j/(S_{i,t-z}/\Sigma S_{i,t-z})_j \tag{11.7}$$

where

z = number of years between each census for which data are used to construct the model
p = population
s = symptomatic indicator

As was shown in Chapter 8, a set of population estimates can be done in a series of six steps, which lead to the estimation version of Equation [11.5], which is algebraically equivalent to equation [11.4]:

$$(P_{i,t+k})^*(P_{i,t}/\Sigma P_{i,t})^*(\Sigma \mathbf{P_{i,t+k}}) = \hat{P}_{i,t+k} \tag{11.8}$$

As discussed by Swanson and Prevost (1985), these equations show that the ratio-correlation model can be viewed as a regression method that uses synthetic estimation (taking a ratio of change for a given "rate" in a parent area and a "censal-ratio" to estimate a current population for area i). Note that the intercept term, a_0, shown in Equation [11.4] serves as a "weight" applied to an estimate of p_i at time t+k ($p_{i,t+k}$) based on the proportion of the population in area i at the time of the last census, t (p_i,$_t$) that is multiplied by the total of the parent area at time t+k ($\mathbf{P}_{t+k}$). The regression coefficient, bj, shown in Equation [11.4] also serves as a weight. In this case it is applied to the "synthetic estimate" based on symptomatic indicator s_j. As Swanson (1980) and Swanson and Prevost (1985) observe, the regression

coefficient in a ratio-correlation model sum to 1.00 (or very nearly so) in virtually every model constructed, which means that as shown in Equation [11.4] the estimate of p_i can be viewed as a weighted average of synthetic estimates based on j symptomatic indicators.

11.2 SPREE

Chambers and Feeney (1977) and Purcell and Kish (1980) proposed structure preserving estimation (SPREE) as a generalization of synthetic estimation in the sense it makes a fuller use of reliable direct estimates. A good example of its use is given by Feeney (1987).

SPREE is a categorical data analysis approach to the problem of small area estimation. It has two general data requirements: The first is that there exist current estimates for the variables of interest by subgroups for the large area; the second is that estimates of variables of interest are available by the same subgroups at the small area level from some previous time period (Griffiths 1996). *SPREE* uses the data from the previous point in time (e.g., the most recent census) to allocate current (e.g., the point in time for which an estimate is being done) at the large area level to the small areas. Thus, the data from the previous time period must be not only be available for the small areas of interest and but they also must be available in the form of cross-classifications with "auxiliary" variables for these same small areas (Griffiths 1996). The data from the previous point in time are known as the association structure and the data for the current point in time at the large area level are known as the allocation structure. The SPREE method allocates the current data (in the allocation structure) to the small area level by retaining the relationship of the data given in the association structure (Feeney 1987).

SPREE essentially uses the method of "iterative proportional fitting of margins" (IP) in a multi- way table, where the margins are "direct estimates" (in the statistical sense, which is that they are from samples or census counts). In a similar vein, Bousfield (1977) described the use of raking to force the marginal totals of a two-way sample table to match census totals and showed how they can be used to generate population estimates. IP is a form of "N-dimensional controlling" (Smith et al. 2001: 260-266).

IP approximates a least squares solution in order to obtain convergence in all n dimensions (Judson and Popoff 2004: 712-71). Smith et al. (2001: 260) observe that there are three main conditions for applying this method: (1) all projections must be greater than or equal to zero; (2), there must be projections for the margins of each controlling dimension (e.g., age and total population); and (3) the sum of all projections over all dimensions must be equal; for example, the sum of the age group projections for the county must be equal to the sum of the total population projections for the census tracts.

In extending IP to SPREE, Berg and Fuller (2009) state that SPREE uses estimators for cell totals and column proportions of a two-way table that preserve

the direct estimators of the marginal totals. Noble et al. (2002) demonstrate that SPREE is maximum likelihood estimation under a generalized linear model in which the interactions in the time point of interest are set equal to the interactions in the Census. If the loglinear expectation function underlying SPREE fails to hold, the SPREE estimators can be severely biased (Griffiths 1996; Zhang and Chambers 2004). Griffiths (1996) develops a composite estimator that is a weighted combination of the direct estimator and the SPREE estimator. Zhang and Chambers (2004) first define a class of loglinear models called generalized linear structural models in which the interactions are proportional to the Census interactions. They then extend the generalized linear structural model to a mixed model with normally distributed random effects.

SPREE may not always come as close to the independent (control) projections as the examples shown here. Raising the level of demographic detail and reducing the geographic scale can cause multiplicative adjustment routines to lose their efficiency because the computations may not change the original values as much as is needed to produce complete convergence.

11.3 RSS (Ranked Set Samples) Method

As an alternative to a method proposed by C. R. Rao (1997), the RSS method was introduced at a conference on population estimates by Sinha and Sinha (1999). Its major use would be to provide an estimate of the total population (or the total number of males or females) of a given domain or area in which census counts (or good estimates) of the total population for some, but not all domains or subareas, are available. Sinha and Sinha (1999) illustrated and tested the RSS method with 1991 data from 145 urban areas in the state of Bihar, India. What they did was use a sample of 27 of these 145 urban areas to show how the RSS method could be used to produce an estimate of the total population of all 145 urban areas.

RSS starts with the fact that "N" (e.g., the 145 urban areas of Bihar) is known and the decision on how big a sample to take along with the "set size," which in turn generate the number of "replications." In their example for the 145 urban areas of Bihar et al. (1999) selected 27 as the sample size with 3 as the "set size", which generated 9 as the number of replications, where $27/3 = 9$. Since $27*3 = 81$, this generated the need to randomly select (with replacement) of 81 of the 145 urban areas of Bihar. The total populations of the 81 urban areas were assembled into 27 rows i.e. sets, with three cities per row.

Sinha and Sinha (1999) then ranked the three urban areas of each row in ascending order of population size: rank "1" was given to the urban area with the lowest population; rank "2" to the urban area with the next-lowest population; and rank "3" to the urban area with the highest population. After doing this ranking to each of the 27 rows, they then made selections from the first three rows by selecting the urban area with the minimum population from the first row; the urban area with the second rank from the second row; and finally the urban area with the highest

population from the third row. They then repeated this process for the remaining rows considering a group of three rows at a time. This resulted in a set of 27 urban areas, which they organized into three groups according to rank, which means that there were nine urban areas in each rank, 1, 2, and 3. They then computed the means of the total populations of the nine urban areas in the "Rank 1", "Rank 2" and the "Rank 3" respectively. They the took the ("grand") mean of these three means and used it as an estimator of the mean population of the 145 urban areas of Bihar. Since they knew that Bihar has 145 urban areas, they multiplied this grand mean by 145 and came up with estimate of the total population of the 145 urban areas of Bihar. Without going into the details, the grand mean they estimated was 77,269.07 ≈ (34,268.2 + 69,615.3 + 127,923.7)/3. By multiplying this grand mean by N (145) they estimated the total 1991 population of the 145 urban areas of Bihar as 11,204,015 ≈ 77,269.07*145. This estimate compares favorably with the census number of 9,905,706, with an absolute difference of 1,298,309 and a relative difference of 13.1%.

11.4 Bayesian Methods

Bayesian inference represents a perspective on statistical inference that provides an alternative to the "frequentist" perspective, which is characterized by hypothesis testing and confidence interval construction (Iversen 1984). Named after Thomas Bayes because his theorem provides the foundation for this perspective, Bayesian methods allow for a systematic introduction of subjective viewpoints into the process of statistical inference (Iverson 1984), which cannot be done under the frequentist perspective.

In the frequentist method, unknown parameters are often, but not always, treated as having fixed but unknown values that are not capable of being treated as random variables. As such, this implies that there is no way that probabilities can be associated with these parameters. In Bayesian inference, all unknown parameters can have probabilities. A simple example of a Bayesian perspective is given in the discussion of deaths as a random variable found in Chapter 9 in conjunction with the view on censal ratio methods provided by Voss et al. (1995). More elaborate examples are found in Bousfield (2002) and Cressie and Dajani (1991).

For our purposes in terms of constructing population estimates, the major feature of Bayesian inference is that it represents a structured system in probabilities can be "up-dated." Within the Bayesian perspective, there are two fundamental ways in which this can be done: (1) the standard Bayes (SB) approach; and (2) the empirical Bayes (EB) approach (Gelman et al. 2004: 115-156). In SB, the "prior" distribution is determined before any data are collected and used; in EB, the prior distribution is estimated from the data. Western (1999) provides descriptions of these two approaches that is aimed at sociologists and includes a general comparison of Bayesian and Frequentist methods and philosophies.

11.5 Summary

Although sample based methods are not generally used by demographers to generate population estimates, we believe that understanding them and their statistical properties provide an important perspective that can be used to inform the methods used by demographers. Clearly, the regression-based trend extrapolation methods discussed in Chapter 6 and the ratio-correlation method discussed in Chapter 8 are embedded in this perspective. Moreover, the "random variable" approach of Voss et al. (1995) to censal ratio estimators also adds an important dimension to understanding these estimators and potential ways to improve their accuracy. In a similar vein, the Bayesian approach to inference provides yet another way that these estimators can be understood and their accuracy improved.

In terms of strengths of the sample based methods that are aimed at generating what the statisticians refer to "direct estimates," they offer a well-understood approach that is less costly than full enumerations along with estimates of their precision. In terms of their weaknesses, the cost of sample surveys often precludes using them to develop usable information for small areas unless they are supplemented by other methods such as synthetic estimation (Ghosh and Rao 1994; Platek et al. 1987; Rao 2003). The RSS method is not costly, given the availability of a "sample" of census counts (or good estimates) for a subset of the domain or area for which an estimate is desired. While it is clearly aimed at developing countries where census counts may be done of selected areas on a regular basis but not for the county as a whole, the RSS method may also work in states or provinces in which annual census counts are done for some areas (e.g., cities and towns) but not for all. Washington and Alberta come to mind in mind in this regards.

Jaffe (1951: 211) notes that while sample surveys are cheaper than full enumerations, "demographic procedures" are cheaper than sample surveys. He also notes that the "direct estimates" resulting from sample surveys can only be used for current estimates since it is impossible to interview a past or future population. He goes to observe that only "demographic procedures" can provide past, current, and future estimates. We note, however, that these same 'demographic procedures' can be improved by using the statistical tools and perspectives that have emerged from sampling. All of this suggests that sample-based methods are used for post-censal estimates, but we also note that given the availability of historical samples and related data they can be used for inter-censal estimates and even pre-censal estimates.

References

Babbie, E. (2009). *The Practice of Social Research*. Florence, KY: Cengage Learning/Wadsworth Publishing.

Berg, E., and W. Fuller. (2009). A SPREE Small Area Procedure for Estimating Population Counts." Proceedings of the Survey Methods Section, Ottawa, Ontario, Canada: Statistical Society of Canada. (http://www.ssc.ca/survey/documents/SSC2009_EBerg.pdf).

Bousfield, M.V. (1977). "Inter-censal Estimation Using a Current Sample and Census Data." *Review of Public Data Use* 5: 6–15

Bousfield, M. V. (2002). "Population Estimation for Census Tracts using Dynamic Models." Paper presented at the Annual Meeting of the Population Association of America, Atlanta, GA.

Chambers, R. and G. Feeney. (1977). Log linear models for small area estimation. Unpublished paper, Australian Bureau of Statistics

Causey, B. (1988). *Evaluation of Census Ratio Estimation and Synthetic Estimation*. Statistical Research Division Report no. Census/SRD/RR/88/15. (http://www.census.gov/srd/papers/pdf/rr88-15.pdf).

Cochran, W. (1977). *Sampling Techniques, 3rd Edition*. New York, NY: Wiley.

Cohen, M. and X. Zhang, (1988). *The Difficulty of Improving Statistical Synthetic Estimation*. Statistical Research Division Report no. CENSUS/SRD/RR-88/12. Washington DC: U. S. Bureau of the Census. (http://www.census.gov/srd/papers/pdf/rr88-12.pdf).

Cressie, N., and A. Dajani. (1991). "Empirical Bayes Estimation of US Undercount Based on Artificial Populations." *Journal of Official Statistics* 7: 57–67.

Dillman, D., J. Smyth, L. Christian. (2008). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 3rd Edition*. New York, NY: Wiley.

Ericksen, E. (1974). "A Regression Method for Estimating Population Changes of Local Areas." *Journal of the American Statistical Association* 69: 867–875.

Ericksen, E. (1973). "A Method for Combining Sample Survey Data and Symptomatic Indicators to obtain Population Estimates for Local Areas." *Demography* 10: 137–160.

Feeney, G. A. (1987). "The Estimation of the Number of Unemployed at the Small Area Level." pp. 198 – 218 in R. Platek, J. N. K. Rao, C. E. Särndal, and M. P. Singh (Eds.) *Small Area Statistics: An International Symposium*. New York, NY: John Wiley and Sons.

Ford, B. (1981). *The Development of County Estimates in North Carolina*. Staff Report AGES 811119. Agriculture Statistical Reporting Service, Research Division. Washington, DC: US Department of Agriculture.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. (2004). *Bayesian Data Analysis, 2nd Edition*. New York, NY: Chapman and Hall/CRC.

Ghosh, M., and J. N. K. Rao. (1994). "Small Area Estimation. An Appraisal." *Statistical Science* 9(1): 55–76.

Gonzalez, M. and C. Hoza, (1978). "Small Area Estimation with Application to Unemployment and Housing Estimates." *Journal of the American Statistical Association* 73 no. 361 (March): 7–15

Griffiths, R. (1996). "Current Population Survey Small Area Estimation for Congressional Districts." *1995 Proceedings of the Statistical Methods Research Section, American Statistical Association*: 314–319. Alexandria, VA: American Statistical Association.

Groves, R., F. Fowler, Jr., M. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. (2009). *Survey Methodology, 2nd Edition*. New York, NY: Wiley.

Hansen, M., W. Hurwitz, and W. Madow. (1953). *Sample Survey Methods and Theory: Volume I, Methods and Applications*. New York, NY: John Wiley and Sons.

Iverson, G. (1984). *Bayesian Statistical Inference*. Quantitative Applications in the Social Sciences, no. 43. Beverly Hill, CA: Sage Publications.

Jaffe, A. J. (1951). *Handbook of Statistical Methods for Demographers: Selected Problems in the Analysis of Census Data, Preliminary Edition, 2nd Printing* (US Bureau of the Census) Washington, DC: US Government Printing Office

Judson, D. and C. Popoff, (2004). "Selected General Methods." pp. 677–732 in J. Siegel and D. A. Swanson (eds.) *The Methods and Materials of Demography, 2nd Edition*. Amsterdam, The Netherlands: Elsevier/Academic Press.

Kish, L. (1965). *Survey Sampling*. New York, NY: John Wiley and Sons.

Levy, P. (1979). "Small Area Estimation-Synthetic and Other Procedures, 1968-1978." pp 4–19 in J. Steinberg (Ed.) *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion. NIDA Research Monograph 24*. Rockville, MD: US Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute on Drug Abuse.

Lowe, T., and M. Mohrman. (2003). *Using a Geographic Information System in a Population Estimate Program: The Pasco Case*. Research Brief no. 21, Washington Office of Financial Management. Olympia, WA: Washington Office of Financial Management.

Morris, C. (1983). "Parametric Empirical Bayes inference: Theory and applications (with discussions)." *Journal of the American Statistical Association*. (78): 47–65.
Noble A., S. Haslett, and G. Arnold. (2002). "Small Area Estimation via Generalized Linear Models." *Journal of Official Statistics* 18(1): 45–60.
Pfeffermann, D. (2002). "Small Area Estimation: New Developments and Directions." *International Statistical Review* 70(1): 125–143.
Platek, R., J. N. K. Rao, C. E. Särndal, and M. P. Singh (Eds.). (1987). *Small Area Statistics: An International Symposium*. New York, NY: John Wiley and Sons.
Purcell, N. J. and L. Kish, L. (1980). "Post-censal Estimates for Local Areas (or Domains)." *International Statistical Review*. 48: 3–18.
Rao, C. R. (1997). *Statistics and Truth: Putting Chance to Work, Second Edition*. World Scientific Publishing Co. Pte. Ltd. Singapore
Rao, J. N. K. (2003). *Small Area Estimation*. New York, NY: Wiley-Interscience.
Salant, P. and D. Dillman, (1994). *How to Conduct your own Survey*. New York, NY: Wiley.
Schaible, W. (1993). "Indirect Estimators, Definition, Characteristics, and Recommendations." *Proceedings of the Survey Research Methods Section, American Statistical Association Vol I: 1-10*. . Alexandria, VA: American Statistical Association (http://www.amstat.org/sections/srms/proceedings/y1993f.html).
Sinha, A. and R. Sinha. (1999). "Estimating Population Total using Ranked Set Samples." Paper presented at the Population Estimates Conference, June 8. Suitland, MD: US Census Bureau.
Smith, S., J. Tayman, and D. A. Swanson. (2001). State and Local Population Projections: Methodology and Analysis. Dordrecht, The Netherlands: Kluwer/Academic Press (Springer)
Steinberg, J. (1979). "Introduction." pp. 1–2 in J. Steinberg (Ed.) *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion. NIDA Research Monograph 24*. Rockville, MD: US Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute on Drug Abuse.
Swanson, D.A. (1980). "Improving Accuracy in Multiple Regression Estimates of County Populations Using Principles from Causal Modeling. *Demography* 17 (November):413–427.
Swanson, D.A. (1981). "Municipal Census Results and Costs for 1981." *Alaska Population Overview 1981*. Juneau, AK: Alaska Department of Labor.
Swanson, D. A., and L. Pol. (2008). "Applied Demography: Its Business and Public Sector Components." in Yi Zeng (ed.) The Encyclopedia of Life Support Systems, Demography Volume. UNESCO-EOLSS Publishers. Oxford, England. (http://www.eolss.net/).
Swanson, D. A. and R. Prevost. (1985). "A New Technique for Assessing Error in Ratio-Correlation Estimates of Population: A Preliminary Note." *Applied Demography* 1 (November): 1–4.
Swanson, D. A., B. Baker, and J. Van Patten. (1983). "Municipal Population Estimation: Practical and Conceptual Features of the Housing Unit Method." Presented at the 1983 Annual Meeting of the Population Association of America, Pittsburgh, PA.
Trochim, W. (2006). "Nonprobability Sampling." Research Methods Knowledge Base (http://www.socialresearchmethods.net/kb/sampnon.php).
US Census Bureau. (2009). *Design and Methodology: The American Community Survey* (ACS DM1). Washington, DC. US Census Bureau.
US Census Bureau. (2008). *A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know*. Washington, DC. US Census Bureau.
US NCHS (U. S. National Center for Health Statistics). (1968). *Synthetic State Estimates of Disability*. PHS Publication No. 1759. US Public Health Service. Washington, DC: US Government Printing Office.
Voss, P. R., Palit, C. D., Kale, B. D., and Krebs, H. J. (1995). "Censal ratio methods". pp.70–89 in N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, and P. R. Voss (Eds.) *Basic methods for preparing small-area estimates*. Madison: Applied Population Laboratory, University of Wisconsin.
Weisstein, Eric W. (2011). "Estimator Bias." From MathWorld–A Wolfram Web Resource. (http://mathworld.wolfram.com/EstimatorBias.html).
Western, B. (1999). "Bayesian Analysis for Sociologists: An Introduction." *Sociological Methods and Research* 28: 7–34.
Zhang, L. and R Chambers. (2004). "Small Area Estimates for Cross-classifications. *Journal of the Royal Statistical Society, B* 66: 479–496.

Chapter 12
Other Methods

This chapter covers methods that are sufficiently different from those discussed in chapters 8 through 11 to warrant a separate chapter. However, it should come as no surprise that pieces of some of these methods are related to the methods already discussed. All of these general approaches are sufficiently complicated that examples cannot be provided here. The idea in this chapter is to provide an overview of the methods and point those of you interested in using any or all of them to resources that provide more details.

In this chapter we first examine *Structural Models* before going on to discuss *Administrative Records* in general and then cover two specific ways in which these records can be used to develop estimates: (1) *Imputation*; and (2) *Dual System Estimators*. We then discuss *Microsimulation (Agent based modeling)* and *Neural Networks*, followed by a discussion of two (sample) survey approaches, *the Grouped Answer Method* and *Social Network Analysis/Snowball Sampling*. We conclude this chapter with a brief discussion of *Spatial Demography*, which like the applied demography principle discussed in Chapter 15, we view as a fundamental element in the conceptual and theoretical foundation upon which population estimation methods rest.

12.1 Structural Models[1]

Because changes in population are not solely determined by demographic factors, but instead depend on the economy, land use rules, transportation systems, and the environment, structural models have been develop because they produce population "determinations" (i.e., estimates and projections) in which these and other factors are taken into account. We describe two general categories of structural models: (1) economic-demographic models; and (2) urban systems models. Economic-demographic models are typically used to determine population and economic

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_12,

estimates and projections for larger geographic areas such as counties, labor market areas, states, and nations (Treyz 1993). Urban systems models focus on small geographic areas such as census tracts and block and typically provide determinations of population, economic activities, land use, and transportation patterns (Hunt and Abraham 2005). In addition to their differences in geographic scale, these two types of models often provide alternative explanations of the causes and consequences of population change. Some structural models contain only a few equations and variables (Korotayez 2005; Mills and Lubuele 1995; Swanson and Beck 1994; US Bureau of Economic Analysis 1995), while others contain huge systems of simultaneous equations with many variables and parameters (Anas and Liu 2007; Hunt and Abraham 2005; Treyz 1995). Our objective is to provide a general introduction and overview of the use of structural models. We do not provide details for building or implementing these kinds of models; such details can be found in Anas and Liu (2007), Miller et al. (1999), Putnam (1983, 1991), SANDAG (1998, 1999), and Treyz (1993).

12.2 Economic Demographic Models

Although most economic-demographic models deal with all three of the components of population change, they typically focus on migration (Treyz 1993). This should not be surprising because western economists typically use equilibrium models in which "rational" actors participate. Thus, economic factors such as job change, unemployment, and wages or income are therefore used to determine migration (Treyz 1993). The empirical evidence suggests that the strongest links are those found with job change (Krieg and Bohara 1999; Treyz 1993, 1995; Treyz et al. 1991).

Migration and population change are also influenced by non-economic factors such as climate, coastal location, life cycle changes, personal characteristics, and social networks (Astone and McLanahan 1994; DaVanzo and Morrison 1978; Fuguitt and Brown 1990; Massey et al. 1987; Murdock et al. 1984). A complete migration model including both economic and non-economic factors, however, is problematic for determining migration or population because the independent variables themselves must be determined. Determinations of these non-economic variables are rarely available, while determinations of economic variables can be obtained from national, state or county-level economic models (Treyz 1993).

We briefly describe three general approaches for designing and implementing economic-demographic models: (1) Econometric models, which use regression methods to determine migration; (2) balancing models, which determine migration as the difference between the supply and demand for labor; and (3) ratio-based models, which typically derive population directly from employment.

Econometric Models. The econometric approach uses equations that determine migration from one or more economic variables. Parameters for these equations

are estimated from historical data using regression techniques. Migration numbers are then made by solving the equation(s) using the values of the independent variable(s). The migration equation(s) are typically integrated into a large economic model that also provides projections of the economic factors.

The most widely used econometric models of migration are "recursive," whereby migration is influenced by the economy, but does not itself influence the economy. Recursive models cannot reflect the full range of interactions between migration and the economy, but nonetheless have proven successful for projecting migration (Clark and Hunter 1992; Greenwood and Hunt 1991; SANDAG 1999; Greenwood 1975; Tabuchi 1985). Recursive models of migration have also been implemented in multiregional migration models (Campbell 1996; Foot and Milne 1989; Isserman et al. 1995). Non-recursive models attempt to capture the joint impacts of migration and the economy on each other. Although they are more complicated and require larger resources than recursive models, non-recursive migration models have been employed (Conway 1990; Mills and Lubuele 1995; Treyz et al. 1993).

Balancing Model. The concept behind the balancing model is straightforward. If labor supply exceeds labor demand, workers migrate out of the area; if labor demand exceeds labor supply, workers migrate into the area. Balancing models are typically less costly to implement and easier to use than econometric models because they do not require large-scale systems of equations, huge amounts of data, or the use of formal statistical procedures. However, they do require numerous computations and assumptions (Murdock and Ellis 1991). Labor demand is often represented by a measure of job opportunities typically projected using export-base models, input-output models, and extrapolation techniques (Greenberg et al. 1978; Murdock et al. 1984). Labor supply is determined by applying labor force participation rates to a population derived from a cohort-component model that assumes zero net migration. The migration of workers is determined by the difference between labor supply and labor demand. The estimated number of economic migrants is leveraged to obtain the number of people migrating with them through assumptions related to characteristics such as marital status and family size.

Population/Employment Ratio. The population/ employment (P/E) model does not use components of population change. Instead, it develops a total population number directly. The P/E model is the easiest and least expensive way to incorporate economic factors into a population determination. The simplest P/E model uses a single ratio representing total population to total employment, holds the ratio constant at its current value, and applies the ratio to an estimate of employment. The "OBERS" model, developed by the US Bureau of Economic Analysis (BEA) in the mid-1960s, was perhaps the most widely used P/E model (US Bureau of Economic Analysis 1995). The approach taken in OBERS divides the population into three age groups: pre-labor pool (less than 18), labor pool (18-64), and post-labor pool (65+). Estimates of the labor pool population are directly related to changes in employment and the pre-labor pool population estimates are tied directly to the labor pool population. Post-labor pool estimates are independent of economic changes.

12.2.1 Urban Systems Models

Urban systems models are used throughout the world to determine the distribution of residential and nonresidential activities within urban or metropolitan areas (Anas and Liu 2007; Hunt and Abraham 2005; US EPA 2000). They are designed to be used for small geographic areas and are used to provide information about a wide range of issues (e.g., air quality, traffic congestion, public transportation). Along with economic factors such as jobs and income, urban systems models include land use characteristics (e.g., zoning, environmental constraints, land value and land supply) and characteristics of the transportation system (e.g., travel times, cost, and distances). As such, urban systems models require considerably more information, time, and resources to implement than economic-demographic models.

Urban systems models vary considerably in their theoretical approaches, mathematical design, data requirements, and ease of implementation, but they typically consist of three major components—regional population and economic estimates, land use and activity, and transportation. The regional estimates are often produced using economic-demographic models. The land use and activity component consists of a complex set of procedures for distributing the regional population and economic estimates into zones within the region. The transportation component provides estimates of transportation system characteristics, such as traffic volumes and speeds on roadways and on public transportation lines.

A fundamental characteristic of urban systems models is the iterative and explicit relationships between land use characteristics, activity location, and the transportation system. The distribution of population in virtually all such models relies on the link between home (residential location) and workplace (employment location). These links are represented by travel probabilities between zones based on time, distance, or cost and commuting patterns (Putnam 1991). Residential location influences the spatial distribution of employment, particularly employment that serves a local population such as retail trade and services. This relationship is implemented by assuming a lag between residential location and location of employment. The transportation system influences land use characteristics that play an important role in determining the location of population and other activities. Thus, urban systems models contain procedures to reconcile the demand for land with its available supply (SANDAG 1998, 1999; and Waddell 2000).

12.2.2 Comments On Structural Models

Structural models - especially urban systems models - require a lot of resources and are difficult to implement. They are typically used in developing projections, but in principle, they can be used to develop estimates. They often require extensive base data, sophisticated modeling skills, and complex statistical procedures and computer programs. Therefore, they are accessible only to a relatively narrow range of

practitioners. In addition, there is no evidence to suggest that structural models provide more accurate population forecasts than other methods and, given their small geographic scale, their forecast accuracy is not likely to be high in many applications. Yet, structural models are used more frequently today than ever before because of their ability to investigate and analyze a wide range of theoretical, planning, and policy questions (Schmidt et al. 1997; Tayman 1996; Treyz 1995).

12.3 Administrative Records

It is clear from previous chapters (i.e., 7, 8, 9, 10) that administrative records (here, we include vital events as a component of administrative records) play an important role in developing population estimates. Some types of records serve directly as symptomatic indicators of population (e.g., the number of Housing units, per the discussions of the Housing Unit Method in Chapter 7, the ratio-correlation method in Chapter 8 and Medicare data in Chapter 10), others serve as indicators of population change (e.g., births and deaths in CM II as described in Chapter 10), and still others serve as direct and indirect indicators (e.g., school enrollment as a symptomatic population indicator in a ratio-correlation model as described in Chapter 8, a censal ratio variable, as described in Chapter 9, and as an indicator change in terms of net migration in CM II, as described in Chapter 10). They also can serve as sample frames (e.g., the discussion of sample-based indicators found in Chapter 14). In short, without administrative records, the range of methods with which we could develop population estimates would largely be confined to the extrapolation and interpolation methods described in Chapters 6 and 17. This would result in a toolkit that was similar to what was used by the US Census Bureau at the beginning of the 20th century, when experimentation with symptomatic indicators for use as censal ratio estimators resulted in estimates that were less satisfactory than those produced with extrapolation models (Shryock and Lawrence 1949).

Administrative records are also used in other ways to develop population estimates. The "Demographic Analysis" (DA) Method used by the US Census Bureau to assess the accuracy of the decennial census at the national level is an estimation technique that relies on administrative records (Robinson et al. 2002; Robinson et al. 1993). DA population estimates are developed for the census date using data that are independent of the census. Births and Deaths are added and subtracted for many years, respectively, and combined with information on immigration and emigration to develop a national population estimate. Because birth and death records are not complete before 1935, data on Medicare for the population aged 65 years and over have been used to supplement the estimates of the elderly population (Robinson et al. 2002). By the time of the 2030 census, the need for the Medicare supplements will be virtually nil. As such, an estimate of the US population in 2030 would be $P_{2030} = (\text{Births}_{1935\text{-}2030}) - (\text{Deaths}_{1935\text{-}2030}) + (\text{Inmigrants}_{1935\text{-}2030}) - (\text{Outmigrants}_{1935\text{-}2030})$.

As described in Chapter 10, the US Census Bureau developed a successful component based method during the 1970s in which migration was estimated using tax return data from the US Internal Revenue Service (Healy 1982; Long 1993). This method required matching addresses on successive years of tax returns and calculating a migration rate based on the total number of exemptions that moved into and out of each area. In aggregate form, these same IRS data are available online (http://www.irs.gov/taxstats/article/0,,id=212683,00.html) and can be used to develop estimates of in and out migration by county and state.

Another administrative records source of migration data is found in the United States Postal Service (USPS) "Change of address" tabulations (Swanson et al. 2009; USPS 2011). These data are widely used by demographic vendors in the private sector to develop population estimates (Martins et al. 2012).

Finally, in addition to the role that administrative records play in the methods we have just described, they have the potential to provide estimates that serve as virtual census counts (Alvey and Scheuren 1982; Kliss and Alvey 1984; Swanson and Walashek 2011), especially when augmented with survey data, record linkage techniques, and modeling and imputation methods (Allison 2001; Fay 2005; Fellegi and Sunter 1969; Judson 2007; Kalton 1983; Liu 2007; 2008; Myrskylä 1991; Peterson 1999; Rubin 2004; Scheuren 1999; Statistics Canada 2009; Statistics Finland 2004; Swanson and Knight 1998; Thomsen and Holmøy 1998; Weinberg 2009). This perspective is not just applicable at the national level. For example, for states with income tax (all but Alaska, Florida, Nevada, South Dakota, Texas, Washington, and Wyoming), the returns could be used in a manner similar to that described by Alvey and Scheuren (1982), Kliss and Alvey (1984) Swanson and Walashek (2011) for federal income tax returns to generate an "administrative records" census. Something similar to this is done in Alaska, which in effect has a 'reverse income tax" in the form of its "permanent fund" distributions (State of Alaska 2008). In addition, as described by Martins et al. (2012), it is pretty clear that private sector vendors have used a wide range of data to develop what amounts to a virtual census count at any given point in time, although the accuracy of any of these records is unknown due to the proprietary interests of the organizations that are involved in developing them.

With this general overview of administrative records, we now turn to two specific methods that can be used with them to develop population estimates: (1) imputation; and (2) Dual system estimation.

12.4 Imputation

Adapting the definition provided by Swanson and Stephan (2004: 762), imputation is a general term used to describe the assignment of values to cases for which one or more variables have missing values due. Four common methods are: (1) deductive imputation, which is based on other information available from the case in question; (2) hot-deck imputation, which is based on information from "closest-matching"

cases; (3) mean-value imputation, which uses means of variables as the source of assignment; and (4) regression-based imputation, in which models are constructed using cases with no missing values and a dependent variable is the one whose missing values will be imputed and the independent variables are those that yield acceptable regression equations. The general idea is that having an imputed value is better than having a missing value, especially if the data set in question is small (e.g., a small sample survey). In addition, as pointed out by Kalton (1983), missing data can produce biased estimates in surveys if not handled appropriately (Swanson 2008, 1986). The US Census Bureau started using imputation in conjunction with the 1950 census (Cresce et al. 2005).

Imputation has largely been used to assign values to variables in censuses and sample surveys for which the respondents provided no information (Fay 1996, 1999; Madow et al. 1983; Rubin 2004; Singh et al. 2001). However, as missing data were encountered in administrative records, analysts turned to imputation methods for assistance in dealing with them (Bye and Judson 2004; Hogan and Cowan 1980). In some case, the missing values were due to confidentiality standards (Raghunathan et al. 2003) and in others, in conjunction with record-matching, which leads us to the next area, Dual Systems Estimators.

12.5 Dual System Estimation

In 1949, C. Chandra Sekar (now known as Chandrasekaran) and W Edwards Deming introduced a system for estimating the total number of births and deaths that is based on the algebra underlying the Chi-squared (χ^2) test in a 2x2 table of cross-classifications. This technique has been refined and modified for a number of uses (Hogan and Cowan 1980; Paradies and Barnes 2005; Popoff and Judson 2004; Wolter 1986), including the estimation of the size of wildlife populations, where it is known as "capture-recapture" (Williams et al. 2002). In demography and statistics, this general method has become known as "Dual System Estimation" (Krótki 1978).

A natural way to show how Dual System Estimation (DSE) works is to start from the $\chi 2$ Test (Norušis 1991: 265-271). Suppose we have two variables, A and B each of which has two values, YES and NO. The χ^2 test is designed to provide a statistical test of whether or not the two variables, A and B, are related. In the test, one has all of the YES and NO values for both A and B, which can be viewed as the sums of each row and column in the 2x2 table in which the "observed" values of A and B would be cross-classified. The χ^2 test proceeds by finding observed values of the two values for each variable and then seeing if the observed values are different than the expected values, where the expected values are what one would see if the two variables had no relationship – were independent. It is the latter, the values expected under the assumption that the two variables are independent, that is the basis of the original DSE method. However, to get to that point, we need to have the "observed" values.

Table 12.1 Hypothetical Survey Results of PAA Members and Their Memberships in ASOCA and ASTATA

		Currently a member of American Sociological Association?		
		Yes	No	**Total**
Currently a Member of the American Statistical Association?	Yes	2,000	5,000	**7,000**
	No	2,000	1,000	**3,000**
	Total	**4,000**	**6,000**	**10,000**

Table 12.2 "Expected" Cross-Classified Values for the Hypothetical Survey of PAA Members

		Currently a member of American Sociological Association?		
		Yes	No	**Total**
Currently a Member of the American Statistical Association?	Yes	E_{11} = $(\mathbf{N_{1j}}* \mathbf{N_{i1}})/ \sum \mathbf{N_{ij}}$ **2,800 = (7,000*4,000)/10,000**	E_{12} = $(\mathbf{N_{1j}}* \mathbf{N_{i2}})/ \sum \mathbf{N_{ij}}$ **4,200 = (7,000*6,000)/10,000**	$\mathbf{N_{1j}}$ **(7,000)**
	No	E_{21} = $(\mathbf{N_{2j}}* \mathbf{N_{i1}})/ \sum \mathbf{N_{ij}}$ **1,200 = (3,000*4,000)/10,000**	E_{22} = $(\mathbf{N_{2j}}* \mathbf{N_{i2}})/ \sum \mathbf{N_{ij}}$ **1,800 = (3,000*6,000)/10,000**	$\mathbf{N_{2j}}$ **(3,000)**
	Total	$\mathbf{N_{i1}}$ **(4,000)**	$\mathbf{N_{i2}}$ **(6,000)**	$\sum \mathbf{N_{ij}}$ **(10,000)**

Suppose that we have surveyed all of the members of the Population Association of America (PAA) and asked each of them two questions that each have the same two responses, "yes" and "no:" (1) Are you currently a member of the American Statistical Association (ASTATA); and (2) Are you currently a member of the American Sociological Association (ASOCA). Suppose, further that there are 10,000 PAA members and that the results of our survey are as found in Table 12.1.

As we can see, 7,000 of the 10,000 PAA members are also ASATA members and 4,000 of them are also members of the ASOCA. Further, we can see that only 2,000 of the 10,000 PAA members are members both of the ASTATA and the ASOCA. In order to conduct a χ^2 statistical test to see if ASTATA membership is independent of ASOCA membership, we would need to find the "expected" values in each of the four cross-classified cells. This would proceed as shown in Table 12.2.

The expected values in the four intersecting cells of a 2x2 table are found by multiplying the sum of the row by the sum of the column corresponding to the intersecting cell and dividing this sum by the overall total. In $Cell_{11}$, of Table 12.2, we see its expected value ($E_{11} = 2{,}800$), is found by multiplying the sum of the "row" total (The number of records in System B, $\mathbf{N_{1j}} = \mathbf{7{,}000}$) by the sum of the "column" total (The number of records in System A, $\mathbf{N_{i1}} = \mathbf{4{,}000}$) that correspond to $Cell_{11}$ and then dividing this product by the overall total number of records in each organization ($\sum \mathbf{N_{ij}} = \mathbf{10{,}000}$). This is repeated for each of the other three

Table 12.3 Hypothetical Estimate of Total PAA members using ASOCA and ASTATA Membership records of PAA members

		PAA member who is an ASOCA Member		
		Yes	No	**Total**
PAA Member who is an ASTATA Member	Yes	O_{11} (2,000)	O_{12} (5,000)	$\sum$**O_{1j} (7,000)**
	No	O_{21} (2,000)	O_{22} (Unknown)	$\sum$**N_{2j} (Unknown)**
	Total	$\sum$**O_{i1} (4,000)**	$\sum$**N_{i2} (Unknown)**	$\sum$**N_{ij} (Unknown)**

cross-classified cells. Essentially what this does is to simultaneously assign the relative frequency distribution of the row margins (7,000, 3,000) and the frequency distribution of the column margins (6,000, 4,000) to the cells. So, in $Cell_{11}$, where we have an expected value of 2,800 concurrent ASTATA and ASOCA members, it is the same relative proportion of its column total of 4,000 that the row total of 7,000 is to 10,000 (2,800/4,000 = 7,000/10,000 = 0.70) and at the same time, it is the same relative proportion of its row total of 7,000 that the column total of 4,000 is to 10,000 (2,800/7,000 = 4,000/10,000 = 0.40). The same type of relationship holds for the "expected" values in the remaining cells. That is, the expected values in the cells have the same relative frequency to their column and row totals that the corresponding column and row totals have to the overall total.

So, by now you may be asking how does DSE actually work? The answer is that it exploits the algebra for finding "expected" values in order to estimate the overall "total," where the overall total corresponds to an estimated number of births, deaths, or people by using two "independent" administrative records systems (or two surveys or a survey and an administrative records system). The underlying assumption is that the administrative records systems are incomplete. If they were complete (e.g., the birth and death registrations systems in Australia, Canada, England, and the United States and the population registry system in Finland), there would be no need for Dual System Estimation. In our hypothetical example, would translate into having no "overall total."

Specifically, what DSE does is to have cross-classified data (e.g., as shown in Table 12.1) and then assume that the two variables are, in fact, independent, so that the unknown overall total can be estimated. To show how this would work in practice, suppose that all of the PAA membership lists were destroyed and the organization was trying to get an idea of how many members there were. A logical place to start would be the membership lists of other organizations to which PAA members tended to belong and for which PAA (along with other memberships) were also recorded. Let us further suppose that it was known that PAA members belong to ASOCA and ASTATA, two organizations that also maintained information on the professional associations to which their members belonged. Suppose that the ASOCA and ASTATA membership lists were searched with the results shown in Table 12.3 for PAA members.

As shown in Table 12.2, we found 7,000 PAA members in the ASTATA membership list and 4,000 PAA members in the ASOCA membership lists, along with 2,000 who have joint membership in both ASOCA and ASTATA.

It might seem that we could simply subtract the joint members (2,000) from the sum of the PAA members found in the two organizations (7,000 + 4,000) to get an estimate of total PAA membership, which would result in an estimate of 9,000 PAA members. Doing so, however, ignores the PAA members who belong neither to ASTATA and ASOCA. Thus, the way forward is to use the equation for finding the "expected" values (Table 12.2) to the single cell for which we have "observed" values and corresponding known row and column totals, (In Table 12.3 this is cell C_{11}), we would have three known values and one unknown value (the overall total, ($\sum \mathbf{N_{ij}}$), which can be solved for the unknown value, viz:

$$E_{11} = (O_{1j}{}^{*}O_{i1})/\sum N_{ij} = \mathbf{2{,}000} = (\mathbf{7{,}000}^{*}\mathbf{4{,}000})/\sum N_{ij}$$

and

$$\sum \mathbf{N_{i2}} = (\mathbf{7{,}000}^{*}\mathbf{4{,}000})/\mathbf{2{,}000} = \mathbf{14{,}000}$$

So, we have an estimate of 14,000 for the previously unknown overall total number of members in the PAA. If, as in our sample survey example, the actual (but unknown) total were 10,000, we would have a relative percent error of 40% with this estimate. While this may be high, an error of 40% would be preferable to having no information at all, given the cost of obtaining it vs. the costs of not having it. This hypothetical example shows that DSE is capable of producing at least reasonable estimates. Note that had there been 2,800 members in $cell_{11}$, we would have an estimate of 10,000 since the 2,800 number would be in accordance with the assumption of independence.

As the preceding example illustrates, one of the major issues for DSE is the need to assume that any two information systems are independent, which for most administrative records systems is not very realistic. For example, it is highly likely that many of records in an income tax system match up to people in a driver's license system. The fact that many systems are so correlated and other issues have led to a great deal of research into DSE, especially since the adoption of DSE by the US Census Bureau as one of the tools to evaluate the accuracy of the decennial census (Hogan 1993; 2000; Hogan and Cowan 1980; Popoff and Judson 2004; Wolter 1986). In turn, the use of DSE by the US Census Bureau and the research into it, provides the method with a substantial theoretical and experiential foundation.

12.6 Micro-Simulation (Agent Based Modeling)

Agent-Based Models (ABMs) collectively represent an individual modeling method that, along with two related approaches, Microsimulation (MSM) and Cellular Automata (CA, also known as Artificial Neural Networks or ANN),

has received attention as a demographic estimation and forecasting tool in the past twenty or so years (Andreassen 1993; Bandyopadhyay and Chattopadhyay 2006; Billari and Prskawetz 2003; Booth 2006; Charette 2010; Clarke and Holm 1987; Griffith et al. 2012; Harding and Gupta 2007; Malenfant et al. 2011; Martel 2010; Sokolova et al. 2006; Van der Gaag et al. 2005; Zinn et al. 2010). This development corresponds to observations made by Smith et al. (2001: 367) that while population projections were primarily made at the national and state levels until the 1970s, they started being routinely made for lower levels of geography such as census tracts and block groups, which, in turn, generated demand for even lower levels of geography such as tax assessor files, block faces, and street segments. They observed that this trend implied that projections would eventually be made for individual addresses, households, and people. Indeed, this observation has been borne out and the reason is largely due to the development of individual modeling methods, including ABM. The same comments made by Smith et al. (2001) in regard to ABM and forecasting apply to ABM and estimation.

What exactly is ABM? According to the International Microsimulation Association (2006), it is closely allied to the other two other individual-level modeling approaches, CA and MSM. In distinguishing these three related approaches, The International Microsimulation Association (2006) describes them as follows.

(1) In a pure CA, all entities are spatially located within a grid of cells, and all entities have only one attribute (alive or dead), with behaviors deterministically dependent upon the state of neighboring cells.
(2) In a pure ABM, the emphasis is on the interaction between individuals, with the main attribute of each individual being the operating characteristics (behavioral rules), which evolve stochastically over time in response to the success or failure of interactions with other individuals.
(3) In a pure MSM, transition probabilities lack evolutionary and spatial dimensions.

The Association (2006) concludes that as microsimulation models add more behavioral and spatial interaction between individual units, as CAs add a growing range of individual attributes and start to incorporate aspatial behaviors, and as ABMs add both space and fiscal/demographic characteristics to their agents, the three approaches move towards a common ground.

As an example, Griffith et al. (2012) discuss an ABM known as "DOMICILE ABM (DOmicile Model Implemented by Calculating In-migration at the Domicile LEvel) that has been specifically designed as a new approach to the generation of population estimates and projections. The model design follows that proposed by Grimm et al. (2006), under the "ODD Protocol" (Overview, Design concepts, and Details), which is rapidly being adopted by the agent based modeling community as a standard ABM format.

Another example of ABM is provided by Malenfant et al. (2011), who describe Demosim, a microsimulation system basically in production by Statistics Canada that is maintained by Modgen, a programming language specially designed by Statistics Canada's Modeling Division to facilitate the development of microsimulation that can be accessed along with documentation at (http://www.

statcan.gc.ca/spsd/Modgen.htm). The starting point for the DEMOSIM is the microdata file for the 20-percent sample of the 2006 census of the population of Canada This database, which includes close to seven million persons with their characteristics, was adjusted to take account of the census net undercoverage.

According to Malenfant et al. (2011), the variables contained in the Demosim initial file can be divided into two major groups. The first consists of the key variables that were projected for public release purposes:

- Age
- Sex
- Place of residence
- Religious denomination
- Visible minority group
- Immigrant status
- Generation status
- Continent/region of birth
- Mother tongue
- Highest level of schooling
- Labour market participation

The second group consists of "support variables," which are variables that are included because they serve to increase the quality of the projection for the variables in the first group. They include the following:

- Marital status
- Province or territory of birth of non-immigrants
- Year of immigration
- Age at immigration
- Aboriginal identity
- Registered Indian status
- Presence and number of children in the home
- Age of youngest child in the home
- Sex of youngest child in the home
- Dates on which diplomas were obtained

Although it has limitations, Malenfant et al. (2011) argue that Demosim represents a powerful and relevant tool. Through microsimulation, a user may easily not only generate multiple characteristics of a population but also take into account differentials in demographic behaviours between groups of the population.

12.7 Neural Networks

Neural Networks (NN) are based on the idea that many complex processes can be adequately modeled in terms of simpler processes (Paik 2000). In essence, NN is based on the processes believed to operate in the human brain, where the complex

tasks performed by the brain can be faithfully modeled in terms of the interactions between such neurons (Churchland and Sejnowski 1992). The interactions take place via synaptic connections, and the electrical resistance of a connection determines the strength of the interaction between the connected neurons. Typically, a single neuron in the memory centers of the brain is connected to approximately 10,000 other neurons in a fully interconnected fashion, but certain portions of the brain have been observed to have a layered structure, which Paik (2000) argues is functionally similar to a nonlinear, multivariate regression model. The vision layer of neurons is analogous to the set of independent (predictor) variables, whereas the motor neurons are analogous to the dependent (response) variables. This is the idea that has led to the development and refinement of NN, especially in regard to statistical methods and the ability to model nonlinear relationships (Bishop (1995), Masters (1993), and Sarle (1994). However, as Paik (2000) observes, for a given data set, an NN may be outperformed by a method with many restrictive and explicit assumptions.

One feature of NN that distinguishes it from standard multiple regression models is that a given model is "trained," in a process that is analogous to learning. A demographic example of this process is found in Tang et al. (2006), who provide an example of various types of "trained" NN models that were evaluated against one another and the ratio-correlation model using ex post facto tests in conjunction with historical census data. They found that properly trained neural networks outperformed the ratio-correlation regression model overall and that among the different NN models they tested, the fuzzy logic NN performed the best.

12.8 The Grouped Answer Method

This (sample) survey-based technique is designed to estimate difficult to count populations, such as illegal immigrants. Introduced by the US Government Accountability Office (US GAO 1998, 1999), it has been refined and simplified (Larson and Droitcour 2012; US GAO 2006). In 2004, as a large scale test, a "grouped answer" question module aimed at estimating the foreign born in the US by immigration status (i.e., legal and otherwise) was included in the General Social Survey (GSS) conducted by the National Opinion Research Center of the University of Chicago (Larson and Droitcour 2012). The GSS question module represents the first time the grouped-answer method has been applied in a household survey of the general population. Data from the GSS test data included foreign-born respondents' answers and interviewers' judgments, and comments written by interviewers, reflecting interviewer observations and statements made by respondents. These results were reviewed by an independent statistical expert and US GAO. The findings suggest that the method has promise and it has been recommended for use by federal statistical agencies in regard to estimating the foreign-born by immigration status (Larson and Droitcour 2012).

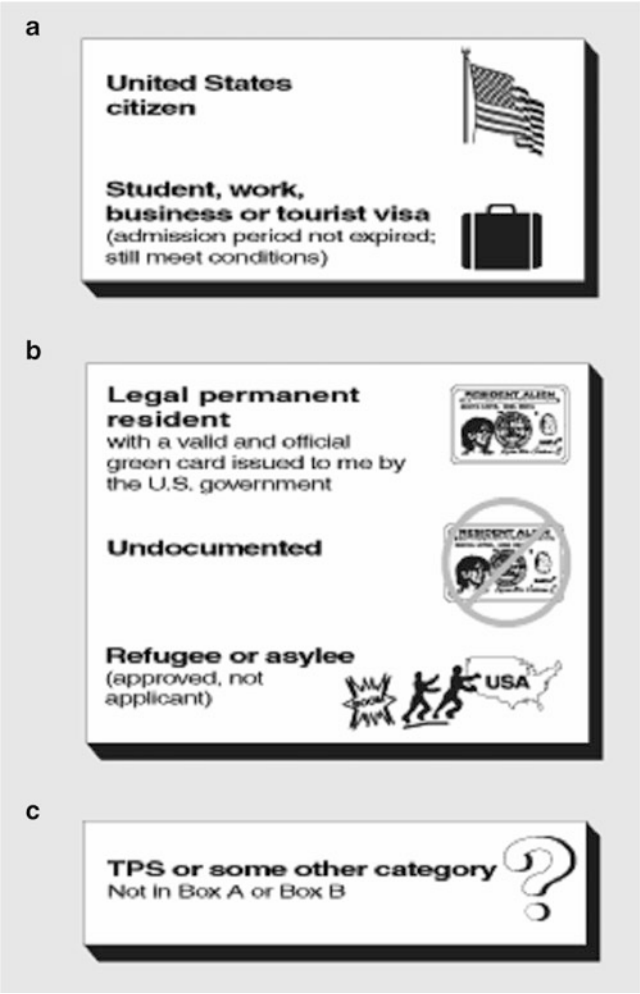


Exhibit 12.1 The Grouped Answers Method for Estimating the Foreign-Born who lack Legal Immigrant Status

Exhibit 12.1 shows a variation of the Grouped Answers method that is designed to estimate the numbers of foreign-born who lack legal immigrant status. The grouped answer method features two or more alternative immigration-status cards, each of which is used with a different sub-sample of foreign-born respondents. Each card features a 3-box set of immigration-status answer categories, A, B, and C; the respondent chooses the box that contains his or her immigration status

Reading across Exhibit 12.1, you can see that: Box A in both card set 1 and card set 2 contains one or two non-sensitive immigration status(es); Box B in both card sets contains the sensitive undocumented status and a variety of other non-sensitive statuses; and Box in both card sets is for some other category, not

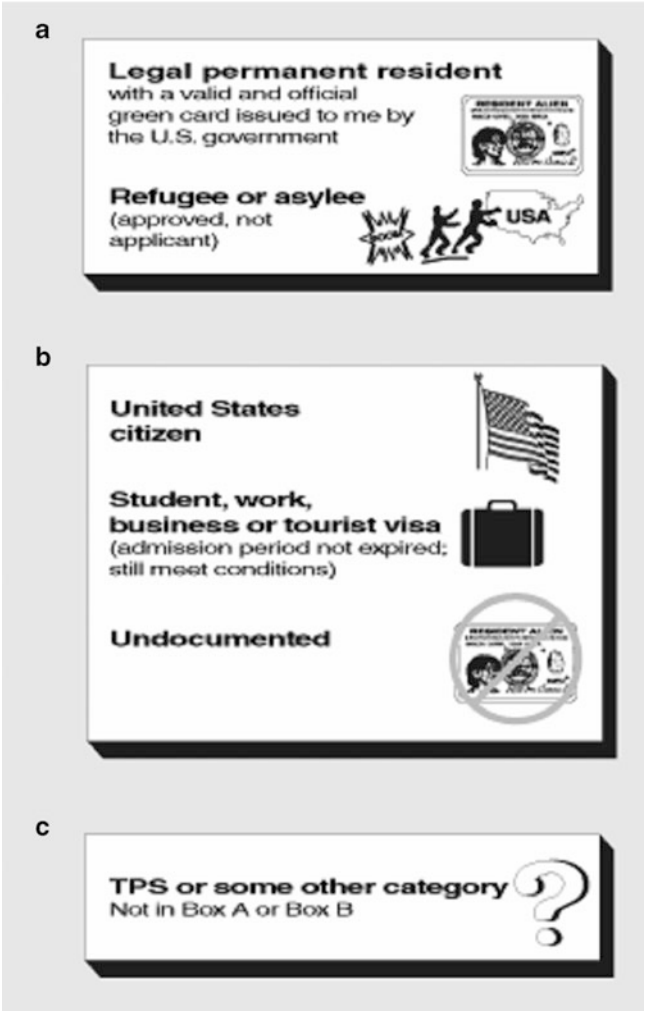


Exhibit 12.1 (continued)

in Box A or Box B. The various cards alternate the non-sensitive statuses appearing in Box A versus Box B; thus, it is possible to obtain direct estimates of each non-sensitive category (Box A) and, via subtraction, to obtain an indirect, "residual" estimate of the sensitive, undocumented status: %Illegal = %A (card set 1) - %B (card set 2) + %C (either card set). No individual respondent is ever associated with the sensitive, undocumented or illegally present status.

Using this technique in conjunction with personal interviews at several construction sites in the Washington, DC metropolitan area that targeted low-skilled immigrant-saturated trades, Golden and Skibniewski (2009) found approximately that of the 896 respondents, 55% were undocumented and "quasi-legal." They cautioned

that it was difficult to extrapolate these findings to all construction sites in the Washington DC metropolitan area because they had no viable sample frame, which left into question how representative the study sites they used were. However, Golden and Skibniewski (2009) argued that their findings were supported by other research.

12.9 Social Network Analysis/Snowball Sampling

Like the Grouped Answer method, using Social Network Analysis to estimate a population of interest is fundamentally a (sample) survey-based tool. The technique used to do this is usually referred to as "snowball sampling" (Goodman 1961). Snowball sampling identifies social networks by asking initial respondents about people they might know who have specific characteristics of interest to the researcher (Palmore 1967). Those identified, in turn, are then themselves asked these same questions. This is referred to metaphorically as snowball sampling because as the process yields more people with the desired characteristics in a manner similar to how a snowball increases in size as it is rolled along, collecting more and more snow.

Because the initial form of snowball sampling lacks a sample frame, it does not have the foundation found in samples selected from a valid frame in a random manner. As such, it was difficult to get an idea of the precision of an estimate yielded by such a snowball sample. However, refinements have been developed to overcome this and other problems and it is now possible to develop estimates of precision and related information for snowball samples designed using these refinements (Heckathorn 1997, 2002; Salganick and Heckathorn 2004).

12.10 Spatial Demography

Voss (2007) argues that virtually all demography was spatially oriented until about the mid 20th century, at which time a paradigm shift occurred that led demography to become more focused on the individual as a subject of study rather the study of demographic attributes aggregated to some level(s) within a geographic hierarchy. However, he notes that three general areas of study continued to be spatially oriented after this shift: (1): urban demography; (2) rural demography; and (3) applied demography (Voss 2007). In terms of the latter, applied demography, Voss observes that research conducted in its subfield of population estimation and forecasting methods blossomed in the 1950s and as it continued, brought a fresh perspective to the analysis of spatial units. Voss believes that this research was greatly enabled by five products that radically changed the field of demography: (1) the Census Bureau's TIGER files; (2) electronic census files; (3) extensive satellite imagery; (4) geographic information system (GIS) software for mapping

and, importantly, for integrating spatially-arrayed data from diverse and disparate georeferenced "layers," and finally (5) the powerful, but affordable computing hardware platforms on which to bring together these various elements. When these elements converged in the early 1990s, Voss believes that they began to alter the way in which spatial demographic research was carried out by generating new and broadly interdisciplinary relationships on campuses and elsewhere that led to the development of new hypotheses and research questions. Examples of these new and invigorated areas of research include spatial interpolation, spatial interaction, and multilevel modeling, among others (Goodchild and Kwan 1978; Mitas and Mitasova 1999; Voss 2007).

One of the areas of interest in the area of population estimation that has contributed to new hypotheses and research questions is spatial dependency, the co-variation of properties within geographic space, which leads to spatial autocorrelation problem (Dubin 1998; Longhi and Nijkamp 2005; Swanson and Tedrow 1984; Voss et al. 2006). Spatially arrayed statistical models such as the ratio-correlation model and its variants can have unstable parameter estimates and yield unreliable significance tests when affected by spatial autocorrelation. However, by viewing spatial dependency as informative rather than problematic, correctly specified spatial regression models can be constructed that capture these relationships (Dubin 1998).

An important concept in demography is the idea of heterogeneity (Vaupel and Yashin 2006), and it should come as no surprise that this concept applies to spatial demography (Goodchild 2009). Neither should it come as a surprise that those building spatially arrayed population estimation models have recognized and attempted to deal with this issue along with temporal heterogeneity (Chu 1974; McKibben and Swanson 1997; Swanson 1980; Tayman and Schafer 1985).

