Environmental and Ecological Statistics

Ganapati P. Patil
Sharad D. Gore
Charles Taillie

Composite Sampling

A Novel Method to Accomplish Observational Economy in Environmental Studies

Springer

Composite Sampling

Environmental and Ecological Statistics

Series Editors

G.P. Patil
Distinguished Professor of Mathematical Statistics
Director, Center for Statistical Ecology and Environmental Statistics
Editor-in-Chief, Environmental and Ecological Statistics
The Pennsylvania State University

Timothy G. Gregoire
J.P. Weyerhaeuser, Jr. Professor of Forest Management
School of Forestry and Environmental Studies
Yale University

Andrew B. Lawson
Division of Biostatistics & Epidemiology
Medical University of South Carolina

Barry D. Nussbaum
Chief Statistician
U.S. Environmental Protection Agency

The Springer series **Environmental and Ecological Statistics** is devoted to the cross-disciplinary subject area of environmental and ecological statistics discussing important topics and themes in statistical ecology, environmental statistics, and relevant risk analysis. Emphasis is focused on applied mathematical statistics, statistical methodology, data interpretation and improvement for future use, with a view to advance statistics for environment, ecology, and environmental health, and to advance environmental theory and practice using valid statistics.

Each volume in the **Environmental and Ecological Statistics** series is based on the appropriateness of the statistical methodology to the particular environmental and ecological problem area, within the context of contemporary environmental issues and the associated statistical tools, concepts, and methods.

Previous Volumes in this Series

VOLUME 1
Landscape Pattern Analysis for Assessing Ecosystem Condition
Glen D. Johnson and Ganapati Patil
VOLUME 2
Pattern-Based Compression of Multi-Band Image Data for Landscape Analysis
Wayne Myers and Ganapati Patil
VOLUME 3
Modeling Demographic Processes In Marked Populations
Edited by David L. Thomson, Evan G. Cooch and Michael J. Conroy
VOLUME 4
Composite Sampling: A Novel Method to Accomplish Observational Economy in Environmental Studies
Ganapati P. Patil, Sharad D. Gore and Charles Taillie (Deceased)

For further volumes:
http://www.springer.com/series/7506

Ganapati P. Patil · Sharad D. Gore ·
Charles Taillie (Deceased)

Composite Sampling

A Novel Method to Accomplish Observational Economy in Environmental Studies

Springer

Ganapati P. Patil
Center for Statistical Ecology
and Environmental Statistics
The Pennsylvania State University
University Park, PA, USA
gpp@stat.psu.edu

Charles Taillie
(Deceased)
Center for Statistical Ecology
and Environmental Statistics
Penn State University
University Park, PA, USA

Sharad D. Gore
Department of Statistics
University of Pune
Pune, India
sdgore@stats.unipune.ac.in

ISBN 978-1-4419-7627-7 e-ISBN 978-1-4419-7628-4
DOI 10.1007/978-1-4419-7628-4
Springer New York Dordrecht Heidelberg London

© Springer Science+Business Media, LLC 2011
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper.

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

To
Lalit, Parimal, Parag, Pawan, Priya
Rekha, Archis, Reshma
Marcella

With love and affection

Acknowledgements

This monograph provides, for the first time, a most comprehensive statistical account of composite sampling as an ingenious environmental sampling method to help accomplish observational economy in a variety of environmental studies. It is primarily based on a 4-year effort on composite sampling in theory and practice under a cooperative research agreement CR-821531 of the Penn State Center for Statistical Ecology and Environmental Statistics with what was known as the Statistical Analysis and Computing Branch of the Division of Environmental Statistics and Information in the Office of Policy, Planning, and Evaluation of the United States Environmental Protection Agency. Several unpublished results and applications appear for the first time in this monograph.

Our appreciation goes to the members of the Statistical Analysis and Computing Branch for their technical interest and administrative support. Our thanks are particularly due to Dr. N. Phillip Ross, the division director, and his distinguished staff statisticians, Herbert Lacayo, Robert O'Brien, John Warren, and Barry Nussbaum, the branch chief, and now the chief statistician of EPA. We should also express our sincere thanks to several of our friends and colleagues for discussions regarding various issues and aspects of composite sampling in environmental studies. They are Marllyn Boswell, George Flatman, Richard Gilbert, Glen Johnson, Henry Kahn, Gianfranco Lovison, Royal Nadeau, Dean Neptune, Charles Rohde, David Schaeffer, Llew Williams, and a few others.

And, finally, our immense appreciation to Barbara Freed and Janet Huff for taking care of it all at the Penn State Center for Statistical Ecology and Environmental Statistics.

Contents

Chapter 1
Introduction

Environmental problems often arise from a suspicion or an allegation that something is wrong somewhere. It can be the suspicion that a municipal waste site is hazardous for neighboring communities; it can be an allegation that a nuclear waste disposal site is unsafe due to possibilities of exposure to radioactive material; or it can be the problem of monitoring industrial effluents, managing a sewage water system, disposal of municipal wastes, treatment of ash from a municipal incinerator, assessing the quality of groundwater, or one of numerous other situations that call for immediate attention from both the authority and the society.

Resolving each of these problems requires evidence in order to conclude one way or the other. The evidence is usually collected in the form of data that come from laboratory analyses of field samples collected from an appropriate site. Statistical considerations can guide the sampling and analysis efforts to make the most effective use of the available resources. That is, it is desired that as many issues be addressed as are possible with the available resources without compromising the quality of the outcome. This entails a carefully planned and implemented sampling protocol, an appropriately developed statistical treatment of the collected data, and a comprehensive interpretation of the results of the statistical procedures applied to the data.

Since sampling at the site of the suspected or alleged violation of environmental safety is the very foundation of any such investigation, it is very important to understand the statistical issues involved in sampling of such sites. Here sampling is assumed to include the stages of selection, acquisition, and quantification of sampling units from a site. It is then quite appropriate to point out the fundamental dilemma of sampling that scientists – both statistical scientists and substantive scientists – face during planning and implementation of a sampling design.

Sampling consists of selection, acquisition, and quantification of a part of the population. While selection and acquisition apply to physical sampling units of the population, quantification pertains only to the variable of interest, which is a particular characteristic of the sampling units. A sampling procedure is expected to provide a sample that is representative with respect to some specified criteria. For example, simple random sampling is known to provide a sample that is representative in that the sample mean is an unbiased estimator of the population mean. In addition to

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_1, © Springer Science+Business Media, LLC 2011

representativeness, it is also expected that a sample be informative. Considerations of desirable criteria for representativeness and informativeness as variously defined usually lead to a desirable sample size of $\bar{n}$ or more. On the other hand, considerations of resources in terms of cost, time, and effort usually lead to an affordable sample size of $\underline{n}$ or less.

A common experience is that $\underline{n} \ll \bar{n}$. Thus, what is desirable is not affordable, and what is affordable is not adequate. How do we deal with this dilemma? One way is to adopt the data quality objectives (DQO) process (see, for example, EPA QAMS, 1991). After discussing the issue of optimizing the sampling design, the DQO process makes the following recommendation: "If it appears that there is no (sampling) design that will meet both the limits of uncertainty (i.e., level of precision) and the budget constraints, then determine whether to compromise by relaxing the limits on uncertainty or other practical constraints or by finding additional funding to achieve the desired limits on uncertainty within the boundaries of the study."

Statistical theory attempts to deal with this situation by exploring additional information, over the entire population, on the variable of interest or an associated variable. This helps stratify the population or do something related in terms of clusters or primary sampling units leading to a reduction in the sample size and hence in the associated cost while maintaining the representativeness, informativeness, and/or precision of the inference and may thus resolve the conflict between cost and precision. However, if it is not possible to have additional information for the entire population, then the existing statistical theory falls short of resolving this conflict. This is where composite sampling approach can help when feasible.

Operationally, composite sampling recognizes the distinction between selection, acquisition, and quantification. In certain applications, it is a common experience that the costs of selection and acquisition are not very high, but the cost of quantification, or measurement, is substantially high. In such situations, one may select a sample sufficiently large to satisfy the requirement of representativeness and precision and then, by combining several sampling units into composites, reduce the cost of measurement to an affordable level. Thus composite sampling offers an approach to deal with the classical dilemma of desirable vs. affordable sample sizes, when conventional statistical methods fail to resolve the problem.

Composite sampling, at least under idealized conditions, incurs no loss of information for estimating the population means. But an important limitation to the method has been the loss of information on individual sample values, such as for example the extremely large value. In many of the situations where individual sample values are of interest or concern, composite sampling methods can be suitably modified to retrieve the information on individual sample values that may be lost due to compositing. In this monograph, we present statistical solutions to these and other issues that arise in the context of applications of composite sampling.

A composite sample is formed by mixing several individual samples or subsamples. The terminology of "sample" that we use in this document, as in "individual sample" and "composite sample," will be that of a physical sampling unit and not a collection of observations as in the statistical sense. A composite sample may be a physical mix of individual sampling units (or subunits) or it may be a batch of

unblended individual sampling units that are subjected to measurement as a group. Composite sampling techniques were developed for engineering applications where it is necessary to test several pieces of equipment simultaneously. If we refer to a group of individual sampling units to be tested simultaneously as a composite sample, then group testing is a subset of composite sampling. Statistically, group testing techniques are equivalent to composite sample techniques.

Individual sampling units can be created in two different ways. In batch or group sampling, the individual sample units exist before sampling occurs. In bulk or integrated sampling, the sampling process creates the sampling units, which may themselves be composite samples formed by the sampling process. Further compositing steps may occur subsequent to the formation of sampling units (see Rohde, 1979). For example, in the sampling of coal that is being passed along a moving conveyor belt, an individual sample is created by dropping a portion of the coal through an opening. This sample is then reduced in particle size prior to testing. Further compositing (forming composite samples), subsampling, and reduction steps may occur before a final subsample is extracted for making measurement.

As far as we know, the original application of composite sampling was the estimation of the prevalence of transmission of plant viruses by insects (see Watson, 1936, for example). In this application, sets of potential insect vectors of disease are allowed to feed upon potential host plants. The transmission rate can be estimated from the number of plants that subsequently become diseased.

Apparently, the next application of composite sampling occurred during World War II. Dorfman (1943) proposed to classify US servicemen as either having or not having been infected with syphilis, by detecting the presence or absence, respectively, of an antigen of the syphilis-causing bacterium in samples of their blood. In this application, composite samples are formed from subsamples of blood samples drawn from the subjects. Composite samples testing positive for the presence of the trait prompt additional tests on aliquots from the original blood samples comprising that composite, until all individual samples are classified. Evidently, it was the 1951 edition of Feller's book that brought this composite sampling problem to the attention of statisticians.

In a third development, composite sampling has become a standard practice in the sampling of soils, biota, and bulk materials when the goal is estimation of the mean, with either a desired standard error or with limits on the cost of sampling, which includes the cost of measurement. Though composite sampling has been most frequently applied in the areas of estimating prevalence, classifying individual samples, and estimating the mean, we anticipate that it may potentially be applicable to other statistical problems requiring a sample. For example, the use of composite sampling has been extended to general estimation and to hypothesis testing (Mack and Robinson, 1985; Messner et al., 1990).

It is important to distinguish between two types of situations in which composite sampling is often used. In the conventional case of sampling from a finite population, sampling units exist prior to sampling. Human populations, items of industrial product, or experimental subjects are examples of this type of setup. Not all populations of interest have this characteristic. In such cases, we can consider

the population as being composed of (hypothetical) sampling units, which are really subsets of the population. Sampling in these cases therefore amounts to choosing a subset of the population. Usually we use the term "finite population" to refer to a population which consists of pre-existing identifiable sampling units and the term "bulk population" to refer to a population in which objects to be selected in a sample are themselves created by the sampling process. In finite populations the individual values may be of interest while in bulk populations they are not, although some measure of variability between samples may serve to measure internal variability of the bulk population.

The term "sampling unit" is conventionally used to refer to an element of a finite population, whereas the term "sample" or "grab" is used to refer to a subset of a bulk population. In the literature on composite sampling, however, these terms are interchangeably used. Rohde (1979) defines the following terms to avoid the confusion between what is selected from a population and what is used to form composite samples; between a sample as selected or formed and a sample as subjected to laboratory analysis or measurement.

Primary Sampling Unit: A unit or an object in the population that is selected by the sampling process.

Secondary Sampling Unit: A portion or aliquot of a primary sampling unit that can be measured or observed. Note that a secondary sampling unit may coincide with a primary sampling unit.

Composite Sample: A mixture of several secondary sampling units.

Laboratory Sample: A subsample of a larger sample unit (either a secondary sampling unit or a composite sample) that is sent to the laboratory for measurement. Again note that a laboratory sample may coincide with a secondary sampling unit (usually a case of finite population sampling without compositing) or a composite sample (a case where the composite sample itself is subjected to measurement).

Primary sampling units may exist before sampling (finite populations) or may be created as a result of sampling (bulk populations). Often only a portion of a primary sampling unit is actually used for laboratory procedures. A composite sample may be a physical blend of secondary (or primary) sampling units or it may only be a conceptual combination, as in group testing. As there are two distinct types of populations, there also are two distinct types of sampling methods.

Batch or Group Sampling: Primary sampling units exist before sampling occurs (finite population sampling). Compositing may be done for convenience or as a cost-saving device. Individual values of the primary sampling units may or may not be of intrinsic interest.