Given this discussion, it should be clear that it is not a huge stretch to view spatial demography as the technical and substantive foundation of population estimation (and forecasting). The most obvious connection is between ratio-correlation and its variants since this method is inherently not only a manifestation of spatial demography in that it deals with population aggregated to geographical levels, but it also is multi-level, substantively embedded in the theoretical issues affecting the distribution and composition of populations over time, affected by spatial autocorrelation, and capable of dealing with both spatial and temporal heterogeneity.

12.11 Summary

Aside from population registry systems (Statistics Finland 2004) and "Master Address Files" (Swanson and Walashek 2011), administrative records systems were not designed to yield population estimates. Because they are correlated with population numbers, they have, however, been extensively explored and exploited for purposes of population estimates, as is clear from this chapter and others in this

book. With the increased difficulty in simultaneously maintaining high response rates for traditional sample surveys and related forms of primary data collection (e.g., censuses) and holding costs down, it is likely that administrative records will loom even more important in the future (Swanson and Walashek 2011).

The two (sample) survey-based methods discussed in this chapter, The Grouped Answers Method and Social Network Analysis/Snowball Sampling, offer ways to estimate populations that are difficult to count. In several regards they are like the methods discussed in Chapter 16 for estimating de Facto populations and those impacted by a disaster and could, in fact, be used for these purposes.

All of the methods discussed in this chapter are largely designed to develop post-censal estimates. However, they can be used for inter-censal estimates. Rarely, if at all, would they be used to develop pre-censal estimates.

Endnote

1. This discussion is adapted from Smith, Tayman, and Swanson (2001: 185-237).

References

Allison, D. (2001). *Missing Data*. Beverly Hill, CA: Sage.

Alvey, W. and F. Scheuren. (1982). "Background for an Administrative Records Census." pp. 47–65 in *Statistics of Income and Related Administrative Record Research*. Washington, DC: US Department of the Treasury, Internal Revenue Service.

Andreassen, L. (1993). "Demographic Forecasting with a Dynamic Stochastic Microsimulation Model." *Discussion Paper no. 85*, Central Bureau of Statistics, Oslo, Norway.

Anas, A., and Y. Liu. (2007). "A Regional Economy, Land Use, and Transportation Model (RELU-TRAN): Formulation, Algorithm Design, and Testing: *Journal of Regional Science* 47 (3): 415– 455.

Astone, N. and S. McLanahan. (1994). "Family Structure, Residential Mobility, and School Dropout: A Research Note." *Demography* 31: 575–584.

Bandyopadhyay, G. and S. Chattopadhyay. (2006). "An Artificial Neural Net Approach to Forecast the Population of India." (http://eprintweb.org/S/authors/nlin/ba/Bandyopadhyay/5).

Billari, F., and Prskawetz, A. (eds.). (2003). *Agent-Based Computational Demography: Using Simulation to Improve Our Understanding of Demographic Behaviour*. Heidelberg: Physica-Verlag.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford, England: Clarendon.

Booth, H. (2006). "Demographic Forecasting: 1980 to 2005 in Review." *Working Papers in Demography, No. 100*. Demography and Sociology Program, Australian National University.

Bye, B., and Judson, D. 2004. "Results from the Administrative Records Experiment in 2000." *Census 2000 Testing, Experimentation, and Evaluation Program Synthesis Report No. 16, TR-16*. Washington, DC: US Census Bureau.

Campbell, P. (1996). Population Projections for States by Age, Color, Race, and Hispanic Origin: 1995 to 2050. Report PPL-47. Washington, DC: U. Census Bureau.

Chandra Sekar, C., and W. E. Deming. (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration." *Journal of the American Statistical Association* 44: 101–115

Charette, C. (2010). “Are Demographers Ready, Willing and Able to Create Microsimulation Models?” Paper presented at the 2010 Applied Demography Conference, January 10–12, San Antonio, Texas.

Chu. S., (1974). “On the Use of Regression Method in Estimating Regional Population.” *International Statistical Review* 42: 17–28.

Churchland, P., and T. Sejnowski, (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Clark, D., and W. Hunter, (1992). “The Impact of Economic Opportunity, Amenities, and Fiscal Factors on Age-Specific Migration Rates.” *Journal of Regional Science* 32: 349–365.

Clarke, M. and Holm, E. (1987). ’Microsimulation Methods in Spatial Analysis and Planning’, Geografiska Annaler. Series B, Human Geography, 69(2): 145–164.

Conway, R. (1990). “The Washington Projection and Simulation Model.” *International Regional Science Revie*w 13: 141–165.

Cresce, A., S. Obenski and G. Chappell. (2005). “Research to Improve Census Imputation Methods: The Plan to Examine Count and Item Imputation.” pp. 2928 -2934 in *Proceedings of the American Statistical Association, Survey Research Methods Section*. Alexandria, VA: American Statistical Association.

DaVanzo, J., and P. Morrison. (1978). “Return and Other Sequences of Migration in the United States.” *Demography* 18: 85–101.

Dubin, R. (1998). “Spatial Autocorrelation: A Primer.” *Journal of Housing Economics* 7: 304–327.

Fay, R. (2005). “Model-Assisted Estimation for the American Community Survey.” *Proceedings of the Joint Statistical Meetings*: 3016–3023. Alexandria, VA: American Statistical Association.

Fay, R. (1999). “Theory and Application of Nearest Neighbor Imputation in Census 2000.” pp. 112– 121 in *Proceedings of the American Statistical Association, Survey Research Methods Section*. Alexandria, VA: American Statistical Association.

Fay R. (1996). "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association* 91: 490–498.

Fellegi, I. and A. Sunter. (1969). “A Theory for Record Linkage. *Journal of the American Statistical Association* 64: 1183–1210.

Foot, D., and W. Milne. (1989). “Multiregional Estimation of Gross Internal Migration Flows.” *International Regional Science Review* 12: 29–43.

Fuguitt, G., and D. Brown. (1990). “Residential Preferences and Population Redistribution.” *Demography* 27: 589–600.

Golden, S. and M. Skibniewski. (2009). “Immigration and Construction: The Makeup of the Workforce in the Washington, DC, Metropolitan Area.” *Journal of Construction Engineering and Management*. 135: 874–880.

Goodchild, M. (2009). “What Problem? Spatial Autocorrelation and Geographic Information Science.” *Geographical Analysis* 4: 411–417.

Goodchild M., and Kwan M., (1978). "Models of Hierarchically Dominated Spatial Interaction." *Environment and Planning A* 10: 1307 – 1317.

Goodman, L.A. (1961). "Snowball sampling.” *Annals of Mathematical Statistics* 32: 148–170

Griffith, C., B. Long, D. A. Swanson, and M. Knight. (2011). “DOMICILE 1.0: An Agent-Based Simulation Model for Population Estimates at the Domicile Level” pp. 345–370 in N. Hoque and D. A. Swanson (Eds.) *Opportunities and Challenges for Applied Demography in the 21st Century*. Dordrecht, Heidelberg, London, and New York: Springer.

Greenberg, N. D. Kreueckeberg, and C. Michaelson. (1978). *Local Population and Employment Projection Techniques*. New Brunswick, NJ: Center for Urban Policy and Research, Rutgers University.

Greenwood, M. (1975). “Simultaneity Bias in Migration Models: An Empirical Investigation.” *Demography* 12: 519–536.

Greenwood, M. and G. Hunt. (1991). “Forecasting State and Local Population Growth with Limited Data: The Use of Employment Migration Relationships and Trends in Vital Rates.” *Environment and Planning A* 23: 987–1005.

Greenwood, M., and G. Hunt. (1991). "Forecasting State and Local Population Growth with Limited Data: The Use of Employment Migration Relationships and Trends in Vital Rates." *Environment and Planning A* 23: 987–1005.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S., Huse, G., Huth, A., Jepsen, J.U., Jørgensen, C., Mooij, W.M., Müller, B., Peer, G., Piou, C., Railsback, S.F., Robbins, A.M., Robbins, M.M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R.A., Vabø, R., Visser, U. and DeAngelis, D.L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling* 198, 115–126 (https://www.slashtmp.iu.edu/public/download.php?FILE=casgriff/48386lUP3AI).

Harding, A. and A. Gupta. (2007). *Modelling our Future: Population Ageing, Social Security and Taxation*. Amsterdam, The Netherlands: North-Holland.

Healy, M. (1982). "Using Administrative Records: Introduction and Basic Procedures." pp. 27 –37 in E. S. Lee and H. F. Goldsmith (Eds.) *Population Estimates: Methods for Small Area Analysis*. Beverly Hill, CA: Sage Press.

Heckathorn, D. (1997). "Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations." *Social Problems* 44: 174–199.

Heckathorn, D. (2002). "Respondent-Driven Sampling II: Deriving Valid Estimates from Chain-Referral Samples of Hidden Populations." *Social Problems* 49: 11–34

Hogan, H. (2000). "Accuracy and Coverage Evaluation: Theory and Application." Paper presented at the joint Statistical Meetings, Indianapolis, IN.

Hogan, H. (1993). "The 1990 Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060.

Hogan, H. (1992). "The 1990 Post-Enumeration Survey: An Overview." *The American Statistician* 46: 261–269.

Hogan, H., and C. Cowan. (1980). "Imputations, Response Errors, and Matching in Dual System Estimation." pp. 263–268 in *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Alexandria, VA: American Statistical Association

Hunt, J. D., and J. Abraham. (2005). Design and Implementation of PECAS: A Generalized System for Allocating Economic Production, Exchange, and Consumption Categories." pp. 253–274 in M. Lee-Gosselin and S. Doherty (eds.), *Integrated Land Use and Transportation Models, Behavioural Foundations*. Toronto, Canada: Elsevier.

International Microsimulation Association. (2006). What is Microsimulation?, (http:www.microsimulation.org).

Isserman, A., D. Plane, P. Rogerson, and P. Beaumont. (1995). "Forecasting Interstate Migration with Limited Data: A Demographic-economic Approach." *Journal of the American Statistical Association* 80: 277–285.

Judson, D. (2007). "Information Integration for Constructing Social Statistics: History, Theory and Ideas Towards a Research Programme." *Journal of the Royal Statistical Society (Series A)* 170 (2): 483–501.

Kalton, G. (1983). *Compensating for Missing Survey Data*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.

Kliss, B. and W. Alvey (eds.). (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects, Volumes I and II*. Washington, DC: Department of the Treasury, Internal Revenue Division, Statistics of Income Division.

Korotayez, A. (2005). "A Compact Macromodel of World System Evolution." *Journal of World-Systems Research* 11: 79–93.

Krieg, R., and A. Bohora. (1999). "A Simultaneous Profit Model of Earnings, Migration, Job Change with Wage Heterogeneity." *The Annals of Regional Science* 33: 453–467.

Krótki, K. J. (1978). *Developments in Dual System Estimation of Population Size and Growth*. Edmonton, Alberta, Canada: University of Alberta Press.

Larson, E. and J. Droitcour. 2012. "The Grouped Answer Method for Estimating Immigration Status: Analysis of Data from the 2004 General Social Survey". pp. 311–334 in N. Hoque and

D. Swanson (Eds.) *Opportunities and Challenges for Applied Demography in the 21st Century*. Dordrecht, Heidelberg, London, and New York: Springer.

Liu, X. (2008). "Using a MAF-Based Frame for Demographic Household Surveys." *Proceedings of the Government Statistics Section, American Statistical Association*: 2864–2871.

Liu, X. (2007). "Comparing the Quality of the Master Address File and the Current Demographic Household Surveys' Multiple Frames." Paper presented at the 2007 Research Conference of the Federal Committee on Statistical Methodology, November 7–10, Arlington, VA.

Long, J. (1993). "Post-censal Population Estimates: States, Counties, and Places." *Population Division Working Paper No. 3*. Washington, DC: US Census Bureau (http://www.census.gov/population/www/documentation/twps0003.html).

Longhi, S. and P. Nijkamp. (2005). *Forecasting Regional Labour Market Developments Under Spatial Heterogeneity and Spatial Autocorrelation*. Tinbergen Institute Discussion Paper, TI 2005-041/3. Amsterdam, The Netherlands, Tinbergen Institute, Amsterdam Free University.

Madow, W., I. Olkin, and D. Rubin. (1983). *Incomplete Data in Sample Surveys, Vol.2*. New York, NY. Academic Press.

Malenfant, É., L. Martel, and A. Lebel. (2011). "An Overview of Demosim, Statistics Canada's Microsimulation Model for Population Projections." forthcoming in N. Hoque and D. A. Swanson (Eds.) *Opportunities and Challenges for Applied Demography in the 21st Century: Selected Papers from the 2010 Conference on Applied Demography*. Dordrecht, Heidelberg, London, and New York: Springer.

Martel, L. (2010). "Population Projections of Ethnocultural Minority Groups in Canada using Demosim, a Microsimulation Model." Paper presented at the 2010 Applied Demography Conference, January 10–12, San Antonio, Texas.

Martins, J., F. Yusuf, and D. A. Swanson. (2012). *Consumer Demographics and Behaviour: Markets are People*. Dordrecht, Heidelberg, London, and New York: Springer.

Massey, D., R. Alarcon, J. Durand, and H. Gonzalez. (1987). *Return to Aztlan: The Social Process of International Migration from Western Mexico*. Berkeley, CA: University of California Press.

Masters, T. (1993). *Practical Neural Network Recipes in C++*. San Diego, CA: Academic Press.

McKibben, J. And D. A. Swanson. (1997). "Linking Substance and Practice: A Case Study of the Relationship Between Socio-Economic Structure and Population Estimation." *Journal of Economic and Social and Measurement* 23: 135–147.

Miller, E., D. Kriger, and J. D. Hunt. (1999). *Integrated Urban Models for Simulation of Transit and Land Use Policies: Guidelines for Implementation and Use*. TCRP Report 48. Transportation Research Board, National Research Council. Washington, DC: National Academy Press.

Mills, E. and L. Lubuele. (1995). "Projecting Growth in Metropolitan Areas." *Journal of Urban Economics* 37: 344–360.

Mitas, L. and H. Mitasova. (1999). "Spatial Interpolation." pp. 481–492 in P. Longley, M. Goodchild, D. Maguire, D. Rhind (Eds.) *Geographical Information Systems: Principles, Techniques, Management and Applications*. New York, NY: Wiley.

Murdock, S. and D. Ellis. (1991). *Applied Demography: An Introduction to Basic Concepts, Methods, and Data*. Boulder, CO: Westview Press.

Murdock, S., F. L. Leistritz, R. Hamm, S. Hwang, and B. Parpia. (1984)."An Assessment of the Accuracy of a Regional Economic-Demographic Model." *Demography* 21: 383–404.

Myrskylä, P. (1991). "Census by Questionnaire – Census by Registers and Administrative Records: The Experience of Finland." *Journal of Official Statistics* 7 (4): 457–474.

Norušis, M. (1991). *The SPPS Gide to Data Analysis for SPSS/PC+, 2nd Edition*. Chicago, IL: SPSS, Inc.

Paik, H. (2000). "Comments on Neural Networks." *Sociological Methods and Research* 28 (4): 425–453.

Palmore, J. (1967). "The Chicago Snowball: A Study of the Flow and Diffusion of Family Planning. pp. 272–363 in D. Bogue (Ed.) *Sociological Contributions to Family Research*. Chicago, IL: Community and Family Study Center, University of Chicago

Paradies, Y. and T. Barnes. (2005). "A New Variant of Dual-record Population Estimation with an Application in Remote Indigenous Communities." *Journal of Population Research* 22: 119–139.

Peterson, I. (1999). "Census Sampling Confusion: Controversy Dogs the Use of Statistical Methods to Adjust US Population Figures." *Science News Online* 155 March 6 (http://www.sciencenews.org/sn_arc99/3_6_99/bob1.htm).

Popoff, C., and D. Judson. (2004). "Some Methods of Estimation for Statistically Underdeveloped Areas." pp. 603–641 in J. Siegel and D. A. Swanson (Eds.). *The Methods and Materials of Demography, 2nd Edition*. New York, NY: Elsevier Academic Press.

Putnam, S. H. (1983). *Integrated Urban Models*. London, England: Pion Press.

Putnam, S. H. (1991). *Integrated Urban Models II*. London, England: Pion Press.

Raghunathan, T., J. Reiter, and D. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16.

Robinson, J. G, A. Adlakha, and K. West. (2002). "Coverage of Population in Census 2000: Results from Demographic Analysis." Paper presented at the Annual Meeting of the Population Association of America, Atlanta, GA.

Robinson, J. G., B. Ahmed, P. Das Gupta, and K. A. Woodrow. (1993). "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis." *Journal of the American Statistical Association* 88: 1061–1071.

Rubin, D. 2004. *Multiple Imputation for Non-response in Surveys*. New York, NY: Wiley Interscience.

Salganick, M., and D. Heckathorn. (2004). "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling." *Sociological Methodology* 34: 193–239.

SANDAG. (1998). *Urban Development Model, Volume 1. Technical Description*. San Diego, CA: San Diego Association of Governments.

SANDAG. (1999). *Urban Development Model, Volume 2. Technical Description*. San Diego, CA: San Diego Association of Governments.

Sarle, W. (1994). "Neural Networks and Statistical Models." pp. 1538–50 in *SAS Institute Inc. Proceedings of the Nineteenth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Scheuren, F. (1999). "Administrative Records and Census Taking." *Survey Methodology* 25 (2): 151–160.

Schmidt, R., C. Barr, and D. A. Swanson. (1997). "Socioeconomic Impacts of the Proposed Federal Gaming Tax." *International Journal of Public Administration* 20: 1675–1698.

Shryock, H., and N. Lawrence. (1949). "The Current Status of State and Local Population Estimates in the Census Bureau." *Journal of the American Statistical Association* 44: 157–173.

Singh, M., M. Hidiroglou, J. Gambino and M. Kovacević. (2001). "Estimation Methods and Related Systems at Statistics Canada." *International Statistical Review* 69: 461–485.

Smith, S., J. Tayman, and D. A. Swanson. (2001). *State and Local Population Projections: Methodology and Analysis*. New York, NY: Kluwer Academic/Plenum Publishers.

Sokolova, M.V., R. J. Rasras and D. Skopin. (2006). "The Artificial Neural Network Based Approach for Mortality Structure Analysis." *American Journal of Applied Science* 3 (2): 1698–1702.

State of Alaska. (2008). *Methods for the Alaska Population Estimates*. Juneau, AK: Alaska Department of Labor and Workforce Development, Research and Analysis Section (http://labor.state.ak.us/research/pop/estimates/data/AKPopEstMethods.pdf).

Statistics Canada. (2009). "Post Conference MODGEN Workshop." 2nd General Conference of the Microsimulation Association. June 11th, Ottawa, Ontario, Canada (http://www.statcan.gc.ca/conferences/ima-aim2009/modgen-eng.htm).

Statistics Finland. (2004). *Use of Registers and Administrative Data Sources for Statistical Purposes: Best Practices of Statistics* Finland. Handbook 45. Helsinki, Finland: Statistics Finland.

Swanson, D. A. (2008). "Applied Demography in Action: A Case Study of "Population Identification." *Canadian Studies in Population* 35: 133–158.

Swanson, D. A. (1986). "Missing Survey Data in End-Use Energy Models: An Overlooked Problem." *The Energy Journal* 7: 149–158.

Swanson, D. A. (1980). "Improving Accuracy in Multiple Regression Estimates of County Populations using Principles from Causal Modeling." *Demography* 17: 413–417.

Swanson, D. A. and D. Beck. (1994). "A New Short-term County Population Projection Method." *Journal of Economic and Social Measurement* 20: 25–50.

Swanson, D. and M. Knight. (1998). *Metromail Wealth Estimation Project Final Report: Recommendations, Summary Findings, and Technical Documentation*. Madison, WI: Third Wave Research Group

Swanson, D. A. and G. E. Stephan. (2004). "Glossary" pp. 751–778 in J. Siegel and D.A. Swanson (Eds.) *The Methods and Materials of Demography, 2nd Edition*. New York, NY: Elsevier Academic Press.

Swanson, D. A., and L. M. Tedrow. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography*, 21 (3): 373–381.

Swanson, D. A. and P. Walashek. (2011). *CEMAF as a Census Method: A Proposal for a Re-Designed Census and an Independent Census Bureau*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Swanson, D. A., J. McKibben, L. Wombold, R. Forgette, and M. Van Boening. (2009). "The Demographic Effects of Katrina: An Impact Analysis Perspective." *The Open Demography Journal* 2: 36–46.

Tabuchi, T. (1985). "Time Series Modeling of Gross migration and Dynamic Equilibrium." *Journal of Regional Science* 25: 65–83.

Tang, Z., C. Leung, and K. Bagchi. (2006). "Improving Population Estimation with Neural Network Models." *Advances in Neural Networks* 3973: 1181–1186.

Tayman, J. (1996). "Forecasting, Growth Management, and Public Policy Decision Making. *Population Research and Policy Review* 15: 491–508.

Tayman, J. and E. Shafer. (1985). "The Impact of Coefficient Drift and Measurement Error on the Accuracy of Ratio-Correlation Population Estimates." *The Review of Regional Studies* 15: 13–23.

Thomsen, I. and A. M. K. Holmøy, (1998). "Combining Data from Surveys and Administrative Records Systems, the Norwegian Experience." *International Statistical Review* 66 (2): 201–221.

Treyz, G. (1993). Regional Economic Modeling: A Systematic Approach to Economic Forecasting and Policy Analysis. Boston, MA: Kluwer Academic Press.

Treyz, G. (1995). "Policy Analysis: Application of REMI economic forecasting and simulation models." *International Journal of Public Administration* 18: 13–42.

Treyz, G., D. Rickman, and G. Shao. (1991). "The REMI Economic-Demographic Forecasting and Simulation Model." *International Regional Science Review* 14: 221–253.

Treyz, G., D. Rickman, G. Hunt, and M. Greenwood. (1993). "The Dynamics of US Internal Migration." *Review of Economics and Statistics* 75: 209–214.

US Bureau of Economic Analysis, (1995). BEA Regional Projections. Washington, DC: US Bureau of Economic Analysis.

US EPA (Environmental Protection Agency). (2000). *Projecting Land-Use Change: A Summary of Models for Assessing the Effects of Community Growth and Change on Land-Use Patterns*. EPA/600/R-00/098. US Environmental Protection Agency, Office of Research and Development, Cincinnati, OH.

U. S. Government Accountability Office (US GAO). (2006). *Estimating the Undocumented Population: A "Grouped Answers" Approach to Surveying Foreign-Born Respondents*. GAO-06-775. Washington, DC: US Government Accountability Office.

U. S. Governmental Accountability Office (US GAO). (1999). *Survey Methodology: An Innovative Technique for Estimating Sensitive Survey Items*. GGD-00-30. Washington, DC: US Government Accountability Office. .

U. S. Government Accountability Office (US GAO). (1998). Immigration Statistics: Information Gaps, Quality Issues Limit Utility of Federal Data to Policymakers. GGD-98-164. Washington, DC: US Government Accountability Office.

United States Postal Service (USPS). (2011). *NCOALink® Systems*.
(http://www.usps.com/ncsc/addressservices/moveupdate/changeaddress.htm).

Van der Gaag, N., J. de Beer, and F. Willekens. (2005). "Combining Micro and Macro Approaches in Demographic Forecasting." Paper presented at the Joint Eurostat-ECE Work Session on Demographic Projections, Vienna, Austria, September 21–23.

Vaupel, J., and A. I. Yashin. (2006). "Unobserved Population Heterogeneity." pp. 271–278 in G. Caselli, J. Vallin and G. Wunsch (Eds.) *Demography: Analysis and Synthesis: A Treatise in Population Studies, Volume 1*. London, England: Academic Press.

Voss, P. (2007). "Demography as a Spatial Social Science." *Population Research and Policy Review* 26: 457–476.

Voss, P., R. Hammer, and S. Friedman. (2006). "County child poverty rates in the US: a spatial regression approach." *Population Research and Policy Review* 25: 369–391.

Waddell, P. (2000). "A Behavioral Simulation Model for Metropolitan Policy Analysis and Planning: Residential Location and Housing Market Components of UrbanSim." *Environment and Planning B* 27: 247–263.

Weinberg, D. (2009). (Reissue): "Currently Planned External Sources of Data for the 2010 Census." 2010 *Census Information Memoranda Series No. 36*. (August 3), Washington, DC: US Census Bureau.

Williams, B., J. Nichols, and M. Conroy. (2002). *Analysis and Management of Animal Populations*. New York, NY: Academic Press.

Wolter, K. (1986). "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346

Zinn, S, J. Gampe, J. Himmelspach, and A. Uhrmacher. (2010). "A DEVS Model for Demographic Microsimulation." Paper presented at the Spring Simulation Multiconference 2010, Orlando, Florida, April 11–15.

Chapter 13
Special Cases and Adjustments[1]

Population estimate methods can be applied in a straightforward manner in many situations, without consideration of any factors beyond those discussed in previous chapters. However, there are also situations in which the basic estimation model should be adjusted to account for special circumstances. Two common adjustments are for international migration and special populations.[2] Failing to account for these factors can lead to unreasonable estimates and can increase estimation errors. Whether any specific set of estimates requires adjustment for these factors is, of course, a question that must be answered on a case-by-case basis. There are also circumstances in which it is desirable to control a set of estimates to an independent estimate. In this chapter we discuss the circumstances in which unadjusted estimates might provide unacceptable results, describe ways for making the necessary adjustments, and describe several techniques for controlling to independent estimates. The adjustments described in this chapter increase the complexity of the estimation process, but we believe enhance the usefulness of the resulting estimates.

13.1 International Migration

In the various component methods described in Chapter 10, the migration element did not distinguish between internal and international migration. However, international migrants often have different characteristics than internal migrants, are influenced by different factors, and exhibit different patterns of change. Consequently, if international immigration is an important component of growth for a particular state or local area, it may be useful to estimate it separately from internal migration.

Table 13.1 shows population change due to internal and international migration for Census regions and states from 2000 to 2009. There is a tremendous amount of variability in the relative contributions of internal and international migration to population change. In the Northeast and Midwest regions, internal migration results

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_13,

Table 13.1 Net Total, Internal, and International Migration, Census Regions and States, 2000–2009

	International	Internal	Total
Region			
Northeast	1,835,442	−2,539,582	−704,140
Midwest	1,158,438	−1,752,191	−593,753
South	3,118,775	3,874,132	6,992,907
West	2,831,515	417,641	3,249,156
State			
Alabama	50,742	85,710	136,452
Alaska	8,308	−9,032	−724
Arizona	272,410	714,354	986,764
Arkansas	36,478	76,445	112,923
California	1,816,633	−1,509,708	306,925
Colorado	144,861	212,822	357,683
Connecticut	112,936	−96,328	16,608
Delaware	19,523	46,524	66,047
District of Columbia	24,179	−41,606	−17,427
Florida	851,260	1,182,974	2,034,234
Georgia	281,998	567,135	849,133
Hawaii	38,951	−33,108	5,843
Idaho	22,121	112,341	134,462
Illinois	403,978	−632,866	−228,888
Indian a	93,367	−21,734	71,633
Iowa	36,329	−52,205	−15,876
Kansas	52,388	−69,962	−17,574
Kentucky	44,314	82,517	126,831
Louisiana	33,046	−318,811	−285,765
Maine	8,079	30,725	38,804
Maryland	191,262	−95,972	95,290
Massachusetts	245,145	−276,768	−31,623
Michigan	168,668	−540,750	−372,082
Minnesota	106,388	−43,962	62,426
Mississippi	17,572	−36,545	−18,973
Missouri	63,420	42,041	105,461
Montana	3,042	39,938	42,980
Nebraska	31,988	−41,144	−9,156
Nevada	110,681	374,762	485,443
New Hampshire	18,373	35,087	53,460
New Jersey	399,803	−459,803	−60,000
New Mexico	47,343	23,215	70,558
New York	839,590	−1,686,583	−846,993
North Carolina	214,573	675,016	889,589
North Dakota	4,568	−19,785	−15,217
Ohio	120,452	−368,203	−247,751
Oklahoma	53,514	39,463	92,977
Oregon	95,484	178,547	274,031
Pennsylvania	176,498	−40,139	136,359

(continued)

Table 13.1 (continued)

	International	Internal	Total
Rhode Island	30,017	−44,649	−14,632
South Carolina	65,869	310,572	376,441
South Dakota	6,545	6,822	13,367
Tennessee	91,508	264,570	356,078
Texas	933,083	848,702	1,781,785
Utah	65,961	52,582	118,543
Vermont	5,001	−1,124	3,877
Virginia	204,219	171,420	375,639
Washington	202,442	238,546	440,988
West Virginia	5,635	16,018	21,653
Wisconsin	70,347	−10,443	59,904
Wyoming	3,278	22,382	25,660

Source: US Census Bureau, Population Division (NST-EST 2009-04).
Release Date: December 2009

in population loss that is not offset by the gains from international migration. Internal migration contributes to 55% of the total net migration in the South region, but only 13% in the West region. Almost half of the states (24) experienced net internal out-migration and net growth due to international migration. In states with positive internal and international migration, internal migration accounts for more of the total net migration in 21 states; the share of internal to total net migration ranges from 51% in South Dakota to 93% in Montana. In the remaining six states, the internal share ranges from 33% in New Mexico to 48% in Texas.

Table 13.2 compares selected demographic characteristics for internal and international migration in the US for 2009-2010. Internal migrants have a greater representation than international migrants in ages under 18 and in ages 55 and older. The two groups of migrants have about the same representation in ages 35 to 54. International migrants have a larger share in the prime migration years of 18 to 34, especially those 25 to 34. Internal migrants tend to have similar concentration of males and females, while international migrants have slightly higher concentration of males (53.3%). As expected, significant differences are seen in the racial and Hispanic origin composition between internal and international migrants. White alone is the predominate race for both internal and international migration, but the White alone share is 21.7 percentage points higher for internal migrants. Black alone has the second highest concentration of internal migrants (14.6%), while Asian alone has the second highest concentration of international migrants (31.3%). Two-thirds of internal migrants are White non-Hispanic compared to 27% of international migrants; international migrants concentrate in the Hispanic (32.2%) and other non-Hispanic (40.5%) groups.

Figure 13.1 shows distinctly different patterns of temporal change due to internal and international net migration in San Diego County from 1980 to 2010. International migration shows a relatively stable pattern of change, ranging from 8,570 to 19,100 per year. Internal migration fluctuates dramatically, generally following

Table 13.2 Characteristics of Internal and International Migrants, United States, 2009–10

Age Group	Internal[a]	International[b]	Difference[c]
<18	23.5%	18.4%	5.1%
18 to 24	19.2%	22.4%	−3.2%
25 to 34	25.1%	31.7%	−6.6%
35 to 44	11.9%	11.4%	0.5%
45 to 54	9.1%	9.3%	−0.2%
55 to 64	6.3%	3.8%	2.5%
65+	4.9%	3.0%	1.9%
	100.0%	100.0%	0.0%
Sex	Internal	International	Difference
Male	50.3%	53.3%	−3.0%
Female	49.7%	46.7%	3.0%
	100.0%	100.0%	0.0%
Race	Internal	International	Difference
White alone	77.1%	55.4%	21.7%
Black alone	14.6%	10.1%	4.5%
Asian alone	4.2%	31.3%	−27.1%
Other races[d]	4.1%	3.2%	0.9%
	100.0%	100.0%	0.0%
Hispanic Origin	Internal	International	Difference
Hispanic	13.3%	32.2%	−18.9%
White non-Hispanic	65.5%	27.3%	38.2%
Other non-Hispanic	21.2%	40.5%	−19.3%
	100.0%	100.0%	0.0%

[a]Movers from different county, same state; different state, same division; different division, same region, and different region
[b]Movers from abroad
[c]Internal - International
[d]Includes American Indian and Alaska Native alone, Native Hawaiian and Other Pacific Islander alone and Other Islander alone, and all race combinations
Source: US Census Bureau, Current Population Survey,
2010 Annual Social and Economic Supplement, Internet release date: May 2011

cycles of employment opportunities and overall economic changes; its range is -40,700 to 50,000 per year. Because of the lack of a relationship between the local economy and international migration, structural economic models of migration (discussed in Chapter 12) often focus on just internal migration.

How can international migration be estimated? Three primary sources of information provide information for estimating international migration for the US and its subnational areas: 1) Department of Homeland Security (DHS); 2) ACS; and 3) the Census Bureau's annual estimates program. These sources are comprehensive in the sense that they provide data for subnational areas covering the entire US, but there are significant differences between them concerning the detail provided and aspects of international migration they estimate. Independent estimates of international migration for particular states and counties are also available from a few state demographic centers (e.g. California, New York).

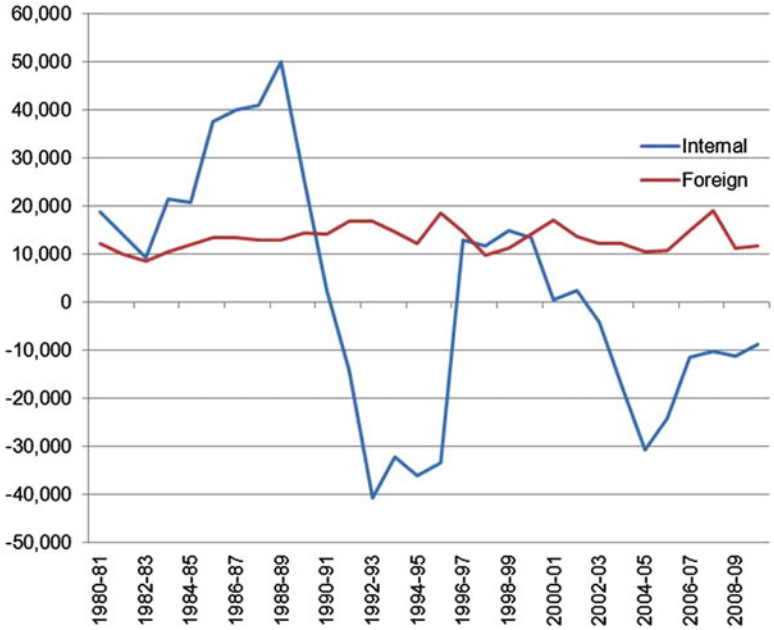


Fig. 13.1 Population Change due to Internal and International Migration, San Diego County, 1980–2010
Source: State of California, Department of Finance, Sacramento, CA
Population Estimates and Components of Change by County- July 1, 1999–2010, with 2010 Census Benchmark. August 2011
Revised County Population Estimates and Components of Change by County, July 1 1990–2000. February 2005
Revised County Population Estimates and Components of Change by County, July 1, 1970–1990. February 1995

The DHS provides annual fiscal year information on legal permanent residents or foreign nationals who have been granted the right to reside permanently in the United States. DHS data identifies new arrivals to an area, as well as the number of persons in an area whose status has been changed to a legal permanent resident. Information is available for the US, states, and core based statistical areas, but not for individual counties.[3] The DHS legal permanent resident profile includes information on sex, age, county of birth, marital status, occupation, and class of admission (family sponsored, employment-base, relative of US citizens, diversity, refugees and asylees, and other); no cross-classifications of these variables are provided. Data from the DHS do not provide a complete picture of international migration. They do not reflect emigration from an area and exclude estimates of undocumented migration. Moreover, the DHS only provides the year of the status change and not when the immigrant came into the US, which may overstate the immigration for a particular year.

The ACS arguably offers the most comprehensive source of information on international migration through its question on previous residence one-year prior. Information is provided for persons who moved to an area and were living abroad

Table 13.3 Alternative Estimates of International Migration, United States and San Diego County, 2009

	Dept. of Homeland Security[a]				
	New Arrivals	Adjustment of Status	Total	ACS[b]	Census[c] Bureau
US	463,042	667,776	1,130,818	1,687,595	854,905
San Diego County	8,367	12,412	20,779	33,613	10,270

[a]Legal Permanent Residents, Fiscal Year 2009
[b]All persons in 2009 whose previous residence one year prior was abroad
[c]Net international migration, 2008–2009
Sources: Department of Homeland Security
US Census Bureau, 2009 ACS
US Census Bureau, 2009 Vintage Population Estimates

during the prior year. The ACS, at least indirectly, does include undocumented migration to the extent they were included in and responded to the survey. A wide range of characteristics are available from the ACS including age, sex, race, Hispanic origin, marital status, educational attainment, and income. The tabulated data from the ACS do not provide cross-classifications of these variables and cover all persons moving from overseas; including, for example, US residents living abroad. ACS Public Use Microdata Sample (PUMS) files can be used to identify the foreign born moving to the US and to cross-classify variables for use in estimation models. Like the DHS profiles, the ACS does not include information on emigration.

The Census Bureau produces estimates of net international migration for the US, states, and counties as a component of change for their annual estimates. While the Census Bureau estimates net international migration by age, sex, race, and Hispanic Origin, only the overall number is provided in their public data. The international migration component combines for parts: 1) net international migration of the foreign born, 2) net migration between the US and Puerto Rico; 3) net migration of natives to and from the US, and 4) net movement of Armed Forces population to and from the United States. One important advantage of the Census Bureau's estimate is it includes both estimates of immigration and emigration for international migration of the foreign born and migration between the US and Puerto Rico.[4]

The Census Bureau estimates foreign born immigration to the US using ACS information on the reported residence of the foreign born population in the prior year. Emigration of the foreign born from the US is estimated using a residual method that ages forward the foreign born population in Census 2000 and compares it to the expected population to the foreign born population estimated by the ACS. State- and county-specific factors based on the foreign born population entering the US within 5-years and with 10-years are used to distribute the US estimates of immigration and emigration. Additional details of the procedures to estimate net international migration are found in US Census Bureau (2010).

Table 13.3 illustrates the differences between the DHS, ACS, and Census Bureau estimates of international migration for the US and San Diego County for the period around 2009. The ACS and DHS estimates of total international migration are larger than the Census Bureau's estimates for both San Diego County and the United States. The ACS estimate is by far the highest of any source. It is double

the Census Bureau's estimate for the US and is more than triple its estimate for San Diego County. The ACS estimate is also 49% higher than the DHS estimate for the US and is 62% higher than the estimate for San Diego County. In both the US and San Diego, approximately 59% of the legal permanent residents in 2009 represent an adjustment of status and not a recent move. A large fraction of status adjustment would impact the use of DHS data for annual estimates. By comparison, the Census Bureau's estimate for San Diego County is 900 persons (9%) less than the estimate of net international migration for the same time period prepared by the State of California (2010).

While these results are illustrative, they suggest the Census Bureau's estimate may be more suitable for population estimation than either the DHS or basic data from the ACS. As noted above, the Census Bureau only provides an estimate of the total net international migration. Demographic characteristics, if needed, can be obtained from the ACS by combining the characteristics of foreign born immigrants and foreign born population entering the US during the past 10 years. These characteristics would be controlled to the Census Bureau total (see the Controlling section later in this chapter). If one chooses to create their own estimates of international migration using DHS, ACS, or other sources, we recommend assuming an emigration rate of 20 percent of the immigration level (see Endnote 4) to better approximate the net effect of international migration on population change.

13.2 Special Populations

A special population is a group of persons located in an area because of an administrative or legislative action (Pittenger 1976: 205). Common types include college students, prison inmates, residents of nursing homes, and military personal and their dependents. Special populations complicate post-censal estimates because their growth and decline are not determined by the same factors that affect change in the general population; consequently, they often follow different growth trends.

Special populations often have different demographic characteristics as well. For example, military personnel and college students are concentrated primarily in the young adult ages, residents of nursing homes are concentrated primarily in the older ages, and the prison population often has a high concentration of males and minorities. These differences can have a substantial impact on the factors used to estimate population.

Another confounding characteristic of special populations is that they often do not age in place like other population groups. Instead, their age structure may remain fairly stable over time. For example, a college town sees a large inflow of people age 17-19 and a large outflow of people age 21-23 every year. Consequently, a substantial proportion of the town's young adult population replaces itself repeatedly rather than aging in place.

Special populations do not cause any particular problems for population estimates if they comprise only a small proportion of the total population or if

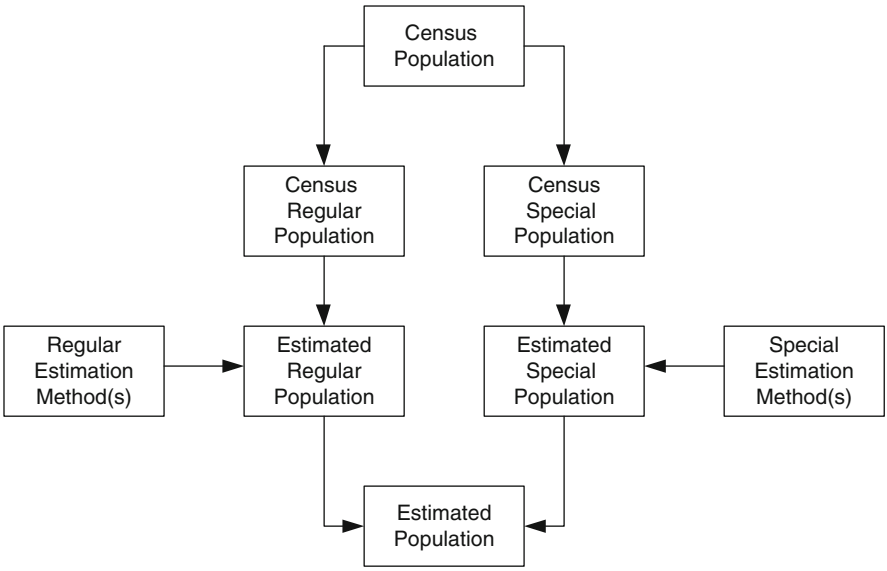


Fig. 13.2 Accounting for Special Populations

their growth rates and demographic characteristics are similar to the rest of the population. In these circumstances no special adjustments are needed. When special populations follow different trends and account for a substantial proportion of the total population, however, adjustments should be made.

Unfortunately, there is no rule of thumb defining "different" or "substantial." Consequently, the analyst must evaluate each situation separately, focusing on the special population's demographic composition, growth trends, components of growth, and—perhaps most important—its share of total population. It is a good bet that special populations will have a significant impact on estimates in areas with large prisons, military installations, colleges, or universities. Nursing homes, boarding schools, and mental institutions can also be important. The impact of special populations is generally greater in small areas than in large areas. For example, a prison may have little impact on the total population of a county, but may comprise the most if not all of the population of a census tract as seen in Table 7.8.

Several steps can be followed to account for special populations when making population estimates (see Figure 13.2). The first is to create the "regular" population in the census data by subtracting the special population from the total population. The second is to estimate the regular population, using the methods described in earlier chapters. The third is to estimate the special population itself, using one of the approaches described below. The final step is to add the estimate of the special population to the estimate of the regular population.

How can special populations be estimated? One approach is to develop a component model for the special population itself, using data and rates for the components of change that pertain specifically to that population (Pittenger 1976: 205). This approach will be useful if the special population accounts for a large proportion

of the total population and if the necessary data are available. Data limitations often make this approach impractical, especially for small areas. A second approach is to use the special population and its characteristics from the last census. This approach will be useful if the size and demographic composition of the special population has been relatively stable over time and is expected to remain so in the post-censal period. It will also be useful if the direction and magnitude of changes are completely unpredictable between the last census and time the estimate is made. Finally, estimates of special populations can be based on information collected from the administrators of facilities such as colleges, prisons, or nursing homes. Combinations of the various approaches can also be used, such as holding the demographic composition of the special population constant while allowing for changes in its total size.

Data availability is a difficult problem for the development of independent estimates of special populations. County-level population and in-migration data for some special population groups can be obtained from the ACS, using either summary files or PUMS data. Data are available for persons residing in group quarters facilities such as military barracks, prisons, and college dormitories, but are not available for military dependents (i.e., spouses and children of military personnel) and for college students not living in dormitories. Estimates for these groups can be made using PUMS data that identify all households headed by military personnel or students or if they are the partner of the head of household, but these estimates may have considerable error. Since PUMS data are available only for counties and subcounty areas with 100,000 residents or more, they cannot be used for small areas.

It is also difficult to obtain mortality and fertility data specific to special populations. In these instances, the analyst may simply have to make an educated guess. Fortunately, the development of birth and death information specific to a special population is generally unnecessary, either because the special population's contribution to local births and deaths is very small or because rates for the special population are similar to rates for the population as a whole. The military population (including dependents) may be an exception. Fertility rates for this group are often higher than for the non-military population. Figure 13.3 illustrates these differences in fertility behavior using child woman ratios (CWR) ages 0-4 for the military and total population in San Diego County.

If the military population comprises a substantial portion of the total population, it may be advisable to account separately for the fertility of this population. Birth certificates in some counties report the military status of parents, providing an excellent source of data on military births. In most counties, however, birth certificates do not include this information. Another possibility is to obtain information on births occurring in military hospitals, either directly from the hospital or from birth certificates. Although useful, this information excludes data on births to military families that did not occur in military hospitals. When complete data on military births are available, the analyst can easily develop estimates of military births and incorporate them directly into the estimation process.

How can estimates of military births be developed for places lacking these types of data? One possibility is to use PUMS data from the most recent census or the ACS

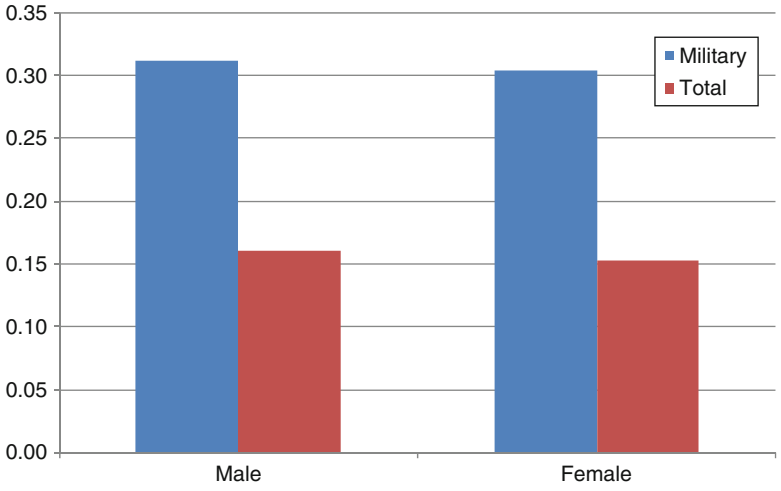


Fig. 13.3 Child- Woman Ratio Ages 0-4, Military and Total Polpulation San Diego County, 2000 Source: US Census Bureau, 2000 PUMS

to calculate the number of children ages 0-4 for the military and total populations. An adjustment factor can be developed by forming a ratio of the military children to the all children. Estimates of military birth rates can then be made by applying this adjustment factor to births reported for the entire population. Military birth rates can be estimated from CWRs for the military and total populations. The ratio of the military CWR to the total CWR applied to the age-specific birth rates (ASBRs) for the entire population will yield an estimate of military ASBRs. This approach assumes that the pattern of ASBRs is the same for military and non-military populations.

Obtaining special population data is even more difficult for subcounty areas than it is for counties. The decennial census and ACS provides subcounty data on total numbers for some types of special populations, but generally does not provide data on their demographic characteristics. One approach to dealing with this problem is to identify a small area—such as a census tract or individual block—in which the entire population belongs to the special population group. In instances like this, data defining the total population also define the special population. Those data can then be used to estimate the characteristics of similar special populations in nearby areas. For example, suppose that estimates are to be made for a census tract containing a prison. Suppose further that two particular blocks within that census tract are identified as containing solely prison inmates. The demographic characteristics of those two blocks can be used as an estimate of the characteristics of the prison population for the entire census tract.

The presence of a special population in an area is a red flag warning the analyst to pay special attention to the data used in the estimation model. If the special population is large enough and differs significantly from the rest of the population in terms of its demographic characteristics and growth rates, it should be accounted for explicitly in the estimation model. Sometimes the necessary data are available for special populations, sometimes they are not. When the necessary data are not

Table 13.4 Alternative Estimates of the Male Population, San Diego County, 2010

	Cohort Component Model		Difference[a]	
Age Group	Basic[b]	Adjusted[c]	Number	Percent
0-4	117,917	117,917	0	0.0%
5-9	114,425	114,437	−12	0.0%
10-14	102,586	102,649	−63	−0.1%
15-19	116,156	121,166	−5,010	−4.3%
20-24	123,614	146,576	−22,962	−18.6%
25-29	122,934	125,756	−2,822	−2.3%
30-34	126,908	112,335	14,573	11.5%
35-39	119,458	116,777	2,681	2.2%
40-44	116,810	113,110	3,700	3.2%
45-49	122,013	117,417	4,596	3.8%
50-54	110,673	107,817	2,856	2.6%
55-59	91,948	90,531	1,417	1.5%
60-64	75,851	75,371	480	0.6%
65-69	52,087	52,117	−30	−0.1%
70-74	37,627	37,638	−11	0.0%
75-79	29,149	29,075	74	0.3%
80-84	21,404	21,256	148	0.7%
85+	19,914	19,533	381	1.9%
County Total	1,621,474	1,621,478	−4	0.0%

[a]Basic - adjusted
[b]No adjustments for uniformed military population.
[c]Separate projections for uniformed military and civilian populations.

directly available, the analyst may have to become particularly creative. Developing reasonable population estimates requires developing reasonable adjustments for special populations.

We use San Diego County, California to illustrate the impact of a special population (uniformed military) on population estimates by age. San Diego has one of the largest concentrations of military personnel in the United States. In 2010, the uniformed military personnel (89,270) accounted for 2.9% of the county's total population. This population is heavily male (91%) and is concentrated in ages 18-29 (69.7%). Given their numbers and age distribution, the uniformed military population is likely to have a substantial impact on the estimates by age for San Diego County.

We developed two alternative sets of 2010 estimates for San Diego County, using the 2000 census as the base. One set used a basic (i.e., unadjusted) cohort-component model and the other used an adjusted model that separated uniformed military personnel from the civilian population. In the basic model, net migration rates were based on the *total* population. In the adjusted model, net migration rates were based on the *civilian* population and were applied solely to that population. An independent estimate was made for the uniformed military population, assuming no change from the 2000 Census. Other than for the net migration rates, the births, survival rates, and overall net migration total used in the two models were identical.

Table 13.4 shows projections of males from the basic and adjusted models. By design the two projections of total population are identical. However, there are

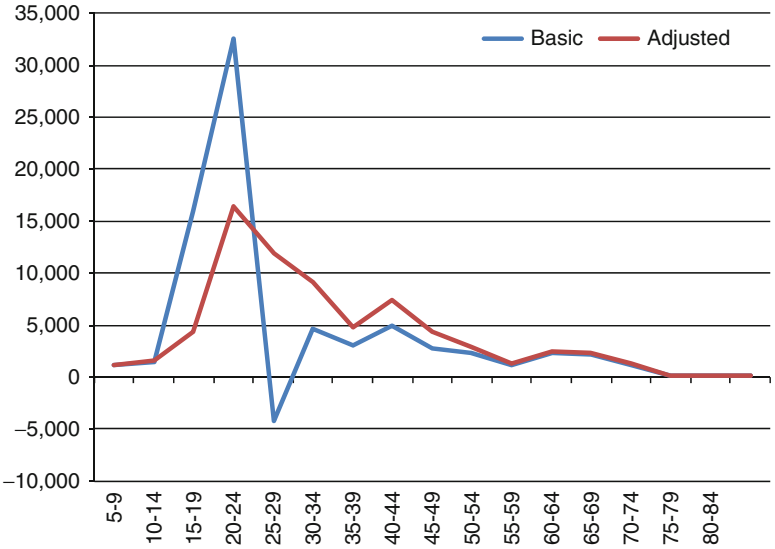


Fig. 13.4 Estimated Male Net Migration by Age, Basic and Adjusted Models, San Diego Country 2000–2010

significant differences in some age groups. For ages 15-19, 20-24, and 25-29, projections from the basic model are lower than those from the adjusted model by 4%, 19%, and 2%. For ages 30-34 to 55-59 the projections from the basic model overstated the population, with the largest divergence in ages 30-34 (12%). Differences between the two models were relatively small for ages under 15 and 60 and above. The large differences in the age composition are due in part because the basic model ages military population at the same rate as the civilian population, while in the adjusted model the military age distribution is unchanged over the post-censal period.

Net migration estimates by age also differ significantly between the two models (see Figure 13.4). The adjusted model shows a smoother and more reasonable pattern of net migration by age, when military migration is removed from the net migration rates. The basic model shows a large, exaggerated net in-migration for ages 15-24 and subsequent net outflow of in ages 25-29. Migration levels for the two models are generally similar after age 35.

13.3 Controlling

Analysts making population estimates often face two distinct but related problems. One is how to make estimates of demographic characteristics (e.g., age, sex, race) match an independent estimate of total population. The second is how to make estimates for a number of geographic areas add up to an independent estimate for a

larger area (e.g., how to make the sum of census tract estimates match a county estimate). Controlling is the term we use to describe this adjustment process.

There are several reasons for controlling estimates. One is the requirement that a set of estimates be consistent with an "official" estimate that has been developed, adopted, or sanctioned by a governmental body or some other decision-making unit. It is also possible that controlling reduces the likelihood of large estimate errors by imposing limits on post-censal changes. This advantage will be especially important for estimates of small areas such as blocks or block groups. Perhaps most important, controlling facilitates the construction of a set of estimates that is consistent across demographic subgroups and geographic areas. That is, estimates of demographic characteristics will sum to estimates of total population and estimates for small geographic areas will sum to estimates for larger geographic areas.

In this section we describe several methods for controlling estimates: 1) single factor; 2) two-factor or plus-minus; and 3) multi- or N-dimensional. We illustrate these methods using 2010 estimates of the population by ethnicity for Major Statistical Areas (MSAs) within San Diego County (SANDAG 2011).[5] In the first illustration, we control the total population estimate by MSA to the total population counted in the 2010 census for San Diego County. In the second illustration, we control the 2010 estimate for each ethnic group in one MSA (Central) to the controlled total population in the same MSA. In the third illustration, we control the estimated population change between 2000 and 2010 for each ethnic group in the Central MSA to the controlled total population change in same MSA. In our final illustration, we control the 2010 estimates by ethnicity in each MSA to the controlled total population in the MSA and to the population by ethnicity for San Diego County counted the 2010 census.

13.3.1 Single Factor Method

The simplest method for controlling to an independent estimate (e.g. total population) is to use a raking procedure based on a single adjustment factor. This factor is computed by dividing the total from the independent estimate by the total from the original estimate. The original estimate for each subgroup or geographic area is then adjusted by multiplying each one by the adjustment factor. The equations for a post-censal time point are:

$$\text{FACTOR} = \text{CNTLE}/\text{E}; \text{ and}$$
$$\text{CE}_{\text{g or c}} = \text{E}_{\text{g or c}} * \text{FACTOR},$$

where E is the original estimate; CNTLE is the independent estimate of the total (i.e., the control total); FACTOR is the adjustment factor; c is the characteristic (e.g. race), and g is the geographic area (e.g. MSA); and CE is the controlled estimate.

Table 13.5 The Single Factor Raking Method: Controlling to the Census Total Population, Major Statistical Areas, San Diego County, 2010

MSA	Original[a]	2010 Estimate Controlled[b]	Difference[c]
Central	660978	634,510	−26,468
East County	23484	22,544	−940
East Suburban	500317	480,282	−20,035
North City	775654	744,594	−31,060
North County East	440472	422,834	−17,638
North County West	437224	419,716	−17,508
South Suburban	386303	370,834	−15,469
County	3,224,432	3,095,314	−129,118
Original County Estimate			3,224,432
2010 Census			3,095,313
Adjustment Factor[d]			0.959956

[a]Estimates from the San Diego Association of Governments (2011)
[b]Original estimate * adjustment factor.
[c]Controlled estimate - original estimate
[d]2010 Census / original County estimate

Table 13.5 shows the original and controlled estimates for the population of MSAs in 2010. The original (uncontrolled) estimate for San Diego County is 3,224,432 and the independent census control total is 3,095,314, yielding an adjustment factor of 3,095,314 / 3,224,432 = 0.959956. This factor represents the proportionate adjustment required to match the control total. In this example, the original population estimate in each MSA is adjusted downward by 4.1 percent. The adjusted (controlled) estimate for each MSA is computed by multiplying the original estimate by 0.959956. The sum of the adjusted MSAs is equal to the 2010 census count of 3,095,314.

In the previous example, estimates of geographic subunits were controlled to an independent estimate of the total population for a large geographic. The same method can be applied to adjust demographic subgroups. For example, estimates of males and females could be based either on a different adjustment factor for each sex or on a single adjustment factor for both sexes. If sex was broken into two racial categories (black and white), estimates for black males, black females, white males, and white females could be based on four separate adjustment factors (one calculated specifically for each race-sex group), on two separate adjustment factors (one for each race), or on a single adjustment factor based on the controlled and uncontrolled projections of the total population. Table 13.6 illustrates the single factor adjustment approach by adjusting estimates by ethnic group to the controlled total population in the Central MSA from Table 13.5.

The choice of the appropriate control group will depend on the availability and reliability of independent estimates for various demographic groups. For demographic groups with similar growth characteristics, it is generally not necessary to develop separate control totals and adjustment factors. The main thing to remember when applying this method is that the sum of the demographic subgroups for which adjustments are made must equal the control total, within rounding error, used in computing the adjustment factor.

Table 13.6 The Single Factor Raking Method: Controlling Ethnicity to the Total Population, Central MSA, San Diego County, 2010

Ethnicty	Original[a]	2010 Estimate Controlled[b]	Difference[c]
Hispanic	270,364	259,537	−10,827
White non Hispanic	203,751	195,592	−8,159
Black non-Hispanic	72,273	69,379	−2,894
Asian & PI non-Hispanic	87,091	83,604	−3,487
Other non-Hispanic[d]	27,499	26,398	−1,101
Central MSA	660,978	634,510	−26,468
Original Central MSA Estimate			660,978
Controlled Central MSA Estimate			634,510
Adjustment Factor[e]			0.959956

[a]Estimates from the San Diego Association of Governments (2011)
[b]Original estimate * adjustment factor
[c]Controlled estimate - original estimate
[d]Includes American Indian and Alaska Native alone, Native Hawaiian and Other Pacific Islander alone and Other Islander alone, and all race combinations
[e]2010 Census / original County estimate

13.3.2 Two Factor (Plus-Minus) Method

The first approach to controlling works well when the adjustments are small to moderate. When adjustments are large, it may produce unsatisfactory results because some demographic subgroups may be adjusted by a larger amount than warranted. In these circumstances, a method that focuses on population change over the estimation period rather than the population in the post-censal year may produce better results.

The basic idea behind the second approach is simple: changes in total population over the estimation period are calculated for both the independent estimate and the original estimate. A ratio of the two projected changes is formed and applied to the change originally projected for each demographic subgroup, producing a set of adjusted changes. These adjusted changes are then added to the census population each demographic subgroup to provide a controlled estimate for the post-censal year.

Although the basic idea behind this approach is simple, its implementation becomes complicated when some subgroups are estimated to increase while others are estimated to decrease. To illustrate this point, suppose that there are only two subgroups. One is estimated to increase by 200 and the other is estimated to decline by 75, implying a total population change of 125. Suppose further that the change for the independent estimate (i.e., the control total) is 100. These numbers produce an adjustment factor of $100/125 = 0.80$ Applying this factor to the changes originally estimated for the two subgroups (200 and –75) produces adjusted changes of 160 and –60. These numbers sum to 100, which is consistent with the change in the independent estimate. However, the adjustment causes the subgroup losing population to lose less than originally estimated. Given that estimated growth for the entire population has been adjusted downward, this does not appear to be a reasonable outcome.

This example points to an important problem with using a simple raking procedure for adjusting estimated population changes; when some demographic subgroups are estimated to increase and others are estimated to decline, a raking procedure based on a single adjustment factor causes both population gains and losses to become larger. This is not a logical outcome. A better outcome would be that adjustments for all demographic subgroups are positive when the overall adjustment is upward and negative when the overall adjustment is downward.

This can be accomplished by using two separate adjustment factors—one for subgroups that are estimated to grow and one for subgroups that are estimated to decrease. This adjustment procedure is known as the plus-minus method (Judson and Popoff 2004: 708-709). The equations for the plus-minus method for a post-censal period are:

$$
\begin{aligned}
\text{CNTLCHG} &= \text{CNTLE} - \text{E}_{\text{cen}} \\
\text{ECHG}_{\text{g or c}} &= \text{E}_{\text{g or c}} - \text{E}_{\text{g or c, cen}} \\
\text{ABSUM} &= |\Sigma\text{ECHG}_{\text{g or c}}| \\
\text{SUM} &= \Sigma\text{ECHG}_{\text{g or c}} \\
\text{POSFACTOR} &= (\text{ABSUM} + (\text{CNTLCHG} - \text{SUM}))/\text{ABSUM} \\
\text{NEGFACTOR} &= (\text{ABSUM} - (\text{CNTLCHG} - \text{SUM}))/\text{ABSUM} \\
\text{If ECHG}_{\text{g or c}} > 0, \text{ then CE}_{\text{g or c}} &= \text{E}_{\text{g or c, cen}} + (\text{ECHG}_{\text{g or c}}{}^{*}\ \text{POSFACTOR}) \\
\text{If ECHG}_{\text{g or c}} < 0, \text{ then CE}_{\text{g or c}} &= \text{E}_{\text{g or c, cen}} + (\text{ECHG}_{\text{g or c}}{}^{*}\ \text{NEGFACTOR}),
\end{aligned}
$$

where CNTLE is the independent estimate of the total (i.e., the control total); E is the estimation variable; CNTLCHG is the change between census point (cen) and post-censal year for the independent estimate; ECHG is the change implied by the original (uncontrolled) estimate for each demographic characteristic (c) or geographic area (g); ABSUM is the sum of the absolute values of uncontrolled changes; SUM is the sum of the uncontrolled changes; POSFACTOR is the adjustment factor for subgroups estimated to increase; NEGFACTOR is the adjustment factor for subgroups estimated to decline; and CE is the controlled estimate of change.

As these equations show, the formulas for the positive and negative adjustment factors are quite similar, differing only by a single sign in the numerator. In fact, if projected changes for all demographic subgroups have the same sign, the plus-minus method produces the same adjustment factor as the single-factor raking procedure. It should also be noted that the sum of the two adjustment factors will always equal two.

Table 13.7 shows the application of the plus-minus method to the ethnic group estimates in the Central MSA. Since the overall adjustment lowers the estimated change, the adjustment factors indicate that population changes for ethnic groups losing population are increased (made smaller) by just over 32% (1.320331), while population changes for ethnic groups gaining population are lowered by the same percentage (0.679669). Comparing the controlled and uncontrolled columns, we see that the adjustment process works as expected. The gains become smaller for ethnic groups estimated to increase and the losses become larger for ethnic groups

Table 13.7 The Plus-Minus Method: Controlling Ethnicity to the Total Population Change, Central MSA, San Diego County, 2010

Ethnic Group	2000 Population	2010 Estimate			2000–2010 Change		Abs. Value
		Original[a]	Controlled[d]	Difference[e]	Original[b]	Controlled[c]	Original Change
Hispanic	223,670	270,364	255,406	−14,958	46,694	31,736	46,694
White non-Hispanic	221,140	203,751	198,182	−5,569	−17,389	−22,958	17,389
Black non-Hispanic	75,275	72,273	71,311	−962	−3,002	−3,964	3,002
Asian & PI non-Hispanic	76,403	87,091	83,667	−3,424	10,688	7,264	10,688
Other non-Hispanic	22,645	27,499	25,944	−1,555	4,854	3,299	4,854
Central MSA	619,133	660,978	634,510	−26,468	41,845	15,377	82,627
Calculation of Plus-Minus Adjustment Factors:							
Sum of Original Changes (SUM)				41,845			
Sum of Abs. Values of Original Changes (ABSUM)				82,627			
2000 Central MSA Population				619,133			
2010 Central MSA Population Estimate				634,510			
2000–2010 Population Change[f] (CNTLCHG)				15,377			
Positive Adjustment Factor[g]				0.679669			
Negative Adjustment Factor[h]				1.320331			
Sum of Adjustment Factors				2			

[a]Estimates from the San Diego Association of Governments (2011)
[b]Original population estimate - 2000 population
[c]Positive original population change * positive adjustment factor or negative original population change * negative adjustment factor
[d]2000 population + controlled population change
[e]Controlled population estimate - original population estimate
[f]2010 Central MSA population estimate - 2000 Central MSA population
[g]Positive adjustment factor = (ABSUM + (CNTLCHG − SUM)) / ABSUM
[h]Negative adjustment factor = (ABSUM - (CNTLCHG - SUM)) / ABSUM

estimated to decline. Of course, if the adjustment increases the overall change, the positive factor will exceed 1.0 and the negative factor will be less than 1.0.

One weakness of the plus-minus method occurs when the difference between the control total (CNTLCHG) and the sum of the uncontrolled projections (SUM) exceeds the sum of the absolute values of the uncontrolled projections (ABSUM). Under this condition, one of the adjustment factors turns negative, reversing the signs of the estimated changes. One solution to this problem is to transform the distribution of estimated changes by adding or subtracting a fixed constant to each value before computing the adjustment factors. The control total also must be modified by the total amount added to or subtracted from the distribution. After the factors are applied, the controlled values are transformed back to the original scale by the amount of the fixed constant. It is relatively easy to find a transformation value that results in positive values for both adjustment factors (e.g., SANDAG 1998).

The examples illustrate how to control estimates of characteristics or geographic areas to an independent control total. In some instances, however, the application may call for controlling to an independent estimate of migration rather than to an independent estimate of total population. This may occur when migration (rather than total population) is the variable of interest or when the focus is on the components of change rather than population per se. The plus-minus method can be used to adjust net migration for subgroups to an independent estimate of net migration. For gross migration models, if there are separate controls for in- and out-migrants the single factor method can be used. If only a net migration control is available, out-migrants are treated as the group with negative changes and in-migrants as the group with positive changes. Examples of controlling migration are found in Smith, Tayman, and Swanson (2001: 253-258).

13.3.3 N-Dimensional Controlling

What if we wanted to make the MSA ethnic group estimates consistent with the census ethnic group counts for San Diego County, but preserve the population totals for each MSA? The methods discussed so far cannot handle this situation. A major problem with single dimensional controlling is making estimates consistent across one dimension makes them inconsistent across another. A procedure is needed that can control across several dimensions simultaneously; this is sometimes called N-dimensional controlling.

N-dimensional controlling is accomplished using the iterative proportions (IP) method, which approximates a least squares solution in order to obtain convergence in all N dimensions (Deming 1943: Chapter 7). This method can handle a wide range of situations. Our illustration covers the situation most commonly encountered when controlling population estimates. There are three main conditions for applying this version of the IP method. First, all estimates must be greater than or equal to zero. Second, there must be totals of each controlling dimension (e.g., ethnic group and total population). If we are controlling MSA estimates to a county

estimate, for example, this condition requires that we have total population controls for each MSA and population controls for each ethnic group for the County. The third condition is that the sum of all totals over all dimensions must be equal; for example, the sum of the ethnic group totals for the county must be equal to the sum of the total population controls for the MSAs.

For purposes of illustration we use two dimensions, one representing a demographic characteristic (ethnic group) and one representing the total population of the seven MSAs in San Diego County. The IP method begins with an initial matrix, whose body contains the population estimates by ethnic group for each MSA. The row controls are the census population by ethnic group for the larger geographic area (the County) and the column controls are the total population for each MSA. These row and column totals are often referred to as marginals. The goal of the IP method is to adjust the matrix so that —when summed horizontally—MSA estimates equal County totals for each ethnic group and—when summed vertically—MSA estimates by ethnic group equal the population control for each MSA.

We achieve this goal by applying a single-factor raking procedure to the original matrix, alternating sequentially between rows and columns. Starting with the rows, we apply a row-specific raking factor to each cell in each row; we repeat this process for all rows. After this step, the sum across the rows matches the County population for each ethnic group. However, the sum down the columns (i.e., all ethnic groups within an MSA) no longer matches the total population control for that MSA. We then apply a column-specific raking factor to each cell in each column. After this step, the sum of the cells in each column matches the total population control of that column, but the sum of the cells in each row no longer matches the County population in that ethnic group. Continuing this sequence of adjustments we eventually arrive at a convergence in which cells in both rows and columns sum to the marginal totals (except for small differences due to rounding).

The rate of convergence is relatively fast, typically requiring less than five cycles of horizontal and vertical adjustments to achieve complete agreement in one dimension and close agreement in the other. It does not matter whether one begins the process by adjusting rows or columns; the results are essentially the same. One can refine the IP method to handle both positive and negative adjustments by using the plus-minus method described earlier to compute two separate adjustment factors to use in the iterative process.

Table 13.8 shows the mechanics of the IP method. The first panel ("Beginning Matrix, First Iteration") shows the initial conditions and the elements needed to apply the IP method. The main body of the matrix is contained in columns 2-8 and the rows for five ethnic groups; the cells of this matrix show the uncontrolled estimates produced by the San Diego Association of Governments (SANDAG 2011). Column 9 shows the "Sum of the MSA" estimates for each ethnic group and Column 10 (Control) shows the row marginals (i.e., the census population by ethnic group for the County). The row labeled "Sum of Ethnic Groups" shows the uncontrolled estimates of total population for each MSA and the row labeled "MSA Control" shows the column marginals (i.e., controlled MSA population from Table 13.5).

Table 13.8 The Iterative Proportions Method: Controlling a Population Projection in Two Dimensions, Major Statistical Areas, San Diego County, 2010

Beginning Matrix, First Iteration

	MSA							County			
Ethnicty	Central	E. County	E. Suburb.	N. City	N.C.E	N.C.W	S. Suburb.	Sum of MSAs[a]	Control	Adjustment[b]	Row Factor[c]
Hispanic	270,364	8,334	109,666	96,201	180,592	106,500	215,621	987,278	991,348	4,070	1.004122
White non-Hispanic	203,751	11,982	319,598	477,346	214,458	269,388	89,872	1,586,395	1,500,047	−86,348	0.945570
Black non-Hispanic	72,273	1,266	25,584	23,140	9,932	18,496	16,620	167,311	146,600	−20,711	0.876213
Asian & PI non-Hispanic	87,091	256	20,526	143,119	19,757	26,593	51,012	348,354	341,562	−6,792	0.980503
Other non-Hispanic	27,499	1,646	24,943	35,848	15,733	16,247	13,178	135,094	115,756	−19,338	0.856855
Sum of Ethnic Groups[d]	660,978	23,484	500,317	775,654	440,472	437,224	386,303			**−129,119**	
MSA Control	634,510	22,544	480,281	744,594	422,834	419,716	370,834				
Adjustment[e]	−26,468	−940	−20,036	−31,060	−17,638	−17,508	−15,469	**−129,119**			
Column Factor[f]	0.959956	0.959973	0.959953	0.959956	0.959957	0.959956	0.959956				

Rows Adjusted, First Iteration

	MSA							County			
Ethnicty	Central	E. County	E. Suburb.	N. City	N.C.E	N.C.W	S. Suburb.	Sum of MSAs[a]	Control	Adjustment[b]	Row Factor[c]
Hispanic	271,478	8,368	110,118	96,598	181,336	106,939	216,510	991,347	991,348	1	1.000001
White non-Hispanic	192,661	11,330	302,202	451,364	202,785	254,725	84,980	1,500,047	1,500,047	0	1.000000
Black non-Hispanic	63,327	1,109	22,417	20,276	8,703	16,206	14,563	146,601	146,60	−1	0.999993
Asian & PI non-Hispanic	85,393	251	20,126	140,329	19,372	26,075	50,017	146,601	146,600	−1	0.999997
Other non-Hispanic	23,563	1,410	21,373	30,717	13,481	13,921	11,292	115,757	115,756	−1	0.999991
Sum of Ethnic Groups[d]	636,422	22,468	476,236	739,284	425,677	417,866	377,362			**−2**	
MSA Control	634,510	22,544	480,281	744,594	422,834	419,716	370,834				
Adjustment[e]	−1,912	76	4,045	5,310	−2,843	1,850	−6,528	**−2**			
Column Factor[f]	0.996996	1.003383	1.008494	1.007183	0.993321	1.004427	0.982701				

Columns Adjusted, First Iteration

	MSA							County			
Ethnicty	Central	E. Country	E. Suburb.	N. City	N.C.E	N.C.W	S. Suburb.	Sum of MSAs[a]	Control	Adjustment[b]	Row Factor[c]
Hispanic	270,662	8,396	111,053	97,292	180,125	107,412	212,765	987,705	991,348	3,643	1.003688
White non-Hispanic	192,082	11,368	304,769	454,606	201,431	255,853	83,510	1,503,619	1,500,047	−3,572	0.997624
Black non-Hispanic	63,137	1,113	22,607	20,422	8,645	16,278	14,311	146,513	146,600	87	1.000594
Asian & PI non-Hispanic	85,136	252	20,297	141,337	19,243	26,190	49,152	341,607	341,56	−45	0.999868
Other non-Hispanic	23,492	1,415	21,555	30,938	13,391	13,983	11,097	115,871	115,756	−115	0.999008
Sum of Ethnic Groups[d]	634,509	22,544	480,281	744,595	422,835	419,716	370,835			**−2**	
MSA Control	634,510	22,544	480,281	744,594	422,834	419,716	370,834				
Adjustment[e]	1	0	0	−1	−1	0	−1	**−2**			
Column Factor[f]	1.000002	1.000000	1.000000	0.999999	0.999998	1.000000	0.999997				

Columns Adjusted, Fourth Iteration

	MSA							County			
Ethnicty	Central	E. Country	E. Suburb.	N. City	N.C.E	N.C.W	S. Suburb.	Sum of MSAs[a]	Control	Adjustment[b]	Row Factor[c]
Hispanic	271,519	8,429	111,599	97,811	180,789	107,921	213,274	991,342	991,348	6	1.000006
White non-Hispanic	191,379	11,335	304,181	453,907	200,797	255,314	83,140	1,500,053	1,500,047	−6	0.999996
Black non-Hispanic	63,130	1,114	22,643	20,463	8,648	16,303	14,298	146,599	146,600	1	1.000007
Asian & PI non-Hispanic	85,038	252	20,309	141,474	19,230	26,202	49,057	341,562	341,562	0	1.000000
Other non-Hispanic	23,444	1,414	21,549	30,938	13,370	13,976	11,066	115,757	115,756	−1	0.999991
Sum of Ethnic Groups[d]	634,510	22,544	480,281	744,593	422,834	419,716	370,835			**0**	
MSA Control	634,510	22,544	480,281	744,594	422,834	419,716	370,834				
Adjustment[e]	0	0	0	1	0	0	−1	**0**			
Column Factor[f]	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.999997				

[a] Sum of the population estimates in each MSA
[b] County control - sum of the population estimates in each MSA
[c] County control / sum of the population estimates each MSA
[d] Sum of the population estimates in each ethnic group
[e] MSA control - sum of the population estimates in each ethnic group
[f] MSA control / sum of the population estimates in each ethnic group

The population control totals are all lower, except for Hispanics, than the sum of the uncontrolled projections. The two numbers in bold print (-129,119) are particularly important. They represent the total amount of the adjustment required in the rows and columns in order to make the projections consistent in both dimensions; they must be equal for the IP method presented here to work properly. The row and column adjustment factors are computed as the ratio of the control total to the sum of the corresponding cells; they are computed separately for each ethnic group and each MSA. All the adjustment factors in the first panel are below 1.0, except for the Hispanics, indicating that downward adjustments are necessary.

The second panel of Table 13.8 ("Rows Adjusted, First Iteration") shows the population estimates by ethnic group for each MSA after we adjusted them to match the row control totals. For example, the adjusted Hispanic population in the N. City MSA is:

$$96{,}201^{*}1.00412 = 96{,}558.$$

After the first set of adjustments all the row adjustment factors are 1.0 (or very close to 1.0), indicating convergence to the population by ethnic group for the County. In addition, the total amount of adjustment now required is close to zero (-2 for both the sum of ethnic groups and the sum of MSAs). However, for individual MSAs the sum of ethnic groups is still inconsistent with the control totals. In fact, four of the seven column adjustment factors have changed from values of less than 1.0 in Panel 1 to values slightly greater than 1.0 in Panel 2.

The third panel ("Columns Adjusted, First Iteration") shows the population projections by ethnic group for each MSA after we adjusted them to match the column control totals. For example, the adjusted Hispanic population in the N. City MSA is now:

$$96{,}558^{*}1.007183 = 97{,}292.$$

All of the column adjustment factors are now 1.0 (or very close to 1.0), indicating convergence to the total population controls for each MSA, but the column adjustments have made the sum of the ethnic group estimates for MSAs inconsistent with the ethnic group controls for the County. However, the differences are much smaller than they were before; which shows that substantial convergence to both marginal totals has occurred after only one full iteration of the process.

The last panel of Table 13.8 ("Columns Adjusted, Fourth Iteration") shows the results after four full iterations. As is evident from this panel, the MSA estimates by ethnic group have now converged (within rounding error) to both the census population by ethnic group and for the County the total population control for each MSA.

The IP method—and other controlling methods—may not always come as close to the independent (control) projections as the examples shown here. Raising the level of demographic detail and reducing the geographic scale can cause multiplicative adjustment routines to lose their efficiency because the computations may not

change the original values as much as is needed to produce complete convergence. For example, integer values less than 5 will not change unless the adjustment is at least 10% (e.g., 5 * 1.09 equals 5 after rounding to the nearest integer). If this occurs with enough frequency, controlling falls short of its intended target.

To handle circumstances where multiplicative adjustments are not adequate, alternative mathematical controlling strategies have been developed (e.g. SANDAG 1998). These involve probabilistic assignment routines and/or iterative schemes that apply small additive adjustments (e.g. $\pm$ 1) to the uncontrolled observations. Private data vendors do some of the most innovative work in this area but—for obvious reasons—are reluctant to reveal their trade secrets.

13.4 Conclusions

The data and techniques used in population estimation methods can usually be described in a fairly simple and straightforward way (although a simple description is a challenge for some of the complex extrapolation and structural models). Applying these methods in the real world of messy data and complicated population dynamics, however, is not always so simple and straightforward. Some data series contain significant errors; others simply do not exist. Some population subgroups react differently to the same events or follow different trends over the same period of time.

In this chapter we described several situations in which a straightforward application of an estimation method might produce misleading results and larger estimation errors. We also described a number of procedures for dealing with these problems. We believe these procedures will help the analyst avoid some common pitfalls and produce better estimates.

Our description of potential problems and solutions is not complete, of course. No description could possibly cover all the circumstances that could potentially affect state or local population change. The analyst must consider many possibilities and be prepared to make adjustments not only for the situations we have described, but for others as well. The development of a good basic model can never replace the need for careful and creative thinking in the construction of population estimates.

Endnotes

1. Adapted from Chapter 11, "Special Adjustments", in S. Smith, J. Tayman, and D. "Swanson. Projecting State and Local Populations: Methodology and Analysis. New York, NY: Kluwer Academic/Plenum Press. 2001.
2. Population estimates can also be impacted by census enumeration errors, particularly those for age and other demographic characteristics. Nationally, census enumeration errors have declined steadily since 1950, except for a small increase between 1980 and 1990 (Robinson, West, and Adlakha 2002; US Census Bureau 2003). Census coverage rates vary considerably from place to place and it is difficult to measure local enumeration errors (Pittenger 1976: 202). To our knowledge, estimates of census enumeration errors are not available for states, counties, or local areas; therefore, they are not usually taken into account when preparing population estimates.

3. Metropolitan and micropolitan statistical areas (metro and micro areas) are geographic entities defined by the US Office of Management and Budget (OMB) for use by federal statistical agencies in collecting, tabulating, and publishing federal statistics. The term "Core Based Statistical Area" (CBSA) is a collective term for both metro and micro areas. A metro area contains a core urban area of 50,000 or more population, and a micro area contains an urban core of at least 10,000 (but less than 50,000) population. Each metro or micro area consists of one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core.
4. The number of emigrants was estimated to be around 200,000 or about 20% of the immigrants in the late 1990s (Martin and Midgley 1999). Separating international and internal migration allows one to make adjustments for the emigration component. Such adjustments are particularly important for subnational areas that have received large numbers of immigrants, because emigration from the US occurs primarily among the foreign born population (Edmonston and Passel 1992).
5. San Diego County is divided into seven MSAs that conform to census tract boundaries. MSAs were constructed so their boundaries remain constant over time, facilitating temporal analysis. They range in size from 23,500 in the sparsely populated eastern half of the County to 775,700 in suburban areas near the center of the County. The average size of the MSAs, excluding smallest, is 533,500.

References

Deming, W. E. (1943). *Statistical adjustment of data*. New York: Dover Publications.

Edmonston, B., & Passel, J. S. (1992). Immigration and immigrant generations in population projections. *International Journal of Forecasting, 8*, 459–476.

Judson, D. H., & Popoff, C. L. (2004). Selected general methods. In J. S. Siegel & D. A. Swanson (Eds.), *The Methods and Materials of Demography, Second Edition* (pp. 677–732). New York: Elsevier Academic Press.

Lowe, T. J. (2001). Understanding Census 2000: Coverage issues and growth trends *Research Brief No. 11*. Olympia, WA: Washington State Office of Financial Management.

Martin, P., & Midgley, E. (1999). Immigration to the United States *Population Bulletin* (Vol. 54). Washington, DC: Population Reference Bureau.

Pittenger, D. B. (1976). *Projecting state and local populations*. Cambridge, MA: Ballinger Publishing Company.

Robinson, J. G., West, K. K., & Adlakha, A. (2002). Coverage of the population census in 2000: Results from demographic analysis. *Population Research and Policy Review, 21*, 19–38.

SANDAG. (1998). *Urban development model, Volume 2: Technical description*. San Diego, CA: San Diego Association of Governments.

SANDAG. (2011). *2010 Demographic Estimates*. San Diego, CA: San Diego Association of Governments.(http://profilewarehouse.sandag.org/).

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

State of California. (2010). California county population estimates and components of change by year, July 1, 2000-2010. Sacramento, CA: Department of Finance.

US Census Bureau. (2003). Technical Assessment of the A.C.E. Revision II. (http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf).

US Census Bureau. (2010). Methodology for the state and county total resident population estimates (Vintage 2009): April 1, 2000 to July 1, 2009. (http://www.census.gov/popest/topics/methodology/2009-st-co-meth.pdf).

Chapter 14
Evaluating Estimates

Population estimates are used for a wide variety of purposes. Businesses use them to develop customer profiles, identify market clusters, and determine optimal site locations. Researchers use them to study development patterns, environmental conditions, and social trends. State and local governments use them to monitor growth trends, the impact of public policies and to estimate the need for schools, roads, parks, public transportation, fire protection, and other goods and services. Producers of estimates use this information to evaluate and improve estimation methodologies. Given these widespread uses of population estimates it is essential to evaluate their error. This chapter provides such an evaluation. We start with a discussion of various statistics that can be used to measure estimate error. We illustrate these measures using 2010 estimates for counties in Washington State and then provide an overview of the empirical evidence, focusing on the effects of differences in estimation methodology, population size, and population growth rate. We conclude the first section of this chapter with a discussion on ways to account for the uncertainty in population estimates.

We have now discussed several approaches to making population estimates that include a variety of models, techniques, special adjustments, and types of data that can be used to produce the desired estimates. Given all the possibilities, how does one go about choosing the specific models, techniques, and data sources to use for a particular set of estimates? Is there a single "best" approach, or at least some that are better than others? Are some approaches better under some circumstances, while others are better under other circumstances? How can we even go about answering these questions? In the final section of this chapter we describe a number of criteria that, in addition to estimation error and uncertainty, can be used to evaluate population estimates. The criteria we believe are most important are the provision of necessary detail, face validity, plausibility, costs of production, timeliness, and ease of application and explanation.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_14,

14.1 Measuring Estimation Error[1]

14.1.1 Defining Estimation Error

We define estimate error (E) as the difference between the population estimate (EST) for a particular geographic area in a particular post-censal year (t) and the actual population (ACT) for the same area and year:

$$E_t = EST_t - ACT_t.$$

For example, if the population of a city had been estimated to be 60,000 in 2010, and the actual population turned out to be 54,000, the estimation error would be 6,000. If the population had been estimated at 48,000, the estimation error would be -6,000. Although not often analyzed, estimation error can also be assessed over the post-censal period by comparing the estimated and observed changes.

Estimation errors are often expressed as percent differences rather than as absolute differences. This specification is useful when measures of relative error rather than absolute error are needed. The use of percent errors is particularly helpful when making comparisons across geographic areas because—without adjustments for population size—errors for places with large populations would dominate the effects of errors for places with small populations. An estimate error of 1,500 has a very different meaning for a place with 2,500 residents than a place with 250,000 residents:

$$\text{Algebraic Percent Error } (ALPE_t) = [(EST_t - ACT_t) / ACT_t]^* 100\,; \text{ and}$$

$$\text{Absolute Percent Error } (APE_t) = | [(EST_t - ACT_t) / ACT_t]^* 100.$$

In the above example, if the population of a city had been estimated to be 48,000 in 2010 and the actual population turned out to be 54,000, the ALPE would be (-6,000 / 54,000) * 100 = -11.1% and the APE would be 11.1%. The ALPE preserves the sign of the percent error; it has a theoretical minimum of -100% and no upper bound, while the APE has a minimum at zero and no upper bound. ALPE and APE represent the individual errors under study, and for a set of geographic areas form the distribution of estimation errors.

Population counts from the decennial census are often used as proxies for the "actual" population of an area. For post-censal or inter-censal years, estimates produced by the Census Bureau, other federal agencies, state and local agencies, or private companies are typically used (e.g., Rynerson and Tayman 1998; Swanson and Tedrow 1984). These proxies are not perfect, of course. Census counts are subject to errors that may be substantial for some places or demographic groups; estimates are subject to even larger errors.[2]

14.1.2 Error Measures

Population estimate error distributions can then be analyzed using a variety of summary measures (e.g., Fonseca and Tayman 1989; Hodges, Wilcox, and Poveromo 2002; Makridakis, Wheelwright and Hyndman 1998: 41-50; Swanson, Tayman and Barr 2000; Tayman 1996). We will describe a number of measures, including the ones most commonly used to evaluate population estimates. The first two measures refer to the average error for a set of n individual estimates:

$$\text{Mean Error (ME)} = \Sigma\, E_t \,/\, n; \text{ and}$$

$$\text{Mean Absolute Error (MAE)} = \Sigma \mid E_t \mid / n.$$

The first measure takes account of the direction of errors; consequently, positive and negative errors offset each other. Measures that account for the direction of the error measure the bias of a set of estimates. In fact, positive and negative values could offset each other completely, resulting in a ME of zero even when individual errors are large. For example, three estimates with errors of 500, 700, and –1,200 would yield a ME of zero. The second measure ignores the direction of the error, so positive and negative errors do not offset each other. Measures that ignore the direction measure the accuracy (or precision) of a set of estimates. The mean absolute error—sometimes called the mean absolute deviation—shows the average difference between estimated and actual populations, regardless of whether the estimates were too high or too low. Using the example cited above, estimates with errors of 500, 700, and –1,200 would yield a MAE of 2,400.

These measures are based on the numerical differences between estimated and actual populations; they do not account for differences in population size. The next two measures account for population size by focusing on percent errors rather than numeric errors:

$$\text{Mean Algebraic Percent Error (MALPE)} = \Sigma\, ALPE_t \,/n; \text{ and}$$

$$\text{Mean Absolute Percent Error (MAPE)} = \Sigma\, APE_t \,/\, n.$$

The MALPE (often called the mean percent error) is a measure of bias in which positive and negative values offset each other. A positive MALPE reflects a tendency for estimates to be too high and a negative MALPE reflects a tendency for estimates to be too low. The MALPE is fairly widely used as a measure of bias (e.g. Rayer 2007; Smith and Sincich 1992; Tayman 1996). The percent of positive errors (%POS) is also used as a measure of bias; a %POS equal to 50% would suggest no bias, values greater than 50% would suggest a positive basis, and values less than 50% would suggest a negative bias.

The MAPE, on the other hand, is a measure of accuracy in which positive and negative values do not offset each other. It shows the average percent difference between estimated and actual populations, regardless of whether the individual estimates were too high or too low. The MAPE is the most commonly used measure of estimate accuracy (Swanson, Tayman and Barr 2000).

Sometimes it is important to use error measures that give more weight to large errors than to small errors; for example, when a large error has a disproportionately large impact on the cost of being wrong. In these situations, the following measures can be used:

$$\text{Mean Squared Error (MSE)} = \Sigma\,(E_t)^2 / n;\ \text{and}$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{[\Sigma\,(E_t)^2 / n]}.$$

Although these two measures are commonly used in general analyses of error (e.g., Armstrong & Collopy 1992; Mahmoud 1987), they are less useful for evaluations of population estimation errors because results for areas with large populations swamp the results for areas with small populations. This problem can be dealt with by using percent errors rather than absolute errors as in the Root Mean Squared Percent Error (RMSPE) (e.g., Keilman 1990; Smith and Sincich 1992; Swanson and Tayman 1995):

$$\text{Root Mean Squared Percent Error} = \sqrt{[\Sigma\,(PE_t)^2 / n]}.$$

14.1.2.1 Robust Measures

Average measures of error have several desirable properties including reliability; ease of use and interpretation; incorporating all of the information; and uniqueness for a set of observations, but they have a major drawback. Arithmetic means are affected by extreme values and in the presence of outliers likely either understate or overstate the error represented by most of the observations in the error distribution, depending on the skewness of the distribution. Averages based are the absolute percent error distribution (APE) are particularly susceptible to outliers, because extreme values typically occur only at the high end of the distribution, and the APE is prone to asymmetry in practice (Emerson and Strenio 1983; Swanson and Tayman 1999). The error distribution of the APEs is often right-skewed because it is bounded on the left by zero and unbounded on the right. Therefore, the MAPE is susceptible to being pulled upward and to overstating the error represented by most of the observations. Swanson, Tayman, and Bryan (2011) have estimated that for every 1% increase in skewness the upward bias of the MAPE increases by approximately 0.7%.

Because of the shortcomings of the average in characterizing an asymmetrical distribution, alternative measures of central tendency are often presented in conjunction with the average. These alternatives are generally referred to as robust or

resistant statistics because they focus on the main body of the data and attempt to minimize the impact of outlying observations (Hampel, Ronchetti, Rousseeuw, and Stahel 1986: 1-18). The median is one such robust statistic, as are the symmetrical MAPE (SMAPE), trimmed mean, and M-estimators (e.g., Goodall 1983; Hodges, Wilcox, and Poveromo 2002; Makridakis and Hibon 1995; Rosenberger and Gasko 1983; Swanson and Tayman 1999). A drawback of the median is it ignores most of the information contained in the error distribution; it is only based on one or two observations. The trimmed mean ignores the most extreme observations in both tails of the error distribution based on a user-defined percentage, typically between five and 10 percent. Swanson and Tayman (1999) found that the SMAPE fell short as an alternative to the MAPE, but that M-estimators provided a valid assessment of accuracy. However, M-estimators lack the intuitive and interpretative qualities of the MAPE and are unfamiliar to many users and practitioners.

Another alternative applies a non-linear transformation to the error distribution of the APEs and then computes an average based on the transformed distribution (Tayman, Swanson, and Barr 1999; Swanson, Tayman, and Bryan 2011). The objective is to find a transformation that creates an error distribution less dominated by the large outlying errors, incorporates all of the information about individual errors, and does not overstate the error represented by most of the observations.

The geometric mean, based on the natural logarithm transformation, might be useful when data are right-skewed:

$$\text{GMAPE} = e^{[(1/n)/\Sigma \ln(x))]}.$$

GMAPE is easy to calculate and interpret, but it has a major drawback; the logarithmic transformation may not be the most appropriate. To address the question of what transformation is most appropriate, Tayman, Swanson, and Barr (1999) introduced the MAPE-R (MAPE-Rescaled). To change the shape of a distribution efficiently and objectively, they use a standardized technique for generating a single, nonlinear function to change the shape of the APE distribution. This technique is based on the power transformation developed by Box and Cox (1964):

$$y(\lambda) = (X^{\lambda} - \lambda)/\lambda \text{ where } \lambda \neq 0; \text{ or}$$

$$y(\lambda) = \ln(x) \text{ when } \lambda = 0,$$

where x is the original APE, y is the transformed APE, and λ is the power transformation constant. One determines λ by finding its value that maximizes the function:

$$ml(\lambda) = -(n/2) \times \ln[(1/n)\sum(y_i - \overline{y})^2] + (\lambda - 1) \times \sum \ln(x_i),$$

where, n is the sample size; y is the transformed observation; $\bar{y}$ is the mean of the transformed observations; x is the original observation (APE). ml(λ) at a local maximum provides the λ that optimizes the probability that the transformed APE

distribution will be symmetrical. Finding λ does not guarantee symmetry, but it represents the λ most likely to yield a symmetrical distribution. The maximum value of ml(λ) is obtained by solving its function for different values of λ between the range of –2 and 2 and identifying the largest resulting Box-Cox value (Draper and Smith 1981: 225-226).

The transformed APE distribution considers all errors, but assigns a proportionate amount of influence to each case through normalization and not elimination; thereby, reducing the otherwise disproportionate effect of outliers on a summary measure of error. The mean of the transformed APE distribution (APE-T) is known as MAPE-T. The APE-T distribution has a disadvantage; the Box-Cox transformation moves the observations into a unit of measurement that is difficult to interpret (Emerson and Stoto 1983: 124). MAPE-T is expressed back into the original scale of the observations by taking its inverse (Swanson and Coleman 2007). The re-expression of MAPE-T is known as MAPE-R:

$$\text{MAPE-R} = [(\lambda)(\text{MAPE} - \text{T} + 1)]^{1/\lambda}.$$

14.1.2.2 Loss Functions

The loss function can be used to quantify low average relative errors and to detect outlying errors (Bryan 1999). A total loss function represents a weighted combination of the numeric error and percent error for a given area or observation (i):

$$L_i = | E_i^{\alpha} * PE_i^{1-\alpha} |,$$

where E is the numeric error, PE is the percent error, and α is the weight, which ranges from 0 to 1. The average of the L_i's provides a summary measure of accuracy, while outliers can be detected from the L_i values themselves. Simplifying the equation algebraically yields:

$$L_i = | EST_i - ACT_i | / ACT_i^{1-\alpha},$$

where EST is the estimate and ACT is the observed value. Now the weight is only applied to the observed value. The relative impacts of the numeric and percent errors on the loss function are determined by the weight; the smaller the weight (α) the greater the influence of the percent error, whose weight is 1-α. The weight can be determined subjectively, but Bryan (1999) has tested and proposed the following function for α based on the range of the observed values:

$$\alpha = \text{In}(ACT_{max} - ACT_{min})/25.$$

The loss function is not well-known to users and practitioners and lacks an intuitive or easy way to judge and the magnitude of error. Insofar as the loss function is a relative measure, it lacks generality. For example, loss functions could not be compared for estimates for different time periods, geographic levels, or variables; it can

only be interpreted in the context of a specific application. The determination of the loss function weight lacks a firm empirical basis and defined set of guidelines. For these reasons, loss functions have been rarely used to evaluate population estimates

14.1.2.3 Allocation Error

The summary measures discussed above are based on the error for a particular geographic area. Another perspective views the misallocation of the estimates across geographic space. This aspect of error is most pertinent for estimation procedures that nest activities from a larger geographic area to smaller geographic areas and use the larger area as a control. Population estimates are often used to distribute resources in a zero-sum fashion making allocation error an important component in measuring the performance of estimates. The Index of Misallocation (IOM), also known as the Index of Dissimilarity (see Chapter 4), measures the extent that the estimates misallocate activities over a set of geographic areas (i) (Duncan, Cuzzort, and Duncan 1961: 83-90; Fonseca and Tayman 1989). Swanson (1981) proposed using the term "Index of Misallocation" when it was used specifically to evaluate this aspect of error for population estimates. Its computation is the same as the Index of Disssimilarity:

$$\text{IOM} = 0.5 * \left[\sum \mid (\text{EST}_i/\text{EST}) - (\text{ACT}_i/\text{ACT}) \mid * 100\right].$$

The IOM compares the percent distributions of an activity across geographic space and measures the percentage that one distribution (i.e. based on the estimates) would have to change to match the other (e.g. based on the census). The IOM ranges from 0 to 100; 0 means no spatial disparity, and 100 means complete disparity between the census and estimates across a set of geographic areas.

14.1.2.4 Relative Error

Estimation errors as measured above are often the main standard for judging the adequacy or quality of a given set of estimates. Estimate error is an important, but it not the only criterion upon which an estimate should be judged. Estimates can also be judged according to their overall "utility," or their value-added in improving the quality of information upon which decisions are based.

To measure the utility, or potential gain in information from an estimate, measures have been developed that compare the errors from a formal model to a no or low-cost naïve model, such as basing the estimate on the last census (e.g., Davis 1994; Harper, Coleman, and Devine 2003). Theil's U-statistic is used extensively to evaluate economic time series models (Theil 1966). Theil's U ranges from 0.0 to 1.0, where the upper bound indicates no improvement over the naive model. The proportionate reduction in error (PRE) also shows the extent to which an estimate

from a formal model (Method a) can improve on an estimate based on a naïve model (Method b) (Swanson and Tayman 1995):

$$\text{PRE} = ((\text{Method b} - \text{Method a})/\text{Method b})^{*}100.$$

In using PRE, one develops a population estimate for the same area using two different methods (a) and (b). The error arising from each methods is defined and measured and the proportionate reduction in error found by the preceding formula. What constitutes Method (a) and Method (b) relative to our discussion of utility? Method (a) results from an estimation technique such as the censal-ratio method, CMII, ratio-correlation and the like, while Method (b) is the estimate resulting from data already at hand through an existing 'count', such as the last census. The estimate resulting from Method (b) is called a 'naive' estimate in that it represents the theoretically (and most often, the practical) maximum error for an estimate because it based on no new knowledge. PRE determines the reduction of error found by using the estimate from Method (a) over the error in the 'naive' estimate from Method (b). A PRE of zero indicates the formal method does not improve on the naïve method. Values between 0 and 100 represent the percentage gain of information from the formal method, while negative vales indicate the formal method performs worse than the naïve method.

14.1.2.5 Error of the Change

All of the error measures discussed above focus on differences in population levels in the post-censal year. This is the approach most commonly used to evaluate population estimation accuracy. An alternative approach focuses on differences between estimated and actual annual growth rates rather than differences between estimated and actual population sizes. Keyfitz (1981), Long (1995), and Stoto (1983) used this approach for evaluating national population projections and Tayman (1996) used it for evaluating census tract projections, but this approach can be applied to estimates as well. Seeing how well the estimated change matches or predicts the observed change is a more rigorous test. A post-censal year estimate includes both activities in the launch year and estimated change, which confounds the measurement of the error caused solely by the estimation method.

Tayman (1996) suggests two approaches for examining the error in estimating the percentage change. A parametric approach uses a regression model that predicts the observed percentage change using the estimated percentage change as the independent variable. A different, and perhaps more useful perspective, is to take a more general look at this relationship using non-parametric cross-tabulation techniques. The percentage change variables are defined by broad categories rather than by their original interval form. These redefined percentage change variables are cross-classified, which facilitates analysis of the conditional relationships within different growth rate categories. Of particular interest are the percentages shown in the diagonal of the table, which represent the success of estimated percentage change category in predicting the same observed percentage change category. If the estimated percentage change category is a perfect predictor, each diagonal cell would contain 100%.

14.1.2.6 Selection Criteria

Given the many different statistics that can be used to measure estimate accuracy and bias, how can one go about choosing the most appropriate measure(s)? A number of researchers have discussed criteria that might be used to select measures of forecast error (e.g., Ahlburg 1995; Armstrong and Collopy 1992; Makridakis 1993), but these criteria are also applicable to measures of estimation error. Several criteria are mentioned frequently. Error measures should be reliable; that is, repeated applications should yield similar results. They should be valid, in the sense that they actually measure what they purport to measure. They should convey as much information about estimate errors as possible and should be easy for the data user to understand. They should be sensitive to differences in error distributions, but should not be unduly influenced by outliers.

The Committee on National Statistics (1980:10-12) defined four accuracy criteria that ideally should be met by post-censal estimates: 1) low average error; 2) low average relative error (disregarding direction of the error; 3) few extreme relative errors; and 4) absence of bias for subgroups. It is generally not possible to produce a set of estimates that will minimize the four criteria simultaneously; the Committee chose to focus on low average relative error and few extreme relative errors, with some attention to low average error or bias.

The MALPE provides a useful way to investigate the tendency for projections to be too high or too low. While not as susceptible to outliers as the MAPE, analysis of estimation bias should also include a robust measure of central tendency along with the % positive measure. While the median is often used because of its familiarity, other robust measures can achieve the aim of the median while preserving most, if not all, of the observations. For example, there are four M-estimators to choose from: 1) Huber, 2) Hampel, 3) Andrews' Wave and 4) Tukey bi-weight. Each estimator varies in their resistance to the impact of outlying observations and relies on a different weighting algorithm. Although the four M-estimators generally yield similar results, we prefer the Tukey bi-weight because it is most resistant over a wider range of distributions and it is the most popular (Goodall 1983; Swanson and Tayman 1999).

The MAPE provides a reasonable measure for evaluating accuracy under a wide variety of circumstances, but it often will overstate the typical error in a set of estimates. Like the MALPE, the MAPE should be accompanied by a robust summary measure of central tendency. We believe that a measure based on a non-linear transformation of the underlying APE distribution is the best robust measure to use. Further, we prefer MAPE-R to GMAPE because of the flexibility in determining the optimal transformation parameter. As Swanson and Coleman (2007) show, GMAPE will always be less than or equal to MAPE-R and thus GMAPE has a greater potential for overstating the accuracy of most of the observations. A comparison of GMAPE and MAPE-R would provide a useful test of the efficacy of alternative non-linear transformations for creating an average measure of accuracy that is not influenced by outlying observations.

Can valid conclusions be drawn when only a few error measures are analyzed? We believe they can in most instances. Although different error measures provide different perspectives on estimation accuracy and bias, error patterns appear to be quite stable and highly correlated across a variety of error measures and the impact

of factors such as population size and growth rate on error is generally about the same regardless of which error measure is used (Davis 1994; Rayer 2007). Because of these similarities, it is not necessary to analyze a wide variety of error measures to achieve valid conclusions.

14.2 Evaluating Post-censal Population Estimates

14.2.1 Error at the Post-censal Time Point

To illustrate the various measures of error just discussed, we present an evaluation of the 2010 population estimates for counties in Washington State. These estimates, were developed from the ratio-correlation model described in Chapter 8, that was calibrated over the 1990-2000 decade and used registered voters, registered automobiles, and school enrollment in grades 1 through 8 as symptomatic indicators. Table 14.1 shows the 2010 census, 2010 estimate, numeric error, and percent error by county. The state estimate used as the control for the ratio-correlation model is quite accurate; it is around 14,000 persons or 0.2% higher than the 2010 census. Douglas and Spokane Counties, highlighted in grey, have virtually identical numeric errors, but the percent error for Spokane (0.9%) is almost 12 times lower than the percent error for Douglas (10.6%). This comparison highlights the point made earlier about the advantage of using percent errors rather than numeric errors when evaluating estimates for areas with varying population sizes. A quick scan shows the outlying errors for Douglas and Steven Counties and to a lesser extent for Wahkiakum County.

Table 14.2 presents summary measures of bias and accuracy, allocation error IOM, and utility (PRE). These estimates show a slight upward bias. The mean error indicates that on average these estimates are high by 358 persons, but this number is difficult to interpret because it lacks the context of the population size. The MALPE, which accounts for population size, indicates an average error of 0.7%. The distribution of the numeric errors and ALPEs is positively skewed being influenced by the three outlying errors, all of which are overestimates. Both the median error (27) and median ALPE (0.0%) are noticeably lower than their average counterparts. There is variation among the other robust measures of bias (based on the percent errors), but they all suggest a slight upward bias to these estimates;

These county estimates have a high level of accuracy. The mean absolute error and MAPE indicate that on average these estimates are within 2,921 persons and 3.2% of the census count. The RMSPE is a percentage point higher than the MAPE, because it gives more weight to the outlying observations. The right skewness in the distribution of the absolute errors and APEs is evident; both the median absolute error (1,365) and Median APE (2.5%) are noticeably lower than their counterparts. There is less variation in the other robust measures of accuracy compared to robust measures of bias, especially removing the trimmed mean and GMAPE. The MAPE-R (2.4%) is in line with the median APE and all M-estimators. The ratio of the MAPE to MAPE-R suggests that the MAPE overstates the absolute percent error representative of most of the observations by 33%.

Table 14.1 Population Estimation Error, Washington State Counties, 2010[a]

	2010		Error	
	Estimate	Census	Number	Percent
Adams	19,550	18,728	822	4.39
Asotin	21,817	21,623	194	0.90
Benton	173,607	175,177	−1,570	−0.90
Chelan	72,480	72,453	27	0.04
Clallam	70,102	71,404	−1,302	−1.82
Clark	435,496	425,363	10,133	2.38
Columbia	4,291	4,078	213	5.22
Cowlitz	99,368	102,410	−3,042	−2.97
Douglas	42,493	38,431	4,062	10.57
Ferry	7,969	7,551	418	5.54
Franklin	73,403	78,163	−4,760	−6.09
Garfield	2,252	2,266	−14	−0.62
Grant	90,485	89,120	1,365	1.53
Gig Harbor	70,516	72,797	−2,281	−3.13
Island	77,546	78,506	−960	−1.22
Jefferson	28,011	29,872	−1,861	−6.23
King	1,921,450	1,931,249	−9,799	−0.51
Kitsap	245,011	251,133	−6,122	−2.44
Kittitas	39,708	40,915	−1,207	−2.95
Klickitat	20,545	20,318	227	1.12
Lewis	74,803	75,455	−652	−0.86
Lincoln	10,668	10,570	98	0.93
Mason	57,962	60,699	−2,737	−4.51
Okanogan	42,180	41,120	1,060	2.58
Pacific	21,780	20,920	860	4.11
Pend Oreille	12,490	13,001	−511	−3.93
Pierce	815,218	795,225	19,993	2.51
San Juan	16,184	15,769	415	2.63
Skagit	116,255	116,901	−646	−0.55
Skamania	11,230	11,066	164	1.48
Snohomish	708,639	713,335	−4,696	−0.66
Spokane	475,286	471,221	4,065	0.86
Stevens	48,519	43,531	4,988	11.46
Thurston	250,590	252,264	−1,674	−0.66
Wahkaikum	4,306	3,978	328	8.25
Walla Walla	61,677	58,781	2,896	4.93
Whatcom	197,493	201,140	−3,647	−1.81
Whitman	42,271	44,776	−2,505	−5.59
Yakima	254,850	243,231	11,619	4.78
Washington State	**6,738,501**	**6,724,540**	**13,961**	**0.20**

[a] Estimates based on the ratio-correlation model from Chapter 8

Table 14.2 Summary Measures of Estimation Error, Washington State Counties, 2010

Bias		Accuracy	
Mean Error	358.0	Mean Absolute Error	2,921.4
Median Error	27.0	Median Absolute Error	1,365.0
MALPE	0.7	MAPE	3.2
% Positive	51.3	RMSPE	4.2
Median ALPE	0.0	Median APE	2.5
Trimmed Mean (5%)	0.6	MAPE-R	2.4
M-Estimators		GMAPE	2.1
Huber's	0.4	Trimmed Mean (5%)	2.9
Tukey's	0.1	M-Estimators	
Hampel's	0.5	Huber's	2.7
Andrew's Wave	0.1	Tukey's	2.5
		Hampel's	2.7
		Andrew's Wave	2.5
Index of Misallocation	0.88		
Proportionate Reduction in Error (PRE)			
	Naïve 1[a]	Naïve 2[b]	
MALPE	94.1	69.4	
MAPE	75.0	50.4	

[a]Use 2000 census
[b]Adjust 2000 census using the proprotionate change in the state population

Table 14.3 Characteristics of the APE and APE-T Distributions, Washington State Counties, 2010

	APE	APE-T
Average	3.2	3.4
Median	2.5	3.5
Standard Deviation	2.7	1.2
Coeff. of Variation	85.9	36.3
Skewness	1.38[a]	−0.03[b]
Minimum	0.04	0.32
Maximum	11.5	5.97
% Errors > 5%	20.5	7.7
% Errors >10%	7.7	0.0

[a]Reject hypothesis of a symmetrical distribution
[b]Accept hypothesis of a symmetrical distribution

These estimates have very little allocation error as evidenced by an IOM of 0.9% and they also have a great deal of utility compared to two naïve estimates of the 2010 population based on the 2000 census (see Table 14.2). The greatest gain in information or knowledge is seen the bias where the ratio-correlation (formal) model lowers the average bias (MALPE) by 94.1% and 69.4%. The formal model also substantially lowers the average error (MAPE) by 75% and 50.4%.

Table 14.3 demonstrates the effect of the Box-Cox transformation and shows selected characteristics of the original APE distribution and the transformed APE distribution (APE-T). First, the APE-T distribution is symmetrical. Its average and median are almost identical, the skewness coefficient is close to zero; and a

statistical test (D'Agostino, Belanger, and D'Agostino 1990) further confirms a symmetrical distribution. Second, the transformation reduces the variation of the error distribution as indicated by the standard deviation and coefficient of variation. The Box-Cox transformation not only compresses very large values, but also increases values greater than one in skewed distributions where λ was relatively small (less than 0.4) (Swanson, Tayman and Barr 2000); the λ for this dataset is 0.294. Third, the percentage of relatively large errors is substantially lower in the distribution of APE-Ts; there are no errors over 10 percent and there are two-thirds fewer errors over 5 percent compared to the distribution of the APEs.

14.2.2 Error of the Change

How do these estimates fare when examining the error of the population change between 2000 and 2010? As noted previously, the State estimate is 14,000 or 0.2% higher than the 2010 census, but the estimated change overstates the observed change (830,419) by 1.7%. So while the method does a very good job at capturing state population change, the differences in the percent errors show that evaluating the change is a more rigorous standard of comparison. Some other examples further illustrate this point. The largest county in the state (King) has an ALPE (-0.5%), but the estimate understates the 2000-10 change by (-5.0%). Gig harbor has an average error (ignoring the sign) of 3.1%, but the estimate misses the observed change by 41%.

How well does the estimated percent change predict the observed percent change (see Figure 14.1)? The estimated percent change is a good predictor of the observed percent change. The regression equation has a moderately strong explained variance of 78% and the slope indicates that a one unit increase in the estimated change results in a 0.9 unit increase in the observed change. Most notably, the estimates do well in picking up extreme changes. For example, Garfield County declined by 6% between the censuses, which is the same percentage decline indicated in the estimates. On the other end of the growth continuum, Franklin, the fastest growing County between the censuses (58.3%), is also the fastest growing county in the estimates (48.8%).

Table 14.4 looks at the relationship between the observed and estimated percent changes from a different (nonparametric) perspective. Along with the cross-tabulation, the table contains several nonparametric statistics that measure the strength of relationship (Siegel 1956: Chapter 9).[3] The estimates are similarly successful in predicting the observed growth rate in all categories, except the most stable counties (5.0% to 9.9%); the observed growth rate category is correctly identified in roughly 60% of the counties. For the most stable counties, the success rate drops to 46% and this category exhibits the most homogeneity in the percentages for adjacent categories. In general, the ratio-correlation model estimates the growth rate pattern fairly well, especially in areas with the greatest change.

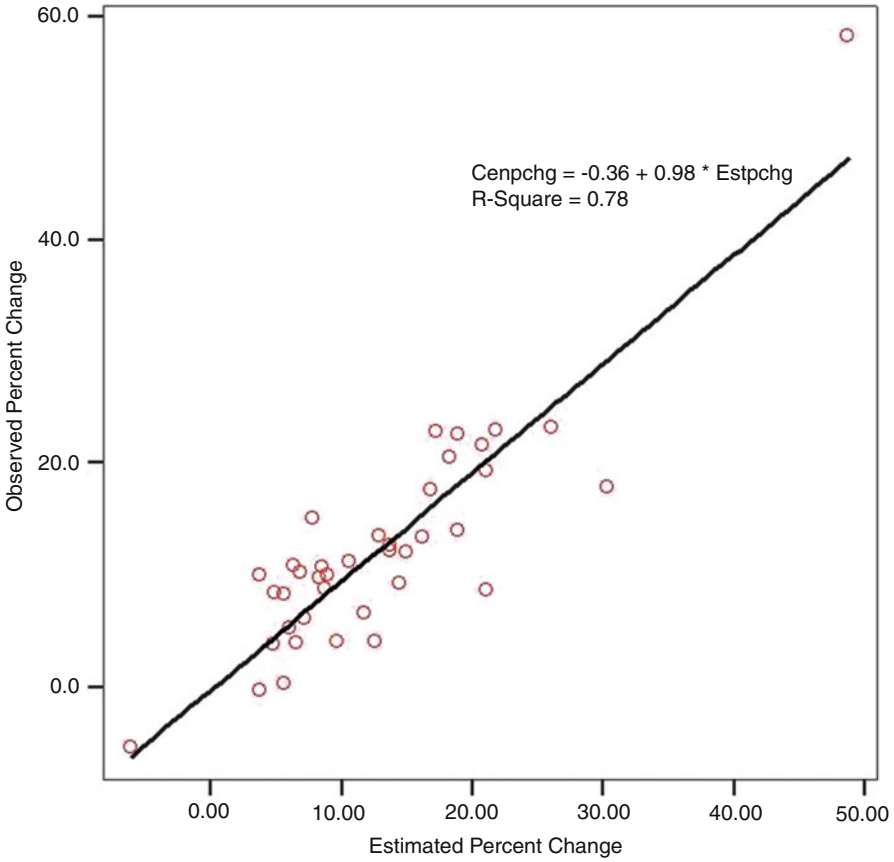


Fig. 14.1 The Relationship between the Observed and Estimated Percent Change, Washington State Counties, 2000-2010

Table 14.4 Cross Classification, Observed and Estimated Change, Washington State Counties, 2000–2010

	Estimate Percent Change (Column Percents)			
Observed Percent Change	<5.0%	5.0% – 9.9%	10.0% – 19.9%	20+%
<5.0%	**60%**	23%	7%	0%
5.0% – 9.9%	40%	**46%**	14%	14%
10.0% – 19.9%	0%	31%	**57%**	29%
20+%	0%	0%	21%	**57%**
	100%	100%	100%	100%
Ordinal Measures of Association				
Somers'd	0.590			
Kendals tau-b	0.596			
Kendals tau-c	0.572			
Gamma	0.776			

14.3 Factors Affecting Estimation Error

To set the stage for the discussion of the factors affecting estimation error, we identify the level of accuracy that can be expected from a set of population estimates (see Table 14.5).[4] We base these expectations on evaluations of 1980 and 1990 estimates of all counties and states conducted by the Census Bureau (Davis 1994 and Long (1993); evaluations of 2000 estimates for all counties, states, census tracts, and block groups conducted by a private data vendor (Hodges, Wilcox, and Poveromo 2002), and an unpublished evaluation of 2010 estimates for all states, counties, census tracts, and block groups conducted by the authors. We also examined a myriad of evaluation studies, most of them pertaining to counties in specific states, to validate the ranges shown in the table.

The expected MAPEs for a 10-year post-censal period range from 1% to 3% for states; from 3% to 6% for counties; from 10% to 13% for census tracts, and from 14% to 19% for block groups. One would expect that the accuracy of population estimates would be greater than the accuracy of population forecasts for a comparable time period, because estimates have access to symptomatic data that track population change. Based on the typical MAPEs for population forecasts for a 10-year horizon (Smith, Tayman, and Swanson 2001: 340), the MAPEs for population estimates are roughly one-half the size of those for population forecasts for states, counties, and census tracts. Estimation errors for any specific set of estimates will be affected by factors such population size, population growth rate, and quality of symptomatic indicators, of course, but we believe these numbers provide reasonable approximation of the accuracy of estimates for these levels of geography.