Bulk or Integrated Sampling: Primary sampling units are created by the sampling process (bulk population sampling). Compositing is often performed by the very mechanism by which primary sampling units are selected. Further compositing may also be performed. Individual values of primary sampling units are of no intrinsic interest, though their variability may provide a measure of internal variability of the population.

The usual goal of composite sampling is to obtain the same information that would have been obtained by measuring the individual samples but at a reduced monetary cost, effort, or data variation. Sometimes the goal is to obtain more

efficient procedures, and sometimes the goal is to obtain information when the individual sample measurements are unavailable.

Practical considerations guide the feasibility and the physical formulation of composite sampling. The physical formulation includes the composite sample design in time or space, the number of individual samples composited, the mixing and subsampling methods, and the retesting procedures. Some information, such as the range and the variance of individual sample measurements, is lost upon compositing. However, some desired information may still be recoverable by certain statistical and/or laboratory procedures. For example, the variance of individual samples can be estimated from the sample variance of multiple composite samples.

If the integrity of the individual sample values is changed by compositing, then composite sampling may not be the desired approach. Changes in the integrity of sample values can occur, for example, if volatile chemicals evaporate upon mixing of samples (Cline and Severin, 1989) or if there is interaction among sample constituents. In the first case, compositing of individual sample extracts may perhaps be a reasonable alternative to mixing individual samples as they are collected.

Another limitation on composite sample techniques is imposed by potential dilution. If an individual sample with a moderate sample value is combined with individual samples having relatively small sample values, then the composite sample value may be small enough to be undetectable. Reporting limits (Rajagopal, 1990, "personal communication") or action levels (Williams, 1990) of hazardous chemical concentrations set by law to be close to the limit of detection eliminates the possibility of composite sampling. When not influenced by regulation, the determination of the composite sample size will often be constrained by the magnitude of the reporting or action level relative to the magnitude of the detection limit. The presence of measurement error further decreases the bound on the composite sample size that is necessary to avoid non-detection (i.e., false-negative) problems.

When a physical blend of individual samples is subsampled, homogeneity of the blend is desired for precise estimation of the mean response. The mixability of sampling units will therefore influence the composite sample formulation. This is usually more of a concern with the sampling of solids than with the sampling of gases or liquids. Gases or liquids usually can be mixed to the level of molecules, but the discreteness of solid material units (e.g., grain kernels, particles of coal or soil) can add complexity to composite sampling. This complexity can be handled substantively by effective subsampling, grinding, and mixing techniques and statistically by assuming that the individual samples contribute to different subsamples from the composite sample in different proportions. These proportions may then be considered as random quantities, varying from subsample to subsample according to a probability distribution.

The ability of the compositing device to blend the individual samples thoroughly may also affect the degree of homogeneity with respect to the variable under study. If the individual samples are themselves heterogeneous, that is, if different portions or increments of the individual samples are characterized by different values, and if the compositing device is unable to eliminate such heterogeneity, then portions of the same individual samples in different subsamples may carry different values.

This situation, which Elder et al. (1980) call "within-increment heterogeneity," can also add complexity to composite sampling.

Often, measurements on multiple attributes are desired. However, if retesting is performed in order to classify individual samples, such as classifying the samples as being above or below an action level in hazardous waste monitoring, it is unclear how to optimize the retesting relative to the different attributes (Schaeffer et al., 1982). For example, should chemicals be tested independently or does there exist dependence in the multivariate information that can be used to improve cost-efficiency? Classifying for multiple attributes remains an open problem in composite sampling.

Particular circumstances may dictate the feasibility or infeasibility of composite sampling. In a nationwide study of the chemical concentrations in human adipose tissue, small amounts of adipose tissue were collected from different subjects. To ensure that enough adipose tissue was available to make a measurement, compositing of tissue from several subjects was forced upon the design (Mack and Robinson, 1985). Conversely, events that are out of control of the field scientists may eliminate composite sampling from constituting a design choice. For example, people whose wells are being tested may demand that their wells be treated as equitably as the wells of their neighbors. Measuring some well samples individually and some well samples solely as part of composites may be seen by some well-owners as unfair and could result in a political decree to measure each well sample individually (Rajagopal, 1990, "personal communication").

The nature of the variable of interest may also dictate the feasibility of composite sampling. For instance, if the interest is in the average height of an individual in a specified population, then grouping several individuals will not reduce the cost of measurement for obvious reasons. On the other hand, if the interest is in the average weight of an individual, then grouping several individuals for weighing provides a measurement on the total weight of an entire group.

Circumstances that presently disqualify composite sampling from being applied may change upon advances in technology. High turn-around time for laboratory results and large labor costs may eliminate optimal sequential retesting designs from consideration. However, retesting designs in the future may be automated and guided by an expert system (Rajagopal, 1990, "personal communication"). Advances in statistical methodology will further extend the utility of composite sampling. Therefore, some of the uncertainties about the relevance of composite sampling today may be obviated tomorrow.

We present here the statistical methodology, implementation strategies, and some reported applications of composite sampling, keeping in view both the substantive scientist and the statistician. The presentation is broadly divided into four parts, each part pertaining to a major theme. The first part (Chapters 2, 3, 4, and 5) covers the case where the interest is in characterizing every individual sample. Except for Chapter 3, this part assumes presence/absence measurements. Chapter 2 discusses the problem of classifying every individual sample as possessing or not possessing a trait. Chapter 3 considers the issue of recovering the extreme individual sample values from composite sample data. Chapter 4 presents statistical methods of

estimating the prevalence of a trait that is detected by presence/absence measurements. Chapter 5 explores a Bayesian approach to the classification problem of Chapter 2.

The second part (Chapters 6, 7, and 8) deals with statistical treatment of composite sample data for continuous measurements. Chapter 6 considers the problem of drawing statistical inference (estimation and testing of hypotheses) on the population mean and variance using composite sample measurements. Chapter 7 models the substantive issue of random weights in composite sampling, that is, the case where the volumes of individual samples that contribute to a composite sample cannot be fixed. Chapter 8 unifies various linear models reported in the literature on composite sampling. The purpose of the unified linear model is to develop a statistical tool to explore the relationship between the individual sample values and composite sample values under all possible complexities in the population structure, sampling protocol, and composite design.

The third part (Chapters 9 and 10) consists of some pertinent issues that affect the performance of composite sampling procedures. Chapter 9 discusses the issue of maintaining data quality through the data quality objectives (DQO) process and by optimizing the composite design. Chapter 10 considers the spatial aspects of composite sampling, where the individual samples represent a spatial point process and compositing may result in altering the sampling interval without possibly reducing the spatial sample support.

The fourth part (Chapters 11, 12, 13, and 14) comprises case studies of composite sampling procedures applied to sampling of different materials and media: solids such as soils and sediments (Chapter 11), liquids such as runoff water, industrial effluents, or milk (Chapter 12), indoor house dust (Chapter 13), and tissue mass (Chapter 14).

Most of this material will appeal to a general scientific reader, though the general framework of Chapter 8 is intended for a reader with a background in linear models. For other reviews of composite sampling, see Rohde (1976, 1979), Elder (1977), Elder et al. (1980), Boswell and Patil (1987), and Garner et al. (1988).

Chapter 2
Classifying Individual Samples into One of Two Categories

2.1 Introduction

Testing of groups for subsequent identification of group members possessing a trait was initiated by Dorfman (1943) to identify US servicemen infected with syphilis. Other reported applications of composite sampling include screening for pollutants (Schaeffer et al., 1982; Rajagopal and Williams, 1989), testing for leaking containers (Sobel and Groll, 1959; Thomas et al., 1973), identifying faulty components in a flow test (Hwang, 1984), identifying active users in a communications system (Hayes, 1978; Berger et al., 1984; Wolf, 1985; Garg and Mohan, 1987), recognizing the pattern of a binary code, and screening experimental factors affecting yield (Hwang, 1984). If the trait is relatively rare, then initially testing composite samples and subsequently only those individual samples that belong to a composite that has tested positive can greatly reduce the required number of tests (see, for instance, Feller, 1968; Garner et al., 1986).

Group testing was originally developed for a binary response variable, where the trait is either present or absent. However, the method can also be used for a (non-negative) continuous response variable where the trait is defined as the exceedance of some specified criterion level c. This latter case requires separate treatment since compositing results in an averaging of the individual values and, consequently, a measurement on the composite can fall below c even while some of the individual values exceed c. By contrast, in the binary response case, absence of the trait in the composite implies absence for all the individual samples. In either case, however, considerable savings can be realized by compositing if the trait is relatively rare.

Individual samples are first collected and prepared for laboratory procedures. Composite samples are then formed and measured. If a composite sample tests positive, then further testing of some or all constituent individual samples must be undertaken in order to classify the individual samples. Testing individual samples, either separately or in the form of composite subsamples of the original composite sample, is called *retesting*. Retesting may thus involve forming further composites as well as making additional laboratory measurements. When the cost of forming composites is negligible compared to the laboratory costs, then the effectiveness

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_2, © Springer Science+Business Media, LLC 2011

of compositing can be characterized by the *relative cost* which is defined as the number of measurements per individual sample classified. Exhaustive testing of all individual samples results in a relative cost of one measurement per sample. In order for composite sampling techniques to be cost-effective, their relative cost must be smaller than 1. Considerable cost savings can be realized when the trait is relatively rare. For example, when the prevalence p of the trait is 0.01, then a simple compositing strategy can result in a relative cost of 20%, or a savings of 80%, in the required number of measurements compared with exhaustive testing of all individual samples.

Consider the case of a continuous response variable. For a measurement such as the concentration of some pollutant in a composite sample, the composite sample measurement is the average of the individual sample values plus a measurement error, if present. For a measurement such as the total pollutant present in a composite sample, the composite sample measurement is the sum of the individual sample values plus a possible measurement error. These two cases are really the same since the concentration in a sample can be obtained from the total amount of the pollutant and the volume of the sample. We will use, as an illustrative formulation, the testing of water wells for the concentration of pollutants. Assume that the analytical measurement is accurate. The measurement on a composite sample is then the average of the individual sample values. Let c be the criterion value or action level. That is, the water from any well with a concentration exceeding c is not potable. Further, assume a detection limit of d. That is, if the concentration of the pollutant in any sample, either individual or composite, does not exceed d, then the laboratory procedure will return a measurement of 0 or an imprecise measurement. If water from one well with a concentration level of c and from $k-1$ other wells with no pollution is mixed to form a composite sample, then the composite sample value will be c/k. In order not to misclassify the polluted well as not polluted, it is necessary that $c/k \geq d$. This implies that the composite sample size k should satisfy $k \leq c/d$. In any application of composite sampling, it is accordingly necessary to place an upper limit on the composite sample size in order to avoid detection limit difficulties. If the criterion level is not at least twice as large as the detection level, then composite sample techniques cannot be used for classification. Conversely, if composite sample techniques are used, then implicitly there is a criterion level $c = kd$ below which classification is undependable. The detection limit for a composite sample of size k is also d, but due to dilution a polluted well with concentration between c and kd may escape detection and be misclassified as unpolluted.

For continuous measurements made without error, if any one individual sample value exceeds c, then the measurement on a composite sample of k individual samples will exceed c/k. Of course, it is possible that none of the individual sample values exceeds c and the composite sample measurement still exceeds c/k. If the composite sample measurement exceeds c/k, then further testing would be undertaken to identify individual sample values that exceed c, even though there may be none. On the other hand, if the composite measurement does not exceed c/k, then none of the individual samples needs be tested further, for none exceeds the value of c.

Section 2.2 discusses the presence/absence case, and Section 2.3 discusses the case of a continuous response variable.

2.2 Presence/Absence Measurements

During World War II, it was feared that some of the US servicemen were infected with syphilis-causing bacteria (*Treponema pallidum*). While the proportion of infected servicemen was not expected to be high, it was necessary to be certain that no infected individual remained undetected. The most commonly used laboratory procedures for the detection of syphilis are carried out on a sample of blood serum (serological tests for syphilis, or STS). The STS are based on detection of one of two substances that appear in blood serum soon after the onset of the disease: syphilis reagin and *treponemal* antibody. It was clear that a large proportion of laboratory procedures would return a negative response, but it was essential to make certain that the blood sample of every individual was subjected to laboratory procedures.

Dorfman (1943) came up with an apparently simple and yet cost-efficient procedure to identify the infected servicemen. The basic argument that Dorfman developed was as follows: Fix a positive integer k, and pool blood samples of k servicemen to be subjected to the STS as a single specimen. Assuming independence among servicemen, the probability that the syphilis bacteria are absent in a pooled sample from k servicemen is $g = 1 - p^k$, where p is the proportion of infected servicemen. Since p is small, g is large and negatively testing composites occur frequently, in which case a single test allows us to correctly classify k servicemen as uninfected.

The cost of collecting and preparing samples for laboratory procedures is constant since the number of individual samples is predetermined. The relative cost of classification can therefore be defined as the expected number of tests divided by the number of samples classified. When the cost of measurement is large and the cost of forming composite samples is relatively small, the relative cost can be reduced by the use of composite sample techniques. In Sections 2.2.1 through 2.2.6, it is assumed that the hierarchical processing of a single composite sample results in the classification of all individual samples eventually making up that composite. The composite sample size is the number of individual samples used to form the composite sample. If the composite sample size does not divide the total number of individual samples to be classified evenly, then near the end of the classification procedure, smaller composite sample sizes must be used. For a relatively small number of samples this "remainder" composite sample size must be taken into account. On the other hand, if the total number of samples is large, then the remainder effect can be ignored. In the limiting case, as the number of individual samples increases indefinitely, the relative cost is the same as the asymptotic relative cost.