14.3.1 Estimation Method

Estimation methods differ in terms of data requirements, mathematical structure, degree of disaggregation, number of variables included, and modeling/statistical skills required. Is there a difference is the performance between estimation methods? A common perception among users of population projections is that complex methods are more accurate than simpler methods, but the vast preponderance of the evidence suggests otherwise (e.g., Smith and Tayman 2003; Smith, Tayman, and Swanson 2001: 307-316). We believe a similar conclusion generally applies for population estimation methods (e.g., Hoque 2010; Poole, Tarver, White, and Gurley 1966; Murdock, Kelley, Jordon, Pecotte and Luedke 2006: 52-56).

Table 14.5 Typical "MAPEs" for Population Estimates by Geographic Area[a]

States	1%	to	3%
Counties	3%	to	6%
Census Tracts	10%	to	13%
Block Groups	14%	to	18%

[a]10-year post-censal period

With the possible exception of extrapolation techniques, no single method that incorporates symptomatic indicators consistently produces more accurate or less biased estimates than any other method.[5] Errors that occur in population estimates are heavily influenced by the quality and availability of symptomatic indicators and the care and rigor taken in implementation, rather than by the mathematics or particular combination of variables contained in the model. For example, it has been shown that in ratio-correlation models error in the symptomatic indicators have the greatest impact on estimation error, while the contribution to error of coefficient instability is minimal (Tayman and Schafer 1985).

There are relatively few comparative evaluations of alternative population estimation methods, especially compared to rich literature for population and other projections. Evaluations of population estimates typically focus on a single method, alternative specifications of a single method, or the individual components of a method. Below we summarize a few studies that have compared multiple methods.

Hoque (2010) evaluated 1990 estimates for counties and places in Texas for four methods: component method II (CMII), ratio-correlation (RC), housing unit (HU), and average of all methods. For counties, RC estimates were the most accurate with a MAPE of 4.8%. CMII and the average had a similar level of accuracy (6.5% and 6.2%), and HU performed markedly poorer (10.3%). In terms of bias in the county estimates, HU and the average had the least bias with MALPEs of 0.6%. The level of bias was double and in the opposite direction for RC and CM II (-1.3%). For places, HU (15.9%) out-performed CM II and the average, whose MAPEs were 19.8% and 18.4%. A similar ranking between the methods was seen for bias; the MALPE ranged from 7.8% for HU to 9.1% for CM II.

Smith and Mandell (1984) evaluated 1980 estimates for counties in Florida for six methods: HU; CM II; RC; Administrative Records (AD); and 2 averages (HU CM II and RC), and (HU, CM II, RC, and AD). The AD method was the same as the CM II, but used tax returns to estimate migration instead of school enrollment. The degree of precision of all methods, except CM II, was similar; the MAPE for CMII was 7.7% compared to 5.2% to 5.7% for the other methods. The lowest MAPE was seen in the average of the four methods. All methods showed a marked downward bias, but the tendency to under-estimate was lowest for the HU method, with a MALPE of -2.9%; the MALPEs for the other methods ranged from -3.9% for RC to -7.1% for CM II. The two average methods had the 4th and 5th highest levels of bias, with MALPEs of -5.1% and -4.6%.

Zitter and Shyrock (1964) evaluated 1950 and 1960 estimates for states for ten methods: CM I; CM II; RC; Composite (COM); Vital Rates (VR); linear trend; geometric trend; and 3 averages (CM II and VR), (CMII and RC), and (COM and RC). For the 1960 estimates, all of the symptomatic methods, except CM 1, were more accurate than the two extrapolation methods. The average percent deviation for the four symptomatic methods had a narrow range from 2.0% for CM II to 2.8% for RC. The three averages performed the best; their average percent deviations ranged from 1.5% to 1.8%. The picture was similar for 1950, but with some differences. First, the evaluation did not include the RC method. Second, the variation of the error was wider for the three symptomatic methods; it ranged from 2.5% for COM to 4.4% for VR. Finally, CM II and COM outperformed

Table 14.6 Estimation Errors by Component, Florida Counties and Subcounty Areas, 2010

					% of absolute errors	
	Component	MAPE	MALPE	%POS	<5%	>10%
Counties	Households	2.6	−1.0	40.3	95.5	0.0
	PPH	2.0	1.1	71.6	97.0	0.0
	GQ	18.4	13.6	80.6	20.9	52.2
Subcounty Areas	Households	7.8	1.6	50.2	75.9	12.4
	PPH	4.0	0.4	57.1	75.4	6.2
	GQ	110.1	86.4	42.3	42.0	52.4

Source: Smith and Cody (2011)

the average method, with average percent deviations of 3.1% and 2.5% compared to 3.5% for the average.

14.3.2 Components of the Housing Unit Method

The housing method is widely used to prepare population estimates. It relies on post-censal estimates of housing stock and households, persons per household, and group quarters population (GQs). Of all post-censal estimation methods, housing unit method components have been the most thoroughly studied, due primarily to the work of Smith and his colleagues at the University of Florida (Smith and Cody 1994, 2004, 2011). We present their most recent evaluation of 2010 estimates for counties and places in Florida (Smith and Cody 2011).

Which component of the HU method can be estimated most accurately? Table 14.6 shows that for counties PPH estimates are somewhat more accurate than household estimates, both considerably smaller than the error for GQs, with MAPEs of 2.0%, 2.5%, and 18.4%. There was a slight tendency for PPH estimates to be too high and household estimates to be too low, but the GQs estimates have a substantial upward bias. The PPH estimates are still the most accurate for subcounty areas, but now the MAPE for households (7.8%) is almost double the MAPE for PPH (4.0%). The large MAPE for GQs occurs because in many places the percent error is based on very small numbers of people. There was a slight tendency for both PPH and household estimates to be too high in subcounty areas. A number of studies have also found errors for households to be greater than errors for PPH (Lowe, Myers, and Weisser 1984; Smith and Cody 1994, 2004; Starsinic and Zitter 1968).

For both counties and subcounty areas, percentage errors for GQs are much larger than percentage errors for households and PPH. Does this mean that GQs error contribute the most to overall estimation error? Synthetic population estimates based on a combination of estimated values and census values show that GQs error contribute the least to population estimation error, because the GQs population generally accounts for a very small proportion of total population. Even with perfect estimates of PPH and GQs, errors in household estimates lead to the largest population estimation errors, averaging 2.4% for counties and 7.6% for

Table 14.7 Population estimation error by population size, Florida Subcounty Areas, 2010

					% of absolute errors	
Size (2000)	N	MAPE	MALPE	%POS	<5%	>10%
<250	23	37.4	1.5	47.8	8.7	69.6
250– 499	23	13.0	3.2	47.8	26.1	56.5
500–999	46	11.3	1.8	47.8	37.0	39.1
1,000–2,499	62	13.5	2.9	61.3	24.2	53.2
2,500–4,999	49	9.8	4.1	59.2	38.8	30.6
5,000–9,999	60	7.6	2.3	43.3	45.0	25.0
10,000 - 14,999	49	5.5	2.2	69.4	55.1	14.3
15,000 - 24,999	34	4.6	1.4	55.9	70.6	2.9
25,000 -49,999	50	4.2	0.6	56.0	70.0	6.0
50,000–99,999	31	2.6	0.5	54.8	83.9	0.0
100,000–199,999	24	3.6	1.2	54.2	75.0	4.2
200,000 +	17	3.3	−0.3	58.8	82.4	0.0

Source: Smith and Cody (2011)

subcounty areas (ignoring the direction of error). With perfect estimates of households and GQs, errors in PPH estimates would have created population estimation errors averaging 1.9% for counties and 3.9% for subcounty areas (ignoring the direction of errors).

14.3.3 Population Size

Many studies covering a variety of estimation methods and geographic areas have found estimate accuracy to improve as population size increases (e.g., Hoque 2010; Long 1993; Smith and Mandell 1984; Smith and Cody 2011). This relationship, however, tends to weaken (or disappear completely) once a certain population size has been reached. The threshold level at which the relationship begins to weaken varies with the size of the geographic unit under consideration. For example, the relationship begins to weaken at a smaller population size for estimates of counties than for estimates of states. It appears that not only does population size matter, but so does the relationship between population size and the size of the geographic area.

A clear relationship between estimation errors and population size is generally found only for measures of accuracy (e.g., MAPE), not for measures of bias (e.g., MALPE). A number of studies have found no consistent relationship between population size and bias (e.g., Davis 1994; Harper, Coleman and Devine 2003; Smith and Cody 2004). Although the evidence is not completely clear-cut, it appears that population size has no predictable impact on the tendency for population estimates to be too high or too low. Similar results regarding population size and accuracy and bias have been found in evaluations of population forecasts (Smith, Tayman, and Swanson 2001: 316-317).

Table 14.7 illustrates the relationship between population size and estimate errors. It shows errors for 2010 estimates for subcounty areas in Florida (Smith and Cody 2011). The direct relationship between accuracy and population size is clearly

Table 14.8 Population estimation error by population growth rate, Florida Subcounty Areas, 2010

					% of absolute errors	
	N	MAPE	MALPE	%POS	<5%	>10%
Growth Rate (2000–2010)						
<−10%	40	29.0	28.8	97.8	2.5	92.5
−10.0%–0%	97	8.5	8.0	90.7	41.2	27.8
0.0%–4.9%	53	5.9	1.1	56.6	64.2	11.3
5.0%–9.9%	48	5.3	1.2	52.1	64.6	8.3
10.0%–14.9%	49	4.1	−0.5	44.9	65.3	6.1
15.0%–24.9%	57	4.2	−2.2	33.3	73.7	10.5
25.0%–49.9%	81	7.3	−4.2	35.8	45.7	23.5
50.0%–99.9%	27	8.3	−5.2	18.5	40.7	29.6
100 + %	16	30.9	−30.0	6.2	12.5	75.0

Source: Smith and Cody (2011)

evident when looking at the MAPEs, which range from 37.4% for areas with less than 250 persons to 3.3% for areas over 200,000. The non-linear, "plateauing" pattern of this relationship is also seen; the MAPE stabilizes at around 50,000 persons. In general, as population size increases the percent of large errors over 10% moves in the opposite direction. The inconsistent relationship between population size and bias is also evident in the pattern of the MALPE, which supports the conclusion that estimation bias is largely unaffected by differences in population size.

14.3.4 Population Growth Rate

Population growth rates have a strong impact on estimate errors. Estimate accuracy is generally greatest for places with small but positive growth rates and decreases as growth rates deviate in either a positive or negative direction from those low levels. That is, there is a U-shaped relationship between estimate accuracy and population growth rates (e.g. Hoque 2010; Smith and Cody 2004; Long 1993). The largest errors are typically found for places that are either growing rapidly or declining rapidly. Bias is also strongly affected by differences in population growth rates (e.g., Hodges and Healy 1984; Smith and Cody 2004). Estimates tend to be too high for places that are losing population and too low for places that are growing rapidly. Similar results regarding population growth rate and accuracy have been found in evaluations of population forecasts (Smith, Tayman, and Swanson 2001: 317-320), but, in terms of bias, forecasts tend to be too low for places losing population and too high for places growing rapidly.

Table 14.8 illustrates the relationship between population growth rate and estimate errors using 2010 population estimates for subcounty areas in Florida (Smith and Cody 2011). The U-shaped relationship between population growth rate and estimation accuracy is evident. By far, the largest MAPEs are in areas that declined by more than 10% or increased by 100% or more. MAPEs are relatively

stable for growth rates between 5% and 25%. The expected pattern of estimation bias and population growth rate is also seen. The highest level of upward bias occurs where areas have declined by more than 10% and the highest level of downward bias occurs where areas have increased by 100% or more. The positive bias continues at decreasing levels until the growth reaches 10% and then turns negative and then increases continually with increases in growth rates (without regard of sign). That is, there is an inverse relationship between population growth rate and estimate bias.

14.4 Accounting For Uncertainty

The almost universal way of obtaining information on population estimation error is through the use of retrospective or post-hoc analysis using the census as the standard of comparison. While these analyses provide a wealth of valuable information about the historical performance of population estimation methods and techniques, they furnish almost no information on the error inherent in the post-censal estimates itself. The main issue is these tests use information not directly relevant for the period for which the estimate is made (Ericksen 1973). The fact that a ratio-correlation estimate had an average error of 2.3 percent in estimating the 2010 population tells little about its performance in estimating the 2015 population. Post-hoc analyses can suggest which technique might or should produce the most accurate estimate and provide some idea of the range of error inherent in the estimate. However, neither of these statements can be made with any certainty or precision because they are inferences based on a time period other than the post-censal period in question (Espenshade and Tayman 1982).

Post-censal estimates are not free of error, but only a single value for the number being estimated is usually presented. This deterministic approach gives the impression that the estimate corresponds to the "true" value of the number in question. Nevertheless, the practice of indicating the direction and magnitude of error in post-censal population statistics is virtually, if not, completely absent in statistical offices (United Nations 1971: 6). Not much has changed in the 40 years since the U.N. report. It is true that now the ACS provides 90% confidence limits for its estimates, but the population and housing unit estimates developed by the Census Bureau used as controls in the ACS are deterministic and treated as if they had no error.

It is useful to provide a direct measure of error and the distributional properties of population estimates. When information is presented in this manner, one can quickly see how much the estimate can be trusted (Keyfitz 1972). Users of post-censal estimates are entitled to the warning constituted by a wide distribution. A single point estimate tends to project a false sense of accuracy into the figures because it implies an unreal deterministic interpretation. As we turn to this discussion, keep in mind that a complementary discussion of this aspect of estimation evaluation is provided in Chapter 8.

14.4.1 Confidence Intervals

Efforts to account for uncertainty in post-censal estimates have focused on developing intervals around a point estimate. These confidence intervals are accompanied by an explicit indication of the probability that a given range will contain the estimated population. Confidence intervals—strictly speaking—apply only to sample data and not to other applications, such as the prediction intervals associated with regression analysis. Prediction intervals focus on random values within the population distribution and not on any specific parameter, as is the often case with confidence intervals. For ease of exposition, we use the term confidence intervals.

14.4.1.1 Sample Surveys

Confidence intervals for post-censal estimates can be based on surveys of households and populations using proper sampling techniques and statistical theory (e.g., Cochran 1977; Hedayat and Sinha 1991; Kish 1965). As described in Chapter 4, confidence intervals are constructed around a point estimate by adding and subtracting the margin of error. The margin of error represents the standard error multiplied by a factor indicating the desired probability or confidence level. The size of standard error is a function of sample size, sampling design, and sample variance. The ACS is the best example of confidence intervals for post-censal estimates based on sample surveys. Swanson, Roe and Carlson (1992) and Swanson, Carlson, Roe, and Williams (1995) used sample surveys to put confidence intervals around the total population and age group estimates in three small rural Nevada communities (Armargosa Valley, Beatty, and Pahrump). Confidence intervals based on sample surveys are expensive and not practical for most post-censal estimation applications

14.4.1.2 Regression Methods

Regression models offer a much more practical alternative for developing confidence intervals around post-censal population estimates, especially the widely used ratio-correlation model. Confidence intervals could easily be integrated into this framework at little or no cost and without requiring additional data collection. The formulae for generating a confidence interval around the post-censal estimate P for a sample size (n) and (k) independent variables are given by Kmenta (1971: 363-364):

$$\text{Upper Limit} = P + \left(t_{n-k,\alpha/2}{}^{*} S_f\right);$$
$$\text{Lower Limit} = P - \left(t_{n-k,\alpha/2}{}^{*} S_f\right),$$

where (t) is the standard score representing the desired confidence level and S_f is the standard error of the prediction.

S_f consists of two parts. One part is that the sample regression will not be the same as the population regression due to sampling error. The second part is due to random error in that the actual value of the point estimate will not lie on the population regression line. As discussed in detail in Chapter 8, a ratio-correlation model is typically based on all counties and not a sample of counties, so the universe is represented by a super population; that is, the observed counties are a representative sample of all possible observations that could have occurred (Swanson and Beck 1994; D'Allesandro and Tayman 1980). Confidence intervals from a statistical model are valid only to the extent that the assumptions underlying that model are valid. The various assumptions and procedures for evaluating the assumptions underlying regression model are found in most texts on regression modeling (e.g., Kmenta 1971; Stock and Watson 2003; Belsley, Kuh, and Welsch 1980). In addition confidence intervals from a statistical model may be influenced by errors in the base data, errors in specifying the model, and errors in estimating the model's parameters.

Swanson and Beck (1994) developed a ratio-correlation model for making short-term (2-year) county population projections for Washington State. They compared the 66% confidence intervals associated with this model to census counts of Washington's 39 counties in 1970, 1980, and 1990. They found the prediction intervals to contain the 1970 census count in 30 counties (77%), the 1980 census count in 24 counties (62%), and the 1990 census count in 31 counties (79%). These results suggest that these 66% prediction intervals provided a reasonably accurate view of short-term forecast uncertainty.

Espenshade and Tayman (1982) developed a time series regression model for making post-censal estimates of the population by age in Florida. The model was based on the concept of an age specific death rate reformulated to express the post-censal estimate in terms of the death rate, number of deaths, and latest census population. Regression models were used to estimate the post-censal death rate and its confidence intervals, which were then translated into confidence intervals for the total population and population by age. Population estimates produced for the State of Florida fell within the 95% intervals in 15 of the 18 age groups, suggesting the intervals provide a reasonable view of post-censal estimation uncertainty. One limitation of this method is that it may be unsatisfactory for substate areas because of a potential small numbers problem in the calculation of age-specific death rates.

14.4.1.3 Other Approaches

Swanson (2008), based on earlier work by him and several colleagues, described a procedure for generating confidence intervals for short-term forecasts based on the cohort-component model. The formal measure of uncertainty takes the form of a "mean square error confidence interval" (MSECI), which is designed to the capture

the uncertainty due to random error in age-specific mortality rates and systematic error (net undercount error in census counts underlying both the base populations). The method was tested by doing an ex post facto comparison of the confidence intervals generated for age-sex projections for Nye County, Nevada against the 2000 census population.

The evaluation revealed that the MSECI appear to be too narrow. In terms of a 66% MSECI, the census count was contained in only 2 of the 17 age groups for males, no age group for females, and one age group for the total population. For the total population, only the total population for males was contained with the 66% MSECI. The census counts for more age groups were contained within the 95% MSECI, but there was still far less than would be expected if the intervals provided a reasonable view of uncertainty. Nye County, Nevada is a very small county with volatile rates of change, and Swanson (2008) suggests that the method may perform better in large, less volatile populations.

A substantial amount of research over the last several decades has dealt with the measurement and evaluation of uncertainty in population forecasts. Much of this research has focused on the development and application of univariate ARIMA time series models (e.g., Alho and Spencer 1997; Pflaumer 1992; Tayman, Smith and Lin 2007). To our knowledge, ARIMA models have not been used to place confidence intervals around post-censal estimates.

ARIMA models, discussed in Chapter 6 along with other complex extrapolation methods, assume that the pattern (structure) of the data does not change over time, that errors are normally distributed with a mean of zero and a constant variance, and that errors are independent of each other (Makridakis, Hibon, Lusk, & Belhadjali 1987). The two main advantages of univariate time series models are: 1) they require only historical data on the population of the area being estimated; and 2) their underlying mathematical and statistical properties provide a basis for developing confidence intervals to accompany the point estimates (Box and Jenkins 1976: Chapter 5; Brockwell and Davis 2002: Chapter 6). Time series models require a fairly long series of historical observations, can be difficult to apply, and require a high level of statistical modeling expertise.

14.4.2 Illustrative Confidence Intervals

In this section, we present two examples of confidence intervals; one based on an ARIMA model and the other based on a ratio-correlation model. Our first example shows 95% confidence intervals for annual population estimates from 2001 to 2010 for Walla Walla County, Washington. These intervals are based on the ARIMA (0,1,0) model developed in Chapter 6. The second example shows 66% confidence intervals around 2010 population estimates for Washington State counties. These intervals are based on the ratio-correlation model developed in Chapter 8.

Table 14.9 contains the point estimates and 95% confidence intervals for Walla Walla County using 20-year (1980-2000) and 40-year (1960-2000) base periods.

Table 14.9 Population Estimates: 95% Confidence Intervals, Walla Walla County, 2001–2010[a]

	20-Year Base Period (1980–2000)			
	Lower Limit	Point Estimate	Upper Limit	Half Width[b]
2001	54,052	55,252	57,082	2.7
2002	53,761	55,640	58,148	3.9
2003	53,595	56,027	59,088	4.9
2004	53,490	56,414	59,968	5.7
2005	53,420	56,801	60,813	6.5
2006	53,374	57,189	61,633	7.2
2007	53,346	57,576	62,436	7.9
2008	53,330	57,963	63,226	8.5
2009	53,324	58,350	64,006	9.2
2010	53,326	58,738	64,779	9.7
	40-Year Base Period (1960–2000)			
	Lower Limit	Point Estimate	Upper Limit	Half Width[b]
2001	54,173	55,252	56,837	2.4
2002	53,923	55,576	57,736	3.4
2003	53,791	55,901	58,516	4.2
2004	53,719	56,226	59,238	4.9
2005	53,683	56,550	59,923	5.5
2006	53,672	56,875	60,583	6.1
2007	53,679	57,199	61,225	6.6
2008	53,701	57,524	61,853	7.1
2009	53,733	57,849	62,470	7.6
2010	53,775	58,173	63,078	8.0

[a]Based on an ARIMA (0, 1, 0) model
[b] (((Upper Limit - Lower Limit) / 2) / Point Estimate) * 100

The half-width expresses the size of the confidence interval in percentage terms. For example, a half-width of 2.7 means there is a 95% chance the post-censal estimate will be within plus or minus 2.7% of 2001 point estimate. As expected, the half-width increases with the length of post-censal period, reflecting the increased uncertainty as one moves further away from the last census. These intervals suggest that in 2010 the probability is 95% that the population of Walla Walla County will range from 53,300 to 64,800. The 2010 census count of 58,781 falls inside of this interval. The longer base period results in slightly narrower intervals; increases in sample size, all things equal, will reduce the variance in post-censal estimates.

Each ARIMA model provides different confidence intervals (e.g., Cohen 1986; Keilman, Pham, and Hetland 2002; Tayman, Smith, and Lin 2007). To illustrate, we ran a second ARIMA (1,1,0) model adding a first order autoregressive term. In 2010, the point estimate from the two models are close (58,740 and 58,980), but the width of the interval is wider under the model with the autoregressive term (13,300 vs. 11,400); a difference of 17%.

Table 14.10 contains 2010 point estimates, 66% confidence intervals, and the 2010 census counts for Washington State counties. The state totals represent the bottom up sum of the counties. The half-widths (not shown) have a fairly narrow

Table 14.10 Population Estimates: 66% Confidence Intervals, Washington State Counties, 2010[a]

				2010	Outside Interval	
	Lower Limit	Point Estimate	Upper Limit	Census	Lower	Upper
Adams	19,223	20,006	20,790	18,728	X	
Asotin	21,379	22,326	23,273	21,623		
Benton	171,103	177,658	184,214	175,177		
Chelan	71,078	74,172	77,265	72,453		
Clallam	68,856	71,738	74,620	71,404		
Clark	429,504	445,660	461,816	425,363	X	
Columbia	4,203	4,391	4,579	4,078	X	
Cowlitz	97,492	101,687	105,883	102,410		
Douglas	41,645	43,484	45,324	38,431	X	
Ferry	7,832	8,155	8,479	7,551	X	
Franklin	72,086	75,116	78,146	78,163		X
Garfield	2,191	2,304	2,418	2,266		
Grant	89,121	92,596	96,071	89,120	X	
Gig Harbor	69,129	72,162	75,194	72,797		
Island	76,105	79,356	82,607	78,506		
Jefferson	27,450	28,665	29,879	29,872		
King	1,886,466	1,966,293	2,046,121	1,931,249		
Kitsap	240,308	250,729	261,151	251,133		
Kittitas	39,129	40,634	42,140	40,915		
Klickitat	20,154	21,024	21,894	20,318		
Lewis	73,452	76,549	79,645	75,455		
Lincoln	10,443	10,917	11,391	10,570		
Mason	57,086	59,315	61,543	60,699		
Okanogan	41,373	43,164	44,955	41,120	X	
Pacific	21,341	22,288	23,236	20,920	X	
Pend Oreille	12,253	12,781	13,309	13,001		
Pierce	802,867	834,243	865,619	795,225	X	
San Juan	15,932	16,562	17,191	15,769	X	
Skagit	114,358	118,968	123,578	116,901		
Skamania	11,053	11,492	11,932	11,066		
Snohomish	698,084	725,177	752,271	713,335		
Spokane	467,506	486,378	505,250	471,221		
Stevens	47,729	49,651	51,573	43,531	X	
Thurston	247,038	256,439	265,839	252,264		
Wahkaikum	4,235	4,407	4,579	3,978	X	
Walla Walla	60,606	63,116	65,626	58,781	X	
Whatcom	194,600	202,102	209,603	201,140		
Whitman	41,263	43,258	45,252	44,776		
Yakima	250,615	260,797	270,980	243,231	X	
Washington State	**6,626,288**	**6,895,760**	**7,165,236**	**6,724,540**		
				% Outside		**38%**

[a]Based on the ratio-correlation model from Chapter 8.

range across counties, ranging from 3.6% to 4.9%. If the confidence intervals portray a reasonably accurate view of uncertainty, the percent of counties where the 2010 census is not contained in its interval should be close to the compliment of the confidence level. For the county intervals, the 2010 census is not contained within the intervals of 38% of the counties, close to the expected 34%. These confidence intervals appear to provide a realistic view of the uncertainty in these 2010 population estimates generated from the ration correlation model.

14.5 Other Evaluation Criteria[6]

Forecast error and uncertainty are important criterion for evaluating population estimates and their models. In this section we describe a number of other criteria relevant to such an evaluation. We discuss criteria we believe are most important: provision of necessary detail, face validity, plausibility, costs of production, timeliness, and ease of application and explanation. After describing these criteria, we consider how they must be balanced against each other when making choices regarding the appropriate methodology to use in any given situation.

14.5.1 Provision of Necessary Detail

Perhaps the most fundamental criterion for evaluating estimates is whether they provide the level of geographic and demographic detail required by the data user. State estimates are of little use to someone needing county estimates. Estimates of total population are of little use to someone needing estimates by age and sex. Estimates for counties are of little use to someone needing estimates for school districts. Many data users need population estimates for states and counties. These needs can be met relatively easily because the geographic boundaries for states and counties generally remain stable over time and many types of data are routinely available at the state and county levels. In addition, the number of states and counties is finite and relatively manageable; there are more than 3,100 counties or county equivalents nationwide, with the largest numbers in Texas (254) and Georgia (159). Most states have fewer than 100 counties or county equivalents.

For subcounty areas, however, the number of potential areas—and even the ways in which those areas might be defined—is virtually endless. Possibilities include cities, townships, census tracts, block groups, blocks, parcels, school districts, traffic analysis zones, and many types of market or service areas. Estimates that meet the needs for geographic detail for all (or almost all) data users would have to be made at the block group or lower levels of geography. Those estimates could then be aggregated to fit the geographic region required by each individual data user. Such a process, of course, would be extremely expensive and fraught with problems of data availability and reliability.

The need for demographic detail also varies from user to user. Some require only total population numbers while others require breakdowns by age, sex, race, and/or ethnicity. Some need age data in single-year age groups; for others, five- or 10-year age groups are sufficient. Some require estimates of specific population subgroups such as college students, military personnel, seasonal residents, and persons with disabilities. Others require estimates by income, education, occupation, poverty status, or other socioeconomic and demographic characteristics. Again, the potential for variation in user needs is virtually limitless.

The needs of the largest number of potential data users can be met (at least theoretically) by making estimates that are highly disaggregated by geographic area and demographic characteristics. Armed with these building blocks, data users can put together estimates that cover the specific geographic areas and demographic characteristics they need. Greater degrees of disaggregation, however, require greater data requirements, have lower data reliability, have higher the costs of production, and have higher the expected degree of error for each detailed category. These are strong incentives against the production of highly disaggregated estimates. As a result, most producers of general-purpose estimates provide estimates that cover only a limited number of geographic areas and demographic categories.

The most basic criterion for judging the potential usefulness of a set of population estimates, then, is whether those estimates provide the level of geographic and demographic detail needed for any particular purpose. If the estimates cannot at least come close to meeting those requirements, they will not be very useful regardless of how well they do with respect to the other evaluation criteria. General-purpose estimates will be able to meet the needs of many data users for many purposes, but some applications will require estimates created specifically for the purposes at hand.

14.5.2 Face Validity and Plausibility

By face validity, we mean the extent to which an estimate uses the best methods for a particular purpose, is based on reliable data, and accounts for relevant factors. Because of the effects of population and geographic size, evaluating face validity is considerably more complex and time-consuming for small areas (e.g., census tracts) than for large areas (e.g., states). The face validity of a method depends primarily on the purposes for which the estimates will be used. All the methods discussed in this book can be used for estimates of total population. For estimates by age group, the analyst must account for shifts in age structure over time; this implies the use of some variant of the cohort-component method. For estimates of the components of growth, the model must distinguish among the effects of fertility, mortality, and migration. Estimates incorporating interactions between economic and demographic variables require the use of structural models. The face validity of a particular model or technique cannot be generalized; rather, it is conditional upon the specific purposes for which the estimates will be used.

Face validity is also affected by the quality of the data used to create the estimates. Although they are not perfect, data from the short form of the decennial census are generally quite accurate, especially for larger geographic regions. Sample data from the ACS are less accurate (especially for small areas), and are subject to greater variations in their reliability. Vital statistics data are highly accurate for states and counties, less accurate for subcounty areas, but vital events data are generally more accurate than school enrollment and tax returns and data for tracking changes in housing units, households, persons per households, and occupancy rates.

The timeliness of input data may also affect face validity. Demographic data vary considerably in terms of time lags and frequency of release. Birth and death data are typically available close to the post-censal time point, but many symptomatic indicators lag one or more years behind. Face validity is also determined by the extent to which the estimation methodology accounts for the impact of relevant factors affecting population change, such as large special populations and international migration. An important part of assessing the face validity of population estimates, then, is evaluating the quality and timeliness of the input data and making adjustments when necessary to correct for apparent errors and to account for relevant factors.

Plausibility is closely related to face validity; in fact, the two may be thought of as opposite sides of the same coin. Face validity focuses on the inputs into the estimation process, whereas plausibility focuses on the outcomes. If an estimate is not based on valid data and techniques, it is not likely to provide plausible or reasonable results. Plausibility, of course, is a subjective concept. Just as beauty is in the eye of the beholder, so too is plausibility. A trend that appears eminently plausible to one observer may seem totally implausible to another.

Are the estimates consistent with current trends and expectations about change? If not, what are the reasons for these differences? Have some special circumstances been overlooked? Were there errors entering data or writing computer programs? Answering questions like these provides one type of "plausibility check. Plausibility checks require a substantial investment of time and effort, especially for highly disaggregate estimates, but have a potentially large pay-off. Given their subjective nature, however, plausibility checks must be viewed as suggestive rather than conclusive. They provide hints and clues, but cannot "prove" that one set of estimates is better than another.

14.5.3 Costs of Production And Timeliness

The costs of production for a set of population estimates are determined primarily by labor costs. A great deal of time must be spent considering all the relevant details involved in producing a set of estimates; collecting, verifying, and cleaning up the input data; putting together an estimation model; and evaluating the plausibility of the results. Other costs (e.g., computer hardware and software, purchases of proprietary data) are typically small in comparison.

Very little research has focused on the costs of producing population projections. Just how high are those costs and how do they vary by method, level of geographic and demographic detail, and frequency of application? Logic and personal experience suggest that costs increase with the degree of methodological complexity, with the level of geographic and demographic detail provided, and with the attention paid to special populations and unique events. However, costs can be expected to decline with the number of times a specific application is repeated; it takes more time to produce a set of estimates for the first time than to repeat the process additional times. Other things being equal, lower costs are preferable to higher costs. Other things, however, are rarely equal. Trade-offs must be made between costs of production and other attributes of population estimates. Assessing the costs of production—and their relationship to other estimation attributes—is central to the evaluation process

Timeliness is the amount of time needed to construct the estimates. This is determined by the scope of the project and by the resources devoted to it. Production time takes on particular importance when a set of customized estimates is created for a specific client. The client (who may be someone within the same organization as the analyst) may require that the estimates be completed within a short (perhaps unreasonably short) period. In some circumstances, production time is a major factor determining the choice of estimation methods. In terms of costs and timeliness, for example, ratio-correlation estimates for a given set of counties are likely to cost less and be produced more quickly than estimates based on the housing unit method. Extrapolation methods are generally the least costly and most timely method, while structural models are generally the most costly and least timely method.

14.5.4 Ease of Application And Explanation

Ease of application is determined by the amount of time and the level of expertise needed to collect, verify, and adjust the input data; develop an estimation model; and generate the desired population estimates. This criterion will be particularly important for analysts with limited training or expertise in the production of population estimates or who face severe time or budget constraints. At the present time, no widely available estimation software package can be implemented quickly and easily. Instead, the analyst generally must develop a set of computer algorithms specifically for the project at hand. We expand on this point in Chapter 18.

Ease of explanation refers to the extent to which data users can be provided with a clear description of the data sources and techniques used in producing the estimates. For some data users, this criterion is irrelevant. They are interested only in the estimates themselves, not in how they were produced. Other data users, however, can truly evaluate (and properly use) a set of estimates only if

they understand how those estimates were made. Indeed, some may have little or no use for estimates based on unknown methods or "black box" models. For those data users, the clearer and more complete the description of the methodology the more valuable the estimates (Rainford & Masser 1987). For example, the housing unit and component methods are easier to explain to users and require less sophisticated statistical and modeling skills then ratio-correlation, structural models, and complex trend extrapolation methods.

14.6 A Balancing Act

All the criteria discussed above are potentially important for choosing the data and techniques that will be used in constructing a set of population estimates or for evaluating a set produced by someone else. The relative importance of each criterion, however, varies according to the purposes for which the projections will be used.

The provision of necessary detail is essential for all purposes. If data for the relevant geographic areas and demographic categories are not available, the estimates clearly will not be very useful. Face validity, plausibility, and timeliness would also seem to be of almost universal importance; exceptions might be when estimates are used simply to illustrate the outcomes of various hypothetical scenarios or to push a particular point of view. Ease of application and costs of production generally do not matter to the data user, but are important to the producer. In fact, these criteria may drive methodological decision making when time is limited or budgets are tight. Ease of explanation is unimportant for some data users, critical for others. Estimation accuracy may be the most important criterion when estimates are used to guide decision making, but is irrelevant when estimates are used for simulations or to push a particular agenda.

Choosing the relevant criteria for evaluating a set of estimates is clearly a balancing act. Some criteria may be much more important than others and decisions based on one criterion may be inconsistent with decisions based on another. Choices must be made regarding which criteria are most important for a particular set of estimates and—when they conflict with each other—which to rank ahead of the other. An optimal estimation strategy can be chosen only after weighing the relative importance of each of the evaluation criteria.

14.7 Conclusions

The first part of this chapter dealt with estimation error. To close this chapter it may be helpful to summarize the empirical evidence regarding accuracy and bias for population estimates. Estimate accuracy generally increases with

population size, but tends to level off once a certain size threshold is reached. Accuracy tends to be greatest for places with slow but positive growth rates and tends to decline as growth rates deviate in either direction from those levels. Estimation bias is unrelated to population size, but estimates tend to be too high for places that are losing population and too low for places that are growing rapidly. For estimates of total population no single model or technique is consistently more accurate than any other. These results have been found in so many circumstances that we believe they can be accepted as general characteristics of population estimation errors.

The last part of the chapter dealt with additional criteria that can be used to evaluate population estimates and their methods. Evaluating population estimates is a two-step process. The first step is to choose the criteria upon which the estimates will be evaluated. Potential criteria include the provision of necessary detail, face validity, plausibility, costs of production, timeliness, ease of application and explanation, and estimate error, utility, and uncertainty. The choice of criteria will depend on the purposes for which the estimates will be used and the constraints imposed on the analyst producing the estimates. For any given purpose some criteria may be very important, some may be moderately important, and a few may be completely unimportant.

The second step is to use these criteria to guide the selection of estimation methods. Simple extrapolation methods are characterized by timeliness, ease of application and explanation, low costs of production, and applicability to very small areas; however, they cannot provide much demographic detail and do not take into account post-censal symptomatic indicators. Complex extrapolation methods share many of these attributes, but typically require more data and modeling expertise and are harder to apply and explain. Component methods are much more costly and less timely, but are capable of providing a rich array of demographic detail. Housing unit methods are also costly and less timely, but provide additional information on housing and household characteristics and are applicable to small geographic areas. Ratio-correlation models are relatively low cost, but provide limited demographic detail and rank low on ease of explanation and application. Structural models are the most data-intensive, time-consuming, and costly, but are capable of providing a variety of inter-related estimates and offer the greatest analytical usefulness.

Again, we are left with a balancing act. The importance of each criterion must be weighed against the importance of all the others, and the characteristics of each method must be weighed against the characteristics of all the other methods. Typically, cost and timeliness must be traded off against richness of geographic and demographic detail. The most fundamental task facing the analyst is to choose the optimal bundle of characteristics based on the resources available and the purposes for which the estimates will be used. This choice will guide the analyst through the selection of estimation methods, the collection of input data, and all the other steps of the estimation process.

Endnotes

1. This section is adapted from Chapter 13 "Forecast Accuracy and Bias", in S. Smith, J. Tayman, and D. "Swanson. Projecting State and Local Populations: Methodology and Analysis. New York, NY: Kluwer Academic/Plenum Press. 2001.
2. Enumeration errors can either raise or lower the error in population estimates, depending on whether they reinforce or offset the differences between the actual and estimates populations. As previously noted, to our knowledge, there is no information on the census enumeration errors for states and local areas. Therefore, most empirical studies do not attempt to adjust for enumeration error when evaluating population estimates; the notable exceptions being Judson, Popoff, and Batutis (2002) and Murdock and Hoque (1995).
3. The cell sizes in this cross-tabulation are small so the results should be viewed in that context (4 cells have zero observations, 5 cells have 1 or 2 observations; 5 cells have 3 or 4 observations; and 2 cells have 5 or more observations).
4. We purposely excluded estimates for places from Table 14.5 because of their great variability in size. The Census Bureau conducted a study of the error in their 2000 estimates by place and minor civil division (Harper, Coleman, and Devine 2003). That study showed an overall MAPE for all places of 12.4%, similar to that for census tracks, with MAPEs ranging from 4.3% for places with more than 100,000 persons to 35.1% for places with few than 100 persons.
5. Extrapolation techniques have been found to underperform most other estimation methods over a 10-year post-censal period (Poole, Tarver, White and Gurley 1966:19; Zitter and Shyrock 1964), but extrapolation techniques have not been evaluated for shorter post-censal periods where they may perform better.
6. This section is adapted from Chapter 12 "Evaluating Projections", in S. Smith, J. Tayman, and D. "Swanson. Projecting State and Local Populations: Methodology and Analysis. New York, NY: Kluwer Academic/Plenum Press. 2001.

References

D'Agostino, R. B., Belanger, A., & D'Agostino, R., B., Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician, 44*(3), 316–321.

Ahlburg, D. A. (1995). Simple versus complex models: Evaluation, accuracy, and combining. *Mathematical Population Studies, 5*, 281–290.

Alho, J., & Spencer, B. D. (1997). The practical specification of the expected error of population forecasts. *Journal of Official Statistics, 13*, 203–225.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*, 69–80.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Box, G. P., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*, 211–252.

Box, G. P., & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting, Second Edition*. Dordrecht, Heidelberg, London, and New York: Springer.

Bryan, T. (1999). *Small area population estimation technique using administrative records and evaluation of results with loss functions and optimization criteria*. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Washington, D.C

Cochran, W. G. (1977). *Sampling techniques, Third Edition*. New York: John Wiley & Sons.
Cohen, J. E. (1986). Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches. *Demography, 23*, 105–126.
Committee on National Statistics. (1980). *Estimating population and income for small areas*. Washington, DC: National Academy Press.
D'Allesandro, F., & Tayman, J. (1980). Ridge regression for population estimation: Some insights and clarification *Staff Document No. 56*. Olympia, WA: Office of Financial Management, State of Washington.
Davis, S. T. (1994). Evaluation of post-censal county estimates for the 1980s *Working Paper No. 5*. Washington, DC: Population Division, US Bureau of the Census.
Draper, N., & Smith, H. (1981). *Applied regression analysis, Second Edition*. New York: John Wiley & Sons.
Duncan, O. D., Cuzzort, R., & Duncan, B. (1961). *Statistical geography: Problems in analyzing areal data*. Glencoe: Free Press.
Emerson, J. D., & Stoto, M. (1983). Transforming data. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pp. 97–128). New York: John Wiley & Sons.
Emerson, J. D., & Strenio, J. (1983). Boxplots and batch comparisons. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pp. 58–96). New York: John Wiley & Sons.
Ericksen, E. P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography, 10*(2), 137–160.
Ericksen, E. P. (1979). Defining criteria for evaluation local estimates *Research Monograph Series, Synthetic Estimates for Small Areas* (Vol. No. 24). Washington, DC: US Department of Health, Education and Welfare.
Espenshade, T. J., & Tayman, J. (1982). Confidence intervals for post-censal state population estimates. *Demography, 19*(2), 191–210.
Fonseca, L., & Tayman, J. (1989). Post-censal estimates of household income distributions. *Demography, 26*(1), 149–160.
Goodall, C. (1983). M-estimators of location: An outline of the theory. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pp. 339–403). New York: John Wiley & Sons.
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: John Wiley & Sons.
Harper, G., Coleman, C., & Devine, J. (2003). Evaluation of 2000 subcounty population estimates *Working Paper Series No. 70*. Washington, DC: Population Division, US Census Bureau.
Hedayat, A. S., & Sinha, B. K. (1991). *Design and inference in finite population sampling*. New York: John Wiley & Sons.
Hodges, K., & Healy, M. K. (1984). *A micro application of a modified housing unit method for tract level population estimates*. Paper presented at the annual meeting of the Population Association of America, Minneapolis, MN.
Hodges, K., Wilcox, F., & Poveromo, A. (2002). *An evaluation of small area estimates produced by the private sector*. Paper presented at the annual meeting of the Population Association of America, Atlanta, Georgia.
Hoque, N. (2010) An Evaluation of Small Area Population Estimates Produced by Component Method II, Ratio-correlation and Housing Unit Methods for 1990. *The Open Demography Journal*, *3*, 18–30.
Judson, D. H., Popoff, C. L., & Batutis, M. J. (2002). An evaluation of the accuracy of US Census Bureau county population estimates. *Statistics in Transition, 5*(2), 205–235.
Keilman, N., Pham, P. Q., & Hetland, A. (2002). Why population forecasts should be probabilistic-Illustrated by the case of Norway. *Demographic Research, 6*, 409–453.
Keilman, N. W. (1990). *Uncertainty in national population forecasting*. Amsterdam: Swets and Zeitlinger.

Keyfitz, N. (1972). On future population. *Journal of the American Statistical Association, 67*, 347–363.

Keyfitz, N. (1981). The limits to population forecasting. *Population and Development Review, 7*, 579–593.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

Kmenta, J. (1971). *Elements of Econometrics*. New York: Macmillan Publishing Co.

Long, J. F. (1993). Post-censal population estimates: States, counties, and places *Working Paper No. 3*. Washington, DC: Population Division, US Bureau of the Census.

Long, J. F. (1995). Complexity, accuracy, and utility of official population projections. *Mathematical Population Studies*, 5, 203–216.

Lowe, T. J., Myers, W. R., & Weisser, L. M. (1984). *A special consideration in improving housing unit estimates: The interaction effect*. Paper presented at the annual meeting of Population Association of America, Minneapolis, MN.

Mahmoud, E. (1987). The evaluation of forecasts. In S. G. Makridakis & S. C. Wheelwright (Eds.), *The Handbook of Forecasting* (pp. 504–522). New York: John Wiley & Sons.

Makridakis, S. G. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting, 9*, 527–529.

Makridakis, S. G., Hibon, M., Lusk, E., & Belhadjali, M. (1987). Confidence intervals: An empirical investigation of the series in the M-competition. *International Journal of Forecasting, 3*, 489–508.

Makridakis, S. G., & Hibon M. (1995). Forecasting accuracy (or error) measures *INSEAD Working Paper Series 95/18/TM*. Fontainebleau, France: INSEAD.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting methods and applications, Third Edition*. New York: John Wiley & Sons.

Murdock, S. H., & Hoque, M. N. (1995). The effect of undercount on the accuracy of small-area population estimates: Implications for the use of administrative data for improving population enumeration. *Population Research and Policy Review, 14*, 251–271.

Murdock, S. H., Kelley, C., Jordan, J., Pecotte, B., & Luedke, A. (2006). *Demographics: A guide to methods and sources of data for demographic analysis in the media, business, and government*. Boulder: Paradigm Publishers.

Pflaumer, P. (1992). Forecasting US population totals with the Box-Jenkins approach. *International Journal of Forecasting, 8*, 329–338.

Poole, R. W., Tarver, J. D., White, D., & Gurley, W. R. (1966). An evaluation of alternative techniques for estimating county population in a six-state area *Economic Research Series 3*. Stillwater, OK: College of Business, Oklahoma State University.

Rainford, P., & Masser, I. (1987). Population forecasting and urban planning practice. *Environmental and Planning A, 19*, 1463–1475.

Rayer, S. (2007). Population forecast accuracy: does the choice of summary measure of error matter? *Population Research and Policy Review, 26*, 163–184.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis* (pp. 297–337). New York: John Wiley & Sons.

Rynerson, C., & Tayman, J. (1998). *An Evaluation of Address-Level Administrative Records Used to Prepare Small Area Population Estimates*. Paper presented at the annual meeting of the Population Association of America, Chicago, IL.

Siegel, J. S. (2002). *Applied demography: Applications in business, government, law, and public policy*. San Diego: Academic Press.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Smith, S. K., & Cody, S. (1994). Evaluating the housing unit method: A case study of 1990 population estimates in Florida. *Journal of the American Planning Association, 60*, 209–221.

Smith, S. K., & Cody, S. (2004). An evaluation of population estimates in Florida: April 1, 2000. *Population Research and Policy Review, 23*, 1–24.

Smith, S. K., & Cody, S. (2011). An evaluation of population estimates in Florida, 2011 *Special Population Reports No. 8*. Gainesville, FL: Bureau of Economic and Business Research.

Smith, S. K., & Mandell, M. (1984). A comparison of population estimation methods: Housing unit versus component II, ratio correlation and administrative records. *Journal of the American Statistical Association, 79*(386), 282–289.

Smith, S. K., & Sincich, T. (1992). Evaluating the forecast accuracy and bias of alternative population projections for states. *International Journal of Forecasting, 8*, 495–508.

Smith, S. K., & Tayman, J. (2003). An evaluation of population projections by age. *Demography, 40*(4), 741–757.

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers.

Starsinic, D. E., & Zitter, M. (1968). Accuracy of the housing unit method in preparing population estimates for cities. *Demography, 5*, 474–484.

Stock, J. H., & Watson, M. W. (2003). *Introduction of Econometrics*. Boston: Addison Wesley.

Stoto, M. (1983). The accuracy of population projections. *Journal of the American Statistical Association, 78*, 13–20.

Swanson, D. A. (1981). Allocation accuracy in population estimates: An overlooked criterion with fiscal implications *Small Area Population Estimates and Their Accuracy, Series GE-41, No. 7* (pp. 13-21). Washington, DC: US Bureau of the Census.

Swanson, D. A. (2008). Measuring Uncertainty in population data generated by the cohort component method. In S. H. Murdock & D. A. Swanson (Eds.), *Applied Demography in the 21st Century*. Dordrecht, Heidelberg, London, and New York: Springer.

Swanson, D. A., & Beck, D. M. (1994). A new short-term county population projection method. *Journal of Economic and Social Measurement, 20*, 25–50.

Swanson, D. A., & Coleman, C. (2007). On the MAPE-R as a measure of cross-sectional estimation and forecast accuracy. *Journal of Economic and Social Measurement, 32*(4), 219–233.

Swanson, D. A., & Tayman, J. (1995). Between a rock and a hard place: The evaluation of demographic forecasts. *Population Research and Policy Review, 14*(2), 233–249.

Swanson, D. A., & Tayman, J. (1999). On the validity of the MAPE as a measure of population forecast accuracy. *Population Research and Policy Review, 18*(4), 299–322.

Swanson, D. A., Roe, L., & Carlson, J. (1992). A variation of the housing unit method for estimating the population of small, rural Areas: A case study of the local expert procedure. *Survey Methodology*, *18*(1), 155–163.

Swanson, D. A., Carlson, J., Roe, L. & Williams, C. (1995). Estimating the population of rural communities by age and gender: A case study of the local expert procedure. *Small Town* May-June, 14–21.

Swanson, D. A., Tayman, J., & Barr, C. F. (2000). A note on the measurement of accuracy for subnational demographic estimates. *Demography, 37*(2), 193–202.

Swanson, D. A., Tayman, J., & Bryan, T. (2011). MAPE-R: A rescaled measure or accuracy for cross-sectional, sub-national forecasts. *Journal of Population Research, 28*, 225–243.

Swanson, D. A., & Tedrow, L. M. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography, 21*(3), 373–381.

Tayman, J. (1996). The accuracy of small area population forecasts based on a spatial interaction modeling system. *Journal of the American Planning Association, 62*, 85–98.

Tayman, J., & Schafer, E. (1985). The impact of coefficient drift and measurement error in the accuracy of ratio correlation population estimates. *The Review of Regional Studies, 15*(2), 3–11.

Tayman, J., Smith, S. K., & Lin, J. (2007). Precision, bias and uncertainty for state population forecasts: An exploratory analysis of time series methods. *Population Research and Policy Review, 26*, 347–369.

Tayman, J., Swanson, D. A., & Barr, C. F. (1999). In search of the idea measure of accuracy for subnational demographic forecasts. *Population Research and Policy Review, 18*(5), 387–409.

Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North Holland Publishing Co.

United Nations. (1971). Methods of estimating total population for current dates *ST*/*SOA*/*Series A*/*No. 10*. New York: Population Division, United Nations Department of Social Affairs.

Voss, P. R., Palit, C. D., Kale, B. D., & Krebs, H. C. (1981). Forecasting state population using ARIMA time series models. Madison, WI: Applied Population Laboratory, University of Wisconsin.

Zitter, M., & Shyrock, H. S. (1964). Accuracy of methods for preparing post-censal estimates for states and local areas. *Demography, 1*(1), 227–241.

Chapter 15
Guidelines for Developing and Presenting Estimates

In assembling our guidelines, we examined ideas in the area of population estimation and also in two areas related to population estimation: (1) sample (statistical) surveys; and (2) population forecasting. In regard to the two related fields, we found ideas that can be applied directly as guidelines for developing and presenting estimates and others that can be easily adapted for this purpose. While, as this book attests, there are substantial areas of overlap among sample surveys, population forecasting, and population estimation, there are important distinctions. And these distinctions lead to different guidelines.

In the area of population estimation, Herman Habermann (2006) describes principles for the US Census Bureau's estimates and projections programs that we find particularly useful in constructing guidelines for the development of population estimates. In regard to sample surveys, we found guidelines issued by the US Office of Management and Budget (2006) that were as detailed and specific as any we had encountered. From the field of population projections we found useful ideas in Habermann (2006), Pittenger (1977) and Smith et al. (2001: 343-360).

Common to all three fields is the idea of error assessment, which suggests that something about error should be contained in guidelines for population estimates. Some methods of population estimation, moreover, share a common inferential framework with sample surveys that renders error assessments very similar. For example, Espenshade and Tayman (1982), Swanson (1989), Roe et al. (1992), and Swanson et al. (1995) have shown how meaningful confidence intervals can be constructed around several different methods used to generate population estimates. It also is the case that some estimation methods are not grounded in the tools of inferential statistics and, as such, have more in common with the more limited assessment framework of forecast error than they do with the framework for sample surveys. So, considering the clear-cut and rigorous standards that can be applied to sample surveys (US OMB 2006), guidelines for population estimates will have generally more in common with the guidelines for population forecasts than with sample surveys. This does not demean the guidelines available for population

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_15,

forecasting. It simply recognizes the many differences between a sample survey and a forecast.

With this in mind, we note in regard to developing population projections, Smith, Tayman, and Swanson (2001: 343) observe that guidelines will not solve every problem that may be encountered. Changes in data collection procedures and definitions are one of the common problems encountered that are difficult to deal with under existing guidelines and procedures. As an example, from 1999 to 2005 the US Census Bureau used annual Medicare enrollment data to estimate the annual net migration of the population aged 65 years and over by county. Because of data consistency problems in the post-2005 Medicare data, Census Bureau now projects the number of Medicare enrollees as of July 1 for each county from 2006 to 2008 based on prior trends in Medicare enrollment (US Census Bureau 2010: A2-A3). Considering the Medicare data example and many other instances where unforeseen circumstances can arise, we nonetheless believe that guidelines for all their limitations are useful.

Although the term "transparency" is not used by Habermann (2006), Pittenger (1977), Smith et al. (2001), and US Office of Management and Budget (2006), it is clear from reading these materials that a traceable trail of documentation is deemed highly desirable. Thus, we believe transparency should be integral to guideless for developing population estimates.

We also believe that the observations made by Swanson and Tayman (1995) in regard to the "irony of forecasting," are largely applicable to estimation. First, as the case with a population forecast, it is virtually impossible to produce an estimate that is without error, yet estimates continue to be done. Second, as identified by Swanson and Tayman (1995) in regard to forecasting, the error expectations among demographers depend mostly on the size of the population and the length forecast horizon. With only a slight wording change, we believe that this statement applies to estimates: The error expectations among demographers in regard to estimates depend mostly on the size of the population and the "estimation horizon." That is, the length of time between the estimate and the nearest census(es) used as a basis for the estimate. It also is worth noting here that it also is far more difficult to estimate the error in a given forecast than it is to estimate the error in a sample survey. However, as we have written earlier, some estimation methods share a common inferential framework with sample surveys. We will take advantage of this in developing our guidelines for population estimates.

With the preceding ideas in hand, the foundation we plan to use for our estimation guidelines lacks only the mortar needed to unify them. For this purpose, we use the "applied demography" principle (Swanson et al. 1996). This principle can be summarized as one that views explanatory power and precision in terms of doing what is necessary to support practical decision-making while minimizing time and resources. Applied to the estimation process, it basically says that, as is the case with a census, a completely accurate estimate is not only unachievable, but also not necessarily a desirable goal (Swanson and Walashek 2011: 4-6). Thus, one should not spend time and resources on trying to achieve the unachievable – a perfect estimate. Instead, one should try to minimize time and cost requirements while

delivering estimates that are sufficiently accurate for their intended use. In this regard we note that accuracy is an important component, but it should not be used as the sole basis for judging the adequacy of a given estimate or set of estimates (Kitagawa 1980; Swanson and Tayman 1995). The result of the preceding is a set of guidelines resulting in a seven step process. We note in advance that each of the first six steps needs to be documented in that they lead to the seventh and final step, which is "presentation and documentation."

15.1 The Seven Step Process

Step 1. Determine What is Needed and the Time and Resources Available to Meet this Need. For estimates being done on a custom basis (e.g., a client has requested a post-censal estimate for a specific area for which no estimates are readily available), the first step in developing population estimates is to determine what exactly is needed and what the time and resource constraints are in meeting this need. In the situation where there is a specific client, this step would be part of the process leading to a final contract.

Clearly identifying these issues is the first step in the documentation trail needed for transparency (and for executing a satisfactory contract). If an estimate of the total population of a given area is needed in less than a week and there is very little in the way of resources available to develop these estimates, then a censal ratio method, an extrapolation method and the Hamilton-Perry Method should be on the list of potential methods to use while more data-intensive methods such as a sample based method, Component Method II, a full-blown cohort-component method, or a ratio-correlation model (and its variants) should be excluded.

When estimates are produced more for general purposes (e.g., annual population estimates produced by the US Census Bureau and Statistics Canada), the process is somewhat different since there is no specific client, but, instead, a range of actual and potential users. It is different in that the decisions underlying estimates of this nature are based on experience and the expectations regarding the primary needs of the majority of data users.

Whether for a specific client or for general use, the applied demography principle should be kept in mind even if time and budget constraints are pre-set. If the time and budget constraints are severe, it may even mean that the estimates may not be sufficient for the task(s) at hand. If the estimate(s) are to be designed for a specific client, it is important that this issue be addressed before a contract is executed. For example, it may be the case that the client wants to have age-race-sex details for extremely small areas (e.g., blocks) that can be done with, say, the Hamilton-Perry Method, but there may insufficient time and money to allow a detailed examination of the block level estimates relative to high levels of quality control (see, e.g., Swanson et al. 2010). In general, the greater the amount of demographic and geographic detail required, and the greater the attention paid to

an area's unique characteristics and special populations, the greater the time and other resources needed to produce the desired estimates.

It is worth noting here that it is much more time-consuming to construct a set of estimates for the first time than it is to repeat the process a second, third, or fourth time. Developing estimates from scratch and collecting, analyzing, and adjusting input data are time-consuming tasks. Updating a set of already-available estimates is much simpler.

What demographic characteristics are needed? Should the estimates be the total population size, or also of its composition. If estimates by age are needed, what are the relevant age groups ? The client will generally be able to answer these questions, but may not realize their importance unless prompted. Knowing the purpose for which the estimates are to be used will help determine the types of characteristics required. Age, sex, race, and ethnicity are the demographic characteristics most commonly included in population estimates. For some purposes, however, estimates of other demographic characteristics or population subgroups may also be needed, such as persons with disabilities. The estimates may also be used for estimates of population-related variables such as the number of households.

Population estimates are often made for well-defined regions such as states or counties that have easily determined boundaries that correspond to the boundaries used for administrative records and generally remain stable over time. For subcounty areas, however, the situation is often very different. Boundaries for subcounty areas are subject to sudden and dramatic changes. It is not uncommon for cities to annex adjoining areas, census tracts to be subdivided, zip codes to be reconfigured, service areas to be redefined, or new school districts to be formed. It is essential to determine if the boundaries of these types of areas have changed during the period for which historical base data have been collected. If they have, the data must be adjusted so that they refer to a geographic area that remains constant over time. If adjustments are not made, the estimates will confound the effects of boundary changes with population changes

Sometimes estimates must be made for regions lacking well-defined boundaries, such as postal delivery areas (zip codes). For example, a client may have only a rough idea of the geographic region making up a service area and lacks even zip code information on customers. In these instances, it is important to establish a clear set of boundaries for which meaningful estimates can be made and relevant data collected. When delineating the boundaries of a geographic area, it will be helpful to match boundaries with those used for the collection of the relevant base data (e.g., cities, census tracts, block groups).

Step 2. Select the Method(s). The results of Step 1 point to the selection of methods to be used. The next step is to choose the one(s) to be used in making the estimates. This choice will depend on the purposes for which the estimates will be used; the level of geographic and demographic detail needed; the amount of time, money, and other resources available; and the availability of relevant input data.

As is indicated elsewhere in this book a wide range of methods can be used for total population estimates. Simple extrapolation and interpolation methods such as linear, exponential, shift-share, and share-of-growth may be sufficient, especially

for inter-censal estimates. However, extrapolation methods can lead to problems for areas that changed rapidly during the base period in which these models were constructed (Swanson et al. 2010).

Estimates that require detailed age data inevitably require some type of cohort approach. Here, the Hamilton- Perry method comes to mind along with the full-blown cohort-component method. The former is capable of producing good inter-censal estimates and reasonable estimates by age for time points not too distant from the census data used for the launch year (see, e. g. chapters 10 and 17). The latter is preferred as the time between the launch year and the target year increases for a post-censal estimate as well as a pre-censal estimate. If the time between the lunch year and the target year is great for a pre-censal estimate, then the Inverse Projection Method is likely to be preferred over using the cohort-component method as a backcasting technique, given the availability of birth and death data (see, chapter 17).

It is important to keep in mind that in choosing an estimation method, no single model or technique is better than all others for all purposes. Each has its own strengths and weaknesses and must be evaluated according to its face validity, timeliness, cost, data requirements, ease of application, and other characteristics. Some of these characteristics are complementary (e.g., low costs, low data requirements, and ease of application typically go together), but others conflict with one another. And of course, it is always prudent to consider the need for error assessment, availability of data, and the deadline for producing estimates, relative to budget constraints

Step 3. Assemble the Data. Without input data of good quality it is virtually impossible to produce a high quality estimate. The quality of the estimate depends a on the quality of the data used to generate it (Habermann 2006; Popoff and Judson 2004; Tayman and Shafer 1985). For a sample-based method, the estimate is generated directly for the time needed. However, for censal ratio methods an additional point in time is required to develop an estimator; for interpolation and extrapolation methods as well as the ratio-correlation and Hamilton-Perry methods, two additional points in time are required. Still more may be required for other component methods. More complex methods such as structural models may require even more data points.

If the data needed for a given method are not adequate then a change in methods is usually in store. In some cases, some sort of analogue (e.g., model) data may be needed. The data needed are usually those closest to the target year (whether post- or pre-censal). Again, we mention that it always is important to assess data quality. Although it is the closest thing to a "gold standard" for demographic data in the United States, the decennial census is not completely error-free. Sometimes these errors are corrected within a year or two after the census, but they often go uncorrected until the following census (or even longer). Many census errors cancel out at higher levels of geography (except for the well-known undercount problem), but for small areas they can be substantial, especially for particular subgroups of the population (e.g., age, sex, and race categories). Thus, it is imperative to assess the

quality of all of the available input data, note and take into account both observed and potential problems, and make appropriate corrections or adjustments, relative to time and resource constraints.

Step 4. Develop the Estimates. Even a custom-made estimate for a specific client is not necessarily a "stand alone" estimate. Up to the level of the world as a whole, all other areas are nested in a geographical hierarchy. Hence, it may be useful to control small-area population estimates to estimates for larger areas even if only for reasons of consistency (see chapter 13).

Step 5. Assess Likely Error(s). In discussing error, we also cover the statistical idea of "uncertainty." An example of the idea of uncertainty is provided in chapter 9 and in chapter 14. In chapter 9, we compare the accuracy of a censal ratio estimator for an estimate for a large population, (King County, Washington) with an estimate for a small population (Garfield County, Washington. In this example, we suggested that the concept of statistical uncertainty was an underlying reason why the estimate for the large population was more accurate in a direct comparison with the 2000 census than was the estimate for the small population. In chapter 14, we discuss the development and use of confidence intervals as a tool for evaluating estimates.

Implicit in our recommendation to assess likely error is the need for an ongoing evaluation program. To assist in this task, we adapt a figure from chapter 5 and display it here as Exhibit 15.1, modified so that methods are classified by whether or not they are rooted in inferential statistics. As Exhibit15.1 suggests, our guidelines focus on methods used to estimate De Jure populations. However, we also will touch on methods used for De Facto populations.

As is the case with population forecasts (Smith, Tayman, and Swanson 2001), we believe it is important to provide the data user with some indication of the uncertainty associated with population estimates. This can be done in several ways, some less formal than others, One example of an informal approach is to construct a range of estimates based on two or more methods, application techniques, or sets of assumptions. Estimates might be developed, for example, using different methods. The main benefit of producing a range of estimates is that it shows the populations generated from different but reasonable models, techniques, or sets of assumptions. The primary limitation is that it does not provide a formal measure of uncertainty. Another way to indicate uncertainty is to construct statistical measures of uncertainty around the estimates. As is discussed in chapter 8, this is a natural component of the ratio-correlation and related regression-based methods. Yet another way to provide an indication of uncertainty is to construct tables summarizing errors from previous estimates for the area(s) being estimated or areas with similar characteristics.

Post-censal and pre-censal population estimates are subject to more error than inter-censal estimates, especially for small places. These errors are caused by our inability to correctly establish the course of mortality, fertility, and migration. We believe it is important to convey this information to the data user. Although it may be disappointing, information on potential estimation errors will give the data user a more realistic view of the estimates and help him/her plan more effectively for the uncertainty inherent in estimates.

Exhibit 15.1 De Jure Estimation Methods*

METHOD	BASIS IN INFERENTIAL STATISTICS? YES	NO	POSSIBLE LINK, DEPENDING ON SPECIFICS
Extrapolation			
Simple		**x**	
Complex, not ARIMA		**x**	
ARIMA			
Ratio	x		
Constant Share		x	
Shift Share		x	
Share-of-Growth		x	
Symptomatic			
Housing Unit			**x**
Censal Ratio			**x**
Regression			
Ratio Correlation	x		
Difference Correlation	x		
Rate Correlation	x		
Lagged Correlation	x		
Component			
Component Method II		x	
Cohort-Component		**x**	
Hamilton-Perry		**x**	
Composite		**x**	
Sample Based	**x**		
Other			x

*Our classification follows generally accepted practices for these methods, including their sources of data.

Finally, we note that while accuracy is important, all of the evidence clearly shows that perfection is not achievable. It was in reaction to this that Swanson and Tayman (1995) suggested that utility should also be used along with accuracy. Why? Because it has been found that population determinations with rather large errors are often useful and in particular they are virtually always more useful than simply assuming that there has been no change in population since the last census (Swanson and Tayman 1995; Swanson et al. 1998). As a means of measuring utility, they suggesting using "Proportional Reduction in Error" (PRE) as a measure to accompany those commonly provided in ex post facto evaluations of error. This measure is simple to construct and interpret:

$$\text{PRE} = [(\text{Error by Method(b)}) - (\text{Error by Method(a)})]/(\text{Error by Method(b)})$$

In using PRE, one develops a population estimate for the same area using two different methods, 'a' and 'b'. The error arising from each of the two methods is defined and measured and the proportionate reduction in error found by using rule (a) as opposed to rule(b) is determined by placing both error measures in the preceding formula. What constitutes method a and method b relative to

our discussion of utility? Method a is the estimation of a given population resulting from an estimation technique such as the censal-ratio method, CMII, ratio-correlation and the like, while Method b is the estimation of the same population resulting from data already at hand through an existing 'count', such as the last census. The estimate resulting from Method b is what can be called a 'naive' estimate in that it represents the theoretically (and most often, the practical) maximum error for an estimate because it based on no new knowledge. By using the PRE formula one can evaluate the reduction of error found by using the 'actual' estimate (derived using method a) over the error in the 'naive' estimate, method b, a number taken from the last census. Thus, this PRE shows the reduction in error (or gain in 'knowledge') due to the particular method b (and its judgments and input data) under an ex post facto evaluation.

Step 6. Review. Reviewing estimates, both internal and external is unquestionably useful. How much reviewing is done of an estimate or given set of estimates depends on many factors, including budget and time constraints, laws and administrative rules and policies, and past practice, and the use(s) to which the estimate(s) will be put, among others. We also suggest that a checklist be developed for purposes of review. This would be analogous to the pre-flight check list used by pilots before putting an airplane into motion. We provide suggestions below for what may be included in such a check list.

As part of internal review, we believe it is always useful to conduct a self-review if nothing else. The first step in a self review is to simply examine the context to see if the estimate is plausible. That is, does it have 'face validity.' If not, then additional internal review is certainly called for. However, even if in a personal review, an estimate has face validity, it generally useful to at least move on to an internal review, which involves others in the same group (some of whom may have worked on the estimate) or same organization.

The estimate should be compared to historical data. is it consistent with the past? If not, then this may indicate an error is present. Does the error remain about the same as the horizon becomes longer? Review may end at this point, but it may be the case that some form of external review is mandated or simply just called for.

By external review, we mean an examination of the results not only by clients, public officials, advisory boards, and various groups of data users (Smith et al. 2001), but also by professional peers (Swanson 2004). In some circumstances, there is no formal external review: Once the results have been reviewed internally, the review process is complete.

If one is preparing estimates for a specific client, we strongly advise that a review of "preliminary estimates" be done by the client and that this process be built into the contract. This also provides an opportunity to discuss the data and methods and in a sense, go through an "assisted personal review."

Step 7. Documentation and Release. In this step, the documentation from each of the preceding six steps needs to be assembled and turned into a report (e.g., a technical appendix) for purposes of transparency. The documentation should be included as part of the report in which the estimate is "released." The documentation should include the checklist used in assembling the estimates

as well as the relevant input data, methods, assumptions, accuracy and utility assessments, and special adjustments such as "controlling" along with the rationale for the decisions made in regard to the selections made in regard to input data, methods, assumptions, accuracy and utility assessments, and special adjustments. Putting this documentation together also provides part of the track record for doing estimates in the future so that lessons learned will then enter into an updated checklist. It also provides the basis for an institutional memory. In the face of personnel turnover, new hires need a transparent, comprehensive description of the estimation methods used previously and evaluations of their performance.

15.2 Summary

As discussed at the beginning of this chapter, neither the seven steps we propose nor any other set of guidelines can answer every question and solve every problem that may be encountered in the development of population estimates. As Smith et al. (2001: 360) observe in regard to population projections, "...very set of circumstances is unique in one way or another...," an observation we believe applies equally well to population estimates. This is the case even with large organizations that produce estimates on a regular basis, where the impact of changes in decennial census concepts, questions, coverage, and definitions cannot always be accommodated under existing procedures (see, e.g., Swanson 2010).

It also may be the case that estimates are produced for a wide range of potential users, some of which have need for more precise estimates than others. The guidelines suggest developing estimates that are appropriately sufficient such that the time and resources necessary to produce them are minimized. In this case, however, it may be that time and resources do not allow for the needs of the users that need more precise estimates. Thus, estimates for these users will not be as sufficient as those for users who can live with less precise numbers. It also is the case that religiously following these guidelines will not necessarily yield accurate population estimates. However, we believe that they can help you make reasonable choices and avoid major mistakes.

References

Espenshade, T. and J. Tayman. 1982. "Confidence Intervals for Post-censal State Population Estimates." *Demography* 19: 191–210.

Habermann, H. 2006. "Research to Improve Population Estimates" Part of a presentation at the Spring (May 18th -19th) 2006 meeting of the Census Advisory Committee for Professional Associations.

Kitagawa, E. 1980. *Estimating Population and Income of Small Areas*. Washington, DC: National Academy Press.

Pittenger, D. 1977. "Population Forecasting Standards: Some Considerations Concerning Their Necessity and Content." *Demography* 14: 363–368.

Popoff, C., and D. Judson. 2004. "Some Methods of Estimation for Statistically Underdeveloped Areas." pp. 603–641 in J. Siegel and D. A. Swanson (Eds.). *The Methods and Materials of Demography, 2nd Edition*. New York, NY: Elsevier Academic Press.

Roe, L. J. Carlson, and D. A Swanson. 1992 "A Variation of the Housing Unit Method for Estimating the Population of Small, Rural Areas: A Case Study of the Local Expert Procedure." *Survey Methodology* 18(1):155–163.

Smith, S., J. Tayman, and D. A. Swanson. 2001. *State and Local Population Projections: Methodology and Analysis*. New York, NY: Kluwer Academic/Plenum Publishers

Swanson, D. A. 2010. "The American Community Survey: Some Considerations Regarding its use as a Substitute for the Sample 'Long Form' in the Decennial US Census." Paper presented at the Workshop on the American Community Survey, held at the annual Conference of the Latin American Population Studies Association (ALAP), Havana, Cuba, November 16th.

Swanson, D. A. 2004. 2004 "Advancing Methodological Knowledge within State and Local Demography: A Case Study." *Population Research and Policy Review* 23: 379–398.

Swanson, D. A. 1989 "Confidence Intervals for Post-censal Population Estimates: A Case Study for Local Areas." *Survey Methodology* 15(2):271–280.

Swanson, D. A., and J. Tayman 1995. "Between a Rock and a Hard Place: The Evaluation of Demographic Forecasts." *Population Research and Policy Review* 14:233–249.

Swanson, D. A. and P. J. Walashek. 2011. *CEMAF as a Census Method: A Proposal for a Re-designed Census and an Independent Census Bureau*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Swanson, D. A., and T. Burch, and L. Tedrow. 1996. "What is Applied Demography?" *Population Research and Policy Review* 15:403–418.

Swanson, D. A., D. Beck, and J. Tayman. 1995 "On the Utility of Lagged Ratio-Correlation as a Short-term County Population Estimation Method: A Case Study of Washington State." *Journal of Economic and Social Measurement* 21:1–16.

Swanson, D. A., A. Schlottmann, and R. Schmidt. 2010. "Forecasting the Population of Census Tracts by Age and Sex: An Example of the Hamilton-Perry Method in Action." *Population Research and Policy Review* 29(1):47–63.

Swanson, D. A., G. Hough,. J. Rodriguez, and C. Clemans. 1998. "K-12 Enrollment Forecasting: Merging Methods and Judgment." *ERS Spectrum* 16:24–31.

Tayman, J. and E. Shafer. 1985. "The Impact of Coefficient Drift and Measurement Error on the Accuracy of Ratio-Correlation Population Estimates." *The Review of Regional Studies* 15:13–23.

US Census Bureau. 2010. Appendix A, Source Notes and Explanations, pp A1–A78 in *State and Metropolitan Area Data Book: 2010*. Washington, DC: US Census Bureau.(http://www.census.gov/compendia/databooks/pdf_version.html).

US Office of Management and Budget (OMB). 2006. *Standards and Guidelines for Statistical Surveys*. Washington, DC: U. S. Office of Management and Budget.

Chapter 16
De Facto Populations and Populations Impacted by Disasters

This chapter deals with three separate estimation topics related by the fact that they are not easily assembled from census data, which in the US and the other countries to which this book is addressed (e.g., Argentina, Australia, Brazil, Canada, England, Ireland, Mexico, and the United States) is based on the concept of a De Jure population. The first issue is that of a de Facto population, which is the concept of people enumerated, estimated, or forecasted where they are found rather than where they usually reside. The second is that of the homeless population and the third, a population impacted by a disaster. For both the homeless and those impacted by a disaster, the underlying concept is that of a De Jure population, but in the case of both, the methods for estimating De Jure populations are rendered virtually useless (Rummel 1991; Smith and McCarty 1996; Swanson 2008; US Department of Housing and Urban Development 2008a). This situation calls into play at least some of the methods for estimating De Facto populations, hence the reason for covering both in this chapter. In addition, having estimates of the De Facto population can play an important role in the plans for coping with disasters.

The idea of a de Facto population also has more than a few nuances. For example, the visitor population in a resort area such as Las Vegas or Honolulu is a De Facto population, but where these visitors are during the day vs. the night also can vary substantially. For example, during the day, visitors to Hawaii may be on beaches while at night they are in their hotels. Similarly, some of the visitors to Las Vegas may be in Death Valley, the Red Rock Natural Conservation area, Lake Mead, or the Grand Canyon during the day, but in hotel rooms during early evening, then in a theater watching a play, then a restaurant, then in casinos, then finally back to their hotel rooms.

Similarly, many of the commuters to the financial district of San Francisco, California for purposes of work may be in Chinatown for lunch. Yet another example is that the population of McAllen, Texas may swell during the winter months with snowbirds from the upper Midwest, who during the day may be at south Padre Island enjoying the beach.

These examples illustrate the fact that the estimation of de Facto populations presents difficulties not found with the estimates of De Jure populations, as is

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_16,

evidenced by some of the colorful names given to these methods – *Demoflush* comes readily to mind (Goldschmidt and Dahl 1976), for example, as one such name that has a cachet not found among the names of De Jure methods (e.g., Component Method II).

For purposes of discussing estimation methods, it is convenient to look at the concept of a De Facto population from three perspectives: (1) daytime population; (2) visitor population; and (3) seasonal population, which we subdivide into (a) the amenity seeking population and (b) migrant workers and their families. One reason for using these three categories is that they correspond roughly to the kinds of estimates (and projections) that are desired for De Facto populations (Akkerman 2000; Happel and Hogan 1987; 2002; Kavanaugh and Lamphere 1989; Las Vegas Convention and Visitors Authority 2011a; Schmitt 1956; 1968; Smith 1989). Another reason is that these categories are important because of the impacts they have on the population numbers of the places were they are found. As examples:

Daytime Population. In the late 1990s, the De Jure population of San Francisco, California was about 750,000; its daytime population was estimated at 1.3 million (San Francisco Planning Department 1997).

Visitor Population. As of the 2010 census, the De Jure population of Clark County, Nevada (metropolitan Las Vegas) was 1,375,765 (US Census Bureau 2011); there were over 37 million visitors to Las Vegas in 2010 (Las Vegas Convention and Visitors Authority 2011b).

Seasonal Amenities Population. The July, 1995 De Jure population of Leelanau County in Michigan's Upper Peninsula was estimated by Becker et al. (1996) to be 18,502; the "second home" (seasonal) population was estimated by them to be 10,937.

Seasonal Migrant Worker Population. In the 2000 Census, the De Jure population of Chelan County, Washington was 66,616 (US Census Bureau 2001); the 2000 population of Migrant Seasonal Farm Workers and their families in this apple-producing county was estimated at 26,382 by Larson (2000)

As we will see in our forthcoming discussion involving definitions of these types of populations, there are more ambiguities involving them then there are in the definition of a De Jure population, and there are plenty in the latter (Cork and Voss 2006).[1] Among other issues, these categories are neither mutually exclusive nor exhaustive. For example, many places have seasonal fluctuations in terms of both what we call visitor populations and what we call seasonal populations. However, our three categories lend themselves to different techniques and in developing our definitions, we will keep these different techniques in mind. We also will use the definition of "Census Day" in terms of our definitions and use the concept of usual residence as a foil to work from. Again, we stress that neither this device nor others will resolve all of the many ambiguities of defining what a population is.

For our purposes, we use the US Census Bureau's (2005) definition of a daytime population, which is the number of people who are present in an area during normal business hours. This is in contrast to the population present during nighttime hours,

which usually corresponds to the De Jure population or residential population. Some areas such as the Financial District of San Francisco, California, contain few residents and have a very small nighttime population. However, San Francisco's Financial District has many thousands of people working in it during weekday while nearby bedroom communities, such as Walnut Creek, California, may have more than 50 percent of its residential population leaving each workday morning to travel to their jobs, with no correspondingly large inflow of workers into the area. Note that this definition is largely based on the traditional idea of the workday being Monday through Friday, which means it does not consider people in San Francisco's Financial District during the day on either a Saturday or Sunday who are waiting to ride the cable cars up Market Street. Such people would be viewed instead as visitors. It would consider, however, workers picking table grapes in California's Central Valley as a daytime population of the vineyard in which they are working. As this example suggests, some or all of these workers may be seasonal in that their usual residence in terms of census residency rules is elsewhere.

We define a visitor population as people who are in a given area on census day for a short period of time that would not be considered their usual place of residence, but who also are not part of the area's daytime population. We introduce the idea of a short period of time to assist in distinguishing a visitor population from a seasonal population. This would include people on vacation staying in a hotel as well as people who are working on assignment for a few days who are staying in a hotel (e.g., conference attendees, salespeople). This follows the temporal dimension described by Happel and Hogan (2002) in their distinction between visitor and seasonal populations. From our definition it is clear we are not looking at visitors to specific attractions, a subject dealt with by Tyrrell and Johnston (2002). Also, we are interested in the number of visitors, not the number of visits, otherwise known as person trips (Leeworthy 1996).

In defining a seasonal population, we begin with the observation by Cork and Voss (2006: 5) that no recent census in the United States has allowed respondents the ability to directly indicate that they believe that address information on their census questionnaire is inaccurate. Respondents have been unable to indicate, for example, that they have received the form at a seasonal home. They also note that unlike the case in the United States, there are other countries that ask questions in their censuses that allow one to determine usual place of residence and seasonal residence information (Cork and Voss 2006: 54).

Happel and Hogan (2002), among others, not only use a temporal dimension to define seasonal population, but also the reasons for travel. As suggested by our earlier examples, this is useful in distinguishing between seasonal effects largely due to amenities (spending the month of July at a second home in Michigan's Upper Peninsula) and those largely due to work (migrant labor). Thus, we distinguish the seasonal population from the visitor population on the basis of time. For those seeking amenities, we view them as being in an areas for more than a couple of weeks, but not more than six months' For the migrant workers, we view them as being in areas for as short as a few days, but also not more than six months.

16.1 Estimating a Daytime Population

Here, we will describe two general approaches that can be used to estimate daytime populations. The first is provided by the US Census Bureau (2005) and the second via remote sensing imagery (Cai et al. 2006; Wicks et al. 1999)

16.1.1 Using (De Jure) Census Data.

The Census Bureau (2005) developed its estimates of daytime populations using information from the 2000 Census "long form" that included data on the employed population, place of work, means of transportation to work, or the other journey to work items. Given the data, the Census Bureau (2005) developed two methods, which are algebraically equivalent to one another. The first method uses "commute to work" information:

$$\begin{aligned}&(\text{estimated daytime population of area i}) = (\text{resident population of area i})\\&+(\text{workers who commute into area i}) - (\text{workers who commute out of area i})\end{aligned} \tag{16.1a}$$

The second uses "place of work" and "place of residence" information:

$$\begin{aligned}&(\text{estimated daytime population of area i}) = (\text{resident population of area i})\\&+(\text{workers working in area i}) - (\text{workers living in area i})\end{aligned} \tag{16.1b}$$

Using [16.1b] we find that as of April 1st (Census Day), 2000, the estimated Daytime population of San Francisco, California is 945,480 (US Census Bureau 2005), where

$$\begin{aligned}&945,458 = (776,733) + (587,300) - (418,553)\\&(S.F.\ resident\ population) + (workers\ working\ in\ S.F) - (workers\ living\ in\ S.F.)\end{aligned}$$

Unfortunately, with the loss of the decennial "long form," the data needed to use these two methods is no longer available and one must turn to the American Community Survey, which while possible to use, presents some challenges not found with the decennial census "long form" (Cork and Voss 2006; Van Auken et al. 2006; Swanson and Walashek 2011). However, countries with census data similar to those needed for methods 1 and 2 would be able to employ either method, respectively (United Kingdom Statistics Authority 2001).

16.1.2 Remote Sensing Imagery

Researchers at the Oak Ridge National Laboratory have developed, LandScan, a method for estimating "ambient" populations at a one kilometer level of resolution using a combination of satellite imagery and GIS (Bhaduri et al. 2007; Cheriyadat et al. 2007), where an ambient population is an average over 24 hours period. The fact that a 24 hour "average" population suggests that daytime populations can be estimated (as well as DeJure and Seasonal populations). Detailed descriptions of Landscan's methods and data can be accessed at http://www.ornl.gov/sci/landscan/index.shtml, along with and data via a registration procedure that can be initiated at this same website. Note that registration is geared toward providing data access to non-commercial users.

16.2 Estimating a Visitor Population

Estimating visitor populations can be done through several methods, the most common of which include counting occupied rooms in hotels and other facilities in combination with an average number per occupied room, and surveys conducted via transportation modes, entry and exit points area, and visitor sites (Leeworthy 1996; Watson et al. 2000). These methods are generally time and resource intensive because in part they rely on surveys, but, even with the use of "administrative records" such as occupied hotel rooms they remain time and resource intensive.

As an example of the time and resource intensity, the Hawaii Tourism Authority (2010: 2) estimates that there were 6,517,054 visitors to Hawaii in 2009, staying an average of 9.33 days. To get these estimates (and other information), the Hawai'i Tourism Authority combined information from three major steps: (1) determining passenger counts on arriving airline flights, foreign and domestic, separating visitors from in-transit passengers, returning Hawai'i residents, and migrants intending to reside in Hawai'i; (2) determining arrivals by cruise ships: Visitors who entered Hawai'i via foreign-flagged cruise ships, derived from the Cruise Visitor survey which covered US flagged and foreign flagged cruise ships; (3) obtaining Cruise ships "Arrivals by Air," derived from the Domestic In-flight and International Departure surveys which sampled only visitor arrivals by air. This figure represented an estimate of visitors staying on cruise ships. These three major steps used data from 10 sources: (1) airline passenger counts (both scheduled and chartered), domestic and foreign; (2) reports by the US Office of Immigration Statistics; (3) reports by the Bureau of Customs and Border Protection, Honolulu Office.; (4) US Customs Declaration Forms; (5) International Intercept Survey, a systematic sample of passengers in the boarding area and walkways at the Honolulu International Airport and the Kahului Airport on Maui; (6) Domestic Survey, the form for which is on the reverse side of the Hawai'i State Department of Agriculture's mandatory Plants and Animals declaration form, which is distributed

to passengers on all flights from the US mainland to Hawai‘i every day of the year; (7) The Island Visitor Survey, from samples taken conducted at departure area of the airports on all the islands; (8) Cruise Visitor Survey, forms for which are distributed to the cabins on the cruise ships; (9) Honolulu International Airport Billing Records, which show the number of passengers on flights from Canada who were pre-cleared in Canada and not included in the INS; and (10) Cruise Passenger Counts: All cruise ships which entered Honolulu, Hilo and Lahaina Harbor for which passenger counts are reported to the Department of Transportation, Harbors Division and the Department of Land and Natural Resources.

As this example for Hawai'i illustrates, the development of visitor population estimates is often time and resource intensive, with a high level of administrative coordination. The example is not dissimilar to methods described elsewhere in this regard (Erkkila 2000; Leeworthy 1996; Tyrrell and Johnston 2002; Watson et al. 2000).

16.3 Estimating a Seasonal Population

16.3.1 The Amenity Seeking Seasonal Population

Some countries have the ability to develop De Facto numbers along with De Jure numbers built directly into their regular census counts, while others are more limited (for a suggested list, see, e.g., Cork and Voss 2006: 303-325). Unfortunately, the United States conducts a census in which De Facto numbers cannot be directly extracted. However, as shown earlier in the section on Daytime Population Estimates, it has collected census information that can be used to develop De Facto estimates. In the case of seasonal populations, of the features of the US decennial census is its classification of vacant housing, which includes those reserved for seasonal, recreational, or occasional use. This can be exploited for purposes of estimating a seasonal population.

To start, here is some background on this classification from the US Census Bureau (2004). First, in order to make the vacation home category consistent over the decades, "seasonal", "held for occasional use", and "for migrant workers" are combined. Second, the "occasional use" category was not used prior to the 1960 census. Third, counts of seasonal and occasional use vacant units are separately provided from 1960 to 1980, but they were combined beginning in 1990 because evidence indicated enumerators had great difficulty determining the difference. Fourth, counts of housing units for migrant workers were included with seasonal units before 1990; for comparability, this housing type was added beginning with the 1990 count of seasonal, recreational, or occasional units. Fifth, separate counts of migratory vacant units are provided beginning with 1990, a number observed to be very small over the decades.

The availability of this information is one of the reasons we made a distinction between the visitor population and the seasonal population. With the preceding data and an estimate of the average number of seasonal persons per seasonal household (SEASONPPH) in hand, the Housing Unit Method (see Chapter 7) can be used to develop an estimate of the total amenity seeking seasonal population of a given area i. To proceed, we need an estimate of SEASONPPH. Although it is dated, the US Census Bureau (1982) produced a report from the 1980 census on non-permanent residents. This report is nicely geared toward seasonal populations, especially those that are amenity seeking. Table C of this report provides Average Persons Per Households for non-permanent households (i.e., SEASONPPH) for selected states, which we can use in conjunction with the Census Bureau's 2004 report on seasonal housing to obtain an estimate of a seasonal population:

$$SEASONP_i = SSMHU_i * PPHSEASON_i \quad (16.2)$$

where

$SEASONP_i$ = Estimated Seasonal Population in area i
$SSMHU_i$ = Seasonal Single and Multiple Housing Units
$PPHSEASON_i$ = Average Number of Persons per Seasonal Household

As an example of the preceding, we develop a seasonal population estimate for Arizona as of April 2000. First, we find that there were 142,601 housing units for seasonal, recreational, and occasional use in Arizona for 2000 (US Census Bureau 2004). Second, we find that the SEASONPPH for Arizona as of April 1980 is 1.84 (Table C, US Census Bureau 1982) and that the median age of persons in non-permanent households is over 65. The latter suggests that the non-permanent households are made up of amenity seeking "snowbirds" (Happel and Hogan 2002). With the preceding in hand, we use equation [16.2] to estimate the seasonal amenity seeking population for the 1999-2000 winter season for Arizona as:

$$262{,}386 = 142{,}601 * 1.84 \quad (16.2)$$

The preceding estimate differs from the 1999-2000 estimate of 273,000 snowbirds in state of Arizona provided by Happel and Hogan (2002), but not by much. The absolute difference is -10,514 and the relative difference is -3.89%.

Our HUM based method as shown in equation [16.2] could be refined, given the availability of information on Recreational Vehicle (RV) parks, which are not part of the permanent housing stock, but should be included because seasonal residents live there. For areas that keep track of RV space inventories, equation [16.2] can be refined as follows

$$SEASONP_i = (SSMHU_i + RVS_i) * PPHSEASON_i \quad (16.3)$$

where

$SEASONP_i$ = Estimated Seasonal Population in area i

$SSMHU_i$ = Seasonal Single and Multiple Housing Units in area i
RVS_i = Recreational Vehicle Spacves in area i
$PPHSEASON_i$ = Average Number of Persons per Seasonal Household in area i

Additional refinements could be made if survey data available. For example, if a survey is done of RV parks that collected data on the occupants, then a separate PPH value for them could be used, along with an estimate of the occupied RV spaces.

There is some ambiguity in the "winter season" 1999–2000 date given for our example estimate for Arizona. As noted by Smith (1989) an accurate enumeration of the entire seasonal population is almost never available. Among other limitations, this means that the empirical relationship between the symptomatic variables and seasonal population is not based on an actual point-in-time census. This means that we have no direct estimate of error. At best, a given estimate can be compared with estimates from other sources in hopes of "triangulating" the seasonal population, keeping in mind that it likely fluctuates over the season in question. These fluctuations leave even such precisely named methods such as "Demoflush" with estimates that are not as precise as the name might suggest.

In concluding this discussion of the amenity seeking seasonal population, we know that there are people who move in combination with seasonal amenity seekers for purposes of employment. For example, many of the people working at lodges and related facilities in national parks only are there for the season, (e.g., summer in Yellowstone and winter in Death Valley). For our purposes, we include them as part of the amenity seeking population and not part of the next seasonal group we examine, the migrant worker population.

16.3.2 Migrant Worker Seasonal Population

This population largely works in agriculture and related areas (e.g., fish canneries in Alaska), and for those that work in services geared toward the amenity seeking seasonal population, we have included them as part of this group, as just stated. Moreover, evidence indicates that the migrant worker seasonal population is decreasing in that people who once moved from place to place following harvests and related seasonal work are becoming permanent year-round residents in agricultural areas (Kandel 2008).

While the data on this population may be skimpy in terms of the Decennial US Census on Population and Housing, this is not the case in regard to the US Census of Agriculture, which was formerly conducted by the US Census Bureau, but is now conducted by the National Agriculture Statistics Service, US Department of Agriculture (http://www.nass.usda.gov/). The US Department of Agriculture (USDA) maintains and analyzes a wealth of data on this population (Kandel 2008) as does US Department of Labor (USDOL), especially in the form of its National Agricultural Workers Surveys (http://www.doleta.gov/agworker/naws.cfm). As an example

of the richness of these data, the 2007 Census of Agriculture shows that in Arizona, 28,754 farmhands were hired, of which 238 were migrant laborers (U.S, Department of Agriculture 2008). Similar data are available for other states and for sub-areas within states via the USDA's "quickstats" service ((http://quickstats.nass.usda.gov).

As we described at the outset of this section, we used information we had about available data and methods to assist in developing our De Facto population categories. Developing estimates of a visitor population is perhaps the most onerous because there are little, if any, publically available data for such a population. At the other end of the spectrum, we have the readily accessible and no-cost data available on the seasonal migrant worker population, courtesy of USDA and USDOL. Very close to the USDA and USDOL information in terms of accessibility and cost, we have the information from the US Census Bureau that can be manipulated to obtain estimates of daytime populations as well as estimates of the seasonal amenity seeking population. We now turn to a related, but distinct task: estimating the immediate effect of disasters on populations.

16.4 Estimating a Homeless Population

In a country such as the United States where the De Jure concept is used to define population, the presence of people who do not live either in permanent resident units or in group quarters (e.g., dormitories, barracks, convents, shelters for the homeless) creates problems for census and estimation purposes. To start with, the US Decennial Census completely went to "mail-out/mail-back" by 1980 as the initial mode of contact (US Census Bureau, n.d.) To implement this method, the "Master Address File" (MAF) was developed, which is a national register of addresses (Swanson and Walashek 2011). As you can guess, the major bulk of census activities are based on the MAF, which returns us to the point made earlier that those not living in permanent units present enumeration problems since where they "reside" is not in the MAF. The US Census Bureau is, of course, well aware of the presence of people not living in permanent units and makes an effort to count them in the decennial census (Glasser 1991; Salo 1990; US Census Bureau, n.d.).

Fortunately, efforts to count the homeless in the United States received a tremendous boost in 1987 when the McKinney-Vento Homeless Act became law in the United States. Among its provisions is the requirement that surveys of the homeless must be done by agencies seeking funding under the Act (US Department of Housing and Urban Development 2008a). The Act was re-authorized in 2009 with the same survey requirement.

Under the charge of the McKinney-Vento Homeless Act, The US Department of Housing and Urban Development (US HUD), needed to define homelessness. In moving toward a definition (US HUD 2008b: 4) observes "residential stability" can be divided into two broad categories of people: (1) those who "literally homeless;" and (2) those who are "precariously housed." The "literally homeless" include people who for various reasons have found it necessary to live in

emergency shelters or transitional housing for some period of time. This category also includes unsheltered homeless people who sleep in places not meant for human habitation (for example, streets, parks, abandoned buildings, and subway tunnels) and who may also use shelters on an intermittent basis" (US HUD 2008b: 4). The "Precariously Housed" refers to "...people on the edge of becoming literally homeless who may be doubled up with friends and relatives or paying extremely high proportions of their resources for rent. The group is often characterized as being at imminent risk of becoming homeless" (US HUD 2008b: 4).

With these definitions in hand, US HUD developed two manuals designed to assist local jurisdictions in meeting the survey requirements of the McKinney-Vento Act. The two manuals are aimed at the two groups composing the "literally homeless," the sheltered homeless (US HUD 2008a) and the unsheltered homeless (US HUD 2008b).

US HUD (2008a) defines the sheltered homeless as adults, children, and unaccompanied youth who, on the night of the count, are living in shelters for the homeless, including: (1) Emergency shelters; (2) Transitional housing; (3) Domestic violence shelters; (4) Residential programs for runaway/homeless youth; (5) Any hotel, motel, or apartment voucher arrangements paid by a public or private agency because the person or family is homeless. US HUD (2008b) defines the unsheltered homeless as the homeless who are not residing in shelters for the homeless and similar facilities. It is designed to produce counts of the unsheltered homeless and their characteristics. This orientation complements the information on those living in

As an example of the type of information that can result from these two manuals, we turn to the 2007 census and survey of the homeless in southern Nevada, which includes Las Vegas (Applied Survey Research 2007). The study was conducted in January, 2007 and included not only counts of both the sheltered homeless and the unsheltered homeless, but estimates of the "precariously housed," which was termed the "hidden homeless" in the study. Using a range of methods geared specifically to enumerating and surveying these three types of homeless population, the study estimated a total homeless population of 11,417, of whom 3,747 were enumerated on the streets, 3,844 in shelters, and the remaining 3,826 were "hidden"(Applied Survey Research 2007: 3). The methods included a systematic two-day canvassing of streets, a canvassing of shelters and institutions, and a general population telephone survey (Applied Survey Research 2007: 67-90). The telephone survey was used as the basis for estimating the "hidden" homeless, "...persons living on private property but in locations that would not be considered "double-ups" as defined by HUD such as tents, cars/vans, unconverted garages, storage sheds, etc. The general population phone survey was a 10-15 minute survey designed to determine if there were people staying in the household who would otherwise be homeless. (Applied Survey Research 2007: 71).

While the 2007 Las Vegas study may be one of the most comprehensive of the homeless counts and surveys, it is not alone. Studies of the homeless abound and it may be the case that a study has already been done for an area of interest to you; if not, the two US HUD manuals and the Las Vegas Report provide the basis for estimating the homeless population in the area of interest to you.

16.5 Estimating the Entire De Facto Population

To our knowledge, nobody has put together a "De Facto Population Equation," which we believe could be a useful tool. To this end, we offer the following equation:

$$D_i = V_i + H_i + A_i + M_i + +REP_i + ND_i + RP_i \quad (16.4)$$

Where i = the area in question

D = De Facto Population
V = Visitor Population
H = Homeless Population
A = Amenity Seeking Seasonal Population
M = Migrant Worker Seasonal Population
ND = Non-Resident "Daytime" Population
RP = Resident (De Jure) Population Present
and
RP = R – RA

where

R = Resident Population
RA = Resident population away

In some areas, there is a large "ND" population and in others, it is virtually zero. For example, a large chunk of the daytime population of San Francisco is composed of people who live elsewhere. Similarly, the Honolulu Census Designated Place (basically, the city of Honolulu), will have a daytime population that commuted in from areas on the island of Oahu, outside of the Honolulu CDP. However, for the entire state of Hawai'i there are virtually no members of a "daytime" population that are from outside of Hawai'i who are not part of either the visitor or seasonal populations.

As an example application, of equation [16.4] we provide an estimate of the De Facto population of 543,665 for Honolulu, Hawai'i, (the Census Designated Place, i.e., the Honolulu CDP) as of April 2000:

$$D_{Honolulu} = V_{Honolulu} + H_{Honolulu} + A_{Honolulu} + M_{Honolulu} + RP_{Honolulu} + ND_{Honolulu}$$
$$636{,}970 = 168{,}101 + 8{,}000 + 14{,}297 + 16 + 353{,}251 + 93{,}305$$

The visitor count of 168,101 is taken from a report by the Hawai'i Department of Business, Economic Development, and Tourism (2000); the homeless estimate of 8,000 is taken from a report done by SMS Research that provided an estimate for 2003, which was delivered to us in a personal communication from the President of SMS Research, Jim Dannemiller (2011) that also provided advice on the likely number in 2000; the amenity seeking seasonal population estimate of 14,297 was derived using the same method described earlier in this chapter for Arizona,

but with data specific to Honolulu, as was the estimated number of 16 for the migrant worker seasonal population. The estimate of 353,251 of the total Honolulu resident population that was present was derived by using statistics on returning residents (60,000) for the month of April, 1999 found in a report by the Hawai'i Department of Business, Economic Development, and Tourism (2001). This number was assumed to apply to April of 2000 and multiplied by the proportion of Hawaii resident who live in Honolulu (60,000* (371,657/ 1,211,537)) to get an estimate of the number of Honolulu residents who were away (18,406), which was subtracted from the total number of residents (371,657) to get the estimate of 353,251 for the total number of residents present.

The estimate of the Daytime population of the Honolulu CDP who are residents from other areas is based on a manipulation of Equation [16.1a], which recall is defined as: (estimated daytime population of area i) = (resident population of area i) + (workers who commute into area i) - (workers who commute out of area i). The preceding equation can be re-arranged to yield (workers who commute into area i) = (estimated daytime population of area i) - (resident population of area i). In the case of the Honolulu CDP, we use the data for daytime population estimates assembled by the US Census Bureau (2005), which shows a daytime population of 464,964 and a De Jure population of 371,657. Thus, we have an estimate of the "ND" population of 93,305 = 464,964 - 371,657.

As is the case with any equation, this one offers the potential to estimate a missing term if the others are available. For example, $H_i = D_i - (V_i + A_i + M_i + ND_i + RP_i)$. Another example of how equation [16.4] might be used would be to take ratios of various elements and then use them to fill in missing terms. For example, if the ratio of the De Facto to the De Jure population was relatively constant (at least during certain seasons or months), this relationship might be used to estimate the total De Facto population, such that a missing piece (e.g., the homeless population) could be estimated. And of course some terms could be combined to make the task of making such estimates more tractable (e.g., the amenity seeking seasonal population could be combined with the migrant worker seasonal population to get a total seasonal population term).

16.6 Estimating a Disaster-Impacted Population

As we mentioned at the outset, estimates of De Facto populations are useful in planning for and coping with a disaster, especially those of daytime populations and seasonal amenity seeking populations. Here, however, we are interested in the impact of a disaster. In this regard we also note that there are two distinct groups of interest: (1) the population remaining in an area in which a disaster occurred; and (2) the population dispersed by the disaster. In regard to the former, the location is generally easy to define (Swanson 2008; Swanson et al. 2007; Swanson et al. 2009) while the latter is less easily defined because of the nature of dispersion (Henderson et al. 2009 Smith and McCarty 1996). Here, we provide an overview of methods used to estimate both groups. We note that these methods, like those used to

estimate visitor and homeless populations are largely time and resource intensive in that all three are ephemeral. One major difference in developing estimates for visitor vs. homeless and disaster impacted populations is that the direct data needed for the latter are usually collected under difficult – even dangerous –circumstances (Applied Survey Research 2007; Swanson et al. 2007). On the plus side, "pre-disaster" data are available (Swanson et al. 2007).

As an example of developing an estimate for the area in which a disaster occurred, we turn to the study of Hurricane Katrina on the Mississippi Gulf Coast (Swanson et al. 2007). As one of nine "social network" post-Katrina research projects funded by the National Science Foundation under the provisions of the SGER program, this study required $96, 212 in funding to accomplish two major tasks:

(1) gather pre- and post-Katrina information on housing and population from 573 targeted census blocks at the epicenter of Katrina's impact on the Mississippi gulf coast that the 2000 census showed as containing people (the "Short Form"); and
(2) employ a random start, systematic selection, cluster sample targeting 126 of these 573 blocks for administration of a 115-item questionnaire (the "Long Form"), such that at least 350 completed questionnaires would be obtained. The Long Form was designed for several purposes, one of which was to collect retrospective information on the roles that social and kinship networks played in determining respondents' success (i.e., the capacity for respondents to sustain their physical and emotional well-being after Hurricane Katrina).

Before Katrina stuck, there were 8,535 (permanent) housing units in the 346 blocks that were canvassed, an increase of nearly 10% over the Census 2000 count of 7,793. Of the 8,555 housing units in study area, 2,227 (27%) were destroyed and 3,997 substantially damaged (47%), leaving 2,261 habitable (26%). There were 2,012 temporary units found the Study Area after Katrina struck, of which 94% were occupied.

There were approximately 16,540 people residing in 6,486 (occupied) permanent housing units in the 346 blocks as of Census 2000. Just prior to the impact of Katrina on August 29th, 2005, there were approximately 7,100 occupied permanent housing units (83% of the total number of permanent housing units) containing 18,105 people in these same 346 blocks After Katrina stuck, the study found approximately 10, 950 people residing in 3,938 permanent and temporary housing units in these same 346 blocks. At the time of Census 2000 and just prior to when Katrina struck, the average number of persons per household (PPH) in the Study Area was 2.55. Subsequent to Katrina the PPH was 2.78.

Thus, for the 346 blocks comprising the study area (Swanson et al. 2007) found that Hurricane Katrina resulted in:

(1) a decline of 7,155 for the household population – a 40% drop from the pre-Katrina household population of 18,105;8 and
(2) an increase of 0.23 persons per household– a 9% increase from the pre-Katrina PPH of 2.55.

The preceding estimates done by Swanson et al. (2007) are consistent with the special estimates of Hancock and Harrison counties that the Census Bureau

released for January of 2006. These estimates were designed to show the impact of Katrina in the 117 counties designated by the Federal Emergency Management Agency (FEMA) as being eligible for individual and public assistance (US Census Bureau 2006).

In a larger study, Swanson et al. (2009) extended their estimates to include New Orleans and other areas of Louisiana directly impacted by Hurricane. They found relative to what had been projected for the zipcode impacted by Katrina, the hurricane had resulted in 311,150 fewer people expected in the absence of its impact. For the 18 zipcodes in Orleans Parish (The city of New Orleans), the impact was a reduction of 203,198 people. As these estimates suggest, the pre-Katrina population was elsewhere. Frey, Singer and Park (2007) found where much of the Pre-Katrina population had moved, at least in terms of the city of New Orleans.

Using data from the 2006 American Community Survey along with data from other sources, such as Census Bureau estimates and Internal Revenue Service migration data (see Chapter 12 for a discussion of these data and how they can be used to estimate migration), Frey et al. (2007) analyzed population change from July 1st of 200 to July 1st, 2005 (pre-Katrina since Katrina struck in August of 2005) with that found for July 1st 2005 to July 1st of 2006 in selected metropolitan areas in Alabama, Louisiana, Mississippi, and Texas to estimate population losses in the impact area and simultaneously estimate gains in terms of nearby receiving areas. The results are not definite, but they are suggestive. For example, Frey et al. (2007) found that Harris County, Texas (where the city of Houston is located) increased its population by 123,000 in 2005–2006. They compared this to the increase of 67,000 people for 2004–2005 and concluded that much of the increase was due to the presence of displaced people from the New Orleans area. Taking into account that some of the people displaced by Katrina went to places far from the impact area, one can get a good picture of the metropolitan areas that were themselves impacted indirectly by Katrina in terms of the movement it caused among the populations it impacted directly.

16.7 Summary

No matter how the pie is sliced, the estimation of a De Facto population in a country that depends on a De Jure concept of population is generally not a task that is easily accomplished. This is true in countries that rely on a population registry system (e.g. Finland) and a regular census (e.g., the United States). As we noted, however, some countries have census information that can be used to develop estimates for daytime and seasonal populations. In this chapter, we have provided examples of how these estimates may be accomplished. In many regards these examples should be viewed as templates that can be adjusted to different situations. For example, where the data are a bit different than those used in our examples, those seeking to develop daytime and seasonal population estimates at least have a starting point so that they can and find the data and make the necessary adjustments to develop

the estimates of these populations. To this end, we believe that the general equation we described for estimating a De Facto population serves as a useful point of reference - or departure.

While it is clear that at the national level, there are countries that have information on international visitors, we are not aware of any county, however, that can easily develop estimates of visitor populations, both domestic and international for subnational areas. As our example shows, in the United States, Hawai'i is virtually unique in this regard since visitors can arrive only by air or sea and because of its economic dependence on visitors, it has developed a sophisticated system for estimating visitors to the state as a whole, and selected subareas.

Like the estimates of visitor populations, those for homeless and disaster impacted populations are time and resource intensive. Some of these needs can be reduced by relying on "off the shelf" methods developed by US HUD (2008a, 2008b) for the homeless and Centers such as the National Hazards Center at the University of Colorado at Boulder or the Disaster Research Center at the University of Delaware for populations impacted by disasters. Along with the "off-the-shelf" methods, there is, of course, a great deal of knowledge and experience in homeless research at US HUD and local jurisdictions seeking its funding for the homeless, and in disaster research at the National Centers, to include methods to estimate the demographic impacts of natural and man-made disasters.

Endnote

1. For those interested in the nuances of defining populations, both De Facto and De Jure, we recommend the book edited by Dan Cork and Paul Voss (2006).

References

Akkerman, A. (2000). "The Diurnal Cycle of Regional Commuter Systems: North Wales, 1991." *Geographical Analysis* 32: 247–266.

Applied Survey Research. (2007). *Southern Nevada Homeless Census and Survey*. Watsonville, CA. Applied Survey Research.

Becker, P.. I. Kincannon, and M. Wyckoff. (1996). Northwest Michigan Seasonal Population Model. Report prepared for the Northwest Michigan Council of Governments. Traverse City, MI: Northwest Michigan Council of Governments.

Bhaduri, B., Bright, E., Coleman, P., Urban, M. (2007). "LandScan USA: A High Resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics." *GeoJournal* 69: 103–117.

Cai, Q., G. Rushton, B. Bhaduri, E. Bright, and P. Coleman. (2006). "Estimating Small-area Populations by Age and Sex Using Spatial Interpolation and Statistical Inference Methods." *Transactions in GIS* 10: 577–598

Cheriyadat, A., Bright, E.A., Bhaduri, B., and D. Potere. (2007). "Mapping of Settlements in High Resolution Satellite Imagery using High Performance Computing." *GeoJournal* 69:119–129.

Cook, T. (1996). “When ERPs Aren’t Enough: A Discussion of Issues Associated with Service Population Estimation. Demography Working Paper 1996/4. Canberra, Australia: Australian Bureau of Statistics (http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/3112.0Main%20Features11996?opendocument&tabname=Summary&prodno=3112.0&issue=1996&num=&view=).

Cork, D. and P. Voss. (2006). *Once, Only Once, and in the Right Place: Residence Rules in the Decennial Census*. Washington, DC: National Academies Press.

Dannemiller, J. (2011). Personal Communication to D. Swanson on homeless estimates done by SMS Research, Honolulu, Hawai’i.

Eldridge, H. (1947). “Problems and Methods of Estimating Post-censal Population.” *Social Forces* 24: 41–46.

Erkkila, D. (2000). “Trends in Tourism Economic Impact Estimation Methods.” pp 235–244 in W. Gartner and D. Lime (Eds.) *Trends in Outdoor Recreation, Leisure, and Tourism*. Oxfordhire, England: CABI Press.

Foley, D. (1954). “Urban Daytime Population: A Field for Demographic-Ecological Analysis *Social Forces* 32: 323–330

Frey, W., A. Singer, and D. Park. (2007). Resettling New Orleans: The First Full Picture from the Census. Washington, DC: The Brookings Institute (http://www.brookings.edu/reports/2007/07katrinafreysinger.aspx).

Glasser, I. (1991). An *Ethnographic Study of Homeless in Windham, Connecticut*. Ethnographic Exploratory Research Report # 17. Center for Survey Research Methods. Washington, DC U. S. Census Bureau.

Goldschmidt, P. and A. Dahl. (1976). "Demoflush: Estimating Population in Seasonal Resort Communities." *Growth and Change* 7: 44–48.

Happel, S. and T. Hogan. (2002). “Counting Snowbirds: The Importance of and the Problems with Estimating Seasonal Populations.” *Population Research and Policy Review* 21: 227–240.

Happel, S., and T. Hogan. (1987). "Estimating the Winter Resident Population of the Phoenix Area" *Applied Demography* 3: 7–8.

Hawai’i Department of Business, Economic Development, and Tourism. (2000). *Annual Visitor Research Report 2000*. Honolulu, HI: Hawai’i Department of Business, Economic Development, and Tourism

Hawai’i Department of Business, Economic Development, and Tourism. (2001). *Hawaii’s Economy* (February). Honolulu, HI: Hawai’i Department of Business, Economic Development and Tourism.

Hawai’i Tourism Authority. (2010). *2009 Annual Visitor Research Report*. Honolulu, HI: Hawai’i Tourism Authority.

Henderson, T. M. Sirois, A. Chia-Chen Chen, C. Airriess, D. A. Swanson, and D. Banks. (2009). “After a Disaster: Lessons in Survey Methodology from Hurricane Katrina.” *Population Research and Policy Review* 28: 67–92

Kandel, W. (2008). *A Profile of Hired Farm workers, a 2008 Update*. Economic Research Report No. 60, Economic Research Service. Washington, DC: U. S. Department of Agriculture.

Kavanaugh, P. and K. Lamphere. (1989). “Estimating Daytime Population in the San Diego Region. *Applied Demography* 4 (Summer): 7–11.

Larson, A. (2000). *Migrant and Seasonal Farm worker Enumeration Profiles Study*. Final Report prepared for the Migrant Health Program, Bureau of Primary Health Care, Health Services and Resources Administration. Washington, DC: U. S Department of Health and Human Services (http://docs.google.com/viewer?a=v&q=cache:UTCxjTe9ufoJ:www.ncfh.org/enumeration/PDF11%2520Washington.pdf+estimate+apples+washington+%22migrant+workers%22&hl=en&gl=us&pid=bl&srcid=ADGEEShAuYUuQsFfey7n9Q_xKXsYul205RBFpmui-THXJD9pFo0MqkWghtzGF-YbspgKFvWAKJtb5bEG7MjcLnwY6MbKZg0FoB1kSM0N-VycUaqIKrk0-AFTHVAutcPrZbdWRW9cyofnrm&sig=AHIEtbQGXA-JZfWQaxVCp6OupPrXLsPnAw).

Las Vegas Convention and Visitors Authority. (2011a). *Las Vegas Visitor Profile Study, 2010*. Las Vegas, NV: Las Vegas Convention and Visitors Authority. (http://www.lvcva.com/press/statistics-facts/visitor-stats.jsp).

Las Vegas Convention and Visitors Authority. (2011b). *Year to Date Visitor Statistics Year End Summary for 2010*. Las Vegas, NV: Las Vegas Convention and Visitors Authority (http://www.lvcva.com/press/statistics-facts/visitor-stats.jsp).

Leeworthy, V. (1996). *Technical Appendix: Sampling Methodologies and Estimation Methods Applied to the Florida Keys/Key West Visitors Surveys*. Silver Spring MD: US Silver Spring, MD: National Oceanic and Atmospheric Administration (http://sanctuaries.noaa.gov/science/socioeconomic/floridakeys/pdfs/vistechappen9596.pdf).

Rummel, R. (1991). *China's Bloody Century*. New Brunswick, NJ: Transaction Publishers

Salo, M. (1990). "1988-89 Exploratory Research on Enumerating Homeless Individuals in Baltimore and Washington." *SRD Research Report Series RSM 2007–29*. Statistical Research Division. Washington, DC" U. S. Census Bureau.

San Francisco Planning Department. (1997). Community Safety Element, San Francisco General Plan, San Francisco, CA: City and County of San Francisco (http://www.sf-planning.org/ftp/general_plan/I8_Community_Safety.htm)

Schmitt. R. (1968). "Travel, Tourism, and Migration." *Demography* 5: 306–310

Schmitt, R. (1956). "Estimating Daytime Populations." *Journal of the American Planning Association* 22(2): 83–85.

Smith, S. K. (1989). "Toward a Methodology for Estimating Temporary Residents." *Journal of the American Statistical Association* 84: 30–436

Smith, S. K. and C. McCarty. (1996). "Demographic Effects of Natural Disasters: A Case Study of Hurricane Andrew." *Demography* 33: 265–275.

Swanson, D.A. (2008). "The Demographic Effects of Hurricane Katrina on the Mississippi Gulf Coast: An Analysis by Zipcode." *Journal of the Mississippi Academy of Sciences* 53: 213–231.

Swanson D.A., R, Forgette, M. Van Boening, C. Holley, and A. Kinnell. (2007). "Assessing Katrina's Demographic and Social Impacts on the Mississippi Gulf Coast." *Journal of the Mississippi Academy of Sciences* 52: 228–42.

Swanson, D. A. and P. Walashek. (2011). CEMAF as a Census Method: A Proposal for a Re-Designed Census and an Independent Census Bureau. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Swanson, D. A., J. McKibben, L. Wombold, R. Forgette, and M. Van Boening. (2009). "The Demographic Effects of Katrina: An Impact Analysis Perspective." *The Open Demography Journal* 2: 36–46.

Tyrrell, T., and R. Johnston. (2002). "Estimating Regional Visitor Numbers." *Tourism Analysis* 7: 33–41.

United Kingdom Statistics Authority. (2001). *Method of Travel to Work - Daytime Population (2001 Census)*. London, England: UK Statistics Authority (http://data.gov.uk/dataset/method_of_travel_to_work_-_daytime_population_2001_census).

US Census Bureau n.d. (no date). *1980 Census Overview* (http://www.census.gov/history/www/through_the_decades/overview/).

US Census Bureau. (2011). DP-1, Profile of General Demographic Characteristics: 2000 Census 2000 Summary File 2 (SF 2) 100-Percent Data, Clark County, Nevada.

US Census Bureau. (2005). *Census 2000 PHC-T-40. Estimated Daytime Population and Employment-Residence Ratios: 2000* (http://www.census.gov/population/www/socdemo/daytime/daytimepop.html).

US Census Bureau. (2004). Historical Census of Housing Tables: Vacation Homes (http://www.census.gov/hhes/www/housing/census/historic/vacation.html).