The asymptotic relative cost is derived for the retesting procedures discussed in Sections 2.2.1 through 2.2.7. The formulation of Section 2.2.7 begins with the finite case and deduces the asymptotic case. In the finite case, an optimal partition of all

individual samples of various sizes is desired. Bush et al. (1984) and Gilstein (1985) discuss optimal partitions in the finite case for the exhaustive retesting procedure. This problem is not discussed in this monograph, as the emphasis here is in the asymptotic relative cost of classification.

2.2.1 Exhaustive Retesting

The original composite sampling procedure of exhaustive retesting is due to Dorfman (1943) and is often referred to as the Dorfman procedure in the literature (see, for instance, Johnson et al., 1991). Exhaustive retesting utilizes two stages of testing. The first stage consists of testing only the composite samples, while the second stage consists of testing all members of positive testing composite samples. Figure 2.1 shows the two stages of the exhaustive retesting procedure.

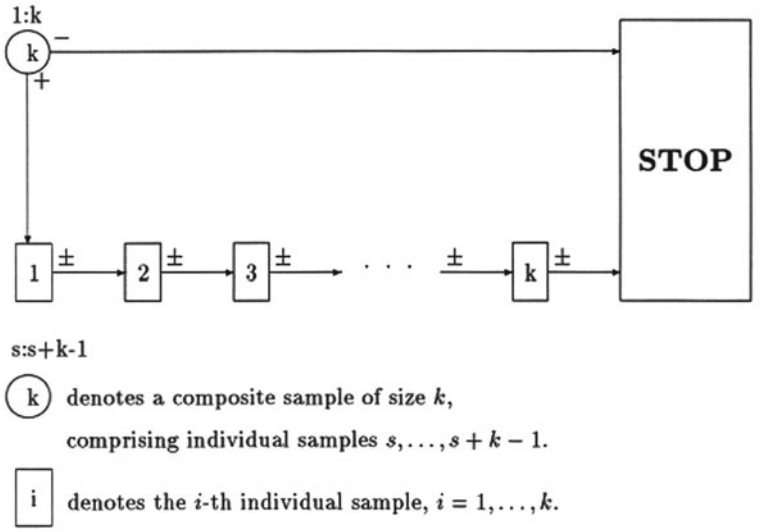

Fig. 2.1 Exhaustive retesting

Our analysis of the method employs a binomial model for the occurrence of the trait. Thus the individual samples are treated as independent trials with a constant probability, p, of possessing the trait. Let $I_1, I_2, \ldots, I_k$ be the k individual sample values, each taking a value of 0 (trait not present) or 1 (trait present). Now $I_1, \ldots, I_k$ are independent and identically distributed with $\Pr[I_i = 1] = p$ which is small for a rare trait. A composite sample formed from these individual samples will have a test result $I = 0$ or $I = 1$, with respective probabilities

$$\Pr[I = 0] = \Pr[I_1 = 0, I_2 = 0, \ldots, I_k = 0] = q^k,$$
$$\Pr[I = 1] = 1 - \Pr[I = 0] = 1 - q^k,$$

where $q = 1 - p$. If the composite sample tests negative, then all the k constituent individual samples are classified as not possessing the trait without any further testing. That is, the classification is achieved with only one test. On the other hand, if the composite sample tests positive, then every constituent individual sample is tested and classified separately. The total number of tests in this case is $k + 1$, with one test for the composite sample and k tests for the k individual samples. In this way, for classifying the k individual samples in a single composite sample, the number of tests, denoted by T_k, is either 1 or $k + 1$. The expected number of tests required is

$$E[T_k] = 1 \cdot q^k + (k+1) \cdot [1 - q^k] = (k+1) - kq^k. \tag{2.1}$$

The ratio of $E[T_k]$, the expected number of tests, to k, the number of individual samples classified, is defined to be the (asymptotic) relative cost, RC, which in this case is given by

$$\text{RC} = 1 + 1/k - q^k. \tag{2.2}$$

For given p, the optimal composite size k can be obtained by minimizing (2.2). Samuels (1978) showed that for all $p \leq 1 - (1/3)^{\frac{1}{3}} \cong 0.307$, the optimal composite sample size is the choice of $1 + [p^{-\frac{1}{2}}]$ or $2 + [p^{-\frac{1}{2}}]$, whichever minimizes the relative cost, where $[x]$ represents the integer part of x (see Table 2.1). The expected number of tests per individual sample classified, when an optimal composite sample size is used, is shown in Fig. 2.2 for selected values of p. Note that p must be sufficiently small to gain most of the benefits of composite sampling (see Table 2.1). For instance, for exhaustive retesting with the optimal k to be twice as efficient as the conventional method of testing individual samples, p must be less than 0.07.

The performance of compositing can also be assessed by the relative savings defined as $\text{RS} = 1 - \text{RC}$. The relative savings is often expressed as a percentage and represents the number of tests per classified item that can be saved, on average, by using compositing instead of individually testing each item. For Dorfman's method, the relative savings becomes

$$\text{RS} = (1 - p)^k - \frac{1}{k}.$$

Notice that the relative savings can be negative indicating that conventional testing would outperform the compositing method under consideration.

Identification of an optimal composite sample size k can yield performance benefits but does require accurate prior information for the value of p. If this prior information is inaccurate then the relative savings can be substantially suboptimal and may even be negative. For this reason it is useful to have some rules of thumb for choosing k when knowledge of p is limited. Ideally the rule would produce near-optimal savings with only a small risk of negative savings. One such rule is called the "1/4/12 rule" for Dorfman's method and divides the possible values of p into

Table 2.1 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for exhaustive retesting

p	k_{opt}	RC[a]	p	k_{opt}	RC
(0.30663, 1.00000]	1	1.000000	(0.00049, 0.00051]	45	(0.044033, 0.044916]
(0.12394, 0.30663]	3	(0.660973, 0.999987]	(0.00047, 0.00049]	46	(0.043129, 0.044033]
(0.06558, 0.12394]	4	(0.487622, 0.660973]	(0.00045, 0.00047]	47	(0.042207, 0.043129]
(0.04112, 0.06558]	5	(0.389372, 0.487622]	(0.00043, 0.00045]	48	(0.041262, 0.042207]
(0.02828, 0.04112]	6	(0.324792, 0.389372]	(0.00041, 0.00043]	49	(0.040296, 0.041262]
(0.02066, 0.02828]	7	(0.278810, 0.324792]	(0.00040, 0.00041]	50	(0.039806, 0.040296]
(0.01577, 0.02066]	8	(0.244410, 0.278810]	(0.00038, 0.00040]	51	(0.038799, 0.039806]
(0.01243, 0.01577]	9	(0.217573, 0.244410]	(0.00036, 0.00038]	52	(0.037771, 0.038799]
(0.01005, 0.01243]	10	(0.196068, 0.217573]	(0.00035, 0.00036]	53	(0.037244, 0.037771]
(0.00830, 0.01005]	11	(0.178510, 0.196068]	(0.00034, 0.00035]	54	(0.036710, 0.037244]
(0.00697, 0.00830]	12	(0.163839, 0.178510]	(0.00033, 0.00034]	55	(0.036169, 0.036710]
(0.00593, 0.00697]	13	(0.151323, 0.163839]	(0.00031, 0.00033]	56	(0.035062, 0.036169]
(0.00511, 0.00593]	14	(0.140635, 0.151323]	(0.00030, 0.00031]	57	(0.034493, 0.035062]
(0.00445, 0.00511]	15	(0.131373, 0.140635]	(0.00029, 0.00030]	58	(0.033915, 0.034493]
(0.00391, 0.00445]	16	(0.123255, 0.131373]	(0.00028, 0.00029]	59	(0.033330, 0.033915]
(0.00346, 0.00391]	17	(0.116037, 0.123255]	(0.00027, 0.00028]	60	(0.032731, 0.033330]
(0.00309, 0.00346]	18	(0.109738, 0.116037]	(0.00026, 0.00027]	61	(0.032121, 0.032731]
(0.00277, 0.00309]	19	(0.103966, 0.109738]	(0.00025, 0.00026]	63	(0.031498, 0.032121]
(0.00250, 0.00277]	20	(0.098827, 0.103966]	(0.00024, 0.00025]	64	(0.030867, 0.031498]
(0.00227, 0.00250]	21	(0.094222, 0.098827]	(0.00023, 0.00024]	65	(0.030219, 0.030867]
(0.00206, 0.00227]	22	(0.089800, 0.094222]	(0.00022, 0.00023]	66	(0.029556, 0.030219]
(0.00189, 0.00206]	23	(0.086054, 0.089800]	(0.00021, 0.00022]	68	(0.028879, 0.029556]
(0.00173, 0.00189]	24	(0.082364, 0.086054]	(0.00020, 0.00021]	70	(0.028184, 0.028879]
(0.00160, 0.00173]	25	(0.079241, 0.082364]	(0.00019, 0.00020]	71	(0.027476, 0.028184]
(0.00148, 0.00160]	26	(0.076237, 0.079241]	(0.00018, 0.00019]	73	(0.026744, 0.027476]
(0.00137, 0.00148]	27	(0.073373, 0.076237]	(0.00017, 0.00018]	75	(0.025992, 0.026744]
(0.00127, 0.00137]	28	(0.070665, 0.073373]	(0.00016, 0.00017]	77	(0.025218, 0.025992]
(0.00118, 0.00127]	29	(0.068134, 0.070665]	(0.00015, 0.00016]	80	(0.024423, 0.025218]
(0.00111, 0.00118]	30	(0.066102, 0.068134]	(0.00014, 0.00015]	82	(0.023596, 0.024423]
(0.00104, 0.00111]	31	(0.063999, 0.066102]	(0.00013, 0.00014]	85	(0.022739, 0.023596]
(0.00097, 0.00104]	32	(0.061821, 0.063999]	(0.00012, 0.00013]	88	(0.021848, 0.022739]
(0.00091, 0.00097]	33	(0.059891, 0.061821]	(0.00011, 0.00012]	92	(0.020919, 0.021848]
(0.00086, 0.00091]	34	(0.058235, 0.059891]	(0.00010, 0.00011]	96	(0.019952, 0.020919]
(0.00081, 0.00086]	35	(0.056529, 0.058235]	(0.00009, 0.00010]	100	(0.018929, 0.019952]
(0.00077, 0.00081]	36	(0.055125, 0.056529]	(0.00008, 0.00009]	106	(0.017848, 0.018929]
(0.00073, 0.00077]	37	(0.053684, 0.055125]	(0.00007, 0.00008]	112	(0.016696, 0.017848]
(0.00069, 0.00073]	38	(0.052201, 0.053684]	(0.00006, 0.00007]	120	(0.015465, 0.016696]
(0.00065, 0.00069]	39	(0.050673, 0.052201]	(0.00005, 0.00006]	130	(0.014118, 0.015465]
(0.00062, 0.00065]	40	(0.049498, 0.050673]	(0.00004, 0.00005]	142	(0.012628, 0.014118]
(0.00059, 0.00062]	41	(0.048293, 0.049498]	(0.00003, 0.00004]	158	(0.010936, 0.012628]
(0.00056, 0.00059]	42	(0.047054, 0.048293]	(0.00002, 0.00003]	183	(0.008940, 0.010936]
(0.00054, 0.00056]	43	(0.046214, 0.047054]	(0.00001, 0.00002]	224	(0.006500, 0.008940]
(0.00051, 0.00054]	44	(0.044916, 0.046214]			

[a]RC: Relative cost

three categories, corresponding to "frequent," "infrequent," and "rare" occurrence of the trait in question. A different value of k is used for each case as follows:

Frequent. If $0.29 < p < 1$, use $k = 1$ (conventional testing).
Infrequent. If $0.01 < p < 0.29$, use $k = 4$.
Rare. If $0 < p < 0.01$, use $k = 12$.

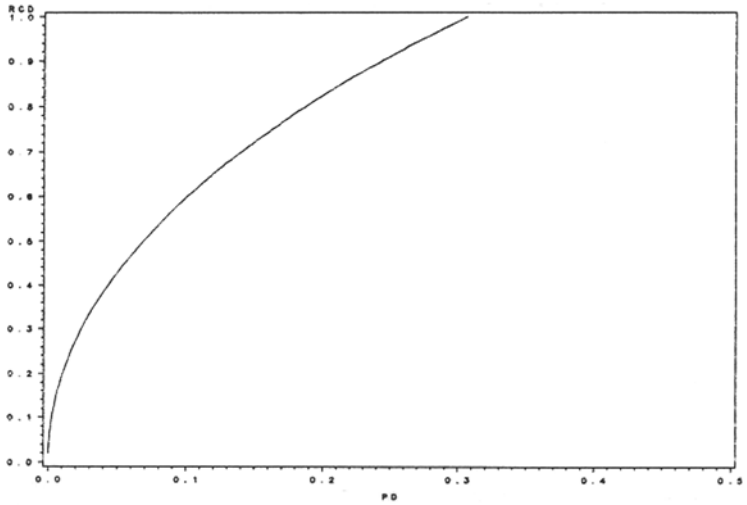

Fig. 2.2 Relative cost for exhaustive retesting using the optimal composite sample size

The following tabulation compares the relative savings achieved by the 1/4/12 rule with the optimal relative savings for several values of p. Relative savings is expressed as percentages.

p	RS (1/4/12)	RS (optimal)
0.275	3	5
0.25	7	9
0.20	16	18
0.15	27	28
0.10	41	41
0.05	56	57
0.025	65	69
0.01+	71	80
0.01−	80	80
0.005	86	86
0.001	90	94

The tabulation supposes that p is correctly classified into the frequent, infrequent, or rare categories. Classification error can occur when p is near a category boundary. In this case, the achieved savings can be somewhat more or somewhat less than indicated in the tabulation.

2.2.2 Sequential Retesting

Sterrett (1957) suggested a modification to exhaustive retesting, and hence this modified procedure is often referred to as the Sterrett procedure. The modification is

motivated by the following observations. When a composite tests positive, then the test result tells us that at least one of the k constituent individual samples possesses the trait. The Dorfman procedure determines which samples have the trait by individually testing each item. Here, Sterrett notes that as soon as an individual sample is found with the trait, then there is no information on whether any of the remaining (untested and therefore unclassified) samples from that composite has the trait. Moreover, the prevalence of the trait among these unclassified samples is still p. It is then natural to argue that compositing the unclassified samples should be more economical than individual testing, even at this stage. This was the observation that led Sterrett (1957) to propose the following: When a composite tests positive, its constituent individual samples are tested sequentially until a positively testing sample is identified. At this stage, all the remaining individual samples (from among the k that formed the original composite) are used to form a new composite sample for testing. If this composite tests negative, then all its constituent individual samples are classified as not possessing the trait, and no more testing is necessary. On the other hand, if this composite tests positive, then the same procedure is repeated, beginning with sequential testing of constituent individual samples until an individual sample is identified as possessing the trait. This procedure continues until all the k individual samples comprising the original composite sample have been classified. Figure 2.3 displays the sequential retesting procedure.

Let T_k be the number of tests required to classify the k individual samples that constitute a composite sample. For small values of k, the expected number of tests can be calculated directly, giving, for instance,

$$E[T_1] = 1 \ \ (T_1 \equiv 1), \ \ E[T_2] = 3 - 2q^2, \ \text{and} \ \ E[T_3] = 5 - q - 2q^2 - q^3.$$

A recurrence formula can be found by conditioning on the number of retests J to find the first positively testing item. Thus, $J = 0$ if the composite sample tests negative; $J = 1$ if the first item retested tests positive, etc. Now

$$\begin{aligned} E[T_k] &= E[E\,(T_k|J)] = \sum_{j=0}^{k} E[T_k|J = j]P[J = j] \\ &= E[T_k|J = 0]q^k + \sum_{j=1}^{k} E[T_k|J = j]q^{j-1}p, \ \ k = 2, 3, \ldots. \end{aligned}$$

Given that the composite sample tests positive, that the first $j - 1$ individual samples test negative, and that the jth individual sample tests positive, the remaining $k - j$ individual sample values are independent Bernoulli random variables with parameter p. We therefore have

$$E[T_k|J = j] = j + 1 + E[T_{k-j}], \quad j > 0,$$

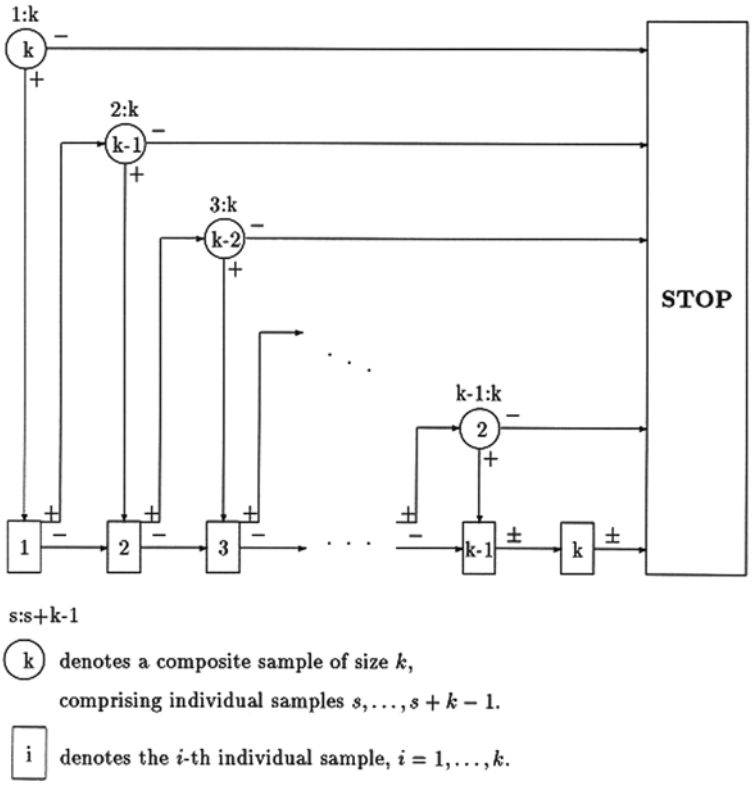

Fig. 2.3 Sequential retesting

where $T_0 = 0$, and hence

$$E[T_k] = q^k + \sum_{j=1}^{k} \left\{ j + 1 + E[T_{k-j}] \right\} q^{j-1} p.$$

So

$$E[T_k] = q^k + \sum_{j=1}^{k} (j+1) q^{j-1} p + \sum_{j=1}^{k} E[T_{k-j}] q^{j-1} p$$

$$= q^k + \sum_{j=0}^{k-1} (j+2) q^j p + \sum_{j=0}^{k-1} E[T_j] q^{k-j-1} p.$$

Also

$$q\,E[T_{k-1}] = q^k + \sum_{j=1}^{k-1} (j+1) q^j p + \sum_{j=0}^{k-2} E[T_j] q^{k-j-1} p, \quad k = 3, 4, \ldots.$$

Then

$$E[T_k] - qE[T_{k-1}] = \left[2p + \sum_{j=1}^{k-1} q^j p\right] + E[T_{k-1}]p.$$

Thus

$$E[T_k] - E[T_{k-1}] = 2p + \frac{qp\left(1 - q^{k-1}\right)}{1-q} = 2p + q - q^k.$$

Further

$$\begin{aligned} E[T_k] - E[T_2] &= \sum_{j=3}^{k} \left\{E[T_j] - E[T_{j-1}]\right\} \\ &= (k-2)(2p+q) - \sum_{j=3}^{k} q^j \\ &= (k-2)(2p+q) - \frac{q^3\left(1-q^{k-2}\right)}{1-q}, \quad k = 2, 3, \ldots. \end{aligned}$$

Substituting the value of $E[T_2]$ gives

$$\begin{aligned} E[T_k] &= 3 - 2q^2 + (k-2)(2p+q) + \frac{1-q^3}{p} - \frac{1-q^{k+1}}{p} \\ &= 2k - (k-3)q - q^2 - \frac{1-q^{k+1}}{p}, \quad k = 2, 3, \ldots. \end{aligned}$$

The (asymptotic) relative cost is

$$\mathrm{RC} = E[T_k]/k = 2 - q + \frac{1}{k}\left[3q - q^2 - \frac{1-q^{k+1}}{p}\right]. \tag{2.3}$$

The optimal composite sample size is tabulated in Table 2.2, and the relative cost using the optimal composite sample size is shown in Fig. 2.4 for selected values of p.

2.2.3 Binary Split Retesting

The exhaustive and sequential retesting procedures presented above have the following limitation to their cost-efficiency. The prevalence of individual samples that possess the trait is sufficiently small to justify the use of composite sampling. However,

Table 2.2 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for sequential retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.30437, 1.00000]	1	1.000000	(0.00064, 0.00066]	56	(0.036512, 0.037090]
(0.21253, 0.30437]	3	(0.827997, 0.999984]	(0.00061, 0.00064]	57	(0.035628, 0.036512]
(0.12774, 0.21253]	4	(0.622819, 0.827997]	(0.00059, 0.00061]	58	(0.035030, 0.035628]
(0.08283, 0.12774]	5	(0.487586, 0.622819]	(0.00057, 0.00059]	59	(0.034418, 0.035030]
(0.05759, 0.08283]	6	(0.397398, 0.487586]	(0.00056, 0.00057]	60	(0.034111, 0.034418]
(0.04224, 0.05759]	7	(0.334221, 0.397398]	(0.00054, 0.00056]	61	(0.033485, 0.034111]
(0.03226, 0.04224]	8	(0.287849, 0.334221]	(0.00052, 0.00054]	62	(0.032847, 0.033485]
(0.02543, 0.03226]	9	(0.252541, 0.287849]	(0.00050, 0.00052]	63	(0.032198, 0.032847]
(0.02055, 0.02543]	10	(0.224784, 0.252541]	(0.00049, 0.00050]	64	(0.031869, 0.032198]
(0.01695, 0.02055]	11	(0.202456, 0.224784]	(0.00047, 0.00049]	65	(0.031200, 0.031869]
(0.01422, 0.01695]	12	(0.184127, 0.202456]	(0.00045, 0.00047]	66	(0.030519, 0.031200]
(0.01210, 0.01422]	13	(0.168814, 0.184127]	(0.00043, 0.00045]	68	(0.029821, 0.030519]
(0.01042, 0.01210]	14	(0.155825, 0.168814]	(0.00042, 0.00043]	69	(0.029466, 0.029821]
(0.00906, 0.01042]	15	(0.144618, 0.155825]	(0.00040, 0.00042]	70	(0.028746, 0.029466]
(0.00796, 0.00906]	16	(0.134999, 0.144618]	(0.00038, 0.00040]	72	(0.028006, 0.028746]
(0.00704, 0.00796]	17	(0.126487, 0.134999]	(0.00036, 0.00038]	74	(0.027249, 0.028006]
(0.00627, 0.00704]	18	(0.118972, 0.126487]	(0.00035, 0.00036]	75	(0.026861, 0.027249]
(0.00562, 0.00627]	19	(0.112299, 0.118972]	(0.00034, 0.00035]	76	(0.026469, 0.026861]
(0.00507, 0.00562]	20	(0.106376, 0.112299]	(0.00033, 0.00034]	78	(0.026070, 0.026469]
(0.00460, 0.00507]	21	(0.101079, 0.106376]	(0.00032, 0.00033]	79	(0.025667, 0.026070]
(0.00419, 0.00460]	22	(0.096254, 0.101079]	(0.00031, 0.00032]	80	(0.025259, 0.025667]
(0.00383, 0.00419]	23	(0.091835, 0.096254]	(0.00030, 0.00031]	81	(0.024840, 0.025259]
(0.00351, 0.00383]	24	(0.087744, 0.091835]	(0.00029, 0.00030]	83	(0.024417, 0.024840]
(0.00323, 0.00351]	25	(0.084020, 0.087744]	(0.00028, 0.00029]	84	(0.023989, 0.024417]
(0.00299, 0.00323]	26	(0.080711, 0.084020]	(0.00027, 0.00028]	85	(0.023550, 0.023989]
(0.00277, 0.00299]	27	(0.077567, 0.080711]	(0.00026, 0.00027]	87	(0.023104, 0.023550]
(0.00258, 0.00277]	28	(0.074758, 0.077567]	(0.00025, 0.00026]	88	(0.022649, 0.023104]
(0.00240, 0.00258]	29	(0.072004, 0.074758]	(0.00024, 0.00025]	89	(0.022187, 0.022649]
(0.00224, 0.00240]	30	(0.069476, 0.072004]	(0.00023, 0.00024]	92	(0.021714, 0.022187]
(0.00210, 0.00224]	31	(0.067194, 0.069476]	(0.00022, 0.00023]	94	(0.021230, 0.021714]
(0.00197, 0.00210]	32	(0.065010, 0.067194]	(0.00021, 0.00022]	95	(0.020735, 0.021230]
(0.00185, 0.00197]	33	(0.062933, 0.065010]	(0.00020, 0.00021]	98	(0.020230, 0.020735]
(0.00174, 0.00185]	34	(0.060973, 0.062933]	(0.00019, 0.00020]	101	(0.019714, 0.020230]
(0.00164, 0.00174]	35	(0.059140, 0.060973]	(0.00018, 0.00019]	104	(0.019182, 0.019714]
(0.00155, 0.00164]	36	(0.057445, 0.059140]	(0.00017, 0.00018]	105	(0.018635, 0.019182]
(0.00147, 0.00155]	37	(0.055900, 0.057445]	(0.00016, 0.00017]	108	(0.018072, 0.018635]
(0.00139, 0.00147]	38	(0.054312, 0.055900]	(0.00015, 0.00016]	113	(0.017495, 0.018072]
(0.00132, 0.00139]	39	(0.052889, 0.054312]	(0.00014, 0.00015]	115	(0.016895, 0.017495]
(0.00125, 0.00132]	40	(0.051428, 0.052889]	(0.00013, 0.00014]	119	(0.016275, 0.016895]
(0.00119, 0.00125]	41	(0.050145, 0.051428]	(0.00012, 0.00013]	126	(0.015629, 0.016275]
(0.00114, 0.00119]	42	(0.049054, 0.050145]	(0.00011, 0.00012]	131	(0.014956, 0.015629]
(0.00108, 0.00114]	43	(0.047710, 0.049054]	(0.00010, 0.00011]	135	(0.014258, 0.014956]
(0.00104, 0.00108]	44	(0.046797, 0.047710]	(0.00009, 0.00010]	141	(0.013520, 0.014258]
(0.00099, 0.00104]	45	(0.045630, 0.046797]	(0.00008, 0.00009]	150	(0.012742, 0.013520]
(0.00094, 0.00099]	46	(0.044435, 0.045630]	(0.00007, 0.00008]	156	(0.011909, 0.012742]
(0.00091, 0.00094]	47	(0.043704, 0.044435]	(0.00006, 0.00007]	169	(0.011026, 0.011909]
(0.00087, 0.00091]	48	(0.042710, 0.043704]	(0.00005, 0.00006]	186	(0.010057, 0.011026]
(0.00083, 0.00087]	49	(0.041694, 0.042710]	(0.00004, 0.00005]	202	(0.008988, 0.010057]
(0.00080, 0.00083]	50	(0.040917, 0.041694]	(0.00003, 0.00004]	228	(0.007782, 0.008988]
(0.00077, 0.00080]	51	(0.040126, 0.040917]	(0.00002, 0.00003]	248	(0.006532, 0.007782]
(0.00074, 0.00077]	52	(0.039321, 0.040126]	(0.00001, 0.00002]	250	(0.005269, 0.006532]
(0.00071, 0.00074]	53	(0.038499, 0.039321]			
(0.00069, 0.00071]	54	(0.037941, 0.038499]			
(0.00066, 0.00069]	55	(0.037090, 0.037941]			