US Census Bureau. (2001). DP-1. Profile of General Demographic Characteristics: 2000 Data Set: Census 2000 Summary File 1 (SF 1) 100-Percent Data Geographic Area: Chelan County, Washington (http://factfinder.census.gov/servlet/QTSubjectShowTablesServlet?_ts=326725807875).

US Census Bureau. (1982). Nonpermanent Residents by States and Selected Counties and Incorporated Places. PC80-S1-6, Supplementary Reports, 1980 Census of Population. Washington, DC: US Census Bureau. (http://babel.hathitrust.org/cgi/pt?id=umn.31951002887576d).

US Census Bureau. (2006). Special Population Estimates for Impacted Counties in the Gulf Coast Area. Washington, DC: U.S. Census Bureau. (http://www.census.gov/Press-Release/www/2005/katrina.htm).

US Department of Agriculture. (2008). *2007 Census of Agriculture*. National Agricultural Statistics Service. Washington, DC: U. S. Department of Agriculture (http://quickstats.nass.usda.gov/results/CCFFADA6-7F38-39CF-89DE-F4B3DC76D359).

US Department of Housing and Urban Development. (2008a). *A Guide to Counting Sheltered Homeless People*. Office of Community Planning and Development. Washington, DC: U. S. Department of Housing and Urban Development.

US Department of Housing and Urban Development. (2008b). *A Guide to Counting Unsheltered Homeless People*. Office of Community Planning and Development. Washington, DC: U. S. Department of Housing and Urban Development.

Van Auken, P. R. Hammer, P. Voss, and D. Veroff. (2006). "The American Community Survey in Counties with 'Seasonal' Populations." *Population Research and Policy Review* 25: 275–292.

Watson, A., D. Cole, D. Turner, P. Reynolds. (2000). *Wilderness Recreation Use Estimation: A Handbook of Methods and Systems*. General Technical Report RMRS-GTR-56. US Department of Agriculture, Forestry Service. Ogden, UT: Rocky Mountain Research Station.

Wicks, J., R. Vincent, D. Swanson, and J. Luiz Pereira De Almeida. (1999). "Population Estimates from Remotely Sensed Data: A Discussion of Recent Technological Developments and Future Research Plans." Presented at the Annual Meeting of the Canadian Population Society, Lennoxville, Quebec, Canada.

Chapter 17
Historical Estimates

In this chapter, we consider methods that can be used to estimate populations at past points in time. Our focus is on the near past rather than the distant past. By this we mean that we primarily examine methods that can be used for inter-censal estimates in areas that have adequate census and vital statistics coverage, but for which there are historical gaps. An example of such a gap would be that if one wants to know the population of the state of Washington in 1938, one must turn to an estimation method. The state has good census (and vital statistics coverage), and its state demographic center started producing annual state level estimates in 1944 (State of Washington 1944). However, there was neither a census conducted in 1938 nor was there an estimate made by either the state's demographic center or the US Census Bureau (Shryock and Lawrence 1949).

To a large degree, this means that we are discussing inter-censal estimates in developed countries, methods for which often overlap with those used to develop post-censal estimates. However, we include a brief discussion of methods that can be used to develop pre-censal estimates. Further, as noted earlier (e.g. chapters 8 and 9), methods used to develop pre-censal estimates overlap to a fair degree methods used to develop population information from incomplete data (Brass et al. 1968; Carrier and Hobcraft 1971; United Nations 1983), which means that they are applicable to less developed countries. To a far lesser extent the methods used to develop inter-censal estimates overlap with incomplete data methods.

Having good historical population information is not just an academic area of interest (Nordyke 1989, Schmitt 1968). It provides a foundation for current estimates and projections (Smith et al. 2001: 172-176; 323-326; 352).

17.1 Inter-censal Methods

Interpolation methods are a well-established technique in the field of demography and have a wide range of uses (Judson and Popoff 2004). Here, we focus on two simple extrapolation methods that were discussed in chapter 6 that can be used as

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_17, © Springer Science+Business Media B.V. 2012

interpolation methods to develop inter-censal estimates. The first is arithmetic and the second is geometric. We look at them both as examples of a broad range of extrapolative techniques that can be used for purposes of inter-censal interpolation.

Arithmetic Interpolation. Recall from chapter 6 that the formula used to determine an arithmetic measure of change is:

$$\Delta = (P_l - P_b)/(y) \tag{17.1}$$

where Δ is the average absolute change, P_l is the population in the launch year, P_b is the population in the base year, and y is the number of years in the base period (i.e., the number of years between the base year, b, and the launch year, l). As noted in chapter 6, it can be used to extrapolate population change in order to obtain a post-censal estimate:

$$P_t = P_l + [(z)(\Delta)] \tag{17.2}$$

where P_t is the population in the target year, P_l is the population in the launch year, and z is the number of years in the post-censal estimation horizon (i.e., the number of years between the target year, t, and the launch year, l), and Δ is the average absolute change computed for the base period.

For purposes of inter-censal estimation, equations [17.1] and [17.2] can be restated as follows, respectively

$$\Delta = (P_{b+y} - P_b)/(y) \tag{17.3}$$

where Δ is the average absolute change, P_{b+y} is the population in the census following the census at time b (the base year), P_b is the population in the base year, and y is the number of years in the inter-censal period (i.e., the number of years between the base year census and the successive census forming the inter-censal period).

$$P_{b+x} = P_b + [(y)(\Delta)] \tag{17.4}$$

where we want to develop an estimate between P_b and P_{b+x}, which represent two successive census years and P_{b+x} is the population in the estimation year, P_b is the population in the base year, x is the number of years in the inter-censal estimation horizon (i.e., the number of years between the estimation year and the base year, b, with $1 \leq x \leq 9$), and Δ is the average absolute change computed for the period between P_b and P_{b+y}.

As an example of arithmetic interpolation, suppose we want to estimate the population of the 39 counties of Washington from 1991 to 1999. Using the 1990 and 2000 census data shown in Table 17.1, each county's average absolute (annual) change is found in right-most column (using equation [17.3], which was used in conjunction with equation [17.4] to develop the annual estimates from 1991 to 1999.

Geometric Interpolation. As discussed in chapter 6, this method assumes that a population will change by the same percentage rate over a given increment of time. The average geometric ratio of population change during the period between the two successive census counts that border the inter-censal period for which estimates are desired can be computed as:

$$r = [(P_{b+y}/P_b)^{(1/y)}] \tag{17.5}$$

where *r* is the average geometric ratio of change, P_{b+y} is the population in the census following the census at time b (the base year), P_b is the population in the base year, and y is the number of years in the base period. An inter-censal estimate using this method can be computed as:

$$P_{b+x} = (P_b)[(r)^x] \tag{17.6}$$

where P_{b+x} is the population in the inter-censal estimate year, P_b is the population in the base year, *r* is the average geometric ratio of change, and x is the number of years in the inter-censal estimation horizon (i.e., the number of years between the estimation year and the base year, b, with $1 \leq x \leq 9$).

As an example of geometric interpolation, we again use the example of the 39 counties of Washington from 1991 to 1999. Using the 1990 and 2000 census data shown in Table 17.2, each county's average (annual) geometric ratio of change is found in right-most column (using equation [17.5], which was used in conjunction with equation [17.6] to develop the annual estimates from 1991 to 1999.

The estimates in tables 17.1 and 17.2 are very similar. For example, the 1995 estimates of Garfield County's population (the smallest in the state) using the arithmetic and geometric methods are 2,323 and 2,336, respectively. For King County, (the county with the largest population in Washington), the estimates using the arithmetic and geometric methods are 1,622,176 and 1,618,103, respectively.

Assessment of Arithmetic and Geometric Backcasting. Clearly, inter-censal estimates cannot go too far astray, but it is difficult to assess their accuracy unless one calibrates the models at censuses twenty years apart and then uses the models to estimate populations for the intervening census, which can then be evaluated against the numbers in the intervening census. The major accuracy issues are sudden population changes that take place in the inter-censal period and then subside. Here, we would be especially concerned about directional changes. This type of change is not common at high levels of geography (e.g., nation, state) and lower levels with large populations (e.g., counties in a major metropolitan area), but one must be cautious about how accurate the estimates are in any given year for lower levels of geography with small populations (e.g., counties in rural areas, census tracts, blockgroups, blocks, and zip codes). One way to gain more precise estimates when sudden population changes are possible is to use information beyond the two census counts forming the inter-censal period of interest. Here, we are thinking of administrative records that are available annually during the inter-censal period, which brings to mind ratio-correlation and its variants.

Table 17.1 1990 and 2000 Census Data for Washington State Counties and Interpolated Population Estimates, 1991 to 1999, using the Arithmetic Method.

	1990 Census	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000 Census	R
WA State	4,866,663	4,969,411	5,072,159	5,174,907	5,277,655	5,380,403	5,483,151	5,585,899	5,688,647	5,791,395	5,894,143	102,748
Adams	13,603	13,886	14,168	14,451	14,733	15,016	15,298	15,581	15,863	16,146	16,428	283
Asotin	17,605	17,900	18,194	18,489	18,783	19,078	19,373	19,667	19,962	20,256	20,551	295
Benton	112,560	115,552	118,543	121,535	124,526	127,518	130,509	133,501	136,492	139,484	142,475	2,992
Chelan	52,250	53,687	55,123	56,560	57,996	59,433	60,870	62,306	63,743	65,179	66,616	1,437
Clallam	56,204	57,002	57,799	58,597	59,394	60,192	60,989	61,787	62,584	63,382	64,179	798
Clark	238,053	248,772	259,490	270,209	280,927	291,646	302,364	313,083	323,801	334,520	345,238	10,719
Columbia	4,024	4,028	4,032	4,036	4,040	4,044	4,048	4,052	4,056	4,060	4,064	4
Cowlitz	82,119	83,202	84,285	85,368	86,451	87,534	88,616	89,699	90,782	91,865	92,948	1,083
Douglas	26,205	26,845	27,485	28,124	28,764	29,404	30,044	30,684	31,323	31,963	32,603	640
Ferry	6,295	6,392	6,488	6,585	6,681	6,778	6,874	6,971	7,067	7,164	7,260	97
Franklin	37,473	38,660	39,848	41,035	42,223	43,410	44,597	45,785	46,972	48,160	49,347	1,187
Garfield	2,248	2,263	2,278	2,293	2,308	2,323	2,337	2,352	2,367	2,382	2,397	15
Grant	54,798	56,788	58,778	60,768	62,758	64,748	66,738	68,728	70,718	72,708	74,698	1,990
Grays Harbor	64,175	64,477	64,779	65,081	65,383	65,685	65,986	66,288	66,590	66,892	67,194	302
Island	60,195	61,331	62,468	63,604	64,740	65,877	67,013	68,149	69,285	70,422	71,558	1,136
Jefferson	20,406	20,995	21,585	22,174	22,763	23,353	23,942	24,531	25,120	25,710	26,299	589
King	1,507,305	1,530,279	1,553,253	1,576,227	1,599,201	1,622,176	1,645,150	1,668,124	1,691,098	1,714,072	1,737,046	22,974
Kitsap	189,731	193,955	198,179	202,402	206,626	210,850	215,074	219,298	223,521	227,745	231,969	4,224
Kittitas	26,725	27,389	28,052	28,716	29,380	30,044	30,707	31,371	32,035	32,698	33,362	664
Klickitat	16,616	16,871	17,125	17,380	17,634	17,889	18,143	18,398	18,652	18,907	19,161	255
Lewis	59,358	60,282	61,206	62,131	63,055	63,979	64,903	65,827	66,752	67,676	68,600	924
Lincoln	8,864	8,996	9,128	9,260	9,392	9,524	9,656	9,788	9,920	10,052	10,184	132
Mason	38,341	39,447	40,554	41,660	42,767	43,873	44,979	46,086	47,192	48,299	49,405	1,106
Okanogan	33,350	33,971	34,593	35,214	35,836	36,457	37,078	37,700	38,321	38,943	39,564	621
Pacific	18,882	19,092	19,302	19,513	19,723	19,933	20,143	20,353	20,564	20,774	20,984	210

Pend Oreille	8,915	9,197	9,478	9,760	10,042	10,324	10,605	10,887	11,169	11,450	11,732	282
Pierce	586,203	597,665	609,126	620,588	632,049	643,511	654,972	666,434	677,895	689,357	700,818	11,462
San Juan	10,035	10,439	10,843	11,248	11,652	12,056	12,460	12,864	13,269	13,673	14,077	404
Skagit	79,545	81,888	84,232	86,575	88,919	91,262	93,605	95,949	98,292	100,636	102,979	2,343
Skamania	8,289	8,447	8,606	8,764	8,922	9,081	9,239	9,397	9,555	9,714	9,872	158
Snohomish	465,628	479,668	493,707	507,747	521,786	535,826	549,866	563,905	577,945	591,984	606,024	14,040
Spokane	361,333	366,994	372,654	378,315	383,975	389,636	395,297	400,957	406,618	412,278	417,939	5,661
Stevens	30,948	31,860	32,772	33,683	34,595	35,507	36,419	37,331	38,242	39,154	40,066	912
Thurston	161,238	165,850	170,461	175,073	179,685	184,297	188,908	193,520	198,132	202,743	207,355	4,612
Wahkiakum	3,327	3,377	3,426	3,476	3,526	3,576	3,625	3,675	3,725	3,774	3,824	50
Walla Walla	48,439	49,113	49,787	50,461	51,135	51,810	52,484	53,158	53,832	54,506	55,180	674
Whatcom	127,780	131,685	135,589	139,494	143,398	147,303	151,208	155,112	159,017	162,921	166,826	3,905
Whitman	38,775	38,972	39,168	39,365	39,561	39,758	39,954	40,151	40,347	40,544	40,740	197
Yakima	188,823	192,199	195,575	198,950	202,326	205,702	209,078	212,454	215,829	219,205	222,581	3,376

Table 17.2 1990 and 2000 Census Data for Washington State Counties and Interpolated Population Estimates, 1991 to 1999, using the Geometric Method.

	1990 Census	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000 Census	R
WA State	4,866,663	4,960,783	5,056,723	5,154,518	5,254,205	5,355,820	5,459,400	5,564,983	5,672,608	5,782,315	5,894,143	1.01934
Adams	13,603	13,862	14,126	14,395	14,669	14,949	15,234	15,524	15,820	16,121	16,428	1.019049
Asotin	17,605	17,880	18,158	18,441	18,729	19,021	19,318	19,619	19,925	20,235	20,551	1.015593
Benton	112,560	115,244	117,993	120,807	123,688	126,637	129,657	132,749	135,915	139,156	142,475	1.023848
Chelan	52,250	53,535	54,851	56,200	57,582	58,997	60,448	61,934	63,457	65,017	66,616	1.024588
Clallam	56,204	56,955	57,715	58,486	59,268	60,059	60,862	61,674	62,498	63,333	64,179	1.013357
Clark	238,053	247,069	256,426	266,138	276,218	286,679	297,537	308,806	320,501	332,640	345,238	1.037874
Columbia	4,024	4,028	4,032	4,036	4,040	4,044	4,048	4,052	4,056	4,060	4,064	1.00099
Cowlitz	82,119	83,143	84,179	85,228	86,290	87,366	88,455	89,557	90,674	91,804	92,948	1.012464
Douglas	26,205	26,784	27,375	27,980	28,598	29,229	29,875	30,535	31,209	31,898	32,603	1.022086
Ferry	6,295	6,385	6,477	6,570	6,665	6,760	6,857	6,956	7,056	7,157	7,260	1.014365
Franklin	37,473	38,519	39,594	40,699	41,835	43,002	44,202	45,436	46,704	48,007	49,347	1.027908
Garfield	2,248	2,262	2,277	2,292	2,306	2,321	2,336	2,351	2,366	2,382	2,397	1.006438
Grant	54,798	56,522	58,301	60,135	62,027	63,979	65,992	68,068	70,210	72,419	74,698	1.031465
Grays Harbor	64,175	64,471	64,768	65,066	65,366	65,667	65,970	66,274	66,579	66,886	67,194	1.004608
Island	60,195	61,245	62,313	63,400	64,506	65,631	66,776	67,941	69,126	70,331	71,558	1.017442
Jefferson	20,406	20,930	21,468	22,020	22,586	23,166	23,761	24,372	24,998	25,640	26,299	1.025695
King	1,507,305	1,528,840	1,550,684	1,572,839	1,595,310	1,618,103	1,641,222	1,664,670	1,688,454	1,712,578	1,737,046	1.014287
Kitsap	189,731	193,583	197,513	201,524	205,615	209,790	214,049	218,395	222,829	227,353	231,969	1.020303
Kittitas	26,725	27,324	27,937	28,564	29,205	29,860	30,529	31,214	31,914	32,630	33,362	1.02243
Klickitat	16,616	16,854	17,096	17,342	17,591	17,843	18,099	18,359	18,623	18,890	19,161	1.014353
Lewis	59,358	60,223	61,101	61,992	62,895	63,812	64,742	65,686	66,643	67,614	68,600	1.014576
Lincoln	8,864	8,988	9,114	9,241	9,370	9,501	9,634	9,769	9,905	10,044	10,184	1.013979
Mason	38,341	39,325	40,335	41,371	42,433	43,523	44,640	45,787	46,962	48,168	49,405	1.025677
Okanogan	33,350	33,925	34,509	35,104	35,709	36,324	36,950	37,587	38,235	38,894	39,564	1.017233
Pacific	18,882	19,082	19,285	19,489	19,696	19,905	20,116	20,330	20,546	20,764	20,984	1.010611
Pend Oreille	8,915	9,163	9,418	9,680	9,950	10,227	10,512	10,804	11,105	11,414	11,732	1.027839

Pierce	586,203	596,766	607,518	618,465	629,609	640,954	652,503	664,260	676,229	688,414	700,818	1.018019
San Juan	10,035	10,380	10,738	11,107	11,490	11,885	12,295	12,718	13,156	13,609	14,077	1.034426
Skagit	79,545	81,626	83,761	85,952	88,200	90,507	92,874	95,303	97,796	100,354	102,979	1.026156
Skamania	8,289	8,435	8,584	8,735	8,889	9,046	9,205	9,368	9,533	9,701	9,872	1.017631
Snohomish	465,628	478,062	490,828	503,935	517,392	531,208	545,393	559,957	574,910	590,262	606,024	1.026704
Spokane	361,333	366,630	372,005	377,458	382,992	388,607	394,304	400,084	405,949	411,901	417,939	1.01466
Stevens	30,948	31,758	32,588	33,441	34,315	35,213	36,134	37,079	38,049	39,045	40,066	1.026158
Thurston	161,238	165,345	169,557	173,877	178,306	182,848	187,506	192,283	197,181	202,204	207,355	1.025474
Wahkiakum	3,327	3,374	3,421	3,469	3,518	3,567	3,617	3,668	3,719	3,771	3,824	1.01402
Walla Walla	48,439	49,074	49,718	50,370	51,030	51,700	52,378	53,065	53,761	54,466	55,180	1.013115
Whatcom	127,780	131,233	134,779	138,421	142,162	146,004	149,949	154,001	158,163	162,437	166,826	1.027023
Whitman	38,775	38,967	39,160	39,354	39,549	39,745	39,942	40,140	40,339	40,539	40,740	1.004956
Yakima	188,823	191,954	195,138	198,374	201,664	205,008	208,408	211,864	215,378	218,950	222,581	1.016584

Rate-Correlation, a Variant of Ratio-Correlation. Another method that can be used for inter-censal estimates is the ratio-correlation method, which was discussed at some length in chapter 8. Here, we will focus on the rate-correlation variant of ratio-correlation and use it as an example since an example of the post-censal use of ratio-correlation was given in chapter 8. To start, recall that in chapter 8, the ratio-correlation method is defined as follows.

$$P_{i,t} = a_0 + \Sigma(b_j)^* S_{i,j,t} + \varepsilon_i \tag{17.7a}$$

where

a_0 = the intercept term to be estimated
b_j = the regression coefficient to be estimated
ε_i = the error term
j = symptomatic indicator ($1 \leq j \leq k$)
i = subarea ($1 \leq j \leq n$)
t = year of most recent census forming the inter-censal period
and

$$P_{i,t} = (P_{i,t}/\Sigma P_{i,t})/(P_{i,t-z}/\Sigma P_{i,t-z}) \tag{17.7b}$$

$$S_{i,j,t} = (S_{i,t}/\Sigma S_{i,t})_j/(S_{i,t-z}/\Sigma S_{i,t-z})_j \tag{17.7c}$$

where

z = number of years between each census for which
data are used to construct the model
p = population
s = symptomatic indicator

Once a ratio-correlation model is constructed, a set of population estimates for time t-x (where $x \leq z$) is developed in a series of six steps. First, $(S_{i,t-x}/\sum S_{i,t-x})_j$ is substituted into the numerator of the right side of equation 17.7c for each symptomatic indicator j and $(S_{i,t-z}/\sum S_{i,t-z})_j$ into the denominator of the right side of equation 17.7c for each symptomatic indicator j, which yields $S_{i,j,t+x}$. Second, the updated model with the preceding substitution of symptomatic data for time t-x is used to estimate $P_{i,t-x}$. Third, $(P_{i,t-z}/\sum P_{i,t-z})$ is substituted into the denominator of $P_{i,t-x}$, which yields $P_{i,t-x} = (P_{i,t-x}/\sum \mathbf{P_{i,t-x}})/(P_{i,t-z}/\sum P_{i,t-z})$, where $\sum \mathbf{P_{i,t-x}}$ represents the independently estimated population of the "parent" area of the i subareas for time t-x (Note that this estimate is given in boldface and is done by a method exogenous to the ratio-correlation model (e.g., a component method)). Fifth, since $P_{i,t-x}$, $(P_{i,-z}/\sum P_{i,t-z})$ and $\sum \mathbf{P_{i,t-x}}$ are all known values, the equation $P_{i,t-x} = (P_{i,t-x}/\sum \mathbf{P_{i,t-x}})/(P_{i,t-z}/\sum P_{i,t-z})$ is manipulated to yield an estimate of the population of area i at time t + k:

$$(P_{i,t-x})^*(P_{i,t-z}/\Sigma P_{i,t-z})^*(\Sigma \mathbf{P}_{i,t-x}) = \hat{P}_{i,t-x} \tag{17.7d}$$

As equation 17.7d shows, it is important to remember that an independent estimate of the population for the "parent" geography ($\sum \mathbf{P_{i,t\text{-}x}}$) of the i subarea is required when using the ratio-correlation model to generate population estimates. The sixth and final step is to effect a final "control" so that the sum of the i subarea population estimates is equal to the independently estimated population for the parent of these i subareas: $\sum P_{i,t\text{-}x} = \sum \mathbf{P_{i,t\text{-}x}}$, which is accomplished as follows:

$$P_{i,t-x} = (P_{i,t-x}/\Sigma P_{i,t-x})^{*}(\Sigma \mathbf{P}_{i,t-x}). \tag{17.7e}$$

In the rate-correlation variant of the ratio-correlation method, population change is viewed from an exponential perspective and a corresponding logarithmic transformation is used on the independent and dependent variables along the length of time between the successive census counts forming the inter-censal period for which estimates are desired. That is, the model develops an exponential estimate of the (annual) rate of change. In developing the model, this means that equation [17.7a] becomes

$$[((\ln(P_{i,t}))/(t-z)] = a_0 + \Sigma(b_j)^{*}[(\ln(S_{i,j,t}))/(t-z))] + \varepsilon_i \tag{17.8a}$$

and equations 17.7b through 17.7d are modified accordingly

As an example of the rate-correlation model, we use the same 1990 and 2000 input data with which the ratio-correlation model was constructed as the example in chapter 8. We then use the rate-correlation model constructed using these data to generate an estimate for 1995. Exhibit 17.1 provides the model. The basic data are found in tables 8.2a through 8.2d.

Exhibit 17.1 Example Rate-Correlation Model

$$[(\mathrm{Ln}(P_{i,t}))/10] = 0.195 + (0.0933^{*}[(\ln(\mathrm{Voters}))/10] + (0.3362^{*}[(\ln(\mathrm{Autos}))10] + (0.3980^{*}[(\ln(\mathrm{Enroll}))/10] + (0.3980^{*}[(\ln(\mathrm{Enroll}))/10] \quad [p<.05][p=0.148][p<.001][p<.001]$$

where

$$P_{i,t} = (P_{i,2000}/\Sigma P_{i,2000})/(P_{i,1990}/\Sigma P_{i,1990})$$
$$S_{i,1,t} = (\mathrm{Voters}_{i,2000}/\Sigma \mathrm{Voters}_{i,2000})/(\mathrm{Voters}_{i,1990}/\Sigma \mathrm{Voters}_{i,1990})$$
$$S_{i,2,t} = (\mathrm{Autos}_{i,2000}/\Sigma \mathrm{Autos}_{i,2000})/(\mathrm{Autos}_{i,1990}/\Sigma \mathrm{Autos}_{i,1990})$$
$$S_{i,3,t} = (\mathrm{Enroll}_{i,2000}/\Sigma \mathrm{Enroll}_{i,2000})/(\mathrm{Enroll}_{i,1990}/\Sigma \mathrm{Enroll}_{i,1990})$$
$$R_2 = 0.789$$
$$\mathrm{adj}\, R_2 = 0.771$$

For the same reasons discussed in regard to the corresponding ratio-correlation model found in chapter 8, we retain the model as shown in Exhibit 17.1 to use for post-censal estimates during the period 1991-1999.

The final "controlled" population estimates are shown in Table 17.3. The appendix shows the results of these steps in detail.

Reverse Demographic Accounting. In a paper given at the annual meeting of the Population Association of America, Jerry McKibben (1988) examined the accuracy of the US Bureau of the Census's 1975 county population estimates for Indiana by comparing them with "expected Census" figures generated by the reverse demographic method. This method develops "expected" 1975 census figures by algebraically subtracting the reported number of net migrants for the period 1975-1980 from the reported 1980 census count and adding to this figure reported deaths and subtracting reported births for the same period.

This technique can be used to generate census quality inter-censal estimates for the mid-decade year (years ending in five) in the United States (and elsewhere, given similar data) when the inter-censal period is bounded by census counts that had the five-year migration question on the long form (1950 to 2000). It also can be used in conjunction with the inter-censal estimation methods discussed previously in this chapter with the mid-decade point estimate found using McKibben's technique serving as a census quality number. This means, for example that the geometric method could be used to estimate populations for the first four years following a decennial census (the years ending in 1, 2, 3, and 4) where the ratio of change is determined over the five year period from the decennial census to the subsequent mid-decade point. Similarly, the geometric method could be used for the last four years (the years ending in 6, 7, 8, and 9) where the geometric ratio of change is determined from the mid-decade point to the subsequent census. Obviously, the arithmetic method could follow a similar path while two rate-correlation models could be constructed for such an inter-censal period, the first for the period from the decennial census to the mid-decade year and the second from the mid-decade year to the subsequent census. The first model would be used to estimate the years ending in 1, 2, 3, and 4, while the second could be used for the years ending in 6, 7, 8, and 9.

17.2 Pre-censal Methods

Backward Extrapolation (Backcasting). This can be done using any model of change. For purposes of exposition, we use again here the arithmetic and geometric models.

Backward Extrapolation (Backcasting) using the Arithmetic Method. For purposes of pre-censal estimation, equations [17.1] and [17.2] can be restated as follows, respectively

$$\Delta = (P_b - P_{b+x})/(y) \tag{17.9}$$

$$P_b = P_{b+x} + [(y)(\Delta)] \tag{17.10}$$

Table 17.3 1995 County Population Estimates for the State of Washington Using the Rate-Correlation Method

Adams	14,635
Asotin	19,174
Benton	130,510
Chelan	60,608
Clallam	62,054
Clark	279,539
Columbia	4,244
Cowlitz	88,693
Douglas	29,514
Ferry	7,192
Franklin	43,487
Garfield	2,280
Grant	63,325
Grays Harbor	68,227
Island	69,264
Jefferson	23,968
King	1,622,180
Kitsap	217,061
Kittitas	30,395
Klickitat	17,829
Lewis	63,917
Lincoln	9,209
Mason	45,805
Okanogan	37,853
Pacific	20,588
Pend Oreille	10,315
Pierce	651,452
San Juan	11,804
Skagit	91,498
Skamania	9,101
Snohomish	523,122
Spokane	399,293
Stevens	34,698
Thurston	178,211
Wahkiakum	3,664
Walla Walla	52,832
Whatcom	147,001
Whitman	41,563
Yakima	210,464
State of Washington	5,396,569

where Δ is the average absolute change, P_{b+y} is the population in the census following the census at time b, P_b is the population in the backcast launch year, and y is the number of years in the inter-censal period (i.e., the number of years between the backcast launch year census and the successive census). A pre-censal estimate (prior to time b) can then be generated using equation [17.9] to find Δ, which then can be used as shown in equation [17.10] to backcast estimates prior to a census at time b.

As an example of arithmetic backcasting, suppose we want to estimate the population of the 39 counties of Washington from 1999 to 1991. Using the 2010 and 2000 census data shown in Table 17.4, each county's average absolute (annual) change is found in right-most column (using equation [17.9], which was used in conjunction with equation [17.10] to develop the annual estimates from 1999 to 1991. For purposes of assessing accuracy, we also show an estimate for 1990 along with the 1990 census population in Table 17.4

Backward Extrapolation (Backcasting) using The Geometric Method For purposes of pre-censal estimation, equations [17.3] and [17.4] can be restated as follows, respectively

$$r = [(P_b/P_{b+y})^{(1/y)}] \tag{17.11}$$

$$P_{b+x} = (P_b)[(r)^x] \tag{17.12}$$

where r is the average (annual) ratio of change, P_{b+y} is the population in the census following the census at time b, P_b is the population in the backcast launch year, and y is the number of years in the inter-censal period (i.e., the number of years between the backcast launch year census and the successive census). A pre-censal estimate (prior to time b) can then be generated using equation [17.11] to find Δ, which then can be used as shown in equation [17.12] to backcast estimates prior to a census at time b.

As an example of geometric backcasting, suppose we want to estimate the population of the 39 counties of Washington from 1999 to 1991. The 2010 and 2000 census data shown in Table 17.5 and each county's average (annual) ratio of change is (found in right-most column, using equation [17.11]) was used in conjunction with equation [17.12] to develop the annual estimates from 1999 to 1991. For purposes of assessing accuracy, we also show an estimate for 1990 along with the 1990 census population in Table 17.5.

Assessment of Arithmetic and Geometric Backcasting. Unlike the situation with inter-censal estimates, one can do an "ex post facto" assessment of the accuracy of pre-censal estimation methods, using the same approach that is used for this type of assessment post-censal methods (e.g., ratio-correlation). This is fortunate because there is much more room for pre-censal estimates to go astray than is the case for inter-censal estimates. It was for these reasons that we provided an assessment of arithmetic and geometric backcasting in tables 17.4 and 17.5, respectively

Generally, both methods show reasonable levels of accuracy at ten years away from the backcasting launch point. The arithmetic backcast has a Mean Absolute Percent Error (MAPE) of 8.70% for the estimates it generated in 1990 while the geometric backcast has a MAPE of 9.37%.

In closing our discussion of the arithmetic and geometric methods as backcasting tools, we note that Bob Schmitt (1977, 1968) appears to have developed estimates for Hawaii from 1832 to 1848 using the geometric method, which most surely was implemented via backcasting. With this, we now turn our discussion to more other forms of developing pre-censal estimates. The Hamilton-Perry Method,

Table 17.4 2010, 2000, and 1990 Census County and State Census Counts along with Backcasted Population Estimates for 1999 to 1990 using the Arithmetic Change Method

	1990 Census	EST 1990	EST - CENSUS	% Difference	ABS % DIFF	1991	1992	1993	1994	1995	1995	1997	1998	1999	2000 Census	2010 Census	R
WA State	4,866,663	5,063,746	197,083	4.05	4.05	5,146,786	5,229,825	5,312,865	5,395,905	5,478,945	5,561,984	5,645,024	5,728,064	5,811,103	5,894,143	6,724,540	−83,040
Adams	13,603	14,128	525	3.86	3.86	13,373	13,143	12,913	12,683	12,453	12,223	11,993	11,763	11,533	16,428	18,728	−230
Asotin	17,605	19,479	1,874	10.64	10.64	17,498	17,391	17,283	17,176	17,069	16,962	16,855	16,747	16,640	20,551	21,623	−107
Benton	112,560	109,773	−2,787	−2.48	2.48	109,290	106,020	102,749	99,479	96,209	92,939	89,669	86,398	83,128	142,475	175,177	−3,270
Chelan	52,250	60,779	8,529	16.32	16.32	51,666	51,083	50,499	49,915	49,332	48,748	48,164	47,580	46,997	66,616	72,453	−584
Clallam	56,204	56,954	750	1.33	1.33	55,482	54,759	54,037	53,314	52,592	51,869	51,147	50,424	49,702	64,179	71,404	−723
Clark	238,053	265,113	27,060	11.37	11.37	230,041	222,028	214,016	206,003	197,991	189,978	181,966	173,953	165,941	345,238	425,363	−8,013
Columbia	4,024	4,050	26	0.65	0.65	4,023	4,021	4,020	4,018	4,017	4,016	4,014	4,013	4,011	4,064	4,078	−1
Cowlitz	82,119	83,486	1,367	1.66	1.66	81,173	80,227	79,280	78,334	77,388	76,442	75,496	74,549	73,603	92,948	102,410	−946
Douglas	26,205	26,775	570	2.18	2.18	25,622	25,039	24,457	23,874	23,291	22,708	22,125	21,543	20,960	32,603	38,431	−583
Ferry	6,295	6,969	674	10.71	10.71	6,266	6,237	6,208	6,179	6,150	6,120	6,091	6,062	6,033	7,260	7,551	−29
Franklin	37,473	20,531	−16,942	−45.21	45.21	34,591	31,710	28,828	25,947	23,065	20,183	17,302	14,420	11,539	49,347	78,163	−2,882
Garfield	2,248	2,528	280	12.46	12.46	2,261	2,274	2,287	2,300	2,314	2,327	2,340	2,353	2,366	2,397	2,266	13
Grant	54,798	60,276	5,478	10.00	10.00	53,356	51,914	50,471	49,029	47,587	46,145	44,703	43,260	41,818	74,698	89,120	−1,442
Grays Harbor	64,175	61,591	−2,584	−4.03	4.03	63,615	63,054	62,494	61,934	61,374	60,813	60,253	59,693	59,132	67,194	72,797	−560
Island	60,195	64,610	4,415	7.33	7.33	59,500	58,805	58,111	57,416	56,721	56,026	55,331	54,637	53,942	71,558	78,506	−695
Jefferson	20,406	22,726	2,320	11.37	11.37	20,049	19,691	19,334	18,977	18,620	18,262	17,905	17,548	17,190	26,299	29,872	−357
King	1,507,305	1,542,843	35,538	2.36	2.36	1,487,885	1,468,464	1,449,044	1,429,624	1,410,204	1,390,783	1,371,363	1,351,943	1,332,522	1,737,046	1,931,249	−19,420
Kitsap	189,731	212,805	23,074	12.16	12.16	187,815	185,898	183,982	182,065	180,149	178,233	176,316	174,400	172,483	231,969	251,133	−1,916
Kittitas	26,725	25,809	−916	−3.43	3.43	25,970	25,214	24,459	23,704	22,949	22,193	21,438	20,683	19,927	33,362	40,915	−755
Klickitat	16,616	18,004	1,388	8.35	8.35	16,500	16,385	16,269	16,153	16,038	15,922	15,806	15,690	15,575	19,161	20,318	−116
Lewis	59,358	61,745	2,387	4.02	4.02	58,673	57,987	57,302	56,616	55,931	55,245	54,560	53,874	53,189	68,600	75,455	−686
Lincoln	8,864	9,798	934	10.54	10.54	8,825	8,787	8,748	8,710	8,671	8,632	8,594	8,555	8,517	10,184	10,570	−39
Mason	38,341	38,111	−230	−0.60	0.60	37,212	36,082	34,953	33,823	32,694	31,565	30,435	29,306	28,176	49,405	60,699	−1,129
Okanogan	33,350	38,008	4,658	13.97	13.97	33,194	33,039	32,883	32,728	32,572	32,416	32,261	32,105	31,950	39,564	41,120	−156
Pacific	18,882	21,048	2,166	11.47	11.47	18,888	18,895	18,901	18,908	18,914	18,920	18,927	18,933	18,940	20,984	20,920	6
Pend Oreille	8,915	10,463	1,548	17.36	17.36	8,788	8,661	8,534	8,407	8,281	8,154	8,027	7,900	7,773	11,732	13,001	−127
Pierce	586,203	606,411	20,208	3.45	3.45	576,762	567,322	557,881	548,440	539,000	529,559	520,118	510,677	501,237	700,818	795,225	−9,441
San Juan	10,035	12,385	2,350	23.42	23.42	9,866	9,697	9,527	9,358	9,189	9,020	8,851	8,681	8,512	14,077	15,769	−169
Skagit	79,545	89,057	9,512	11.96	11.96	78,153	76,761	75,368	73,976	72,584	71,192	69,800	68,407	67,015	102,979	116,901	−1,392
Skamania	8,289	8,678	389	4.69	4.69	8,170	8,050	7,931	7,811	7,692	7,573	7,453	7,334	7,214	9,872	11,066	−119
Snohomish	465,628	498,713	33,085	7.11	7.11	454,897	444,166	433,435	422,704	411,973	401,241	390,510	379,779	369,048	606,024	713,335	−10,731
Spokane	361,333	364,657	3,324	0.92	0.92	356,005	350,677	345,348	340,020	334,692	329,364	324,036	318,707	313,379	417,939	471,221	−5,328
Stevens	30,948	36,601	5,653	18.27	18.27	30,602	30,255	29,909	29,562	29,216	28,869	28,523	28,176	27,830	40,066	43,531	−347

(continued)

Table 17.4 (continued)

	1990 Census	EST 1990	EST - CENSUS	% Difference	ABS % DIFF	1991	1992	1993	1994	1995	1995	1997	1998	1999	2000 Census	2010 Census	R
Thurston	161,238	162,446	1,208	0.75	0.75	156,747	152,256	147,765	143,274	138,784	134,293	129,802	125,311	120,820	207,355	252,264	−4,491
Wahkiakum	3,327	3,670	343	10.31	10.31	3,312	3,296	3,281	3,265	3,250	3,235	3,219	3,204	3,188	3,824	3,978	−15
Walla Walla	48,439	51,579	3,140	6.48	6.48	48,079	47,719	47,359	46,999	46,639	46,278	45,918	45,558	45,198	55,180	58,781	−360
Whatcom	127,780	132,512	4,732	3.70	3.70	124,349	120,917	117,486	114,054	110,623	107,192	103,760	100,329	96,897	166,826	201,140	−3,431
Whitman	38,775	36,704	−2,071	−5.34	5.34	38,371	37,968	37,564	37,161	36,757	36,353	35,950	35,546	35,143	40,740	44,776	−404
Yakima	188,823	201,931	13,108	6.94	6.94	186,758	184,693	182,628	180,563	178,498	176,433	174,368	172,303	170,238	222,581	243,231	−2,065
Sum of Counties	4,866,663	5,063,746				4,783,623	4,700,584	4,617,544	4,534,504	4,451,465	4,368,425	4,285,385	4,202,345	4,119,306	5,894,143	6,724,540	
		MEAN (n=39)	5,053	5.56	8.70												

Table 17.5 2010, 2000, and 1990 Census County and State Census Counts along with County Backcasted Population Estimates for 1999 to 1990 using the Geometric Change Method

	1990 Census	EST 1990	EST - CENSUS	% Difference	ABS % DIFF	1991	1992	1993	1994	1995	1995	1997	1998	1999	2000 Census	2010 Census	R
WA State	4,866,663	5,166,290	299,627	6.16	6.16	5,234,834	5,304,289	5,374,664	5,445,973	5,518,229	5,591,443	5,665,629	5,740,798	5,816,965	5,894,143	6,724,540	0.986906044
Adams	13,603	14,410	807	5.94	5.94	14,601	14,793	14,988	15,186	15,386	15,589	15,795	16,003	16,214	16,428	18,728	0.98698222
Asotin	17,605	19,532	1,927	10.95	10.95	19,632	19,732	19,832	19,933	20,035	20,137	20,240	20,343	20,447	20,551	21,623	0.99492811
Benton	112,560	115,878	3,318	2.95	2.95	118,297	120,767	123,288	125,862	128,490	131,173	133,911	136,707	139,561	142,475	175,177	0.979548983
Chelan	52,250	61,249	8,999	17.22	17.22	61,766	62,287	62,812	63,342	63,876	64,415	64,958	65,506	66,059	66,616	72,453	0.991635847
Clallam	56,204	57,685	1,481	2.64	2.64	58,304	58,929	59,561	60,200	60,845	61,498	62,158	62,824	63,498	64,179	71,404	0.989388915
Clark	238,053	280,206	42,153	17.71	17.71	286,116	292,150	298,311	304,603	311,027	317,587	324,285	331,124	338,107	345,238	425,363	0.979345401
Columbia	4,024	4,050	26	0.65	0.65	4,051	4,053	4,054	4,056	4,057	4,058	4,060	4,061	4,063	4,064	4,078	0.999656163
Cowlitz	82,119	84,360	2,241	2.73	2.73	85,182	86,012	86,850	87,696	88,550	89,413	90,284	91,163	92,051	92,948	102,410	0.990352423
Douglas	26,205	27,659	1,454	5.55	5.55	28,117	28,584	29,058	29,539	30,029	30,527	31,033	31,548	32,071	32,603	38,431	0.983688486
Ferry	6,295	6,980	685	10.89	10.89	7,008	7,035	7,063	7,091	7,119	7,147	7,175	7,203	7,232	7,260	7,551	0.996077695
Franklin	37,473	31,154	-6,319	-16.86	16.86	32,621	34,156	35,764	37,447	39,209	41,055	42,987	45,010	47,129	49,347	78,163	0.955049658
Garfield	2,248	2,536	288	12.79	12.79	2,521	2,507	2,493	2,479	2,465	2,451	2,438	2,424	2,411	2,397	2,266	1.005636002
Grant	54,798	62,610	7,812	14.26	14.26	63,725	64,860	66,015	67,191	68,387	69,605	70,845	72,107	73,391	74,698	89,120	0.982501856
Grays Harbor	64,175	62,022	-2,153	-3.35	3.35	62,521	63,024	63,531	64,041	64,556	65,075	65,599	66,126	66,658	67,194	72,797	0.992022908
Island	60,195	65,225	5,030	8.36	8.36	65,832	66,445	67,064	67,688	68,318	68,954	69,596	70,244	70,898	71,558	78,506	0.990776129
Jefferson	20,406	23,153	2,747	13.46	13.46	23,450	23,751	24,055	24,364	24,676	24,992	25,313	25,637	25,966	26,299	29,872	0.987341731
King	1,507,305	1,562,372	55,067	3.65	3.65	1,579,018	1,595,842	1,612,844	1,630,028	1,647,395	1,664,947	1,682,687	1,700,615	1,718,734	1,737,046	1,931,249	0.989457865
Kitsap	189,731	214,267	24,536	12.93	12.93	215,975	217,696	219,431	221,180	222,943	224,719	226,510	228,315	230,135	231,969	251,133	0.992093528
Kittitas	26,725	27,203	478	1.79	1.79	27,764	28,337	28,921	29,517	30,126	30,747	31,381	32,028	32,688	33,362	40,915	0.979798911
Klickitat	16,616	18,070	1,454	8.75	8.75	18,176	18,283	18,391	18,499	18,607	18,717	18,827	18,938	19,049	19,161	20,318	0.994154131
Lewis	59,358	62,368	3,010	5.07	5.07	62,965	63,567	64,176	64,790	65,410	66,036	66,668	67,306	67,950	68,600	75,455	0.990520822
Lincoln	8,864	9,812	948	10.70	10.70	9,849	9,885	9,922	9,959	9,996	10,034	10,071	10,109	10,146	10,184	10,570	0.996286717
Mason	38,341	40,212	1,871	4.88	4.88	41,049	41,903	42,774	43,664	44,572	45,500	46,446	47,412	48,398	49,405	60,699	0.979622918
Okanogan	33,350	38,067	4,717	14.14	14.14	38,214	38,362	38,510	38,659	38,808	38,958	39,109	39,260	39,412	39,564	41,120	0.99614993
Pacific	18,882	21,048	2,166	11.47	11.47	21,042	21,035	21,029	21,022	21,016	21,010	21,003	20,997	20,990	20,984	20,920	1.000305507
Pend Oreille	8,915	10,587	1,672	18.75	18.75	10,696	10,807	10,918	11,031	11,145	11,260	11,376	11,493	11,612	11,732	13,001	0.98978195
Pierce	586,203	617,619	31,416	5.36	5.36	625,474	633,428	641,484	649,642	657,905	666,272	674,745	683,327	692,017	700,818	795,225	0.987441833
San Juan	10,035	12,567	2,532	25.23	25.23	12,710	12,855	13,002	13,150	13,300	13,452	13,606	13,761	13,918	14,077	15,769	0.9887138
Skagit	79,545	90,715	11,170	14.04	14.04	91,873	93,045	94,232	95,435	96,653	97,886	99,135	100,400	101,681	102,979	116,901	0.987399822
Skamania	8,289	8,807	518	6.25	6.25	8,908	9,010	9,114	9,218	9,324	9,431	9,540	9,649	9,760	9,872	11,066	0.988647444
Snohomish	465,628	514,856	49,228	10.57	10.57	523,319	531,921	540,664	549,550	558,583	567,765	577,097	586,582	596,224	606,024	713,335	0.983829021

(continued)

Table 17.5 (continued)

	1990 Census	EST 1990	EST - CENSUS	% Difference	ABS % DIFF	1991	1992	1993	1994	1995	1995	1997	1998	1999	2000 Census	2010 Census	R
Spokane	361,333	370,682	9,349	2.59	2.59	375,156	379,685	384,268	388,907	393,602	398,353	403,162	408,029	412,954	417,939	471,221	0.988072532
Stevens	30,948	36,877	5,929	19.16	19.16	37,184	37,494	37,806	38,121	38,438	38,758	39,081	39,407	39,735	40,066	43,531	0.991739781
Thurston	161,238	170,441	9,203	5.71	5.71	173,815	177,256	180,766	184,344	187,994	191,716	195,511	199,382	203,330	207,355	252,264	0.980586531
Wahkiakum	3,327	3,676	349	10.49	10.49	3,691	3,705	3,720	3,734	3,749	3,764	3,779	3,794	3,809	3,824	3,978	0.996059565
Walla Walla	48,439	51,800	3,361	6.94	6.94	52,128	52,459	52,791	53,126	53,463	53,802	54,143	54,487	54,832	55,180	58,781	0.99369813
Whatcom	127,780	138,366	10,586	8.28	8.28	140,978	143,640	146,352	149,116	151,931	154,800	157,722	160,700	163,735	166,826	201,140	0.98146887
Whitman	38,775	37,068	-1,707	-4.40	4.40	37,420	37,775	38,133	38,495	38,861	39,229	39,602	39,978	40,357	40,740	44,776	0.990598288
Yakima	188,823	203,684	14,861	7.87	7.87	205,499	207,331	209,178	211,042	212,923	214,821	216,735	218,666	220,615	222,581	243,231	0.991167188
Sum of Counties	4,866,663					5,246,645	5,314,404	5,383,166	5,452,950	5,523,773	5,595,653	5,668,611	5,742,666	5,817,836	5,894,143	6,724,540	
		MEAN (n=39)	8,031	8.10	9.37												

for example, which can be used in reverse to generate pre-censal estimates. We then follow the discussion of the Hamilton-Perry Method with one on inverse projection.

Another Backcasting Tool: The Hamilton-Perry Method in Reverse. In chapter 10, we discussed the Hamilton-Perry Method, which is a variant of the cohort-component method that has far less intensive input data requirements (Hamilton and Perry 1962; Smith et al. 2001: 153-158; Swanson et al. 2010). Running it in reverse to obtain pre-censal estimates requires data from the two earliest censuses so that we can move a population by age (and sex) backwards, from time t to time t-k using reverse cohort-change ratios (RCCR). The formula for a RCCR is:

$$ {}_{n}\mathrm{RCCR}_{i,}x = {}_{n}\mathrm{P}_{i,x-k,t-k} \,/\, {}_{n}\mathrm{P}_{i,}x,t \tag{17.13a} $$

where

${}_{n}P_{i,x\text{-}k,t\text{-}k}$ is the population aged x-k to x-k + n in area i at the earliest census (t-k),

${}_{n}P_{i,x,t}$ is the population aged x to x + n in area i at the census at time t, which follows the earliest census (t-k) for area i,

and k is the number of years between the earliest census and the one that follows it

The basic formula moving a population into the past to do an estimate (or a backcast) is:

$$ {}_{n}\mathrm{P}_{i,}x-k,t-k = ({}_{n}\mathrm{RCCR}_{i,x})^{*}({}_{n}\mathrm{P}_{i,x,t}) \tag{17.13b} $$

where

${}_{n}P_{i,x\text{-}k,t\text{-}k}$ is the population aged x-k to x-k-n in area i at earliest census (t-k)

$$ {}_{n}\mathrm{RCCR}_{i,x} = {}_{n}\mathrm{P}_{i,x\text{-}k,t\text{-}k}/{}_{n}\mathrm{P}_{i,x,t} $$

and

${}_{n}P_{i,x,t}$ is the population aged x to x+n in area i at the earliest census(t),

One advantage of RCCRs is that we can backcast age groups 0-4 and 5-9 (from those aged 10-14 and 15-19, ten years later, respectively). This is not possible for the forward-looking Hamilton-Perry Method, which generally employs Child Woman Ratios. A backcast of the oldest age group also is more straightforward than in the projection of it in the forward looking version of the Hamilton-Perry Method. For example, if the final closed age group is 80-84, with 85+ as the terminal open-ended age group in the census following the earliest census, then calculations for the $\mathrm{RCCR}_{i,x+}$

$$ \mathrm{RCCR}_{i,}85+ = \mathrm{P}_{i,75+,t}/\mathrm{P}_{i,85+,t+k} \tag{17.14a} $$

The formula for estimating the population 75+ of area i for the year t-k is:

$$ \mathrm{P}_{i,75+,t-k} = (\mathrm{P}_{i,75+,t}/_{i,85+,t+k})^{*}\mathrm{P}_{i,75+,t} \tag{17.14b} $$

Table 17.6a provides an example of a backcast from a Reverse Hamilton-Perry Method. It uses 1930 and 1920 age-sex census data on Native Hawaiians in Hawai'i to develop RCCRs and then backcasts the 1920 Native Hawaiian population to 1910 to generate population estimates by age and sex for Native Hawaiians in Hawai'i. Because this group was counted in the 1910 census, we can compare the "pre-censal" estimates of them to the enumerated numbers to get an idea of the method's accuracy. These comparisons are found in Table 17.6b. The method underestimated the total population of Native Hawaiians in 1910 by 930 people (-3.27%). For the estimates by age group for both sexes combined, the MAPE is 7.10%.

When one considers the use of a CCR (and its inverse, the RCCR), it is easy to see that the survivorship ratio found in a life table is a CCR (and the inverse of a survivorship ratio is an RCCR). The survivorship rates computed from the "$_nL_x$" column (Years lived in a given age interval) of a life table are equivalent to the CCRs calculated for age groups of a specific width, while the survivorship rates computed from the "T_x" column (Years lived at this and all subsequent ages) are equivalent to the CCRs calculated for open-ended terminal age groups. The relationship between survivorship rates calculated from T_x and CCRs calculated from open-ended, terminal age intervals brings up a way in which the reverse Hamilton-Perry Method can be used to estimate a total pre-censal population. The way to proceed is to first consider the relationship between T_0 and T_x in a life table as follows:

$$_xS_0 = T_x/T_0 \tag{17.15a}$$

where

$_xS_0$ = the survivorship rate from birth to the open-ended terminal age group, x
T_x = Years lived in the open-ended, terminal age group
T_0 = Years lived at birth and all subsequent age groups
Re-arranging the terms in Equation [17.15], we see that

$$T_x = {}_xS_0 * T_0 \tag{17.15b}$$

and, further that

$$T_0 = (T_x/{}_xS_0) \tag{17.15c}$$

The preceding equations suggest that an RCCR can be constructed such that a total pre-censal population can be estimated. First, note that

$$RCCR_{k+} = P_{0+,t}/P_{k+,t+k} \tag{17.16a}$$

The formula for estimating the total population 0+ of area i for the year t-k is:

$$P_{0+,t-k} = RCCR_{k+} * P_{k+,t} \tag{17.16b}$$

Table 17.6a Estimation of 1910 Native Hawaiians using Reserve Hamilton-Perry method and 1930 to 1920 Cohort Change Ratios

Age in 1930	MALES	FEMALES	Age in 1920	MALES	FEMALES	CCR MALE	CCR FEMALE	Age in 1910	Estimated 1910 Males	Estimated 1910 Females	Estimated 1910 Total
10–14	1,161	1,222	0–4	1,266	1,298	1.0904	1.0622	0–4	1,223	1,282	2,506
15–19	1,127	1,071	5–9	1,219	1,209	1.0816	1.1289	5–9	1,180	1,242	2,422
20–24	952	1,031	10–14	1,122	1,207	1.1786	1.1707	10–14	1,223	1,290	2,513
25–29	760	915	15–19	1,091	1,100	1.4355	1.2022	15–19	1,380	1,273	2,653
30–34	728	794	20–24	1,038	1,102	1.4258	1.3879	20–24	1,098	1,137	2,235
35–39	748	876	25–29	961	1,059	1.2856	1.2089	25–29	1,196	1,059	2,255
40–44	710	642	30–34	770	819	1.0853	1.2757	30–34	665	856	1,521
45–49	631	625	35–39	930	876	1.4744	1.4017	35–39	1,255	1,036	2,291
50–54	553	522	40–44	613	671	1.1080	1.2853	40–44	573	598	1,170
55–59	466	399	45–49	851	739	1.8250	1.8513	45–49	878	718	1,596
60–64	371	296	50–54	517	465	1.3947	1.5719	50–54	633	486	1,119
65–69	266	202	55–59	481	388	1.8070	1.9202	55–59	669	413	1,081
70–74	153	118	60–64	454	309	2.9710	2.6200	60–64	377	296	673
75+	197	166	65+	673	485	3.4162	2.9217	65+	601	459	1,060
TOTAL	8,822	8,879	TOTAL	11,986	11,727			TOTAL	12,951	12,144	25,095

Table 17.6b Comparison of Estimates of 1910 Native Hawaiians using Reverse Hamilton-Perry method to the 1910 Census

Estimated 1910 Males	Estimated 1910 Females	Estimated 1910 Total	CENSUS 1910 MALES	CENSUS 1910 FEMALES	CENSUS 1910 Total	DIFFERENCE MALE	DIFFERENCE FEMALE	DIFFERENCE TOTAL	% DIFF MALE	%DIFF FEMALE	%DIFF TOTAL	ABS% DIFF MALE	ABS% DIFF FEMALE	ABS% DIFF TOTAL	1910 AGE
1,223	1,282	2,506	1,368	1,345	2,713	−145	−63	−207	−10.56	−4.68	−7.65	10.56	4.68	7.65	0–4
1,180	1,242	2,422	1,253	1,256	2,509	−73	−14	−87	−5.82	−1.14	−3.48	5.82	1.14	3.48	5–9
1,223	1,290	2,513	1,307	1,221	2,528	−84	69	−15	−6.40	5.66	−0.57	6.40	5.66	0.57	10–14
1,380	1,273	2,653	1,343	1,314	2,657	37	−41	−4	2.72	−3.11	−0.16	2.72	3.11	0.16	15–19
1,098	1,137	2,235	1,129	1,138	2,267	−31	−1	−32	−2.76	−0.11	−1.43	2.76	0.11	1.43	20–24
1,196	1,059	2,255	1,123	1,090	2,213	73	−31	42	6.47	−2.84	1.88	6.47	2.84	1.88	25–29
665	856	1,521	837	947	1,784	−172	−91	−263	−20.52	−9.61	−14.73	20.52	9.61	14.73	30–34
1,255	1,036	2,291	1,043	1,006	2,049	212	30	242	20.30	2.97	11.79	20.30	2.97	11.79	35–39
573	598	1,170	734	734	1,468	−161	−136	−298	−21.96	−18.58	−20.27	21.96	18.58	20.27	40–44
878	718	1,596	841	734	1,575	37	−16	21	4.38	−2.14	1.34	4.38	2.14	1.34	45–49
633	486	1,119	638	604	1,242	−5	−118	−123	−0.75	−19.58	−9.91	0.75	19.58	9.91	50–54
669	413	1,081	611	438	1,049	58	−25	32	9.43	−5.74	3.09	9.43	5.74	3.09	55–59
377	296	673	407	244	651	−30	52	22	−7.29	21.34	3.44	7.29	21.34	3.44	60–64
601	459	1,060	800	520	1,320	−199	−61	−260	−24.84	−11.79	−19.70	24.84	11.79	19.70	65+
12,951	12,144	25,095	13,434	12,591	26,025	−483	−447	−930	−3.60	−3.55	−3.57	3.60	3.55	3.57	TOTAL
												10.30	7.81	7.10	MAPE
												6.88	5.17	3.46	MEDAPE

As an example, we again turn to historical data on the Native Hawaiian population in Hawai'i. Here we will use 1930 and 1910 data for Native Hawaiians to estimate a RCCR for age group 20+ and then apply this RCCR to the Native Hawaiian population aged 20+ in 1910 in order to estimate the total number of Native Hawaiians in 1890. As shown in Table 17.6a, there are 13,120 Native Hawaiians age 20 years and over in 1930, while in 1910 there are 25,095 Native Hawaiians in total, of whom 15,001 are aged 20 and over. From these data, we find that:

$$\mathrm{RCCR}_{20+,1910} = 25{,}095/13{,}120 = 1.9127$$

And that

$$\mathrm{P}_{0+,1890} = 1.9127^{*}15{,}001 = 28{,}693$$

So, our estimate of the Native Hawaiian population of Hawai'i in 1890 is 28, 693. This estimate is 16.7 percent less than the number reported by Schmitt (1977: 25) from the Hawaiian Kingdom's 1890 census, which is 34,436. Given that migration of Native Hawaiians was not a major factor in its population change (see, e.g., Schmitt 1968: 183), it appears that mortality rates were dramatically higher for this population between 1890 and 1910 than it was between 1910 and 1930, which, in fact, the available evidence suggests is the case (Nordyke 1989; Schmitt 1968; Schmitt 1977). The correct RCCR_{20+} for estimating the total number of Native Hawaiians in 1890 from those aged 20 and over in 1910 is 2.2956 = 34,436/15,001.