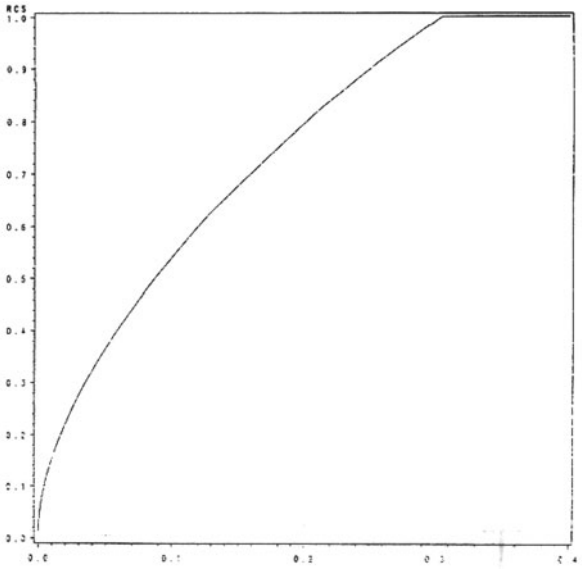

Fig. 2.4 Relative cost for sequential retesting using the optimal composite sample size

after a composite sample tests positive, both of these retesting procedures recommend exhaustive testing of the constituent individual samples (at least initially), which may not be very cost-effective. Also, note that the conditional prevalence of the trait among individual samples that comprise a composite that tests positive is $p^+ = \frac{p}{1-q^k}$, where p is the prevalence of the trait in the population and k is the composite sample size. It is clear that $p^+ > p$, and hence the optimal composite sample size corresponding to the prevalence p^+ will clearly not be larger than k, which is optimal corresponding to the prevalence p. It may therefore be more reasonable to form subcomposites of a positive testing composite rather than resort to exhaustive testing. With this modification and assuming a binomial model, Gill and Goltlieb (1974) proposed that positive testing composites be divided into two subcomposites of as equal sizes as possible, and positive testing subcomposites be recursively tested after dividing into two further subcomposites. Figure 2.5 describes the binary split retesting procedure of Gill and Gottlieb. Examples of the binary split retesting procedure for $k = 4$ and $k = 8$ are also presented in Figs. 2.6 and 2.7.

Let T_k be the number of tests required for classifying the k individual samples in a composite. For small values of k, the expectations can be calculated directly, giving

$$E[T_1] = 1, \quad E[T_2] = 3 - 2q^2, \text{ and } E[T_3] = 5 - 2q^2 - 2q^3.$$

When retesting is required, aliquots of the k individual samples are composited into two groups of sizes k_1 and $k_2 = k - k_1$, where $k_1 = k_2 = k/2$ if k is even and

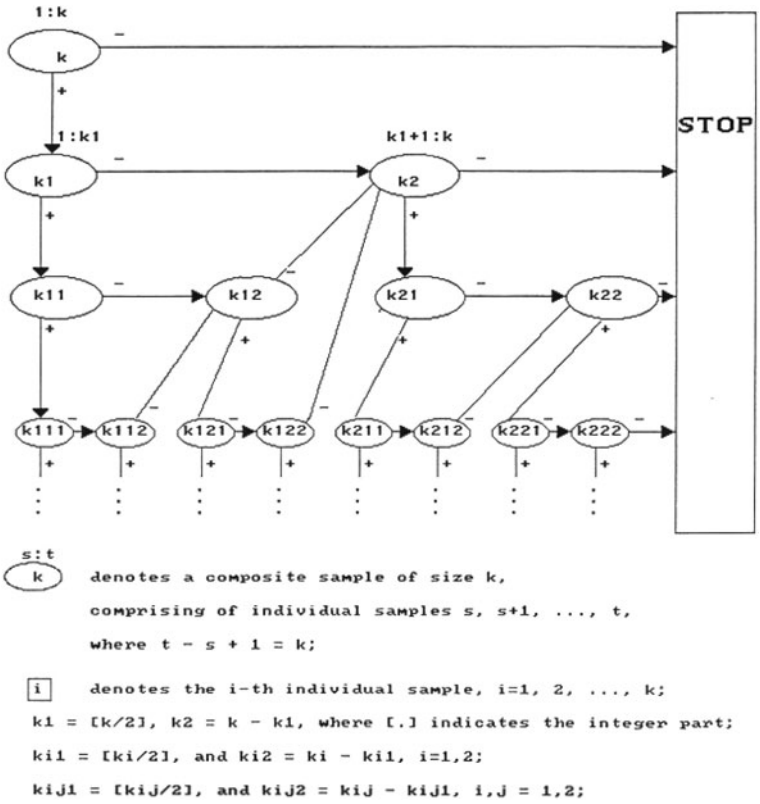

Fig. 2.5 Binary split retesting

where $k_1 = (k-1)/2$ and $k_2 = (k+1)/2$ if k is odd. Let "split" indicate that the composite of size k tests positive. Note that "split" means at least one of the k items has the trait. Now

$$\begin{aligned} E[T_k] &= E\left[T_k \mid \text{split}\right] + E\left[T_k \mid \text{not split}\right] \\ &= \left(1 + E\left[T_{k_1} + T_{k_2} \mid \text{split}\right]\right) \Pr\left[\text{split}\right] + 1 \cdot \Pr\left[\text{not split}\right] \\ &= E\left[T_{k_1} + T_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] + 1. \end{aligned}$$

Now consider two separate composites of sizes k_1 and k_2, and let "split" indicate that at least one of the composites tests positive. Note that "split" means that at least one of the $k_1 + k_2 = k$ items tests positive. Then

$$\begin{aligned} E\left[T_{k_1} + T_{k_2}\right] = {} & E\left[T_{k_1} + t_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] \\ & + E\left[T_{k_1} + T_{k_2} \mid \text{not split}\right] \Pr\left[\text{not split}\right]. \end{aligned}$$

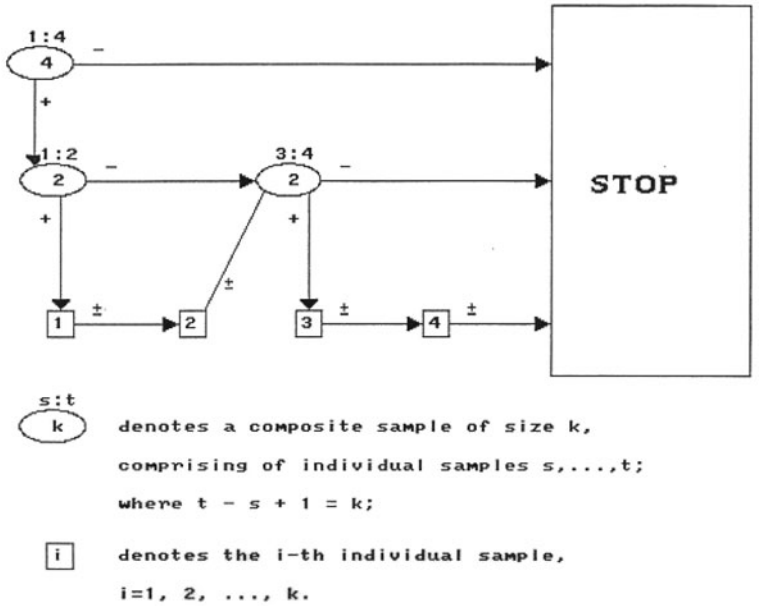

Fig. 2.6 An example of binary split retesting with $k = 4$

That is,

$$E\left[T_{k_1}\right] + E\left[T_{k_2}\right] = E\left[T_{k_1} + t_{k_2} \mid \text{split}\right] \Pr\left[\text{split}\right] + 2q^k.$$

Subtracting from the above expression for $E\,[T_k]$ now yields

$$E\,[T_k] = E\left[T_{k_1}\right] + E\left[T_{k_2}\right] + 1 - 2q^k. \tag{2.4}$$

This formula can be used to recursively generate $E\,[T_k]$ for any given k. For example, $E[T_2] = 2E[T_1] + 1 - 2q^2$ and $E[T_3] = E[T_1] + E[T_2] + 1 - 2q^3$. In this way, we obtain

$$\begin{aligned}
E[T_2] &= 3 - 2q^2, \\
E[T_3] &= 5 - 2q^2 - 2q^3, \\
E[T_4] &= 7 - 4q^2 - 2q^4, \\
E[T_5] &= 9 - 4q^2 - 2q^3 - 2q^5, \\
E[T_6] &= 11 - 4q^2 - 4q^3 - 2q^6, \\
E[T_7] &= 13 - 6q^2 - 2q^3 - 2q^4 - 2q^7, \\
E[T_8] &= 15 - 8q^2 - 4q^4 - 2q^8, \\
E[T_9] &= 17 - 8q^2 - 2q^3 - 2q^4 - 2q^5 - 2q^9,
\end{aligned}$$

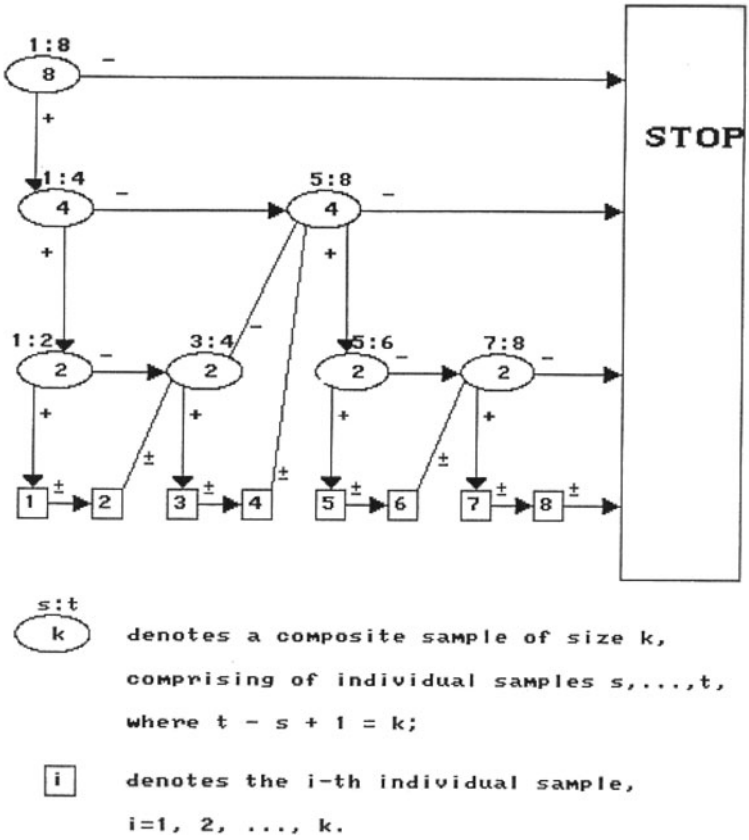

Fig. 2.7 An example of binary split retesting with $k = 8$

$$E[T_{10}] = 19 - 8q^2 - 4q^3 - 4q^5 - 2q^{10},$$
$$E[T_{11}] = 21 - 8q^2 - 6q^3 - 2q^5 - 2q^6 - 2q^{10},$$
$$E[T_{12}] = 23 - 8q^2 - 8q^3 - 4q^6 - 2q^{12}.$$

The asymptotic relative cost can be calculated by dividing the expected number of tests by the corresponding composite sample size. The optimal composite sample size is tabulated in Table 2.3, and the relative cost of classification with the binary split procedure using the optimal composite sample size is shown in Fig. 2.8 for selected values of p.

2.2.4 Curtailed Exhaustive Retesting

Many of the retesting strategies can be improved upon when the laboratory procedure is free of testing error. It is also necessary to have some freedom to schedule

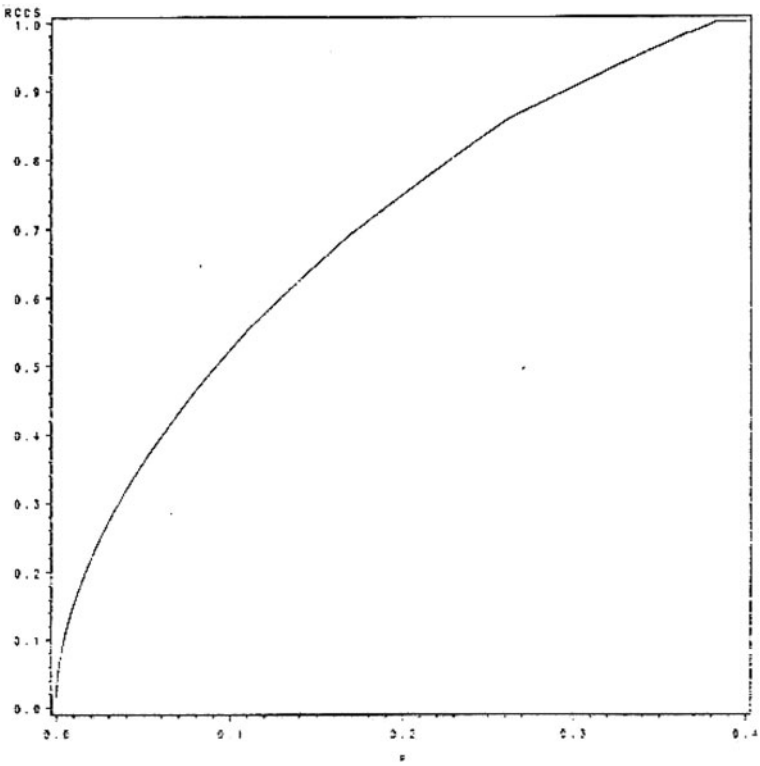

Fig. 2.8 Relative cost for binary split retesting using the optimal composite sample size

Table 2.3 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for binary split retesting

p	k_{opt}	RC
(0.29289, 1.00000]	1	1.00
(0.15910, 0.29289]	2	(0.792883, 0.999995]
(0.08299, 0.15910]	4	(0.555524, 0.792883]
(0.04239, 0.08299]	8	(0.360729, 0.555524]
(0.02142, 0.04239]	16	(0.222719, 0.360729]
(0.01077, 0.02142]	32	(0.132800, 0.222719]
(0.00540, 0.01077]	64	(0.077175, 0.132800]
(0.00267, 0.00540]	128	(0.043552, 0.077175]

the laboratory tests sequentially. In general, whenever a composite tests positive, the items comprising that composite must be subjected to retesting, either individually or in groups. Now suppose the items making up the composite can be partitioned into two subsets and we know, from our retesting, that none of the items in one of the subsets has the trait. Then, without testing, we know that at least one of the items in the second subset does have the trait. The avoidance of the test on this second subset is referred to as *curtailment.* Curtailment comes at a price when testing errors

(specifically false positives) are possible. When retesting is not curtailed, the items in positively testing groups undergo retesting which reduces the false-positive rate. This advantage of compositing is lost with curtailed retesting. The effects of testing error are examined in greater detail in Section 2.2.8.

Now consider the exhaustive retesting (Dorfman) procedure and suppose the individual item values in a positively testing composite are denoted by $X_1, \ldots, X_k$. Let these items be retested in sequential order and suppose after the first $k-1$ retests that $X_1 = X_2 = \cdots = X_{k-1} = 0$. Then we know that the last item must have the trait ($X_k = 1$) without testing and the items can be completely classified with only $k - 1$ instead of k retests (see Fig. 2.9). Let T_k be the number of tests required to completely classify a composite of size k using curtailed exhaustive retesting. Writing $q = 1 - p$ as above, T_k takes three possible values: 1, k, and $k + 1$. But

$$\begin{aligned} \Pr[T_k = 1] &= q^k, \\ \Pr[T_k = k] &= q^{k-1}p, \\ \Pr[T_k = k+1] &= 1 - q^k - q^{k-1}p = 1 - q^{k-1}. \end{aligned}$$

Thus

$$\begin{aligned} E[T_k] &= q^k + kq^{k-1}p + (k+1)(1 - q^{k-1}) \\ &= k + 1 - kq^k - q^{k-1}p. \end{aligned} \tag{2.5}$$

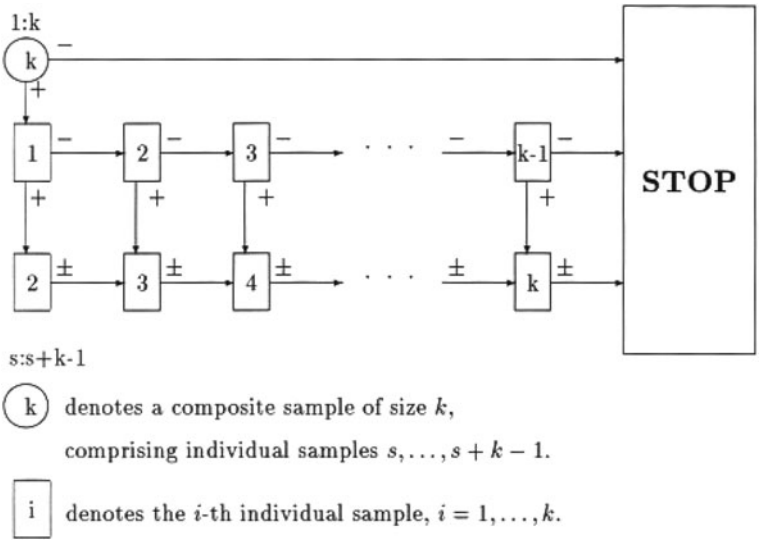

Fig. 2.9 Curtailed exhaustive retesting

The (asymptotic) relative cost is $\mathrm{RC} = E\left[T_k\right]/k = 1 - q^k + \frac{1}{k}\left[1 - q^{k-1}p\right]$. For selected values of p, the optimal composite sample size for the curtailed exhaustive retesting procedure is tabulated in Table 2.4 and the corresponding relative cost is shown in Fig. 2.10.

Table 2.4 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed exhaustive retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.38196, 1.00000]	1	1.000000	(0.00051, 0.00054]	44	(0.044905, 0.046202]
(0.20261, 0.38196]	2	(0.783386, 0.999993]	(0.00049, 0.00051]	45	(0.044022, 0.044905]
(0.10291, 0.20261]	3	(0.583770, 0.783386]	(0.00047, 0.00049]	46	(0.043119, 0.044022]
(0.06010, 0.10291]	4	(0.457106, 0.583770]	(0.00045, 0.00047]	47	(0.042198, 0.043119]
(0.03906, 0.06010]	5	(0.373965, 0.457106]	(0.00043, 0.00045]	48	(0.041253, 0.042198]
(0.02733, 0.03906]	6	(0.315871, 0.373965]	(0.00041, 0.00043]	49	(0.040288, 0.041253]
(0.02017, 0.02733]	7	(0.273231, 0.315871]	(0.00039, 0.00041]	50	(0.039297, 0.040288]
(0.01549, 0.02017]	8	(0.240669, 0.273231]	(0.00038, 0.00039]	51	(0.038792, 0.039297]
(0.01226, 0.01549]	9	(0.214956, 0.240669]	(0.00036, 0.00038]	52	(0.037764, 0.038792]
(0.00994, 0.01226]	10	(0.194156, 0.214956]	(0.00035, 0.00036]	53	(0.037238, 0.037764]
(0.00822, 0.00994]	11	(0.177008, 0.194156]	(0.00034, 0.00035]	54	(0.036704, 0.037238]
(0.00691, 0.00822]	12	(0.162633, 0.177008]	(0.00033, 0.00034]	55	(0.036163, 0.036704]
(0.00589, 0.00691]	13	(0.150415, 0.162633]	(0.00031, 0.00033]	56	(0.035056, 0.036163]
(0.00508, 0.00589]	14	(0.139900, 0.150415]	(0.00030, 0.00031]	57	(0.034488, 0.035056]
(0.00443, 0.00508]	15	(0.130814, 0.139900]	(0.00029, 0.00030]	58	(0.033910, 0.034488]
(0.00389, 0.00443]	16	(0.122720, 0.130814]	(0.00028, 0.00029]	59	(0.033325, 0.033910]
(0.00345, 0.00389]	17	(0.115686, 0.122720]	(0.00027, 0.00028]	60	(0.032727, 0.033325]
(0.00308, 0.00345]	18	(0.109404, 0.115686]	(0.00026, 0.00027]	61	(0.032117, 0.032727]
(0.00276, 0.00308]	19	(0.103645, 0.109404]	(0.00025, 0.00026]	63	(0.031495, 0.032117]
(0.00249, 0.00276]	20	(0.098514, 0.103645]	(0.00024, 0.00025]	64	(0.030864, 0.031495]
(0.00226, 0.00249]	21	(0.093915, 0.098514]	(0.00023, 0.00024]	65	(0.030216, 0.030864]
(0.00206, 0.00226]	22	(0.089714, 0.093915]	(0.00022, 0.00023]	66	(0.029553, 0.030216]
(0.00188, 0.00206]	23	(0.085749, 0.089714]	(0.00021, 0.00022]	68	(0.028876, 0.029553]
(0.00173, 0.00188]	24	(0.082298, 0.085749]	(0.00020, 0.00021]	70	(0.028181, 0.028876]
(0.00159, 0.00173]	25	(0.078932, 0.082298]	(0.00019, 0.00020]	71	(0.027473, 0.028181]
(0.00147, 0.00159]	26	(0.075926, 0.078932]	(0.00018, 0.00019]	73	(0.026742, 0.027473]
(0.00137, 0.00147]	27	(0.073326, 0.075926]	(0.00017, 0.00018]	75	(0.025990, 0.026742]
(0.00127, 0.00137]	28	(0.070623, 0.073326]	(0.00016, 0.00017]	77	(0.025216, 0.025990]
(0.00118, 0.00127]	29	(0.068096, 0.070623]	(0.00015, 0.00016]	80	(0.024421, 0.025216]
(0.00111, 0.00118]	30	(0.066067, 0.068096]	(0.00014, 0.00015]	82	(0.023594, 0.024421]
(0.00104, 0.00111]	31	(0.063967, 0.066067]	(0.00013, 0.00014]	85	(0.022738, 0.023594]
(0.00097, 0.00104]	32	(0.061793, 0.063967]	(0.00012, 0.00013]	88	(0.021847, 0.022738]
(0.00091, 0.00097]	33	(0.059865, 0.061793]	(0.00011, 0.00012]	92	(0.020918, 0.021847]
(0.00086, 0.00091]	34	(0.058211, 0.059865]	(0.00010, 0.00011]	96	(0.019951, 0.020918]
(0.00081, 0.00086]	35	(0.056507, 0.058211]	(0.00009, 0.00010]	100	(0.018929, 0.019951]
(0.00077, 0.00081]	36	(0.055104, 0.056507]	(0.00008, 0.00009]	106	(0.017847, 0.018929]
(0.00073, 0.00077]	37	(0.053665, 0.055104]	(0.00007, 0.00008]	112	(0.016695, 0.017847]
(0.00069, 0.00073]	38	(0.052183, 0.053665]	(0.00006, 0.00007]	120	(0.015465, 0.016695]
(0.00065, 0.00069]	39	(0.050657, 0.052183]	(0.00005, 0.00006]	130	(0.014118, 0.015465]
(0.00062, 0.00065]	40	(0.049483, 0.050657]	(0.00004, 0.00005]	142	(0.012628, 0.014118]
(0.00059, 0.00062]	41	(0.048280, 0.049483]	(0.00003, 0.00004]	158	(0.010936, 0.012628]
(0.00056, 0.00059]	42	(0.047041, 0.048280]	(0.00002, 0.00003]	183	(0.008940, 0.010936]
(0.00054, 0.00056]	43	(0.046202, 0.047041]	(0.00001, 0.00002]	224	(0.006500, 0.008940]

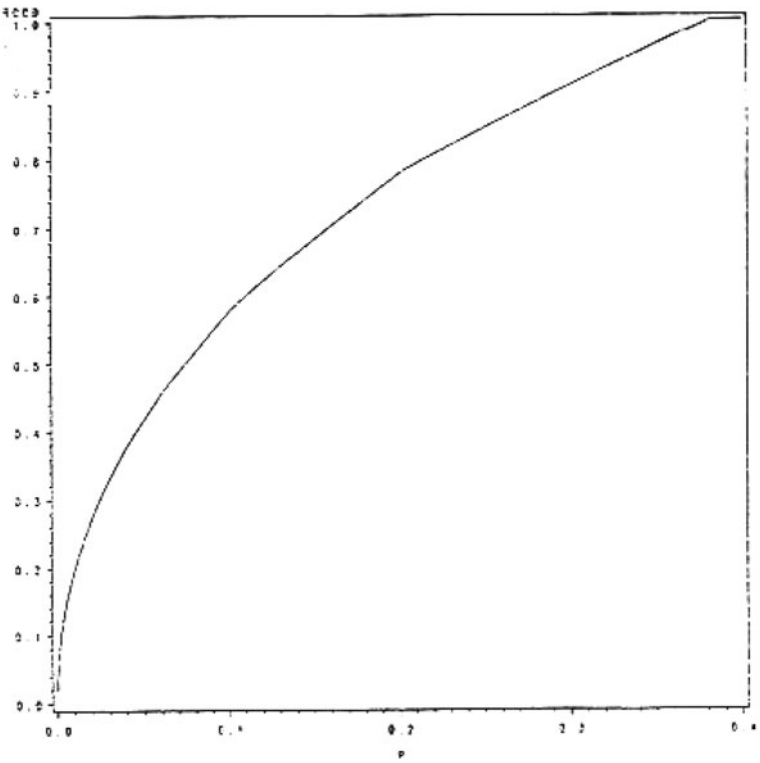

Fig. 2.10 Relative cost for curtailed exhaustive retesting using the optimal composite sample size