In summary, the Reverse Hamilton-Perry appears to be capable of working well going back ten years, but backcasting to points in time beyond ten years from a census is subject to higher levels of error. This is not surprising for populations going through dramatic changes in the components of change that are summarized in the form of cohort change ratios and reverse cohort change ratios. As a means of accounting more directly for changes in these components, we turn to a tool specifically developed for this purpose, inverse projection (Barbi et al. 2004; Lee 1985).

Inverse Projection. Since Lee (1974) first developed Inverse Projection (IP), the method has undergone revisions and has been used to construct population histories of countries, cities, and parishes or missions, covering a wide range of demographic histories (McCaa 2001). Moreover, the evidence suggests that with very little input data, IP can deliver accurate demographic estimates (Lee 1985; McCaa 2001, 1989; and McCaa and Vaupel 1992). McCaa (2001) cautions, however, that as is the case with most demographic techniques, as input data quality and detail increase, so do the accuracy and quality of the estimates.

The IP variations differ in their use of models to supplement missing or unknown parameters, but in all cases it provides refined mortality and fertility rates, as well as age composition from crude birth and death rates and an initial estimate (or census) of the total population (McCaa 2001). In order to operate reasonably well, IP needs

relatively accurate vital events data. Like, the cohort-component method, it does need age data, but they need not be detailed and may be estimated for the year in which an IP is launched. Unlike, the cohort-component method, however, which uses age-specific rates to generate counts, IP uses counts of vital events to estimate age-specific rates. To generate an estimate, IP requires a count of annual births and deaths from the launch year to the target year and an estimate of the size of launch year population. In the absence of reliable empirical age data for the launch year, model age structures of mortality, fertility, migration, and, most importantly, the initial population may be relied upon (McCaa 2001).

In the initial IP approach developed by Lee (1974), only the total population is required at the launch year along with counts of births and deaths for each IP cycle (e.g., annual, five year, ten year). Where a count of the total population is known at one or more of the cycles, the total derived from the IP is subtracted from the observed total to obtain an estimate of net migration, which is apportioned equally among the intervening IP cycles.

The age structure of the population at the launch year is needed, but this can be obtained from a model if it is not available otherwise. Moreover, even if the launch year age structure is inaccurate, its effect on accuracy diminishes and at some point becomes inconsequential due to ergodicity - the condition in which a dynamic process "forgets" the state in which it started, which in this case is the launch year age structure (Lee 1985; and Wachter 1986). This has many implications, including the fact that within a century even a very peculiar age structure is "forgotten" as birth, death and migration rates displace it. However, this is not the case with the size of a population, which exerts a powerful influence for a long time (McCaa and Vaupel 1992).

Borrowing from McCaa's (2001) description, we here describe Lee's (1974) original exposition of inverse projection. It starts when an arbitrary set of age specific death rates from a single parameter family of life tables is applied to the launch year population. Multiplying these age-specific rates by the numbers of people in the corresponding age groups at the launch year provides estimates of the age-specific deaths for the initial inverse projection cycle. When the age specific-deaths are summed over all age groups and divided by the total number of actual deaths, an adjustment factor is produced, the normalized death ratio, or "k." This adjustment factor, k, is a measure of the discrepancy between the observed number of deaths and the total number resulting from the inverse projection. This same adjustment factor, k, is then used to adjust the age specific deaths so that they correspond to the independently generated total number of deaths. This is accomplished in several steps. First, the age specific death rates found in the model life table used in step one are subtracted from an adjacent model life table in the same family (e.g., the "West model life tables). The result is a domain of mortality variability at each age. This domain is used to interpolate to an adjusted age-specific death rate. The calculation is accomplished by simply multiplying k and the mortality domain at each age and adding the result to the first approximation. This yields a final estimate of deaths for each age group. When summed, they equal the total deaths for the period.

Net migration by age is apportioned crudely as a function of age-specific rates and a level parameter for each period. Where migration is a minor factor relative to mortality, this solution – required due to a lack of data – produces acceptable results.

The number of births for each interval is determined by the stream of vital events. What is lacking is a summary fertility statistic which takes into account the age structure of the population. A solution common to all inverse projection algorithms is the estimation of the gross reproduction ratio from the number of births, the age structure of the population, and normalized age patterns of fertility.

McCaa (2001) notes that the "k" adjustment factor also can be used as a measure of goodness of fit where model age data are required. Because k is an indicator of the age-standardized intensity of mortality relative to the age structure of the initial population and to the pair of age-specific mortality schedules used to inverse project the population to the next period, the sum of squares of this ratio (K^2) is a useful measure of goodness of fit, when the birth and death series, and initial population size are held constant. Note that this does not directly account for migration.

As a means of dealing with populations not "closed" to migration, back projection was developed by Wrigley and Schofield for their massive reconstruction of the population of England, 1541–1871 (Wrigley and Schofield 1981). This technique is designed to develop migration estimates from a terminal age structure by backcasting the population against the flows of births and deaths (McCaa 2001). Lee (1993) criticized this approach on several grounds including the fact that it ignores ergodicity as well as the fact it "resurrects people who have died into the oldest age group, an attempt that is, in practice, hypersensitive to error." It should be noted, however, that reasonable estimates of net migration can be derived from backcasting as long as the input data are reasonably accurate (Al-Jiboury and Swanson 1988).

The debate over IP versus back projection led to the development of Generalized Inverse Projection, which exploits whatever data are available as well as a broad range of assumptions or constraints, including components derived from back projection (McCaa 2001; Oeppen 1993a, 1993b). Generalized inverse projection uses a standard method of demographic accounting and a standard non-linear optimization algorithm to overcome a range of empirical and theoretical problems.

Fortunately, McCaa (2001) has developed "POPULATE," an online IP tool that implements a range of variations of the IP approach. For those interested in using this tool to develop pre-censal estimates, we strongly recommend not only reading McCaa's essay (2001) and the materials associated with it, but also running the example data sets, all of which can be found at http://www.hist.umn.edu/~rmccaa/populate/index.htm.

17.3 Summary

Population projections seem to be involved with a fair amount of debate and conflict (D'Allesandro 1987; Moen 1984; Smith et al. 2001; Swanson and Tayman 1995). Similarly, historical population estimates, especially pre-censal

ones going relatively far back in time, appear to generate intense debates. Stannard (1989) takes great issue with estimates of the number of people provided by Schmitt (1968, 1977) and by Nordyke (1989) in Hawai'i prior to the arrival of Captain Cook and general contact with Europeans. The crux of the debate is that Stannard believes that the number of pre-contact Hawaiians was much higher than the numbers estimated by Schmitt and by Nordyke. Similarly, Thornton (1987) believes that most of estimates of Native American population numbers prior to European contact also are too low, as does Stannard (1992). Less contentious, but as lively, have been the debates over inverse projection and backcasting (McCaa 2001; Oeppen 1993a, 1993b). Even less contentious are the inter-censal estimates, which have less room for error than do pre-censal estimates. The importance of accurate historical information is reflected in these debates. Among other functions, the information sets the context for social agreements on both our present and our future (Mead 1929). For these and a myriad of other reasons, the ability to develop estimates of historical populations, both near-term and those in the distant past, are an important component in the estimation toolkit.

References

Al-Jiboury, A. and D. A. Swanson. 1988. "Inter-censal Net Migration Among the Three Major Regions of Iraq., 1957-1977." *Janasamkhya* 6: 93–125.

Barbi, E., S. Bertino, and E. Sonnino (eds,). 2004. Inverse Projection Techniques: Old and New Approaches. Dordrecht, Heidelberg, London, and New York: Springer.

Brass, W., A. Coale, D. Heisel, F. Lorimer, A. Romaniuc, and E. van de Walle. 1968. *The Demography of Tropical Africa*. Princeton, NJ: Princeton University Press.

Carrier, N. and J. Hobcraft. 1971. *Demographic Estimation for Developing Countries*. Population Investigation Committee, London School of Economics. London: London School of Economics.

D'Allesandro, F. 1987. Should applied demographers take out liability insurance? *Applied Demography* 3 (Fall): 1–3.

Hamilton, C. H, and J. Perry. 1962. A short method for projecting population by age from one decennial census to another. *Social Forces* 41: 163–170.

Judson, D. and C. Popoff. 2004. "Selected General Methods." pp. 677-732 in J. Siegel and D. A. Swanson (Eds.) *The Methods and Materials of Demography, 2nd Edition*. Amsterdam, Netherlands, Elsevier Academic Press.

Lee, R. 1974. "Estimating Series of Vital Rates and Age Structures from Baptisms and Burials: A New Technique, with Applications to Pre-Industrial England." *Population Studies* 28: 495–512.

Lee. R. 1985 "Inverse Projection and Back Projection: A Critical Appraisal and Comparative Results for England, 1539–1871," *Population Studies*, 39, pp. 233–248.

Lee, R. 1993. "Inverse projection and demographic fluctuations: a critical assessment of new methods," pp. 7-28 in D. Reher and R. Schofield (eds.) *Old and New Methods in Historical Demography*. Oxford, England, Clarenton Press.

McCaa, R. 1989. "Populate: A Microcomputer Projection Package for Aggregative Data Applied to Norway, 1736–1970," Annales de Démographie Historique, pp. 287–298.

McCaa, R. 2001. "An Essay on Inverse Projection." (http://www.hist.umn.edu/~rmccaa/populate/ipessay.htm).

McCaa, R. and J. W. Vaupel. 1992. "Comment la projection inverse se comporte-t-elle sur des données simulées?," in Alain Blum, Noël Bonneuil et Didier Blanchet (eds.) Modèles de la démographie historique, Institut National d'Etudes Démographiques, Paris, pp. 129–146.

McKibben, J. 1988. "Evaluating the Accuracy of County Population Estimates Using the Reverse Demographic Accounting Method." Paper presented at the annual meeting of the Population Association of America, New Orleans, LA.

Mead, G. H. 1929. "The Nature of the Past." pp. 235-242 in John Coss (ed.) *Essays in Honor of John Dewey*. New York, NY: Henry Holt & Co. (http://www.brocku.ca/MeadProject/Mead/pubs2/papers/Mead_1929d.html).

Moen, E. 1984. "Voodoo Forecasting: Technical, Political, and Ethical Issues Regarding the Projection of Local Population Growth." *Population Research and Policy Review* 3: 1–25.

Nordyke, E. 1989. *The Peopling of Hawai'i, 2nd Edition*. Honolulu, HI; University of Hawai'i Press.

Oeppen, J. 1993a. "Back Projection and Inverse Projection: Members of a wider Class of Constrained Projection Models." *Population Studies* 47: 245–67.

Oeppen, J. 1993b "Generalized inverse projection," pp. 29-39 in D. Reher and R. Schofield (eds.) *Old and New Methods in Historical Demography*. Oxford, England: Clarendon Press.

Shryock, H., and N. Lawrence. 1949. "The Current Status of State and Local Population Estimates in the Census Bureau." *Journal of the American Statistical Association* 44: 157–173.

Schmitt, R.C. 1968. *Demographic Statistics of Hawai'i, 1778-1965*. Honolulu, HI: University of Hawai'i Press.

Schmitt, R.C. 1977. *Historical Statistics of Hawai'i*. Honolulu, HI: University of Hawai'i Press.

Smith, S., J. Tayman, and D. Swanson. 2001. *State and Local Population Projections: Methodology and Analysis*. New York, NY" Kluwer Academic/Plenum Publishers.

Stannard, D. 1992. *American Holocaust: The Conquest of the New World*. Oxford, UK: Oxford University Press.

Stannard, D. 1989. *Before the Horror: The Population of Hawai'i on the Eve of Western Contact*. Honolulu, HI: Social Science Research Institute, University of Hawai'i.

State of Washington. 1944. *Report of the Census Board (Created by Chapter 81, Laws of 1943) For the Years 1943 and 1944*. Olympia, WA: State Printing Plant.

Swanson, D. A., J. Schlottmann, and B. Schmidt. 2010. "Forecasting the population of census tracts by age and sex: An example of the Hamilton-Perry method in action." *Population Research and Policy Review* 29(1): 47–63.

Swanson, D. A. and J. Tayman. 1995. "Between a Rock and a Hard Place: The Evaluation of Demographic Forecasts." *Population Research and Policy Review* 14: 233–249.

Thornton, R. 1987. *American Holocaust and Survival: A Population History since 1492*. Norman, OK: University of Oklahoma Press.

United Nations. 1983. *Manual X. Indirect Techniques for Demographic Estimation*. New York. United Nations.

Wachter, K.1986. "Ergodicity and Inverse Projection." *Population Studies* 40: 275–287.

Wrigley, E. and R. Schofield. 1981. *The Population History of England, 1541–1871: A Reconstruction*. Cambridge, MA: Harvard University Press.

Chapter 18
Future Directions in Population Estimation

Where estimation methods primarily differ is in their specific variables, data sources, and the ways in which those variables and data sources are related to each other. Future changes in the field of population estimation will therefore stem from changes in the availability of historical data, the tools for organizing and manipulating those data, our understanding of how different variables interact to determine population change, and our ability to build new models or develop new methods based on these new insights. The inspired analyst will incorporate factors not previously considered in developing estimates, or will put them together in creative new ways.

What recent developments might change the way we make population-estimates? Where is current research headed? Are any "paradigm shifts" imminent? In this chapter, we describe several promising new developments in the field of population estimation, make some suggestions regarding areas needing further research, and offer several predictions regarding future developments. We also discuss several recent changes in the scope of estimates and some of the challenges we see on the horizon.

We distinguish between two types of developments pertaining to population estimates. Technological developments are those affecting the availability of input data and the computing tools used to organize and manipulate those data. Methodological developments are those affecting the models used to formulate relationships among input data and estimate those relationships into the post-censal period. Put another way, technological developments affect the resources we have to work with, whereas methodological developments reflect new ways of using those resources. Although this distinction is not always clear-cut, we believe it helps clarify the developments discussed in this chapter.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0_18,

18.1 Technological Developments

Three technological developments—greater data availability, expanding computing power, and the growth of Geographic Information Systems (GIS)—are transforming the way we make population estimates. These developments have already had a substantial impact and promise to have an even greater impact in the coming years.

18.1.1 Data Availability

Sources of mortality, fertility, and migration data have evolved gradually over the past several decades. Perhaps the most important improvement was the development of annual migration estimates based on IRS records, which first became available during the late 1970s. These data provide an alternative to school enrollment for developing migration estimates. Other than that, data on the components of population growth have changed very little over the last half century. As seen throughout this book, administrative records play an important role in the development of population estimates. Without them population estimation methods would largely be confined to extrapolation and interpolation methods described in Chapters 6 and 17.

While it is difficult to predict whether a new source or type of administrative record is on the horizon, the way these data are used and combined are likely to evolve. For example, Swanson and Walashek (2011) describe a Census-Enhanced Master Address File (CEMAF) system built on a combination of four elements: (1) administrative records; (2) the continuously updated Master Address File; (3) survey data; and (4) modeling and imputation techniques. CEMAF could deliver population estimates that are timely, comprehensive, and internally consistent and also estimates of housing and demographic and socio-economic characteristics for the national and subnational areas. We also believe that the use of parcel files will continue to escalate for organizations that require spatially intensive population and housing estimates.

The biggest sea change in data that will impact population estimation is the American Community Survey (ACS) discussed in Chapter 3. Now fully implanted, the ACS has replaced the census long-form. It is designed to provide accurate and timely social, demographic, and economic indicators on a "continuous measurement" basis for federal, state, and local governments, and businesses. Annual characteristics are available for many of geographic areas where they were not previously available. The ACS migration data may be particularly useful because there are so few alternative sources of migration data. The ACS is not without its share of issues (e.g., small sample sizes, implausible temporal changes, interpretation of multi-year samples, and different residency rules from the census). These issues limit its effectiveness as a replacement for the census long form and its use in post-censal population estimation. But currently the ACS is the only game in town

for providing detailed social, demographic, and economic characteristics. It is safe to say that considerable research will be directed along two major themes: 1) a more comprehensive understanding of the issues and their implications in using the ACS; and 2) strategies for improving the efficacy of the ACS, such as those suggested by Rogers, Little, and Raymer (2010) for dealing with its migration information.

Although there have been some recent changes in the sources of data used for making population estimates, there have been major changes in the formats in which those data are available. For centuries, hard copy (i.e., printed reports) was the only available format. This format was slow, cumbersome, and expensive. Recent decades have seen the development of computer tapes, diskettes, CDs, and the Internet. Many types of data are now available instantly and for free, simply at the click of a mouse. Information technology has revolutionized our ability to access and use data for population estimates and a variety of other purposes. It takes no great leap of faith to anticipate substantial further improvements along these lines.

18.1.2 Computing Capabilities

Before 1950, most population estimates were generated manually using pencils and paper or rudimentary mechanical devices for doing arithmetic operations. The dawn of the computer age not only made it possible to produce population estimates much more quickly and easily and for a wider range of geographic areas and demographic categories, but allowed the incorporation of large, complex datasets into the estimation process (e.g., parcel files, IRS tax returns). Increases in computing capabilities will undoubtedly influence future developments in the field of population estimation, just as they have in other fields where estimates are built on complex models driven by masses of data (e.g., meteorology).

In the early years of the electronic era, automated estimation programs were written using a high-level programming language (usually FORTRAN) and were run through a centralized computer. These programs required hundreds of punch cards and were slow, inefficient, and often frustrating. The advent and widespread use of personal computers, spreadsheet programs, and statistical analysis packages has revolutionized the production, analysis, and evaluation of population estimates. Spreadsheets and statistical analysis packages are simpler to use, more tolerant or errors, and require less training than formal programming languages. They have many built-in functions and macros to perform repetitive tasks, facilitating the creation, display, printing, and graphing of data (e.g., Klosterman, Brail, and Bossard 1993). Greater computing power has also made possible the development of powerful desktop GIS applications. These developments have provided the technical infrastructure needed to support ever more complicated, spatially-intensive, and data-intensive estimation methods.

Distributed computing environments have made a comeback in recent years, along with new software and programming applications. Many analysts now work in networked computer environments, making it easy to share information and obtain

access to centralized sources of information. Relational database management systems, modern computing platforms, and user-friendly interfaces are being used more and more frequently for population estimates. These new technologies greatly facilitate the development, analysis, and distribution of population estimates.

Relational database storage and retrieval systems make it easier to manage, maintain, document, and verify information. Modern computing platforms contain structured and modular programming algorithms. These platforms are easier to maintain, can handle a wide range of computational algorithms, and permit a detailed documentation of computer code. Finally, a well-designed, user-friendly GUI interface can tie the entire estimation system together (e.g., data, results, and computer programs). This frees one from having to perform a number of tedious, time-consuming chores, makes it easier to explain estimation methods to data users, and greatly facilitates the analysis and review of the estimates.

The Internet, too, influences the production of population estimates. We believe "estimation tool kits" will likely become widely available on the Internet in the years ahead, especially with the increasing use of the "cloud." These tool kits will contain a variety of computational algorithms and have sophisticated data management and output capabilities, enabling analysts to apply their own data and methods and develop reports, charts, graphs, and data files. This will substantially reduce the time required to put together the appropriate software for constructing population estimates.

The Internet has already made it quicker, easier, and less costly to obtain the data needed for population estimates. However, this is still mostly a "manual" process that requires locating information site-by-site. Looking ahead, we expect this process to become automated, perhaps through a common, easy-to-use interface that locates and integrates all the relevant data sources, regardless of their file structure, format, or location on the Internet.

18.1.3 Geographic Information Systems (GIS)

Most demographic analyses have a strong geographic component. For example, the population estimates analyzed in this book referred to specific geographic areas such as states, counties, census tracts, or parcels. Many other data-intensive fields (e.g., urban planning, marketing, epidemiology, and environmental science) also have close relationships with geography. As discussed in Chapter 2, it is not surprising, then, that a computer-based methodology has come into widespread use for the geographic display and analysis of geo-spatial data.

A GIS provides the tools for linking spatial data with non-spatial data. It allows one to organize data from one or more sources into a variety of geographic areas (e.g., census tracts, ZIP codes, market areas) and helps data users visualize spatial relationships. GIS provides a methodology for quickly and efficiently organizing, analyzing, and displaying a large amount of spatial data. The ability to geocode (i.e., assign an observation of a variable to a specific geographic area) data such as

building permits, electric customers, tax records, births, deaths, school enrollment, and other symptomatic indicators of population change will greatly enhance our ability to make demographic estimates for very small units of geography. GIS also makes it possible to analyze historical demographic trends for very small geographic areas. These capabilities will facilitate the refinement of current estimate methods and possibly the development of new ones.

18.2 Methodological Developments

The technological developments discussed above help improve the quality and quantity of the data and tools available for making population estimates. Will these developments lead to better population estimates? That depends primarily on improvements in how those tools are used. Methodological advances are likely in at least four areas: 1) synthetic population and household estimates, 2) spatial regression models, 3) remote sensing, and 4) measuring uncertainty. All of these developments have the potential to raise the overall utility of population estimates; some of them may reduce estimate error as well.

18.2.1 Synthetic Populations and Households

For many years population estimates were made primarily at the national and state levels. In recent decades estimates have been carried out at progressively lower levels of geography. Estimates now are routinely made for subcounty areas such as census tracts, block groups, and traffic analysis zones. There is a growing demand for estimates for even smaller areas such as tax assessor parcels and block faces. Taking this trend to its logical conclusion implies the development of estimates for individual households and people, also known as synthetic estimates. Synthetic households are comprised of one or more individuals; each has an associated set of characteristics (e.g., age, sex, income, labor force and marital status). Synthetic estimates not only have value in their own right, but support the expanding array of micro-simulation models being developed and used today (e.g., Davidson, Donnelly, Vovsha, Freedman, Ruegg, Hicks, Castiglione and Picado 2007; Statistics Canada 2011; Waddell, Borning, Noth, Freier, Becke, and Ulfarsson 2003). We believe estimates for very small areas—including estimates of individuals and households—will become increasingly common.

Making estimates at this level may seem like science fiction, but several approaches for estimating individual households and people have already been constructed. Most of the existing procedures for synthesizing population and housing data are based on the Iterative Proportional Fitting (IP) method, discussed in Chapter 13, where the number of individuals in each cell of the cross-classification table is estimated (Muller and Axhausen 2011). Beckman, Baggerley, and McKay

(1996) were the first to apply IP to the problem of generating synthetic populations. The unadjusted data cells are referred to as the "seed" cells, and the selected totals are referred to as the "marginal" totals. IP uses a sample dataset, most often PUMS, to establish the correlation between the dimensions or attributes under the condition that the mapping of the summations in lower dimensions should fit the margins given by census/ACS data.

The direct application of IP for synthesis of data requires specific conditions to work efficiently. The marginal column totals and the marginal row totals must add up to the same value, and the marginal cell values cannot be zero. Guo and Bhat (2007) alleviated the zero-cell-value problem and the inability to control the statistical distributions of both household- and individual-level attributes. Pritchard and Miller (2009) also improved the IP method by allowing many more attributes per agent through a Monte Carlo simulation based on a sparse list-based data structure. IP-based methods are based on pre-defined categories of individuals. A different type of categorization would yield a different classification table, which would also change the end results of the analysis. This phenomenon is called the modifiable attribute cell problem (MACP). Otani, Sugiki, Vichiensan, and Miyamoto (2011) have proposed a method for dealing with the MACP that determines the best combination of categories. The best cell organization is one that minimizes the number of cells in the table with respect to the key output variable that has been defined and used as an evaluation criterion.

18.2.2 Spatial Regression Models

As discussed in Chapter 8, regression methods are widely used for post-censal population estimates, most notably the ratio-correlation method. Over the years a number of modifications to the basic ratio-correlation model have been tested including alternative measurements of the variables (i.e. differences and natural logarithms), use of dummy variables and stratification, and use of the average of estimates from single variable regression models. Swanson and Tedrow (1984) suggested that the logarithmic transformation improved estimate accuracy because that transformation may reduce the effects of spatial autocorrelation. They suggested that it would be useful to examine spatial correlation issues involving regression models for population estimates. To our knowledge, this issue still has not been investigated.

Although spatial statistics have been applied to numerous fields in the last few decades, it has drawn demographers' attention only recently (Chi and Zhu 2008). Spatial autocorrelation can be loosely defined as a similarity (or dissimilarity) between two values of an attribute that are nearby spatially (Griffith 1987: 9). With positive spatial autocorrelation, high or low values of an attribute tend to cluster in space whereas with negative spatial autocorrelation, locations tend

to be surrounded by neighbors with very different values. Spatial autocorrelation can be measured by various indexes, the most well-known being Moran's I statistic (Moran 1948). Spatial autocorrelation, like temporal autocorrelation, violates standard statistical techniques that assume independence among observations. Regression analyses that do not compensate for spatial dependency can have unstable parameter estimates and yield unreliable significance tests (Anselin and Griffith 1988). Despite its complexity, spatial demographic analysis has in recent years become more accessible for demographers to explore, due to the development of user-friendly spatial data analysis software packages.

18.2.3 Remote Sensing

Remote sensing and GIS have been used to estimate population, particularly for large areas. Wu, Qui, and Wang (2005), provide a comprehensive review of the use of remote sensing for population estimation before 2005. Early remote sensing applications involved manually counting the number of houses using aerial photos (Lo 1986), which was followed by automatic approaches with satellite remote-sensing imagery to estimate population density (Lo 1995). The launch of the IKONOS satellite in 1999, with very-high-resolution sensors, offered new opportunities to investigate urban physical configurations at a fine spatial scale. In addition, airborne Light Detection and Ranging (LiDAR) has become widely used for deriving high-resolution vertical information in urban areas. A recent issue of the International Journal of Remote Sensing focused on recent developments in population estimation using remote-sensing and GIS technologies (Wang and Wu 2010). Below we summarize a few papers from this special issue to illustrate the current state-of-the practice.

Dong, Ramesh, and Nepali (2010) combined LiDAR, Landsat Thematic Mapper, and a parcel data set for a study area in Denton, Texas. Using census blocks as samples, building count, building area, and building volume were calculated from a digital surface model, zonal statistics, and 2000 census data using a set of OLS and geographically weighted regression models. They found that the population count was often overstated when population density of the census block is low ($<$300 persons km^{-2}) and was always understated when the population density is high ($>$3,500 persons km^{-2}). They concluded the minimum estimation error of -23% was too imprecise for small area population estimation and suggested the low accuracy was caused by the lack of high resolution LiDAR and image data, which made it difficult to separate trees and buildings.

Lu, IM, Quackenbush, and Halligan (2010) using higher resolution data than the previous study, estimated the population of census blocks in a study area of Denver, Colorado. This research examined the utility of QuickBird imagery and LiDAR data using two approaches: area-based and volume-based. Residential-building footprints were first delineated from the remote-sensing data using image segmentation and machine-learning decision-tree classification. Regression

analysis was used to model the relationship between population and the area or volume of the delineated residential buildings. Both approaches resulted in the successful performance for estimating population with high accuracy (coefficient of determinations = 0.80 – 0.95; root-mean-square errors = 10 – 30 people; relative root-mean-square errors = 0.10 – 0.30). The area-based approach was slightly better than the volume-based approach because the residential areas of the study sites are generally homogeneous (i.e. single houses), and the volume-based approach is more sensitive to classification errors. LiDAR-derived shape information such as height greatly improved population estimation compared to population estimation using only spectral data.

Deng, Wu, and Wang (2010) explored the feasibility of incorporating GIS, remote-sensing, and demographic data into the housing unit method to estimate the population in census blocks in Grafton, Wisconsin. Two major components of the housing unit method, housing unit counts and persons per household (PPH), were obtained by modeling their relationships with demographic, geographic, and spatial factors using a sequence of OLS regression models. The results indicated that spatial factors derived from remote sensing using the dasymetric-mapping method, GIS datasets, and demographic information can significantly improve the accuracy of small-area population estimation.

18.2.4 Measuring Uncertainty

Population estimates cannot provide perfectly accurate estimates of current or past populations (see Chapter 14 for a detailed discussion of forecast accuracy). The uncertainty inherent in population estimates has been accounted for by developing probabilistic intervals that provide an explicit statement of the level of error expected to accompany a specific point estimate. However, the practice of indicating the direction and magnitude of error in post-censal population statistics is still virtually, if not, completely absent in statistical offices

We believe it is important to provide a direct measure of error and the distributional properties of population estimates, and that future research will focus increasingly on the measurement of uncertainty in population estimates. The ratio-correlation method offers a ready-made, low cost approach to developing measures of estimate uncertainty that is greatly underutilized. Perhaps, probabilistic intervals can also be developed for the housing unit method by treating its components as stochastic and using margins of error from the ACS to estimate their uncertainty. Although such research may not directly improve estimate accuracy, it will enhance our understanding of the uncertainty inherent in population estimates, particularly how uncertainty varies by population size and growth rate, geographic region, length of the post-censal period, estimation method, and perhaps other factors as well. We hope that private companies and federal, state, and local government agencies will start producing explicit probabilistic intervals to accompany their estimates.

18.3 Scope of Estimates

Not only are changes occurring in the technology and methodology of population estimates, but in their scope as well. In particular, estimates are being made for smaller and more varied units of geography and for a broader array of demographic characteristics. This trend is driven by market and other demands for such estimates and by technological changes making such estimates possible. The demand for small-area population estimates is increasing. In the public sector estimates are the basis for such purposes as constructing budgets, developing transportation systems, planning for schools, and determining the optimal location of public facilities. Private sector uses include financial planning, site analyses, sales forecasting, target marketing, and new product introduction. Many of these uses require estimates for very small levels of geography and/or for very detailed population and other characteristics.

Trends in marketing illustrate the growing demand for greater geographic and demographic detail (Martins, Yusuf, and Swanson 2011; Pol and Thomas 1997: 36-37). In the 1960s, mass marketing was the order of the day. The market was seen as a homogeneous mass of consumers, each one pretty much like any other. Over time, this concept gave way to the notion of target marketing in which specific products or brands are directed toward specific types of consumers. In response to the development of even more refined demographic clusters, target marketing has given way to micro-marketing, which focuses on characteristics at the household level (Verdino 2010). Micromarketing grew to prominence in the 1990s, as personal computers allowed easier segmentation and dissemination of information to customers. Micromarketing attempts to pinpoint very small groups of potential customers and focus on their buying patterns. Armed with information from an ever-growing number of consumer data bases, micro-marketers are targeting customers down to the block or even the household level. As was noted in Chapter 12, microsimulation methods appear to be ideally suited for this use.

We believe the demand for detailed estimates for very small areas will continue to grow. Technological changes—such as improvements in GIS and the greater availability and variety of geocoded data bases—will allow practitioners to meet that demand more easily than they could in the past. The development of synthetic population and household estimates will be particularly useful for these purposes. The trend toward greater geographic and demographic and other characteristic detail in estimates is likely to continue and perhaps to accelerate.

18.4 Some Challenges

Earlier in this chapter, we mentioned that more and more databases are being developed, containing more and more information on more and more people. Although these databases are extremely useful for many purposes, their growth

and widespread availability raise serious questions regarding the confidentiality of data. A backlash against the collection and utilization of data—based on privacy and confidentiality concerns—could have serious repercussions for all data users (Anderson and Seltzer 2009; El-Badry and Swanson 2007). In particular, it could threaten the development of linked data systems based on geographic or personal identifiers if ethical standards are in fact or in principle compromised. This would be a major blow for the production of population estimates. Seltzer (2010) provides important ideas regarding the balance between the maintenance of ethical standards and professional integrity and the generation and availability of data. His ideas suggest that developing new data sources and maintaining access to current sources may become an increasingly difficult challenge for producers and consumers of population estimates.

The widespread availability of demographic data and software will make it possible for more people to make population estimates, and to make them faster, cheaper, with greater demographic detail, and for a wider variety of geographic areas than ever before. This trend will be generally beneficial for data users and will raise the overall usefulness of population estimates. However, it will likely increase the number of estimates based on poor quality data, inadequate attention to detail, vested political or economic interests, and a poor understanding of the causes of population growth and demographic change. This proliferation of estimates will almost certainly be confusing. Data users will face a broader array of options than ever before and will have to do more homework in order to make the best choices. As elsewhere in a market-driven economy, caveat emptor (let the buyer beware) will be the order of the day.

Endnote

1. Portions of this chapter are adapted from Chapter 15 "New Directions in Population Estimation Research", in S. Smith, J. Tayman, and D. "Swanson. *Projecting State and Local Populations: Methodology and Analysis*. New York, NY: Kluwer Academic/Plenum Press. 2001.

References

Anderson, M. and W. Seltzer. 2009. Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues. *Journal of Privacy and Confidentiality* 1: 7–52;

Anselin, L., & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis. *Papers in Regional Science, 65*, 11–34.

Beckman, R. J., Baggerley, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research A, 30*, 415–429.

Chi, G., & Zhu, J. (2008). Spatial regression models for demographic analysis. *Population Research and Policy Review, 27*, 17–42.

Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., & Picado, R. (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. *Transportation Research Part A: Policy and Practice, 41*(5), 464–488.

Deng, C., Wu, C., & Wang, L. (2010). Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information. *International Journal of Remote Sensing, 31*(21), 5673–5688.

Dong, P., Ramesh, S., & Nepali, A. (2010). Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data. *International Journal of Remote Sensing, 31*(21), 5571–5586.

El-Badry, S. and D. A. Swanson. 2007 "Providing Census Tabulations to Government Security Agencies in the United States: The Case of Arab-Americans." *Government Information Quarterly* 24(2): 470–487

Griffith, D. A. (1987). *Spatial autocorrelation: A primer*. Washington, DC: Association of American Geographers.

Guo, J. Y., & Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record, 2014*, 92–101.

Klosterman, R. E., Brail, R. K., & Bossard, E. G. (1993). *Spreadsheet models for urban and regional analysis*. New Brunswick: Rutgers University, Center for Urban Policy Research.

Lo, C. P. (1986). Accuracy of population estimation from medium-scale aerial photography. *Photogrammetric Engineering and Remote Sensing, 52*, 1859–1869.

Lo, C. P. (1995). Automated population and dwelling unit estimation from high-resolution satellite images: A GIS approach. *International Journal of Remote Sensing, 16*, 17–34.

Lu, Z., IM, J., Quackenbush, L., & Halligan, K. (2010). Population estimation based on multi-sensor data fusion. *International Journal of Remote Sensing, 31*(21), 5587–5604.

Martins, J., F. Yusuf, and D. A. Swanson. (2011). *An Introduction to Consumer Demographics and Behaviour: Markets are People*. Dordrecht, Heidelberg, London, and New York: Springer.

Moran, P. A. (1948). The interpolation of statistical maps. *Journal of the Royal Statistical Society, Series B, 10*, 243–251.

Muller, K., & Axhausen, K. W. (2011). *Population synthesis for microsimulation: State of the art*. Paper presented at the 90th TRB Meeting, Washington, DC

Otani, N., Sugiki, N., Vichiensan, V., & Miyamoto, K. (July 2011). *Modifiable attribute cell problem in population synthesis for land-use microsimulation*. Paper presented at the 12th International Conference on Computers in Urban Planning and Urban Management, Lake Louise, AB, Canada.

Pol, L. G., & Thomas, R. K. (1997). *Demography for business decision making*. Westport: Quorum Books.

Pritchard, D. R., & Miller, E. J. (2009). *Advances in agent population synthesis and application in an integrated land use / transportation model*. Paper presented at the 88th TRB Meeting, Washington, DC

Rogers, A., Little, J., & Raymer, J. (2010). *The indirect estimation of migration: Methods for dealing with irregular, inadequate, and missing data*. Dordrecht, Heidelberg, London, and New York: Springer.

Seltzer, W. 2010. The role of ethics in a Federal Statistical Agency. Presentation given to the US Bureau of The Census, June 9th, Suitland, MD.
(https://pantherfile.uwm.edu/margo/www/govstat/integrity.htm).

Smith, S. K., Tayman, J., & Swanson, D. A. (2001). *State and local population projections: Methodology and analysis*. New York: Kluwer Academic/Plenum Publishers (Springer).

Statistics Canada. (2011). Microsimulation. (http://www.statcan.gc.ca/microsimulation/index-eng.htm).

Swanson, D. A., & Tedrow, L. M. (1984). Improving the measurement of temporal change in regression models used for county population estimates. *Demography, 21*(3), 373–381.

Swanson, D. A., & Walashek, P. A. (2011). *CEMAF as a census method: A proposal for a redesigned census and independent US Census Bureau*. Springer Briefs in Population Studies. Dordrecht, Heidelberg, London, and New York: Springer.

Verdino, G. (2010). *MicroMarketing: Get big results by thinking and acting small* New York: McGraw Hill.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M., & Ulfarsson, G. (2003). Microsimulation of urban development and location choices: Design and implementation of UrbanSim. *Networks and Spatial Economics, 3*(1), 43–67.

Wang, L., & Wu, C. (2010). Population estimation using remote sensing and GIS technologies. *International Journal of Remote Sensing 31*(21), 5569–5570.

Wu, S. S., Qui, X., & Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. *GIScience and Remote Sensing 42*, 80–96.

Demographic and Statistical Glossary[1]

Adjusted R-Squared A regression goodness-of-fit statistic that is "corrected" for the degrees of freedom. It is interpreted as the fraction of variation in the dependent variable that is explained by the regression model (independent variables and constant).

Adjusted Rate (See STANDARDIZATION).

Administrative Records Data collected by governmental (and sometimes private) organizations for taxation, registration, fee collection, and other administrative purposes that indirectly provide demographic information. These data are used by demographers for analyses, estimates, projections, and the evaluation of data specifically collected for demographic purposes (See also ADMINISTRATIVE RECORDS METHOD).

Administrative Records Method In the United States, a member of the family of component methods for estimating population that relies on a past census, vital statistics data, and migration data derived from tax returns (See also POPULATION ESTIMATE).

Age The length of time that a person has lived. A distinction is made between completed age and exact age, with completed age usually defined in terms of the last birthday and exact age being the exact time since birth. Conventions for determining age vary somewhat between cultures and countries.

Age Distribution (See POPULATION COMPOSITION).

Aggregation The process of assembling individual elements into summary form for purposes of presentation or analysis. For example, to assemble census records for individuals in a given area into a summary for the area as a whole.

Aggregation Bias A type of distortion that can result by attributing relationships found among summaries to the individual elements from which the summaries were obtained.

Allocation The assignment of values to cases for which "item non-response" is found in a sample survey or census. Many allocation methods are available, including automated algorithms (See also IMPUTATION, NON-RESPONSE, and SUBSTITUTION).

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0, © Springer Science+Business Media B.V. 2012

Allocation Error The extent to which estimates misallocate population over a set of geographic areas such as counties. The Index of Dis-similarly can be used to measure this type of error.

Alpha (α) The probability of committing a Type I error. A Type I error occurs when a researcher rejects a true null hypothesis. Alpha is also called the level of significance.

Alternative Hypothosis The hypothesis that complements the null hypothesis; usually it is the hypothesis that the researcher is interested in proving. The alternative hypothesis is generally denoted as Ha.

American Community Survey (ACS) In the United States, an on-going household survey conducted by the Census Bureau on a "rolling" geographic basis that is designed to provide demographic characteristics for counties, places, and other small areas. It may replace the long-form in the 2010 census.

Analysis of Variance A statistical technique that uses the F-test to determine whether there is a significant difference in the means of two or more independent groups.

Annexation In the United States, the legal act of adding territory to a governmental unit, usually an incorporated place, through the passage of an ordinance, court order, or other legal action.

Arithmetic Mean (See Mean) - The average of a group of numbers. The mean is a common measure of central tendency.

AT-Risk Population The persons to whom an event can potentially occur. In the form of the population at the middle of a given period, such as a year, it is used as an approximation of "Person-years lived"(See also EXPOSURE, PERSON-YEARS LIVED, and PROBABILITY).

Autocorrleation A problem that arises in regression analysis when the data occur over time and the error terms are correlated; also called serial correlation.

Autoregression A multiple regression forecasting technique in which the independent variables are time-lagged versions of the dependent variable.

Balancing Equation A term attributed to A. Jaffe that describes the basic population relation: $P_t = P_0 + I - O$, where P_t equals a given population at time $= 0 + t$, P_0= the given population at time =0, I = the number of persons entering the population through birth and immigration between time=0 and time $= 0 + t$, and O = the number of persons exiting the population through death and emigration between time=0 and time=0 + t (See also COHORT-COMPONENT METHOD, COMPONENT METHOD, ERROR OF CLOSURE, and RESIDUAL METHOD).

Base Period In a population estimate, this is the period between the initial year for which data are used to generate the estimate and the last year, which is known as the launch year. (See also LAUNCH YEAR, ESTIMATION HORIZON, and TARGET YEAR; and POPULATION ESTIMATE).

Baseline Survey A collection of data used for subsequent comparison or control.

Bayes Rule An extension of the conditional law of probabilities discovered by Thomas Bayes that can be used to revise probabilities.

Bayesian A way of doing inferential statistics based on ideas developed by Thomas Bayes. Some Bayesian techniques incorporate extensions of formal logic and others incorporate subjective probabilities.

Beta (β) The probability of committing a Type II error. A Type II error occurs when a researcher fails to reject a false null hypothesis.

Bias The deviation of an estimate or set of estimates from the correct value(s) in one direction (i.e., above or below the correct value(s)).

Bimodal Data sets that have two modes.

Binomial Distribution A discrete distribution which gives the probability of observing X successes in a fixed number (n) of independent Bernoulli trials.

Birth This refers only to a live birth and is referenced by the issuance of a birth certificate in the United States. As defined by the World Health Organization and allied organizations, a live birth is the complete expulsion or extraction from its mother of a product of conception, irrespective of the duration of pregnancy, which, after such separation, breathes or shows any other evidence of life such as beating of the heart, pulsation of the umbilical cord, or definite movement of voluntary muscles, whether or not the umbilical cord has been cut or the placenta is attached; each product of such a birth is considered live-born. According to this definition, the period of gestation, or the state of life or death at the time of registration, are not relevant. The US standard contains this definition plus a statement recommended by the American College of Obstetricians and Gynecologists to assist in the determination of what should be considered a live birth: "Heartbeats are to be distinguished from transient cardiac contractions; respirations are to be distinguished from fleeting respiratory efforts or gasps." (See also ABORTION and FETAL LOSS).

Block In the United States, the lowest level of geography for which census data are compiled. It is a typically a city block, but specifically is a small area bounded on all sides by identifiable features (e.g., roads, rivers, and city limits) that does not cross the boundaries of a given census tract. Each block is numbered uniquely within census tracts (See also BLOCK GROUP, BLOCK NUMBERING AREA, CENSUS TRACT, and CENSUS GEOGRAPHY).

Block Group In the United States, a cluster of blocks within a census tract that have the same first digit in their identifying numbers (See also BLOCK, BLOCK NUMBERING AREA, CENSUS TRACT and CENSUS GEOGRAPHY).

Block Numbering Area In the United States, these were used in the1990 census as the framework for grouping and numbering blocks in counties that did not have census tracts and provided coverage only for the block-numbered portion of a county. Starting with the 2000 Decennial census all US counties have census tracts. (See also BLOCK, BLOCK GROUP, CENSUS TRACT, and CENSUS GEOGRAPHY).

Bounds The error portion of the confidence interval that is added and/or subtracted from the point estimate to form the confidence interval.

Censal-Ratio Method A set of population estimation techniques found within the "Change in Stock Method" family that uses crude rates (e.g., birth and death) as measured at the most recent census date(s) and post-censal administrative

records. For example, a population estimate for 2002 can be obtained by dividing reported deaths for 2002 by the crude death rate measured in 2000 or by a crude death rate projected from 2000 to 2002. Often a series of Censal-Ratio estimates are averaged together. D. Swanson and R. Prevost showed in 1985 that the Ratio-Correlation Method is algebraically equivalent to a weighted average of censal-ratio estimates in which regression slope coefficients serve as weights (See also CHANGE IN STOCK METHOD, POPULATION ESTIMATE, RATIO-CORRELATION METHOD, and WEIGHTED AVERAGE).

Censored A condition affecting time-ordered data because the time frame for which data are collected does not cover the entire time span over which an event of interest may occur (e.g., a pregnancy at future point beyond the time frame in which data were collected). "Left-Censored" is used to describe the period preceding the data collection time frame and "right-censored," the subsequent period.

Census The count of a given population (or other phenomena of interest) and record its characteristics, done at a specific point in time and usually at regular intervals by a governmental entity for the geographic area or subareas under its domain (See also CENSUS COVERAGE, CENSUS DEFINED RESIDENT, POPULATION, POPULATION ESTIMATE, and SAMPLE).

Census Coverage An estimate of how complete a census was of a given population (See also COVERAGE ERROR, NET CENSUS UNDERCOUNT ERROR and TRUE POPULATION).

Census Coverage Error (See COVERAGE ERROR).

Census County Division In the United States, a statistical subdivision of counties in states established cooperatively by the Census Bureau and local groups in which minor civil divisions (e.g., townships) are not suitable for presenting census data (See also CENSUS GEOGRAPHY).

Census Defined Resident The concept of defining persons counted in a census in order to count each and every person once and only once. One of two counting bases is used: (1) De Jure, which attempts to locate persons at their usual residence; and (2) De facto, which counts people where they are found. The US Decennial Census is based on the De Jure method. (See also CENSUS, DE FACTO POPULATION, DE JURE POPULATION, DOMICLE, RESIDENCE, and USUAL RESIDENCE).

Census Designated Place (CDP) In the United States, a concentration of population enumerated during the decennial census in an area lacking legal boundaries, but recognized by the residents (and others) as a distinctive area with a name. A CDP is defined cooperatively by local officials and the Census Bureau. CDPs have been used since the 1980 census; from 1940 to 1970, they were called Unincorporated Places. (See also CENSUS GEOGRAPHY).

Census Error (See COVERAGE ERROR).

Census Geography In the United States, this refers to the hierarchical system of geographic areas that is used in conjunction with each decennial census. It consists of two major components: (1) areas defined by political or administrative boundaries (e.g., states, counties, townships, and cities.); and (2) areas defined by "statistical" boundaries (e.g., block, census designated place, census

tract). The areas so defined are used for analytical, political, and administrative purposes. Any country conducting a census uses some type of census geography. (See also BLOCK, CENSUS COUNTY DIVISION, CENSUS DESIGNATED PLACE, CENSUS TRACT, CITY, COUNTY. METROPOLITAN AREA).

Census Tract In the United States, this is the lowest level of "statistical geography" found in the decennial census designed to be homogenous with respect to population and economic characteristics (note that blocks and block groups, while at a lower level, are not designed with respect to population or economic homogeneity). Once established it is designed to be consistent in its boundaries for a long period of time. Starting with the 2000 census, all areas in the United States are tracted. (See also BLOCK, BLOCK GROUP, BLOCK NUMBERING AREA; and CENSUS GEOGRAPHY).

Central City Within the US Census Bureau's geography system, the core area in a metropolitan area. However, in other contexts, it is usually viewed as the concentrated inner area of a city consisting of business districts and urban housing.

Central Limit Theorem A theorem that states that regardless of the distribution of a population, the sample means and sample proportions will be normally distributed as long as the sample sizes are large.

Change in Stock Method A family of techniques for estimating population that is based on the measuring the total change in population since the last census rather than the components of change. Examples include the censal-ratio method, housing unit method, and the ratio-correlation method (See also COMPONENT METHOD, CENSAL-RATIO METHOD, HOUSING UNIT METHOD, and POPULATION ESTIMATE).

Chebyshev's Inequality A theorem stating that at least 1 - 1/k values will fall within +/− k standard deviations of the mean regardless of the shape of the distribution.

Chi-Squared Distribution A continuous distribution determined by the sum of the squares of k independent, normally distributed random variables.

Chi-Squared Goodness of Fit Test A statistical test used to analyze probabilities of multinomial distribution trials along a single dimension; compares expected, or theoretical, frequencies of categories from a populations distribution to the observed, or actual, frequencies from a distribution.

Chi-Squared Test of Independence A non-parametric statistical test used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent.

City In the United States, a type of incorporated place (See also CENSUS GEOGRAPHY).

Class Midpoint For any given class interval of a frequency distribution, the value halfway across the class interval; the average of the two class endpoints.

Closed Population A population for which in and out migration is minimal, if at all. For example, the population of the world as a whole is "closed," whereas the population of New York City is not.

Cluster Sampling A type of random sampling in which the population is divided into non-overlapping areas or clusters and elements are randomly sampled from the areas or clusters.

Coefficient of Determination The proportion of variability of the dependent variable accounted for or explained by the independent variable in a regression model.

Coefficient of Skewness A measure of the degree of skewness that exists in a distribution of numbers; compares the mean and the median in light of the magnitude of the standard deviation.

Coefficient of Variation The ratio of the standard deviation to the mean, expressed as a percentage.

Cohort A group of people who experience the same demographic event during a particular period of time such as their year of marriage, birth, or death. Cohorts are typically defined on the basis of a initiating signal event (e.g., birth), but they also can be defined on the basis of a terminating signal event (e.g., death). (See also COHORT ANALYSIS, COHORT EFFECT, COHORT MEASURE, and PERIOD).

Cohort Change Ratio (see HAMILTON-PERRY METHOD).

Cohort-Component Method A projection technique that takes into account the components of population change, births, deaths, and migration, and a population's age and sex composition, (See also BALANCING EQUATION, COMPONENT METHOD, and POPULATION PROJECTION).

Component Method In general, this refers to any technique for estimating population that incorporates births, deaths, and migration. Also known as a "Flow Method" (See also BALANCING EQUATION, CHANGE IN STOCK METHOD, COMPONENT METHOD I, COMPONENT METHOD II, and POPULATION ESTIMATE).

Component Method I A component method of estimating population that uses the relationship between local and national school enrollment data to estimate the net migration component. (See also COMPONENT METHOD, COMPONENT METHOD II, and POPULATION ESTIMATE).

Component Method II A component method of estimating population that uses the relationship between expected (survived) and actual local school enrollment data to estimate the net migration component. (See also COMPONENT METHOD, COMPONENT METHOD I, and POPULATION ESTIMATE).

Components of Change There are four basic components of population change: births, deaths, in-migration, and out-migration. The excess of births over deaths results in natural increase, while the excess of deaths over births results in natural decrease. The difference between in- and out-migration is net migration. In an analysis of special characteristics or groups, the number of components is broadened to include relevant additional factors (e.g., aging, marriages, divorces, annexations, and retirements), depending on the group (See also BALANCING EQUATION).

Composite Method A technique for estimating total population that is based upon independent estimates of age or age-sex groups that are summed to obtain the total population (See also POPULATION ESTIMATE).

Consolidated Metropolitan Statistical Area (See METROPOLITAN AREA).

Continuous Distributions Distributions constructed from continuous random variables.

Controlling The act of adjusting a distribution to an independently derived total value (See also CONTROLS).

Controls Independently derived estimates of a "total value" to which distributions are adjusted for purposes of improving accuracy, reducing variance and bias, or maintaining consistency. Controls can be univariate (one-dimensional) or multivariate (n-dimensional). Many methods may be used, including those that take account of whether the distributions have only positive values or both positive and negative values. (See also CONTROLLING, ITERATIVE PROPORTIONAL FITTING and PLUS-MINUS METHOD).

Correlation A measure of the degree of relatedness of two or more variables.

Covariance The variance of X and Y together.

Coverage Error In principle, this refers to the difference between the "true population" and the number reported in a set of data such as a census, survey, or set of administrative records. In practice, it is the difference between an *estimate* of the true number and the number reported in a set of data such as a census, survey, or set of administrative records (See also CENSUS, NET CENSUS UNDERCOUNT ERROR, TOTAL ERROR, and TRUE POPULATION).

County In the United States, a type of governmental unit that is the primary administrative subdivision of every state except Alaska and Louisiana (See also CENSUS GEOGRAPHY).

County Equivalent In the United States, a geographic entity that is not legally recognized as a county but referred to by the Census Bureau as the equivalent of a county for purposes of data presentation. Boroughs and certain statistically defined areas are county equivalents in Alaska and parishes are county equivalents in Louisiana (See also COUNTY and CENSUS GEOGRAPHY).

Critical Value The value that divides the non-rejection region from the rejection region.

Crude Rate A rate that relates a demographic event to the total population and makes no distinction concerning different exposure levels to the event. Examples include the Crude Birth Rate, Crude Death Rate, Crude Divorce Rate, Crude Marriage Rate, and Crude Rate of Natural Increase (See also AGE-SPECIFIC RATE, GENERAL RATE and RATE).

Current Population Survey(CPS) In the United States, a sample survey conducted monthly by the Census Bureau designed to represent the civilian non-institutional population that obtains a wide range of socio-economic-demographic data (See also CIVILIAN NON-INSTITUTIONAL POPULATION).

Curve A mathematical function, usually continuous and otherwise "well-behaved" that can be used as a model for a demographic process such as the change in the size of a population over time. Examples include the Exponential, Geometric, Gompertz, Linear, Logistic, and Polynomial.

Curve-Fitting The process of finding a mathematical function that serves as a model for a given demographic process.

Data Aggregation Compounding primary data into an aggregate to express data in summary form. National income is an example of aggregate data.

Data Linkage (see MATCHING).

Death The permanent disappearance of all evidence of life at any time after live birth has taken place. The loss of a member of a population, as recorded by a death certificate.

Decrement The exit of an individual or set of individuals from a "population" of interest, where the population is often defined by a model. In the case of a model such as the standard life table, such an exit would be due to death (See also INCREMENT and INCREMENT-DECREMENT LIFE TABLE).

De Facto Population A census concept that defines an enumerated person on the basis of his or her actual location at the time of the census (See also CENSUS DEFINED RESIDENT and DE JURE POPULATION).

De Jure Population A census concept that defines an enumerated person on a basis other than his or her actual location at the time of the census. The most common basis is the person's usual place of residence at the time of a census. (See also CENSUS DEFINED RESIDENT and DE FACTO POPULATION).

Degrees of Freedom A mathematical adjustment made to the size of the sample; used along with alpha to locate values in statistical tables.

Demographic Accounting The process of analyzing the change in a population using "stocks" (e.g., conditions such as the number of people in a given age-sex group) and "flows" (e.g., events such as births and deaths by age and sex) to show how the flows affect stocks over time. Ideally the stocks and flows should be measured without error and form mutually exclusive and exhaustive categories.

Demographic Analysis Generally, this refers to the methods of examination, assessment, and interpretation of the components and processes of population change, especially births, deaths, and migration. In the United States, it also refers to a specific method of estimating net census undercount using the components and process of population change.

Demographics A popular term for demography also used to represent demographic data and the application of demographic data, methods, and perspectives to activities undertaken by non-profit organizations, businesses, and governments (See also DEMOGRAPHY).

Demography The study of population, typically focused on five aspects: (1) size; (2) geographic distribution; (3) composition; (4) the components of change (births, deaths, migration); and (5) the determinants and consequences of population change. This term is usually used to refer to human populations, but it also is used to refer to non-human, particularly wildlife, populations. (See also DEMOGRAPHICS, FAMILY DEMOGRAPHY, HOUSEHOLD DEMOGRAPHY, ORGANIZATIONAL DEMOGRAPHY, and POPULATION).

Dependent Variable In regression analysis, the variable that is being predicted.

Descriptive Statistics Statistics that have been gathered on a group to describe or reach conclusions about that same group.

Difference-Correlation Method (See RATIO-CORRELATION METHOD).

Direct Estimation The measurement of demographic phenomena using data that directly represent the phenomena of interest.(See also INDIRECT ESTIMATION).

Direct Standardization The adjustment of a summary rate (e.g., the crude death rate) for a population in question found by computing a weighted average of group-specific rates (e.g., age specific death rates) for the population in question, where the weights consist of specific groups (e.g., the proportion in each age group) found in a "standard" population. This procedure is designed to produce a summary rate that controls for the effects of population composition (e.g., age) and is usually used for purposes of comparison with directly standardized rates for other populations computed using the same standard population. To standardize a crude death rate by the direct method, multiply the age-specific death rates for the population in question by the age-specific proportions in a standard population and sum the products. (See also INDIRECT STANDARDIZATION, STANDARD POPULATION, and STANDARDIZATION).

Discrete Distributions Distributions constructed from discrete random variables.

Diurnal Fluctuation For a given area, the change in its De Facto population over the course of a day (i.e., a 24 hour period) (See also DE FACTO POPULATION).

Domicile A person's fixed, permanent, and principal home for legal purposes (See also HOUSEHOLD, HOUSING UNIT, RESIDENCE and USUAL RESIDENCE).

Dual Residence The state of having two usual places of residence over a given period of time, which must be resolved in a De Jure census through the use of a set of procedures designed to count persons once and only once.

Dual-Systems Estimation Estimation of the true number of events or persons by matching the individual records in two data collections systems (See also MATCHING).

Dummy Variable Another name for a qualitative or indicator variable; usually coded as 0 or 1 and represents whether or not a given item or person possesses a certain characteristic.

Durbin-Watson Test A statistical test for determining whether a significant autocorrelation is present in a time-series regression model.

Emigrant A resident of a given country who departs to take up residence in another country (See also DOMESTIC MIGRATION, FOREIGN MIGRATION, and MIGRATION).

Enumeration The act of counting the members of a population in a census.

Enumeration District The area assigned to an enumerator during a census or survey of a given area.

Error Of Closure The difference between the change in population implied by census counts at two different dates and the change implied by an estimate not dependent on both census counts. This also can refer to a term added to the demographic balancing equation to account for errors in the components of change that cause them not to exactly match the change in measured independently for the population to which they apply. (See also BALANCING EQUATION and RESIDUAL METHOD).