2.2.5 Curtailed Sequential Retesting

The sequential testing method can also be curtailed by avoiding a test of the last item whenever the first $k-1$ items of a composite test negative (see Fig. 2.11). The expected number of tests can be obtained by modifying the argument of Section 2.2.2. Direct calculation gives $E[T_1] = 1$; $E[T_2] = 3 - q - q^2$; and $E[T_3] = 5 - 2q - q^2 - q^3$. Let J be defined as in Section 2.2.2. Then

$$\begin{aligned} E[T_k] &= \sum_{j=0}^{k} E\left[T_k | J = j\right] P\left[J = j\right] \\ &= q^k + \sum_{j=1}^{k-1} E\left[T_k | J = j\right] q^{j-1} p + k q^{k-1} p \\ &= q^k + k q^{k-1} p + \sum_{j=1}^{k-1} \left\{j + 1 + E\left[T_{k-j}\right]\right\} q^{j-1} p \end{aligned}$$

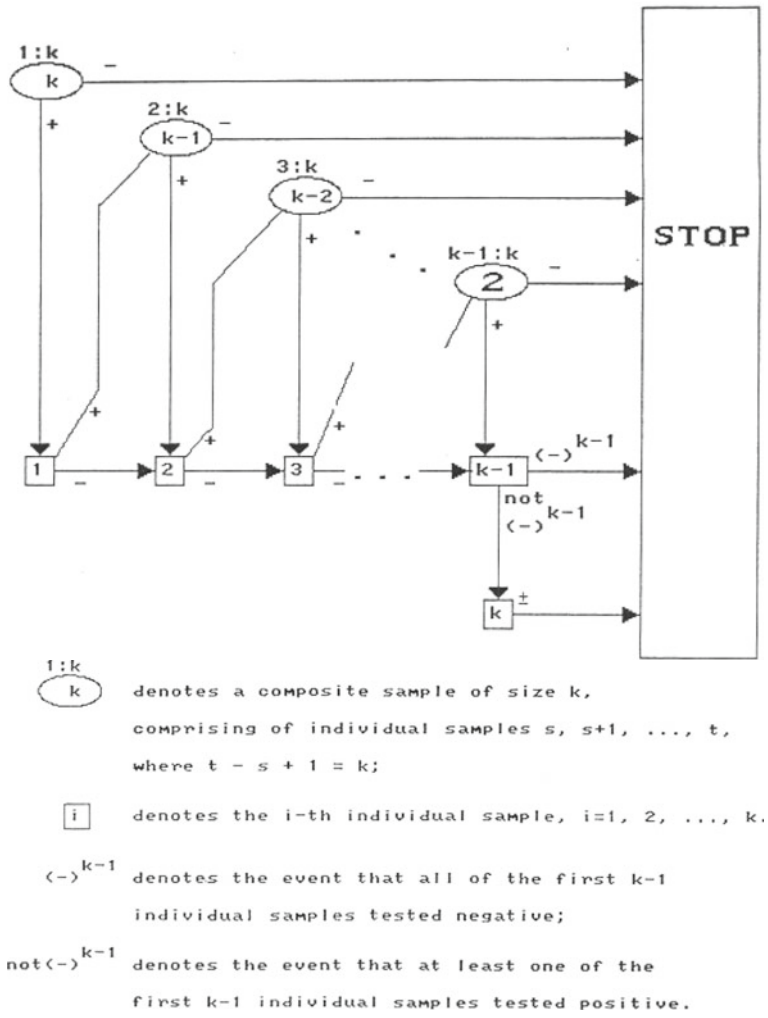

Fig. 2.11 Curtailed sequential retesting

$$= q^k + kq^{k-1}p + \sum_{j=0}^{k-2}(j+2)q^j p + \sum_{j=1}^{k-1} E\left[T_j\right] q^{k-j-1}p.$$

Similarly,

$$qE\left[T_{k-1}\right] = q^k + (k-1)q^{k-1}p + \sum_{j=0}^{k-3}(j+2)q^{j+1}p$$

$$+ \sum_{j=1}^{k-2} E\left[T_j\right] q^{k-j-1}p$$

$$= q^k + (k-1)q^{k-1}p + \sum_{j=1}^{k-2}(j+1)q^j p$$
$$+ \sum_{j=1}^{k-2} E\left[T_j\right] q^{k-j-1}p, \quad k = 3, 4, \ldots.$$

Thus, again

$$E\left[T_k\right] - qE\left[T_{k-1}\right] = q^{k-1}p + 2p + \sum_{j=1}^{k-2} q^j p + E\left[T_{k-1}\right] p$$

or

$$E\left[T_k\right] - E\left[T_{k-1}\right] = q^{k-1}p + 2p + pq\left(1 - q^{k-2}\right)/(1-q)$$
$$= q^{k-1}p + 2p + q\left(1 - q^{k-2}\right), \quad k = 3, 4, \ldots.$$

The number of tests takes any value (except 2) from 1 to $(2k-1)$. The average number of tests for classifying k individual samples is therefore given by

$$E\left[T_k\right] = k(2-q) + 2q - \left(1 - q^{k+1}\right)/p \tag{2.6}$$

and

$$\mathrm{RC} = 2 - q + \frac{1}{k}\left[2q - \frac{1-q^{k+1}}{p}\right]. \tag{2.7}$$

For example, if $k = 5$ and $p = 0.15$, then the expected number of tests is 3.30 and the relative cost is 0.66. So, sequential retesting would require only 66% as many tests as conventional testing of all individual items. By comparison, curtailed exhaustive retesting gives RC = 0.74 for $k = 5$ and $p = 0.15$. For these parameters, the curtailed sequential retesting procedure is more efficient than the curtailed exhaustive retesting procedure.

The optimal composite sample size for the curtailed sequential retesting procedure is tabulated for selected values of p in Table 2.5. Figure 2.12 shows the relative cost of classification when the optimal composite sample size is used for the curtailed sequential retesting procedure.

Table 2.5 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed sequential retesting

p	k_{opt}	RC	p	k_{opt}	RC
(0.38196, 1.00000]	1	1.000000	(0.00064, 0.00066]	56	(0.036501, 0.037078]
(0.26101, 0.38196]	2	(0.857449, 0.999993]	(0.00061, 0.00064]	57	(0.035618, 0.036501]
(0.16832, 0.26101]	3	(0.689891, 0.857449]	(0.00059, 0.00061]	58	(0.035020, 0.035618]
(0.10986, 0.16832]	4	(0.551025, 0.689891]	(0.00057, 0.00059]	59	(0.034409, 0.035020]
(0.07506, 0.10986]	5	(0.448910, 0.551025]	(0.00056, 0.00057]	60	(0.034102, 0.034409]
(0.05381, 0.07506]	6	(0.374861, 0.448910]	(0.00054, 0.00056]	61	(0.033477, 0.034102]
(0.04022, 0.05381]	7	(0.320154, 0.374861]	(0.00052, 0.00054]	62	(0.032839, 0.033477]
(0.03109, 0.04022]	8	(0.278530, 0.320154]	(0.00050, 0.00052]	63	(0.032191, 0.032839]
(0.02471, 0.03109]	9	(0.246079, 0.278530]	(0.00048, 0.00050]	64	(0.031529, 0.032191]
(0.02008, 0.02471]	10	(0.220107, 0.246079]	(0.00047, 0.00048]	65	(0.031193, 0.031529]
(0.01664, 0.02008]	11	(0.199027, 0.220107]	(0.00045, 0.00047]	66	(0.030513, 0.031193]
(0.01400, 0.01664]	12	(0.181486, 0.199027]	(0.00044, 0.00045]	67	(0.030166, 0.030513]
(0.01194, 0.01400]	13	(0.166741, 0.181486]	(0.00043, 0.00044]	68	(0.029815, 0.030166]
(0.01030, 0.01194]	14	(0.154162, 0.166741]	(0.00042, 0.00043]	69	(0.029460, 0.029815]
(0.00897, 0.01030]	15	(0.143279, 0.154162]	(0.00040, 0.00042]	70	(0.028740, 0.029460]
(0.00789, 0.00897]	16	(0.133894, 0.143279]	(0.00038, 0.00040]	72	(0.028001, 0.028740]
(0.00699, 0.00789]	17	(0.125616, 0.133894]	(0.00036, 0.00038]	74	(0.027244, 0.028001]
(0.00623, 0.00699]	18	(0.118237, 0.125616]	(0.00035, 0.00036]	75	(0.026856, 0.027244]
(0.00559, 0.00623]	19	(0.111699, 0.118237]	(0.00034, 0.00035]	76	(0.026464, 0.026856]
(0.00504, 0.00559]	20	(0.105800, 0.111699]	(0.00033, 0.00034]	78	(0.026066, 0.026464]
(0.00457, 0.00504]	21	(0.100521, 0.105800]	(0.00032, 0.00033]	79	(0.025663, 0.026066]
(0.00417, 0.00457]	22	(0.095828, 0.100521]	(0.00031, 0.00032]	80	(0.025255, 0.025663]
(0.00381, 0.00417]	23	(0.091421, 0.095828]	(0.00030, 0.00031]	81	(0.024837, 0.025255]
(0.00350, 0.00381]	24	(0.087471, 0.091421]	(0.00029, 0.00030]	83	(0.024413, 0.024837]
(0.00323, 0.00350]	25	(0.083896, 0.087471]	(0.00028, 0.00029]	84	(0.023986, 0.024413]
(0.00298, 0.00323]	26	(0.080459, 0.083896]	(0.00027, 0.00028]	85	(0.023547, 0.023986]
(0.00276, 0.00298]	27	(0.077320, 0.080459]	(0.00026, 0.00027]	87	(0.023101, 0.023547]
(0.00257, 0.00276]	28	(0.074516, 0.077320]	(0.00025, 0.00026]	88	(0.022647, 0.023101]
(0.00239, 0.00257]	29	(0.071768, 0.074516]	(0.00024, 0.00025]	89	(0.022185, 0.022647]
(0.00224, 0.00239]	30	(0.069404, 0.071768]	(0.00023, 0.00024]	92	(0.021711, 0.022185]
(0.00209, 0.00224]	31	(0.066961, 0.069404]	(0.00022, 0.00023]	94	(0.021228, 0.021711]
(0.00196, 0.00209]	32	(0.064778, 0.066961]	(0.00021, 0.00022]	95	(0.020733, 0.021228]
(0.00185, 0.00196]	33	(0.062879, 0.064778]	(0.00020, 0.00021]	98	(0.020228, 0.020733]
(0.00174, 0.00185]	34	(0.060923, 0.062879]	(0.00019, 0.00020]	101	(0.019713, 0.020228]
(0.00164, 0.00174]	35	(0.059094, 0.060923]	(0.00018, 0.00019]	104	(0.019180, 0.019713]
(0.00155, 0.00164]	36	(0.057403, 0.059094]	(0.00017, 0.00018]	105	(0.018633, 0.019180]
(0.00147, 0.00155]	37	(0.055862, 0.057403]	(0.00016, 0.00017]	108	(0.018071, 0.018633]
(0.00139, 0.00147]	38	(0.054276, 0.055862]	(0.00015, 0.00016]	113	(0.017494, 0.018071]
(0.00132, 0.00139]	39	(0.052856, 0.054276]	(0.00014, 0.00015]	115	(0.016894, 0.017494]
(0.00125, 0.00132]	40	(0.051398, 0.052856]	(0.00013, 0.00014]	119	(0.016274, 0.016894]
(0.00119, 0.00125]	41	(0.050117, 0.051398]	(0.00012, 0.00013]	126	(0.015628, 0.016274]
(0.00113, 0.00119]	42	(0.048805, 0.050117]	(0.00011, 0.00012]	131	(0.014955, 0.015628]
(0.00108, 0.00113]	43	(0.047686, 0.048805]	(0.00010, 0.00011]	135	(0.014258, 0.014955]
(0.00104, 0.00108]	44	(0.046774, 0.047686]	(0.00009, 0.00010]	141	(0.013519, 0.014258]
(0.00099, 0.00104]	45	(0.045608, 0.046774]	(0.00008, 0.00009]	150	(0.012742, 0.013519]
(0.00094, 0.00099]	46	(0.044415, 0.045608]	(0.00007, 0.00008]	156	(0.011909, 0.012742]
(0.00091, 0.00094]	47	(0.043685, 0.044415]	(0.00006, 0.00007]	169	(0.011026, 0.011909]
(0.00087, 0.00091]	48	(0.042692, 0.043685]	(0.00005, 0.00006]	186	(0.010057, 0.011026]
(0.00083, 0.00087]	49	(0.041677, 0.042692]	(0.00004, 0.00005]	202	(0.008988, 0.010057]
(0.00080, 0.00083]	50	(0.040902, 0.041677]	(0.00003, 0.00004]	228	(0.007782, 0.008988]
(0.00077, 0.00080]	51	(0.040111, 0.040902]	(0.00002, 0.00003]	248	(0.006532, 0.007782]
(0.00074, 0.00077]	52	(0.039307, 0.040111]	(0.00001, 0.00002]	250	(0.005269, 0.006532]
(0.00071, 0.00074]	53	(0.038486, 0.039307]			
(0.00069, 0.00071]	54	(0.037929, 0.038486]			
(0.00066, 0.00069]	55	(0.037078, 0.037929]			