Estimation Horizon In a population estimate, the period between the launch year and the target year (See also BASE PERIOD, LAUNCH YEAR, and POPULATION PROJECTION).

Estimate (See POPULATION ESTIMATE).

Ethnicity A common cultural heritage that sets a group apart on the basis of national origin, ancestry, language, religion, and similar characteristics. In the US decennial census, ethnicity is self-identified (See also RACE).

Event A change in condition or status (e.g., single to married).

Expected Value The long-run average of occurrences; sometimes referred to as the mean value.

Ex Ante This literally means "Before the event" in Latin. In the context of population estimates, it is usually used to mean that a set of estimates was developed that is not tested against a set of corresponding census numbers. However, the estimates may be compared to other information such as forecasts or other estimates. Such a comparison is known as an Ex Ante Test of Accuracy (See EX POST FACTO).

Ex Post Facto This literally means "After the fact" in Latin. In the context of population estimates, it is usually used to mean that a set of estimates was developed for a year in which a census was done either in advance of the census or without using the census numbers to inform the estimates. The estimates are then compared to the census numbers after the latter are made available. This comparison is known as an Ex Post Facto Test of Accuracy. (See EX ANTE)

Exponential Distribution A continuous distribution closely related to the Poisson distribution that describes the times between random occurrences.

Exponential Smoothing A forecasting technique in which a weighting system is used to determine the importance of previous time periods in the forecast.

Extinct Generations A technique introduced by P. Vincent in the early 1950s that is designed to estimate the number of extremely old persons in a population at a given date by cumulating deaths (to include, as needed, reported, estimated, and projected deaths) to given cohorts to the point where all members of the given cohorts have expired.

Extrapolation The process of determining (estimating or projecting) values that go beyond the last known data point in a series (e.g., the most recent census or estimate). It is typically accomplished by using a mathematical formula, a graphic procedure, or a combination of the two. (See also INTERPOLATION).

F Distribution A distribution based on the ratio of two random variances; used in testing two variances and in analysis of variance.

F Value The ratio of two sample variances, used to reach statistical conclusions regarding the null hypothesis; in ANOVA, the ratio of the treatment variance to the error variance.

Family In the United States, defined by the Census Bureau as those members of a household who are related through blood, adoption, or marriage (See also HOUSEHOLD).

Fertility The reproductive performance of a woman, man, couple, or group. Also a general term for the incidence of births in a population or group. One of the components of population change (See also COMPONENTS OF CHANGE and FECUNDITY).

Fips Code In the United States, one of a series of codes issued by the National Institute of Standards and Technology for the identification of geographic entities. FIPS stands for "Federal Information Processing Standards."

Flow Method (See COMPONENT METHOD).

Forecast (See POPULATION FORECAST).

Forward Survival Rate A type of rate that expresses survival of a population group from a younger age to an older age. Where a survival rate is not further labeled, forward survival is to be assumed (See also REVERSE SURVIVAL RATE and SURVIVAL).

Forward-Reverse Survival Method A technique used in both estimating intercensal populations and net migration between two censuses in which an "average" is taken between the results of using forward and reverse survival rates to age and "young" a given population, respectively, over the period between the two censuses (See also FORWARD SURVIVAL RATE, REVERSE SURVIVAL RATE, and SURVIVAL).

Frequency Distribution A summary of data presented in the form of class intervals and frequencies.

General Rate A rate that relates a demographic event to a set of people in a given population generally thought to be exposed to the event of interest, but one for which no distinction is made regarding different exposure levels to the event. A GENERAL RATE is distinguished from a CRUDE RATE because of the former's attempt to limit the population at risk to those actually exposed to the event in question, typically on the basis of age. Examples include the General Activity Rate, General Divorce Rate, General Enrollment Rate, and the General Fertility Rate. (See also AGE-SPECIFIC RATE, CRUDE RATE and RATE).

Geocoding The assignment of geographic or spatial information to data, such as coordinates. It is the most fundamental operation in the development of a "GIS" - Geographic Information System (See also GEOGRAPHIC INFORMATION SYSTEM).

Geographic Information System (GIS) A chain of operations involving the collection, storage, manipulation, and display of data referenced by geographic or spatial coordinates (e.g., coded by latitude and longitude).

GIS (See GEOGRAPHIC INFORMATION SYSTEM).

Gravity Model A model (borrowed from classical physics) based on the hypothesis that movement (migration, commuting, retail purchasing, etc.) between two areas is directly related to the population size of each area and inversely related to the distance between the two areas.

Gross Migration The sum of in-migration and out-migration for a given area (See also MIGRATION, and NET MIGRATION).

Group Quarters In the United States, a term used by the Census Bureau for places in which people reside that are not considered as "housing units." Such places include prisons, long-term care hospitals, military barracks, and school and college dormitories. (See also HOUSING UNIT and HOUSEHOLD POPULATION).

Growth Rate Often used as a general expression to describe the rate of change in a given population, even one that is declining (See also RATE and RATE OF CHANGE).

Hamilton-Perry Method A technique developed by H. Hamilton and J. Perry used in population projections that refers to a type of survival rate calculated for a cohort from two censuses. It includes not only the effects of mortality, but also the effects of net migration and relative census enumeration error (See also SURVIVAL RATE).

Head of Household A "marker" for a household, its type and structure. It is usually defined as the principal wage-earner or provider for a multi-person household, or, alternatively, is a person in whose name the housing unit is rented or owned. Persons living alone also are designated as heads of households. In principle, the number of households is equal to the number of household heads (See also HOUSEHOLD).

Headship Rate Usually defined as the proportion of the (household) population who are "heads" of households. (i.e., divide the number of households by the household population), often by age. It is often used in conjunction with population projections to obtain household projections (See also HEAD OF HOUSEHOLD, HOUSEHOLD and POPULATION PROJECTION).

Heteroskedasticity The condition that occurs when the error variances produced by a regression model are not constant.

Hispanic A person of Spanish or Latin American origin (also known as "Latino"). In the U. S. decennial census, persons of Hispanic origin are self-identified. Persons of Hispanic origin may be of any race (See also ETHNICITY and RACE).

Homeless Person Member of a population without a home or an official address usually found in shelters, on the streets, in vacant lots or vacant buildings.

Homoskedsticity The condition that occurs when the error variances produced by a regression model are constant.

Horizon (See PROJECTION HORIZON).

Hot Deck Imputation (See IMPUTATION).

Household Either a single person or a group of people making provision for food and other essentials of living, occupying the whole, part of, or more than one housing unit or other provision for shelter. The definitions vary by country. (See also DOMICLE, FAMILY, GROUP QUARTERS, HEAD OF HOUSEHOLD, HOMELESS PERSON, HOUSEHOLD POPULATION, and HOUSING UNIT).

Household Population Members of a population living in households, (as opposed to those who are homeless or living in group quarters - e.g., prisons, long-term care hospitals, military barracks, and school and college dormitories) (See also GROUP QUARTERS, HOMELESS PERSONS, HOUSEHOLD, HOUSING UNIT).

Housing Unit Generally a shelter intended for "separate use" by its occupants, such that there is independent access to the outside and the shelter is not a group quarters. A housing unit may be occupied or vacant. (See also DOMICLE, FAMILY, GROUP QUARTERS, HOMELESS PERSONS, and HOUSEHOLD).

Housing Unit Method A population estimation technique found within the "Change in Stock Method" family that uses current housing unit counts, vacancy estimates, and estimates of the number of persons per household to estimate the total household population, to which can be added an estimate of the group quarters population to obtain an estimate of the total population (See also CHANGE IN STOCK METHOD, HOUSEHOLD, HOUSING UNIT, GROUP QUARTERS, and POPULATION ESTIMATE).

Hypothesis Testing A process of testing hypotheses about parameters by setting up null and alternative hypotheses, gathering sample data, computing statistics from the samples, and using statistical techniques to reach conclusions about the hypothesis.

Immigrant Residents of a given country entering another country in order to take up permanent residence (See also DOMESTIC MIGRATION, FOREIGN MIGRATION, and MIGRATION).

Immigration (see FOREIGN MIGRATION).

Impairments Chronic health conditions involving abnormalities of body structure and appearance, the most common being chronic sensory and musculoskeletal conditions.

Imputation In a sample survey or census, a general term used to describe the assignment of values to cases for which one or more variables have missing values due to "non-response." Four common methods are: (1) deductive imputation, which is based on other information available from the case in question; (2) hot-deck imputation, which is based on information from "closest-matching" cases; (3) mean-value imputation, which uses means of variables as the source of assignment; and (4) regression-based imputation, in which models are constructed using cases with no missing values and a dependent variable is the one whose missing values will be imputed and the independent variables are those that yield acceptable regression equations (See also ALLOCATION, NON-RESPONSE, and SUBSTITUTION).

Incidence Rate The frequency with which an event, such as a new case of illness, occurs in a population at risk to the event over a given period of time.

Increment The entry of an individual or set of individuals into a population of interest, where the population of interest is often defined by a model. In the case of a model of nuptiality, such an entry would be marriage (See also DECREMENT and INCREMENT-DECREMENT LIFE TABLE).

Independent Variable In regression analysis, the predictor variable.

Index of Dis-similarity In the context of population estimates, this measure is used to determine the accuracy of a set of estimates for a set of geographic areas (e.g., counties) in an ex post facto test of accuracy. It also is known as the Index of Mis-allocation. It provides the percent of the estimated populations that would need to be re-allocated in order to match the corresponding census numbers for the same geographic areas. (see EX POST FACTO)

Indirect Estimation The measurement of demographic phenomena using data that do not directly represent the phenomena of interest. (See also DIRECT ESTIMATION).

Indirect Standardization The adjustment of a summary rate (e.g., the crude death rate) for a population in question found in part by computing a weighted average of group-specific rates (e.g., age-specific death rates) of a "reference" population, where the weights are the specific groups (e.g., proportion in each age group) of the population in question. This procedure is designed to produce a summary rate that controls for the effects of population composition (e.g., age) and is usually used for purposes of comparison with indirectly standardized rates for other populations computed using the same reference population. To standardize a crude (death) rate by the indirect method, first multiply the age-specific-(death) rates in the reference population by the population in the corresponding age groups of the population in question and sum the products to get the "expected" total (deaths) for the population in question. Then divide the expected total (deaths) into the total reported (deaths) for the population in question and multiply this ratio by the crude (death) rate of the reference population (See also DIRECT STANDARDIZATION and STANDARDIZATION).

Inferential Statistics Statistics that have been gathered from a sample and used to reach conclusions about the population from which the sample was taken.

Inflation-Deflation Method A technique that compensates for census coverage error by adjusting the demographic composition of the population of interest, but not its total number. It is sometimes used in conjunction with the cohort-component method of population projection, with the population in the launch year subject to "inflation" and the subsequent projection(s) subject to a compensating "deflation." It also is employed in the preparation of the official estimates of the population of the United States by age, sex, race, and ethnicity (Hispanic and non-Hispanic) (See also COVERAGE ERROR, COHORT-COMPONENT METHOD, LAUNCH YEAR, and POPULATION ESTIMATE).

In-Migrant A person who takes up residence within a "migration-defined" receiving area (the destination) after leaving a residence at a location outside of the receiving area (the origin), but one within the same country. For most countries, the destination and origin must be in different areas as defined by a political, administrative, or statistically-defined boundary. In the US, the destination must be in a different county than the origin for a person to be classified as an in-migrant by the Census Bureau (See also DESTINATION, IMMIGRANT, IN-MIGRATION RATE, MIGRANT, MIGRATION, MOVER, NET MIGRATION, NON-MIGRANT, ORIGIN, and OUT-MIGRANT).

In-Migration (See IN-MIGRANT).

In-Migration Rate The ratio of the number of in-migrants to a receiving area (the destination) over a given period to any one of a number of measures of the population of the receiving area, including the population at the end of the period, the population at the beginning of the period, and so on. Sometimes

the denominator is formed by using an approximation of the population at risk of migrating, e.g., the national population outside of the destination (See also DESTINATION, IN-MIGRANT, MIGRATION, NET MIGRATION RATE, and OUT-MIGRATION RATE).

Inter-censal The period between two successive censuses.

International Migration The movement across an international boundary for the purpose of establishing a new permanent residence (See also DOMESTIC MIGRATION).

Interpolation The calculation of intermediate values for a given series of numbers. It is typically accomplished by using a mathematical formula, a graphic procedure, or a combination of the two. It typically imparts or even imposes a regularity to data and can, therefore, be used for smoothing, whether or not the imposed regularity is realistic (See also EXTRAPOLATION and SMOOTHING).

Interquartile Range The range of values between the first and the third quartile.

Interval Estimate A range of values within which it is estimated with some confidence the population parameter lies.

Iterative Proportional Fitting A method for adjusting a multi-way distribution to a set of independently derived total values that approximates a least-squares approach. (See also CONTROLLING, CONTROLS and PLUS-MINUS METHOD).

J-Index A measure of the intrinsic growth of a population in a generation developed by A. J. Lotka that approximates the net reproduction rate and that, in turn, is approximated by the Replacement Index. When divided by the mean length of a generation it yields an estimate of the intrinsic rate of increase (See also MEAN LENGTH OF A GENERATION, NET REPRODUCTION RATE, REPLACEMENT INDEX and INTRINSIC RATE).

Jump-Off Year (See LAUNCH YEAR).

Karup-King Method A technique used to interpolate between given points or to subdivide groups. It is based on a polynomial osculatory formula (See also INTERPOLATION).

Kurtosis The amount of peakedness of a distribution.

Latino (See HISPANIC).

Launch Year The year in which a population estimate is launched, typically the year of the most recent census. Sometimes referred to as the "Jump-Off" year, it is the starting point of the estimation horizon (See also BASE PERIOD, ESTIMATION HORIZON, TARGET YEAR; and POPULATION ESTIMATE).

Left-Censored (See CENSORED).

Least Squares The process by which a regression model is developed based on calculus techniques that attempt to produce a minimum sum of the squared error values.

Level of Significance The probability of committing a Type I error; also know as alpha.

Life Table A statistical model comprised of a combination of age-specific mortality rates for a given population. A period life table (Also known as a cross-sectional life table) is constructed using mortality and age data from a single point in time; a generation life table (also known as a cohort life table) is based on the mortality of an actual birth cohort followed over time (to its extinction). A complete life table contains mortality information for single years of age, while an abridged table contains information by age group (See also GENERATION LIFE TABLE, INCREMENT-DECREMENT LIFE TABLE, PERIOD LIFE TABLE, LIFE EXPECTANCY and SURVIVAL RATE).

Life Table Functions The fundamental elements of a life table, to include the number surviving to a given age, the number of deaths to those surviving to a given birthday before they reach a subsequent birthday, the probability of dying before reaching a subsequent birthday for those who survived to a given birthday, the number alive between two birthdays, and the years of life remaining for those who survive to a given birthday (including birth). Life table functions can be interpreted in two ways: (1) as a depiction of the lifetime mortality experience of a cohort of newborns; and (2) as a stationary population that would result from a fixed mortality schedule and a constant number of annual births equal to the constant number of annual deaths resulting from the fixed mortality schedule (See also LIFE TABLE).

Logistic Curve A mathematical model that depicts an "S-Shaped" curve indicative of three stages of population change: (1) an initial period of slow growth; (2) a subsequent period of rapid growth; and (3) a final period in which growth slows and comes to a halt (See also DEMOGRAPHIC TRANSITION).

Long Form In the United States, the decennial census form given on a sample basis (approximately 1 in 6 households) that is designed to collect a wide range of population and housing data. The data collected go well beyond the basic information collected in the short form, which is given to the remaining households. Note, however, that the questions on the short form are contained in the long form (See also SHORT FORM).

Longevity (See LIFE SPAN).

Major Civil Division A "primary" subnational political area established by law or a related process. (See also CENSUS GEOGRAPHY and MINOR CIVIL DIVISION).

Master Address File (MAF) In the United States, the set of records maintained by the Census Bureau for purposes of conducting the decennial census. It is intended to represent the geographic location of every housing unit.

Matched Groups A group constructed on a case-by-case basis through matching of sets of records according to a limited number of characteristics.

Matched Pairs Data or measurements gathered from pairs of items or persons that are matched on some characteristic or from a before-and-after design and then separated into different samples; also called paired data or related measures.

Matched Pairs Test A t test to test the differences in two related or matched samples; sometimes called the t test for related measures or the correlated t –test.

Matching (of Records) Assembly of data in a common format from different sources but pertaining to the same unit of observation, (e.g., a person, household, or an event such as death). Also known as Record Matching and Data Linkage (See also DUAL SYSTEMS ESTIMATION).

Mean This refers to the arithmetic average. It is calculated by summing the values of a given variable (e.g., total county population) and then dividing the sum by the number of cases (e.g., the number of counties).

Mean Absolute Deviation The average of the absolute values of the deviations around the mean for a set of numbers.

Mean Absolute Error The average of the absolute values of errors of a set of estimates.

Mean Absolute Percentage Error (MAPE) In the context of population estimates, this refers to the arithmetic average of absolute percent differences between a set of estimate and corresponding census numbers. It is frequently used in an ex post facto test of accuracy. (see also EX POST FACTO, MEAN ALGEBRAIC PERCENT ERROR (MALPE) MEAN ABSOLUTE PERCENT ERROR RECALCULATED (MAPE-R) and Median Absolute Percent Error (MEDAPE)).

Mean Absolute Percentage Error Recalculated (MAPE-R) Because the MAPE is usually impacted by extreme errors, MAPE-R was developed as a way to mitigate the effects of extreme errors without losing significant information about the errors. (see also EX POST FACTO and MEAN ABSOLUTE PERCENT ERROR (MAPE)).

Mean Algebraic Percent Error (MALPE) Unlike MAPE, which uses absolute values, this measure uses the direction of errors in that it preserves both negative and positive errors. MALPE is rarely impacted by extreme errors. MAPE-R (see also EX POST FACTO and MEAN ABSOLUTE PERCENT ERROR (MAPE)).

Mean Error The average of all the errors of in a set of estimates.

Mean Square Error The average of all errors squared in a set of estimates data.

MEDAPE to MAPE RATIO This can be used as a descriptive tool to help judge the influence of outliers on the MAPE. A ratio of 1.0 indicates that outlying observations are not influencing the MAPE. Ratios above 1.0 indicate the potential magnitude of the overstatement of the typical error by the MAPE. A Ratio below 1.0 might suggest the existence of a left-skewed distribution; in this case, the MAPE potentially understates the typical error. (see also MAPE, MEDIAN ABSOLUTE PERCENT ERROR (MEDAPE) and SKEWNESS).

Median The middle value in an ordered array of numbers.

Median Absolute Percent Error (MEDAPE) This is the percent error which falls in the middle of the error distribution; half of the absolute percent errors are larger and half are smaller. MEDAPE is useful when the objective is to highlight the "typical" error and ignore the effects of outlying errors. One drawback of the MEDAPE is that it ignores most of the information contained in the error distribution; it is only based on one or two observations. (see also MAPE, MAPE-R, and MEDAPE to MAPE RATIO).

Metric Data Interval and ratio level data; also called quantitative data.

Metropolitan Area In the United States, this refers to a family of specific census geographies intended to represent a large population nucleus and aggregations thereof. Specific types of include "Primary Metropolitan Statistical Area" and "Standard Consolidated Statistical Area." (See also CENSUS GEOGRAPHY, PRIMARY METROPOLITAN STATISTICAL AREA, and STANARD CONSOLIDATED STATISTICAL AREA).

Migrant A person who makes a relatively permanent change of residence from one country, or region within a country (an origin), to another (the destination) during a specified (migration) period. For most countries, the change must be across a political, administrative, or statistically-defined boundary for a person to be classified as a migrant. In the US, the origin and destination must be in different counties for a person to be classified as a migrant (See also DESTINATION, EMIGRANT, IMMIGRANT, IN-MIGRANT, MIGRATION, MOVER, NON-MIGRANT, ORIGIN, and OUT-MIGRANT).

Migration A general term for the incidence of movement by individuals, groups or populations seeking to make relatively permanent changes of residence. One of the components of population change (See also ASYLEE, COMMUTING, COMPONENTS OF CHANGE, DESTINATION, DOMESTIC MIGRATION, EMIGRANT, FOREIGN-BORN, GROSS MIGRATION, IMMIGRANT, IN-MIGRANT, INTERNALLY DISPLACED PERSONS, INTERNATIONAL MIGRATION, MIGRANT, MOBILITY, MOVER, NATIVE, NET MIGRATION, NON-MIGRANT, ORIGIN, OUT-MIGRANT, and REFUGEE).

Migration Stream A group of migrants with a common origin and destination over a given period. (See also COUNTERSTREAM).

Military Population Persons who are members of the armed forces.

Military Dependent Population Persons who are dependents of members of the armed forces.

Minor Civil Division A "secondary" subnational political area established by law or a related process (See also CENSUS GEOGRAPHY and MAJOR CIVIL DIVISION).

Mobility, Geographic Any move resulting in a change of residence (See also DOMESTIC MIGRATION and MIGRATION).

Mobility Rate The ratio of the number of movers over a given time period to the population at risk of moving over the same period (See also IN-MIGRATION RATE, MIGRATION, and OUT-MIGRATION RATE).

Mobility Status A Classification of people based on their residential locations at the beginning and end of a given time period.

Model A generalized representation of a demographic process, set of demographic relationships, pattern of mortality, fertility, migration, or marriage, or method of population estimation or projection.

Mover A person who reports in a census or survey that he or she lived at a different address at an earlier date (e.g., five years before the census or survey). In the US, a mover is classified by the Census Bureau as a person who changed residence, but within the same county (See also MIGRATION).

Moving Average When an average of data from previous time periods is used to forecast the value for ensuing time periods and this average is modified at each

new time period by including more recent values not in the previous average and dropping out values from the more distant time periods that were in the average. It is continually updated at each new time period.

Multicollinearity A problematic condition that occurs when two or more of the independent variables of a multiple regression model are highly correlated.

Multiple Regression Regression analysis with one dependent variable and two or more independent variables.

Multi-Regional Analysis An analysis of multi-regional systems in which spatial and demographic factors are linked.

Multi-State Life Table An extension of the standard life table in which multiple transitions between states are possible and the transitions are expressed in terms of transition probabilities between states (See also DECREMENT, INCREMENT, and INCREMENT-DECREMENT LIFE TABLE).

Natural Increase The excess of births over deaths in a population is defined as natural increase; an excess of deaths over births is defined as natural decrease.

Net Census Undercount Error The estimated level of coverage error in a census computed by algebraically adding estimated overcounts and estimated undercounts for population groups (e.g., age-sex-race) and summing them. (See also COVERAGE ERROR, NON-RANDOM ERROR and TRUE POPULATION).

Net Migration The difference between the number of in-migrants and the number of out-migrants for a given area (e.g., a county) over a given period of time: Net = In - Out (See also GROSS MIGRATION, IN-MIGRANT, MIGRATION, NET MIGRATION RATE, and OUT-MIGRANT).

Net Migration Rate The ratio of net migration for a given area (e.g., a county) over a given period to any one of a number of measures of the population of the area, including the population at the end of the period, the population at the beginning of the period, and so on. Sometimes the denominator is formed by using a population outside of the area (e.g., the national population outside of the county) (See also IN-MIGRATION RATE, MIGRATION, NET MIGRATION, and OUT-MIGRATION RATE).

Net Number of Migrants (See Net Migration).

Nonlinear Regression Multiple regression models in which the models are nonlinear, such as polynomial models, logarithmic models, and exponential models.

Nonmetric Data Nominal and ordinal level data; also called qualitative data.

Non-Metropolitan Population The number of people living outside large urban settlements. In the US, this represents the population outside Metropolitan Statistical Areas (See also CENSUS GEOGRAPHY).

Non-Migrant In a census or survey, an individual who resided in an area both at the beginning and end of the designated migration period. Alternatively, an individual who has neither migrated into nor migrated out of his or her area of residence. (See also IN-MIGRANT, MIGRATION, MOVER, NET MIGRATION and OUT-MIGRANT).

Nonparametric Statistics A class of statistical techniques that make few assumptions about the population and are particularly applicable to nominal and ordinal level data.

Non-Random Sampling Sampling in which not every unit of the population has the same probability of being selected into the sample.

Non-Random Error All errors not due to the effects of random sample selection (i.e., random error). It can occur both in a sample survey and in a population census. Examples include non-response, incorrect answers by a valid respondent, answers given by a non-valid respondent, as well as coding and other processing errors. Statistical inference can only be used to estimate random error, not non-random error (See also NET CENSUS UNDERCOUNT ERROR, NON-RESPONSE, POPULATION, RANDOM ERROR, SAMPLE, and TOTAL ERROR).

Non-Rejection Region Any portion of a distribution that is not in the rejection region. If the observed statistic falls in this region, the decision is a fail to reject the null hypothesis.

Non-Response Missing data on a form used in a survey or census due to a number of reasons, including the refusal of a respondent to answer, the inability to locate a potential respondent, the inability of a respondent (or informant) to answer questions, or the omission of answers due to a clerical or some other form of error. Total non-response refers to a case (i.e., an observation) in which all variables have missing values and item non-response refers to a case in which fewer than all variables have one or more missing values. Imputation is often used to estimate values for cases in which they are missing (See also IMPUTATION, NON-RANDOM ERROR, NON-RESPONDENT).

Non-Response Error (See NON-RESPONSE).

Non-Respondent In a sample survey or census, a respondent who refuses to be interviewed, or is otherwise unable to take part (See also NON-RESPONSE).

Normal Distribution A widely known and much-used continuous distribution that fits the measurements of many human characteristics and many machine-produced items.

Null Hypothesis The hypothesis that assumes the status quo - that the old theory, method, or standard is still true; the complement of the alternative hypothesis.

Observed Significance Level Another name for the p-value method of testing hypotheses.

Odds Ratio As defined for a dichotomous variable, the ratio of the proportion of the population having a characteristic of interest to the proportion not having the characteristic. For example, the percent of the population to the percent not in poverty The logarithm of the odds ratio is termed a logit. (See also LOGIT).

One-Tailed Test A statistical test wherein the researcher is interested only in testing one side of the distribution.

Open-Ended Interval A class interval in a distribution of grouped data that is not bounded on one end. For example, in a distribution of data on income, the highest income class may be given as $100,000 or more; in a life table the last age interval may be given as 85 years and over. In a longitudinal analysis, the period between the most recent occurrence of an event of interest (e.g., a live birth) and a subsequent time point . For example, in a survey of birth histories, the period

between the second birth and the survey would constitute an open-ended interval for a woman reporting two births, whereas the periods between her first and second birth would be a closed interval.

Open Interval (See OPEN-ENDED INTERVAL).

Origin The place of residence that a migrant left at the start of a given (migration) period (See also DESTINATION, MIGRANT, and MIGRATION).

Osculatory Interpolation An interpolation method that involves combining higher-order polynomial formulas into one equation, designed to provide a smooth junction between two adjacent groups of data (e.g., age group 5-9 and age group 10-14). (See also INTERPOLATION).

Outliers Data points that lie apart from the rest of the points.

Out-Migrant A person who leaves his or her residence in a "migration-defined" sending area (the origin) to take up residence at a location outside of the sending area (the destination), but one within the same country. For most countries, the origin and destination must be in different areas as defined by a political, administrative, or statistically-defined boundary In the US, the origin must be in a different county than the destination for a person to be classified as an out-migrant by the Census Bureau (See also DESTINATION, EMIGRANT, IN-MIGRANT, MIGRANT, MIGRATION, MOVER, NET MIGRATION, NON-MIGRANT, ORIGIN, and OUT-MIGRATION RATE).

Out-Migration (See INTERNAL MIGRATION).

Out-Migration Rate The ratio of the number of out-migrants from a sending area (the origin) over a given period to some measure of the population of the sending area, including the population at the beginning of the period, the population at the end of the period, and so on. (See also ORIGIN, OUT-MIGRANT, IN-MIGRATION RATE, MIGRATION, and NET MIGRATION RATE).

Overcount In a census, this can be due to counting some people more than once, counting people in a census who are not members of the population in question, or a combination of both (See also NET CENSUS UNDERCOUNT ERROR and UNDERCOUNT).

Own-Child Method A census or survey-based method for measuring fertility that uses counts of children living with their mothers.

P-Value Method A method of testing hypotheses in which there is no preset level of alpha. The probability of getting a test statistic at least as extreme as the observed test statistic is computed under the assumption that the null hypothesis is true. This probability is called the p value, and it is the smallest value of alpha for which the null hypothesis can be rejected.

Paired Data Data gathered from pairs of items or persons that are matched on some characteristic or from a before-and-after design and then separated into different samples; also called matched pairs data or related measures.

Parameter A descriptive measure of the population.

Parametric Statistics A class of statistical techniques that contain assumptions about the population and that are used only with interval and ratio level data.

Partial Migration Rate The number of in-migrants from a particular origin to a given destination relative to the population of either the origin or destination.

Participation Rate The proportion of a population or segment of a population with a certain characteristic, usually social or economic, e.g., the proportion aged 10-14 who are enrolled in school.

Percent (See PROPORTION).

Percentiles Measures of central tendency that divide a group of data into 100 parts.

Period Measure A summary measure of data collected during a brief period of time (usually one year) that typically represent more than one cohort (See also COHORT MEASURE and PERIOD ANALYSIS).

Place of Residence (See USUAL RESIDENCE).

Plus-Minus Method A "controlling" technique that attempts to compensate for both increasing and decreasing subsets of a population of interest by using two separate adjustment factors. For example, one might use the plus-minus method in adjusting post censal population estimates of census tracts to an estimate of the county containing the tracts, if some tracts show growth since the last census and others show decline (See also CONTROLLING, CONTROLS, and ITERATIVE PROPORTIONAL FITTING).

Point Estimate An estimate of a population parameter constructed from a statistic taken from a sample.

Population In the demographic sense, the "inhabitants" of a given area at a given time, where inhabitants could be defined either on the De Facto or De Jure basis (but not a mixture of both). Note that the concept of "area" can be generalized beyond the geographical sense to include, for example, formal organizations. In the statistical sense, the term "population" refers to the entire set of persons (or phenomenon) of interest in a particular study, as compared to a sample, which refers to a subset of the whole (See also CENSUS, DE FACTO POPULATION, DE JURE POPULATION, DEMOGRAPHY, SAMPLE, and SPECIAL POPULATION).

Population In the statistical sense, A collection of persons, objects, or items of interest.

Population At Risk (See AT-RISK POPULATION).

Population Composition The classification of members of a population by one or more characteristics such as age, sex, race, and ethnicity. It can be presented in either absolute or relative numbers. "Population distribution" and "population structure" are often used as synonyms (See also POPULATION DISTRIBUTION).

Population Decrease Reduction in the number of inhabitants in an area.

Population Density Number of persons per unit of land area.

Population Distribution Usually used to refer to the location of a population over space at a given time, but sometimes used as a synonym for population composition (See also POPULATION COMPOSITION).

Population Dynamics Changes in population size and structure due to fertility, mortality, and migration, or the analysis of population size and structure in these terms.

Population Estimate An approximation of a current or past population of a given area at a given time, or its distribution and composition, in the absence of a complete enumeration, ideally done in accordance with one of two standards for defining a population, De Facto or De Jure (See also ADMINISTRATIVE RECORDS METHOD, CENSAL-RATIO METHOD, CENSUS, CENSUS DEFINED RESIDENT, CHANGE IN STOCK METHOD, COMPONENT METHOD, COMPOSITE METHOD, DE FACTO POPULATION, DE JURE POPULATION, HOUSING UNIT METHOD, POPULATION PROJECTION, RATIO-CORRELATION METHOD, RATIO ESTIMATION, SYNTHETIC METHOD, and VITAL RATES METHOD).

Population Forecast An approximation of the future size of the population for a given area, often including its composition and distribution. A forecast usually is one of a set of projections selected as the most likely representation of the future (See also POPULATION ESTIMATE and POPULATION PROJECTION).

Population Projection The numerical outcome of a particular set of implicit and explicit assumptions regarding future values of the components of population change for a given area in combination with an algorithm. Strictly speaking, it is a conditional statement about the size of a future population (often along with its composition and distribution), ideally made in accordance with one of the two standards used in defining a population, De Facto or De Jure (See also BASE PERIOD, CENSUS, CENSUS DEFINED RESIDENT, COHORT-COMPONENT METHOD, DE FACTO POPULATION, DE JURE POPULATION, LAUNCH YEAR, POPULATION FORECAST, POPULATION ESTIMATE, PROJECTION HORIZON, and TARGET YEAR).

Population Register An administrative record system used by many countries (e.g., China, Finland, Japan, and Germany) that requires residents to register their place of residence, usually at a local police station. By itself, such a system provides limited demographic information (e.g., total population), but where it can be matched to other administrative record systems (e.g., tax, social and health care services), the result is often a system that provides a wide range of longitudinal and cross-sectional demographic information.

Population Size The number of persons inhabiting a given area at a given time. (See also CENSUS and POPULATION).

Power The probability of rejecting a false null hypothesis.

Power Curve A graph that plots the power values against various values of the alternative hypothesis.

Prevalence The number of persons who have a given characteristic (e.g., disease, contraceptive use, impairment, labor force participation) in a given population at a designated time or who had the characteristic at any time during a designated period, such as a year (See also PREVALENCE RATE).

Prevalence Rate The proportion of persons in a population who have a particular disease or attribute at a specified time (point prevalence) or at any time during a designated period, such as a year (period prevalence). (See also PREVALENCE).

Primary Metropolitan Statistical Area In the United States, a census-based piece of geography defined by the Office of Management and Budget that is comprised of a central city and county and adjoining counties linked to the central city by social and economic interactions that meet prescribed standards. (See also CENSUS GEOGRAPHY, METROPOLITAN AREA, and STANDARD CONSOLIDATED AREA).

Probability A ratio in which the numerator consists of those in a population experiencing an event of interest (e.g., death) over a specified period of time, while the denominator consists of the at-risk population. (See also AT-RISK POPULATION, PROPORTION, RATE, and RATIO).

Probit A mathematical transformation, often used in event history analysis, for "linearizing" the cumulative normal distribution of a variable of interest. The probit unit is $y = 5 + Z(p)$, where $p =$ the prevalence of response at each dose level and $Z(p) =$ the corresponding value of the standard cumulative normal distribution (See also EVENT HISTORY ANALYSIS and LOGIT).

Projection (See POPULATION PROJECTION).

Projection Horizon In a population projection, the period between the launch year and the target year (See also Base Period, Launch Year, and Target Year; and POPULATION PROJECTION).

Proportion A ratio used to describe the status of a population with respect to some characteristic (e.g., married), where the numerator is part of the denominator. When multiplied by 100, a proportion is known as a "percent." (See also PROBABILITY, RATE, and RATIO).

Public Use Microdata Sample (Pums) In the United States and elsewhere this usually refers to a hierarchically-structured data set that contains individual, family, and household information in a given record and for which confidentiality is maintained by deleting identifying information. It is typically obtained by sampling from census records.

Quartiles Measures of central tendency that divide a group of data into four subgroups or parts.

Race In theory, classification of the members of a population in terms of biological ancestry, in which a range of physical characteristics, such as hair structure, cephalic index, and so on, is employed to assign persons to one category or another (one of three principal races or unclassified). In demographic practice, classification of the members of a population in terms of socially constructed definitions of membership in categories in which skin color and other characteristics, including national ethnic affiliations, may be the basis of assignment by census or survey enumerators or by self-enumeration. In the US decennial census, persons are self-identified by race (See also ETHNICITY).

Raking (See CONTROLLING).

Random Error The difference between a statistic of interest (e.g., mean age) found in a sample unaffected by non-random error and its corresponding parameter (e.g., mean age) found in the population from which the sample was drawn. Random error can only occur in a sample, never in a population. It is often

referred to as sample error or sampling error (See also NON-RANDOM ERROR, POPULATION, SAMPLE, and TOTAL ERROR).

Rate Technically, this type of ratio is the same as a probability. However, the term is often applied to the type of ratio known as a proportion, as in the case of "vacancy rate," which is the ratio of unoccupied housing units to all housing units. It is also applied to other types of ratios in which the denominators are not precisely the "at-risk populations," as is the case of the crude birth rate (See also AT-RISK POPULATION, PROBABILITY, PROPORTION, and RATIO).

Rate-Correlation Method (See RATIO-CORRELATION METHOD).

Rate of Change The change of population during a given period express as a rate. The rate may relate to the entire period, in which case the denominator is usually the initial population. Alternatively, it may be an average annual rate, in which case the rate may assume annual compounding, continuous compounding, or some other function (See also POPULATION CHANGE).

Rate of Natural Increase The result of subtracting the crude death rate from the crude birth rate. For a population closed to migration it provides the rate of increase (or the rate of decrease if the crude death rate exceeds the crude birth rate) (See also CRUDE BIRTH RATE, CRUDE DEATH RATE, and INTRINSIC RATE).

Ratio A single number that expresses the relative size of two other numbers - i.e., a quotient, which is the result of dividing one number by another. (See also PROBABILITY, PROPORTION, and RATE).

R-Squared The coefficient of multiple determination; a value that ranges from 0 to 1 and represents the proportion of the dependent variable in a multiple regression model that is accounted for by the independent variables.

Random Sampling Sampling in which every unit of the population has the same probability of being selected for the sample.

Random Variable A variable that contains the outcomes of a chance experiment.

Range The difference between the largest and the smallest values in a set of numbers.

Ratio-Correlation Method A regression-based subnational population estimation technique found within the "Change in Stock Method" family. Introduced by R. Schmitt and A. Crosetti in the early 1950s: (1) the dependent variable consists of the ratio formed by dividing the most recent population proportion for a set of sub-areas (e.g., proportion of a state population in each of its counties at the most recent census) by the population proportion for the same subareas at an earlier time (i.e., the previous census); and (2) the independent variables consist of corresponding ratios of proportions for symptomatic indicators of population (e.g., school enrollment, automobile registrations, births, deaths) available from administrative records. Variations of the Ratio-Correlation Method include the Difference-Correlation Method introduced by R. Schmitt and J. Gier in 1966 and the Rate-Correlation Method introduced by D. Swanson and L. Tedrow in the 1984 (See also CENSAL-RATIO METHOD, CHANGE IN STOCK METHOD, POPULATION ESTIMATE, and WEIGHTED AVERAGE).

Ratio Estimation A set of techniques used to estimate population based on ratios across geographic areas, variables, or both. (See also POPULATION ESTIMATE).

Record Linkage (See MATCHING).

Record Matching (See MATCHING).

Reference Population (See STANRARD POPULATION).

Regression The process of constructing a mathematical model or function that can be used to predict or determine one variable by any other variable.

Rejection Region If a computed statistic lies in this portion of a distribution, the null hypothesis will be rejected.

Residence The place where a person lives. Defined differently in different censuses, but often interpreted as "usual residence," which is the case in the US decennial census based on the De Jure method (See also CENSUS, CENSUS-DEFINED RESIDENT, DE JURE, DOMICLE, USUAL RESIDENCE).

Residential Mobility A change of residence, either in the same city or town, or between cities, states, countries, or communities.

Residual The difference between the actual Y value and the Y value predicted by the regression model; the error of the regression model in predicting each value of the dependent variable.

Residual Method A technique that estimates inter-censal net migration for a given area by subtracting from the most recent census count, the algebraic sum of inter-censal births and deaths added to the population counted at the preceding census. Resulting estimates are confounded by differences in net census under-count error (See also BALANCING EQUATION, ERROR OF CLOSURE, and NET MIGRATION).

Response Variable The dependent variable in a multiple regression model; the variable that the researcher is trying to predict.

Return Migration A move back to point of origin, whether domestic or foreign (See also MIGRATION).

Reverse Record Check A technique used to estimate census coverage error that attempts to match a sample drawn from a reliable source of records independent of the census with data collected in the census. For example, a reverse record check may attempt to match a sample of births over a 10-year period with children under 10 in the census, or a sample of enrollees under Medicare with the elderly population in the census (See also CENSUS and COVERAGE ERROR).

Reverse Stream (See COUNTERSTREAM).

Reverse Survival Method Any method of estimating population involving backward "survival" of a population to an earlier date (See also SURVIVAL RATE).

Right-Censored (See CENSORED).

Robust Describes a statistical technique that is relatively insensitive to minor violations in one or more of its underlying assumptions.

Rural Population Usually defined as the residual population after the urban population has been identified (See also URBAN POPULATION).

Rural-Urban Migration The migration from rural to urban areas, both internal and international.

Sample A subset of a population (in the statistical sense) for which data are typically collected in a "survey," which is a way of providing respondents with questions to be answered (e.g., through personal interviews, telephone interviews, mail-out/mail-back questionnaires). Samples may also be selected from administrative and other records such that interviews are not needed because data are taken directly from the records themselves (e.g., from Medicare files). Samples may be defined in a number of ways, but if statistical inference is to be used, a sample's elements should have a known probability of selection, or at least a reasonable approximation thereof, so that "random error" can be estimated (See also CENSUS, NON-RANDOM ERROR, POPULATION, and RANDOM ERROR).

Sample Error (See RANDOM ERROR).

Sample Proportion The quotient of the frequency at which a given characteristic occurs in a sample and the number of items in the sample.

Sample Size Estimation An estimate of the size of sample necessary to fulfill the requirements of a particular level of confidence and to be within a specified amount of error.

Sample Space A complete roster or listing of all elementary events for an experiment.

Sampling Error (See RANDOM ERROR).

School-Age Population Children if school age, usually defined by the ages for which school attendance is compulsory, which varies from country to country and sometimes with a given country.

Seasonal Adjustment A statistical modification to a data series to reduce the effect of seasonal variation (See also SEASONAL VARIATION).

Seasonal Variation Seasonal differences in the occurrence of data collected over time and reported at least quarterly (See also SEASONAL ADJUSTMENT).

Self-Enumeration A method of conducting a census or sample survey in which respondents fill out questionnaire themselves, usually in connection with a mail-out/mail-back design for distributing and retrieving the questionnaires.

Serial Correlation A problem that arises in regression analysis when the error terms of a regression model are correlated due to time-series data; also called autocorrelation.

Short Form In the United States, the decennial census form asking a limited range of basic population and housing questions and distributed to about five-sixths of the households, with the so-called "long form" being distributed to the remaining households. Note, however, that the questions on the short form are contained in the long form, so in effect all households receive the short form (See also LONG FORM).

Simple Random Sampling The most elementary of the random sampling techniques; involves numbering each item in the population and using a list or roster of random numbers to select items for the sample.

Skewness The lack of symmetry of a distribution of values (See also MAPE).

Small Area The subdivisions of the primary political subdivisions of a country. In the United States, counties and their subdivisions are usually considered small areas, although some limit the term to subcounty areas such as census tracts, block groups, and blocks and the areas that can be aggregated from them (See also CENSUS GEOGRAPHY).

Smoothing The adjustment of data to eliminate or reduce irregularities and other anomalies assumed to result from measurement and other errors. A common application of smoothing procedures is in connection with single-year-of-age data that appear to be affected by age heaping (See also AGE-HEAPING and INTERPOLATION).

Special Population Population groups identified separately for purposes of a census and or sample survey because of their distinctive living arrangements, such as college students, prison inmates, residents of nursing homes, and military personnel and their dependents. Special populations usually are characterized by components of change very different from the broader populations in which they are found, sometimes because of laws or regulations governing them. (See also COMPONENTS OF CHANGE and POPULATION).

Standard Consolidated Area In the United States, a combination of Primary Metropolitan Statistical Areas, with a total population of at least 1,000,000, established by the Office of Management and Budget. (See also CENSUS GEOGRAPHY, METROPOLITAN AREA, and PRIMARY METROPOLITAN STATISTICAL AREA).

Standard Deviation The square root of the variance.

Standard Error of the Estimatie - A standard deviation of the error of a regression model.

Standard Error of the Mean The standard deviation of the distribution of sample means.

Standard Error of the Proportion The standard deviation of the distribution of sample proportions.

Standard Metropolitan Statistical Area (See PRIMARY METROPOLITAN AREA).

Standard Population A "reference" population used for purposes of analyzing a population of interest. Also, specifically, a population whose age distribution is employed in the calculation of standardized rates by the direct method. (See also DIRECT STANDARDIZATION and STANDARDIZATION).

Standardization In the demographic sense, the adjustment of a summary rate (e.g., the crude death rate) to remove the effects of population composition (e.g., age), usually done to compare rates across populations with different compositions. There are two general types of standardization, direct and indirect. The type selected is dependent on the data available for the population(s) of interest (See also DIRECT STANDARDIZATION, INDIRECT STANDARDIZATION, POPULATION COMPOSITION, STANDARD POPULATION, and STANDARDIZED RATE).

Standardization In the statistical sense,

Standardized Normal Distribution Z distribution; a distribution of Z scores produced for values from a normal distribution with a mean of 0 and a standard deviation of 1.

Standardized Rate A rate that has been subjected to standardization. (See also STANDARDIZATION).

Statistic A descriptive measure of a sample.

Statistics A science dealing with the collection, analysis, interpretation, and presentation of numerical data.

Stepwise Regression A step-by-step multiple regression search procedure that begins by developing a regression model with a single predictor variable and adds and deletes predictors one step at a time, examining the fit of the model at each step until there are no more significant predictors remaining outside the model.

Stratified Random Sampling A type of random sampling in which the population is divided into various non-overlapping strata and then items are randomly selected into the sample from each stratum.

Stocks and Flows A **stock** (or "level variable") in this broader sense is some entity that is accumulated over time by inflows and/or depleted by outflows. Stocks can only be changed via flows. Mathematically a stock can be seen as an accumulation or integration of flows over time - with outflows subtracting from the stock. Stocks typically have a certain value at each moment of time - e.g. the number of population at a certain moment. A **flow** (or "rate") changes a stock over time. Usually we can clearly distinguish inflows (adding to the stock) and outflows (subtracting from the stock). Flows typically are measured over a certain interval of time - eg. the number of births over a day or month.

Subjective Probability A probability assigned based on the intuition or reasoning of the person determining the probability.

Substitution In a sample survey or census, the process of assigning values for a case in which there is "total non-response." Many substitution methods are available, including automated algorithms (See also ALLOCATION, IMPUTATION, and NON-RESPONSE).

Suburban A popular term referring to the residential area surrounding a central city. Such an area may follow the transportation lines and be dependent on the central city both economically and culturally but, increasingly, such areas are becoming the equivalent of central cities to suburbs of their own. (See also URBAN FRINGE).

Survey (See SAMPLE).

Sum of Squares Error The sum of the residuals squared for a regression model.

Sum of Squares X The sum of the squared deviations about the mean of a set of values.

Survival Rate A rate expressing the probability of survival of a population group, usually an age group, from one date to another and from one age to another. A survival rate can be based on life tables or two censuses. When based on two censuses, the rate includes not only the effects of mortality, but also the effects of

net migration and relative census enumeration error. (See also FORWARD SURVIVAL RATE, HAMILTON-PERRY METHOD, LIFE TABLE, SURVIVAL, and SURVIVORSHIP FUNCTION).

Synthetic Method A member of the family of ratio estimation methods that is used to estimate characteristics of a population in a subarea (e.g., a county) by reweighting ratios (e.g., prevalence rates or incidence rates) obtained from survey or other data available at a higher level of geography (e.g., a state) that includes the subarea in question. (See also POPULATION ESTIMATE, RATIO ESTIMATION and WEIGHTED AVERAGE).

T Distribution A distribution that describes the sample data in small samples when the standard deviation is unknown and the population is normally distributed.

T Statistic The computed value of t used to reach statistical conclusions regarding the null hypothesis in small-sample analysis.

Target Year In a population estimate, the final year for which a population is estimated, the end point of the estimation horizon (See also BASE PERIOD, LAUNCH YEAR, and ESTIMATION HORIZON; and POPULATION ESTIMATE).

Temporary Migration A type of migration, both internal and international, in which the duration of stay is temporary. Data for temporary migration are not normally included in the official data on internal or international migration and are usually obtained from a special sample survey.

Tiger (See TOPOLOGICALLY INTEGRATED GEOGRAPHIC ENCODING AND REFERENCING SYSTEM).

Time Series Data Data gathered on a given characteristic over a period of time at regular intervals.

Topologically Integrated Geographic Encoding and Referencing System. (TIGER) A digital database of geographic features (e.g., roads, rivers, political boundaries, census statistical boundaries, etc.) covering the entire United States. It was developed by the US Census Bureau to facilitate computerized mapping and areal data analysis. (See also GEOGRAPHIC INFORMATION SYSTEM).

Total Error In a sample, the theoretical sum of random error and non-random error, which in practice can at best only be roughly approximated because of the difficulty of estimating non-random error. Also known as Total Sample Error. In a census, total error is comprised solely of non-random error (see also NON-RANDOM ERROR, RANDOM ERROR, and TRUE POPULATION).

Trend Long-run general direction of a business climate over a period of several years.

Trend Extrapolation (See EXTRAPOLATION).

True Population In theory, the population that would be counted if there were no errors in a census. In practice, it is a value representing the theoretical actual number for the population at a given date, which cannot be precisely measured, but which can be roughly approximated by adjusting a census for net census undercount error (See also CENSUS and NET CENSUS UNDERCOUNT ERROR).

Truncation Bias Distortion of results due to the systematic omission from an analysis of values that fall below or above a given range.

Turnover A term sometimes employed to refer to the sum of the components of change during a period, i.e., births plus deaths plus immigrants/in-migrants plus emigrants/out-migrants.

Two-Tailed Test A statistical test wherein the researcher is interested in testing both sides of the distribution.

Type I Error An error committed by rejecting a true null hypothesis.

Type II Error An error committed by failing to reject a false null hypothesis.

Undercount In a census, the omission of valid members of the population in question (See also NET CENSUS UNDERCOUNT ERROR and OVERCOUNT).

Under-Enumeration (See UNDERCOUNT).

Under-Registration The omission of persons or events from a registration system or other administrative record system.

Uniform Distribution A relatively simple continuous distribution in which the same height is obtained over a range of values; also called the rectangular distribution.

Unincorporated Place (See CENSUS DESIGNATED PLACE).

Urban Fringe The densely settled area surrounding the core city of an urbanized area. Sometimes population referred to as the suburban area (See also SUBURBAN).

Urban Population Usually defined as a large population in a densely-packed area that meets criteria derived from geographic, social, and economic factors, which, in turn, may vary by country (See also RURAL POPULATION).

Urbanization Growth in the proportion of persons living in urban areas; the process whereby a society changes from a rural to an urban way of life.

Usual Residence The place where one usually eats and sleeps, a concept associated with a De Jure census (See also CENSUS, CENSUS-DEFINED RESIDENT, DE JURE, DOMICILE, LABOR FORCE, and RESIDENCE).

Variance The average of the squared deviations about the arithmetic mean for a set of numbers.

Variance Inflation Factor A statistic computed using the R-squared value of a regression model developed by predicting one independent variable of a regression analysis by other independent variables; used to determine whether there is multicollinearity among the variables.

Vital Events Births, deaths, fetal losses, abortions, marriages, annulments, divorces–any of the events relating to mortality, fertility, marriage, and divorce recorded in registration systems (See also VITAL STATISTICS).

Vital Rates Method A censal-ratio method of population estimation introduced by D. Bogue in the 1950s that uses crude birth and crude death rates (See also CENSAL-RATIO METHOD and POPULATION ESTIMATE).

Vital Records (See VITAL STATISTICS).

Vital Statistics Data on births, deaths, fetal losses, abortions, marriages, and divorces usually compiled through registration systems or other administrative record systems (See also VITAL EVENTS).

Weighted Average Usually an arithmetic mean of an array of specific rates or ratios, with variable weights applied to them representing the relative distribution of the populations on which the rates or ratios are based. More generally, a summary measure of a set of numbers (absolute numbers or ratios), computed as the cumulative product of the numbers and a set of weights representing their relative importance in the population. An unweighted average is one in which each number in the set has the same weight (e.g., 1 or 1/n, where n is the total set of numbers) (See also CENSAL RATIO METHOD and SYNTHETIC METHOD).

Z Distribution A distribution of Z scores; a normal distribution with a mean of 0 and a standard deviation of 1.

Z Score The number of standard deviations a value (X) is above or below the mean of a set of numbers when the data are normally distributed.

Zip Code Administrative areas set up by the US Postal Service as postal delivery areas and used for marketing and related purposes in the United States. They have fluid boundaries that do not correspond to any established political area or statistical area of the decennial census but may approximate some small areas defined by the census (See also CENSUS GEOGRAPHY).

A Demography Timeline Relevant to Population Estimates

1661 Giovanni Battista RICCIOLI: "De verisimili hominum numero," *Geographiae et Hydrographiae Reformatae* (scholarly estimate of the earth's population and in various nations)

1741 Johann SÜSSMILCH: *Die Göttliche Ordnung in den Veränderungen des menschlichen Geschlechts, aus der Geburt, dem Tode und der Fortplanzung desselben erwissen* (The Divine Order. . ., most painstaking estimate of world population to his time, editions: 2nd, 1761; 3rd, 1765)

1748 Swedish law requiring national compilation of parish vital statistics records

1753 Robert WALLACE: *The Numbers of Man in Ancient and Modern Times* (French ed., 1760)

1755 Benjamin FRANKLIN: *Observations Concerning the Increase of Mankind, Peopling of Countries, etc*.

1786 Pierre Simon LAPLACE proposes a censal ratio method for estimating population

1790 Mar 1 In accordance with Article 1, Section 2 of the US Constitution, the first United States census began; the world's first continuous, periodic national census

1801 periodic census begins in England and France

1802 LAPLACE's censal ratio method used to estimate the population of France.

1850 US Census collected individual-level data for the first time.

1850 ff. Otto L. HÜBNER: *Geographisch-statistiche Tabellen* (until 1919)

1895 Edwin CANNAN: "The Probability of a Cessation of the Growth of Population in England and Wales during the Next Century" *Economic Journal* (An early use of the cohort-component method of population projection)

1911 E.C. SNOW: "The Application of the Method of Multiple Correlation to the Estimation of Post-censal Population," *Journal of he Royal Statistical Society* (first known use of linear regression to estimate population)

1938. Henry S. SHRYOCK: "Methods of estimating post-censal population," *Journal of the American Statistical Association*.

1940 US census first employed sampling, 1 in 20 households received "supplemental questions," based on experience gained from unemployment surveys done in the 1930s and the first "Current Population Survey," which itself preceded the 1940 census; participants include W . Edwards DEMING, Philip HAUSER, Morris HANSEN, William HURWITZ, William MADOW, who collectively would make important contributions to the theory and practice of sampling

1941. Margaret HAGOOD. *Statistics for Sociologists.*

1946 US Congress enacts federal programs that use statistical formulas in conjunction with census and other data for funding purposes; sets the stage for a tremendous expansion of the use of such formulas from the 1950s to the 1990s

1947. Hope Tisdale ELDRIGDGE: "Problems and methods of estimating post-censal population," *Social Forces*.

1949. US Census Bureau, "Illustrative examples of two methods of estimating the current population of small areas," *Current Population Reports*, Series P-25, No. 20.

1950. Don BOGUE: "A technique for making extensive population estimates," *Journal of the American Statistical Association*.

1954 Robert C. SCHMITT and Albert H. CROSETTI: "Accuracy of the Ratio-Correlation Method for Estimating Post-censal Population," *Land Economics* (seminal work on subnational population estimation)

1962. C. Horace HAMILTON and Josef PERRY: "A short method for projecting population by age from one decennial census to another," *Social Forces*.

1967 United Nations: *Methods of Estimating Basic Demographic Measures from Incomplete Data, Manual IV*

1968 William BRASS et al.: *The Demography of Tropical Africa* (Important methodological developments for dealing with missing and defective demographic data)

1971 Henry S. SHRYOCK, Jacob S. SIEGEL, and ASSOCIATES: *The Methods and Materials of Demography* (Important comprehensive text book and reference work on demography)

1978. Maria GONZALEZ and Christine. HOZA: "Small area estimation with application to unemployment and housing estimates," Journal of the American Statistical Association (Seminal paper on synthetic estimation method)

1979. Robert FAY and Roger HERRIOT: "Estimates of income for small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*.

1980 Evelyn KITAGAWA et al.(editors): *Estimating Population and Income for Small Places* (Comprehensive treatment of small area population estimation)
1982 Everett LEE and Harold GOLDSMITH (editors): Population Estimates: *Methods for Small Area Analysis*.
1983 United Nations: *Indirect Techniques for Demographic Estimation, Manual X*
1987. Statistics Canada, *Population Estimation Methods, Canada*.
1987. P. PLATEK, J.N.K. RAO, C. SÄRNDAL, and M.P. SINGH (editors): *Small Area Statistics: An international Symposium*.
1995. N. RIVES, W. SEROW, A. LEE, H. GOLDSMITH, and P. VOSS (editors): *Basic Methods for Preparing Small-Area Population Estimate*s.
1996 American Community Survey initiated, starting the process that could lead to collecting in a "continuous measurement" sample survey of the long form of the US Census
2002. J. SIEGEL: *Applied Demography: Applications to Business, Government, Law, and Public Policy*.

Endnote

1. Sources of information used in compiling this Glossary /Demography Timeline include the following publications:

Easton, J. and J. McColl (n.d.). *Statistics Glossary v1.1* (http://www.cas.lancs.ac.uk/glossary_v1.1/main.html).
Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.
Swanson, D. A. 2012. *Learning Statistics: A Manual for Sociology Students*. San Diego, CA: Cognella Academic Publishing.
Swanson, D.A. and G. E. Stephan. 2004. Glossary and Demography Timeline. pp 751–786 in J. Siegel and D.A. Swanson (eds.) *The Methods and Materials of Demography 2nd Edition*. New York, NY: Elsevier Academic Press.

Index

Note: The entries that appear more than 100 times will only show up once in the Index, for the page on which the subject appears. You also should note that entries in the Glossary are not included in the index of subjects.

D.A. Swanson and J. Tayman, *Subnational Population Estimates*,
The Springer Series on Demographic Methods and Population Analysis 31,
DOI 10.1007/978-90-481-8954-0,