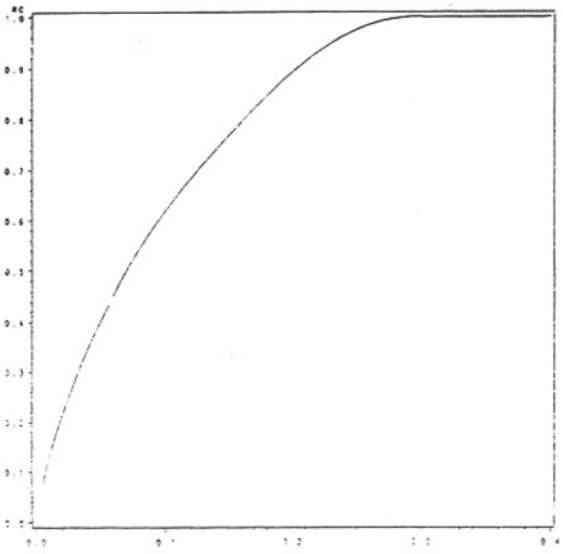

Fig. 2.12 Relative cost for curtailed sequential retesting using the optimal composite sample size

2.2.6 Curtailed Binary Split Retesting

The binary split retesting procedure of Gill and Gottlieb (2.4) can be curtailed so that only one of the two composite subsamples needs to be tested after certain binary splits. If the first of the two subcomposites tests negative, then the other subcomposite would test positive and the test need not be carried out. That is, the second composite is tested only when the first one tests positive. This can save up to half of all retesting efforts and costs. The following recursion formula can be obtained in the same way as that obtained for the uncurtailed procedure:

$$E\left[T_k\right] = E\left[T_{k_1}\right] + E\left[T_{k_2}\right] + 1 - q^{k-1} - q^k. \tag{2.8}$$

When k is even, then $k_1 = k_2 = k/2$. If k is odd, then the expected number of tests is smaller with $k_1 = (k-1)/2$ rather than with $k_1 = (k+1)/2$. With this choice for the value of k_1, the recursion formula in (2.8) can be used iteratively to obtain $E[T_k]$ for any given k. For example,

$$\begin{aligned}
E[T_2] &= 2E[T_1] + 1 - q - q^2 = 3 - q - q^2,\\
E[T_3] &= E[T_1] + E[T_2] + 1 - q - q^3 = 5 - 2q - q^2 - q^3,\\
E[T_4] &= 7 - 2q - 3q^2 - q^4,\\
E[T_5] &= 9 - 3q - 3q^2 - q^3 - q^5,\\
E[T_6] &= 11 - 4q - 2q^2 - 3q^3 - q^6,
\end{aligned}$$

$$
\begin{aligned}
E[T_7] &= 13 - 4q - 4q^2 - 2q^3 - q^4 - q^7, \\
E[T_8] &= 15 - 4q - 6q^2 - 3q^4 - q^8, \\
E[T_9] &= 17 - 5q - 6q^2 - q^3 - 2q^4 - q^5 - q^9, \\
E[T_{10}] &= 19 - 6q - 6q^2 - 2q^3 - 3q^5 - q^{10}, \\
E[T_{11}] &= 21 - 7q - 5q^2 - 4q^3 - 3q^5 - q^{11}, \\
E[T_{12}] &= 23 - 8q - 4q^2 - 6q^3 - 2q^5 - q^6 - q^{12}.
\end{aligned}
$$

The (asymptotic) relative cost can be calculated by dividing the expected number of tests by the corresponding composite sample size. Table 2.6 shows the optimal composite sample size corresponding to selected values of p for the curtailed binary split procedure, while Fig. 2.13 shows the relative cost of classification for this procedure when used with the optimal composite sample size.

Table 2.6 Optimal composite sample size (k_{opt}) and the corresponding relative cost (RC) for curtailed binary split retesting

p	k_{opt}	RC
(0.38196, 1.00000]	1	1.00
(0.26101, 0.38196]	2	(0.857449, 0.999993]
(0.16582, 0.26101]	3	(0.685100, 0.857449]
(0.10106, 0.16582]	5	(0.513090, 0.685100]
(0.08439, 0.10106]	7	(0.458100, 0.513090]
(0.06817, 0.08439]	9	(0.397968, 0.458100]
(0.06640, 0.06817]	10	(0.391045, 0.397968]
(0.05054, 0.06640]	11	(0.325611, 0.391045]
(0.04316, 0.05054]	13	(0.292382, 0.325611]
(0.03346, 0.04316]	19	(0.243240, 0.292382]
(0.02563, 0.03346]	21	(0.200881, 0.243240]
(0.02170, 0.02563]	27	(0.177679, 0.200881]
(0.01709, 0.02170]	37	(0.148125, 0.177679]
(0.01689, 0.01709]	38	(0.146811, 0.148125]
(0.01659, 0.01689]	42	(0.144806, 0.146811]
(0.01282, 0.01659]	43	(0.119011, 0.144806]
(0.01090, 0.01282]	53	(0.105004, 0.119011]
(0.00842, 0.01090]	75	(0.085534, 0.105004]
(0.00643, 0.00842]	85	(0.069102, 0.085534]
(0.00546, 0.00643]	107	(0.060614, 0.069102]
(0.00426, 0.00546]	149	(0.049464, 0.060614]
(0.00422, 0.00426]	150	(0.049086, 0.049464]
(0.00414, 0.00422]	170	(0.048319, 0.049086]
(0.00321, 0.00414]	171	(0.039262, 0.048319]
(0.00249, 0.00321]	213	(0.031876, 0.039262]
(0.00244, 0.00249]	214	(0.031355, 0.031876]
(0.00238, 0.00244]	234	(0.030722, 0.031355]
(0.00186, 0.00238]	235	(0.025168, 0.030722]
(0.00151, 0.00186]	245	(0.021335, 0.025168]
(0.00148, 0.00151]	246	(0.021003, 0.021335]

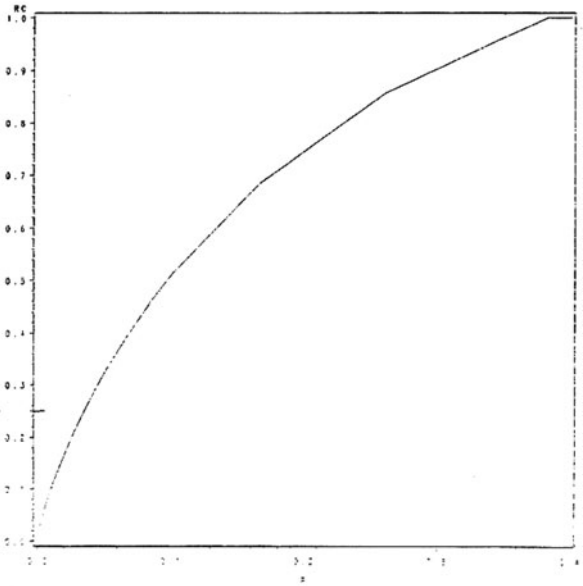

Fig. 2.13 Relative cost for curtailed binary split retesting

2.2.7 Entropy-Based Retesting

A source of inefficiency with hierarchical models such as binary splitting is that we may be required to test subcomposites of smaller and smaller sizes; but these sizes may be substantially less than the optimal size. In binary splitting, for example, suppose an original composite of size k has tested positive and has been split into two subcomposites of sizes k_1 and k_2. When the first subcomposite also tests positive, then we have no information regarding the status of the second composite of size $k_2 < k$. Instead of testing this subcomposite, we would be better off to return its k_2 items to the pool of unclassified items and select a composite of size k from that pool. This is the essence of the entropy-based retesting method due to Hwang (1984).

Suppose we have a large collection of items to be classified. At each stage of the entropy method, this collection is divided into three disjoint parts: those items that have been classified as positive, those that have been classified as negative, and the remaining unclassified pool. Let k be a fixed composite sample size and write k^* for the smaller of k and the current size of the unclassified pool. The entropy-based retesting method is described below. Notice that whenever items are put back into the unclassified pool, we have no information about these items. Throughout, then, the unclassified pool can be looked upon as a collection of independent Bernoulli trials with an unchanging parameter p.

Step 1. If $k^* = 0$, then exit; otherwise, select a subset of size k^* from the unclassified pool and test the resulting composite. If the test is negative, classify all k^* items as negative and go to Step 1. Otherwise, go to Step 2 with $k' = k^*$.

Step 2. Here we have a set of k' items whose corresponding composite has tested (or would test) positive. If $k' = 1$, classify the item as positive and go to Step 1. Otherwise, split the k' items into two disjoint subsets, of sizes $k_1 \leq k_2$, with k_1 and k_2 as nearly equal as possible. Form a composite of size k_1 from the first of these subsets and test the composite.

Step 2a. If the test is negative, classify all k_1 items as negative and go to Step 2 with the remaining $k' = k_2$ items.

Step 2b. If the test is positive, return the remaining k_2 items to the unclassified pool and go to Step 2 with $k' = k_1$ items.

Notice that the algorithm differs from binary retesting only in Step 2b, where items are returned to the unclassified pool. Starting from a given composite in Step 1, processing of that composite results in at most one item being classified as positive; along the way, several items may be classified as negative.

Let the k^* items in the composite at Step 1 be arranged in sequence from left to right and let the binary splitting maintain this sequential arrangement. If J is the position of the first item in the sequence that possesses the trait, then processing of the composite results in classifying the first $J - 1$ items as negative, classifying the Jth item as positive, and returning the remaining $k^* - J$ items to the unclassified pool. Consistent with earlier sections, we write $J = 0$ when a starting composite contains no item with the trait. In this case all k^* items are classified as negative. Write T_m for the number of tests required to fully classify a pool of m items using composites of size k.

Near the end of the classification procedure, smaller composite sample sizes may have to be used. For the finite case, when the number of unclassified individual samples drops below k, the composite sample size being used, all the remaining individual samples are used to form the next composite sample. The following cases develop the asymptotic relative cost and give several finite relative costs for various composite sample sizes.

Case 1. Composites of Size $k = 2$ Are Used. Processing each composite requires one or two tests, so that

$$\begin{aligned} E[T_m] &= E[T_m|J=0]\,q^2 + E[T_m|J=1]\,p + E[T_m|J=2]\,qp \\ &= \left\{1 + E\left[T_{m-2}\right]\right\} q^2 + \left\{2 + E\left[T_{m-1}\right]\right\} p + \left\{2 + E\left[T_{m-2}\right]\right\} qp \\ &= q^2 + 2p + 2qp + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right] \\ &= 2 - q^2 + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right], \quad m = 2, 3, \ldots. \end{aligned}$$

Since $T_1 = 1$ and $T_0 = 0$, we obtain

$$E[T_2] = 2 - q^2 + p = 3 - q - q^2,$$

$$
\begin{aligned}
E[T_3] &= 2 - q^2 + p\left(3 - q - q^2\right) + q \\
&= 5 - 3q - q^2 + q^3, \\
E[T_4] &= 2 - q^2 + pE[T_3] + q[T_2] \\
&= 2 - q^2 + p\left(5 - 3q - q^2 + q^3\right) + q\left(3 - q - q^2\right) \\
&= 7 - 5q + q^3 - q^4, \\
E[T_5] &= 9 - 7q + q^2 - q^4 + q^5.
\end{aligned}
$$

We observe that $E[T_1] - E[T_0] = 1$, and

$$
\begin{aligned}
E[T_2] - E[T_1] &= 2 - q - q^2, \\
E[T_3] - E[T_2] &= 2 - 2q + q^3, \\
E[T_4] - E[T_3] &= 2 - 2q + q^2 - q^4, \\
E[T_5] - E[T_4] &= 2 - 2q + q^2 - q^3 + q^5.
\end{aligned}
$$

Recall that $E[T_m] = 2 - q^2 + pE\left[T_{m-1}\right] + qE\left[T_{m-2}\right]$, $m = 2, 3, \ldots$. So, $qE\left[T_{m-1}\right] = 2q - q^3 + pqE\left[T_{m-2}\right] + q^2E\left[T_{m-3}\right]$. Subtracting and simplifying gives

$$
E[T_m] - E\left[T_{m-1}\right] = p(2 - q^2) + q^2\left\{E\left[T_{m-2}\right] - E\left[T_{m-3}\right]\right\}, \quad m = 3, 4, \ldots.
$$

This difference is the average number of tests needed to classify one item when there are m items to be classified. For m large, this difference is essentially a constant independent of p. That is, for large m, the asymptotic relative cost (RC) can be found by solving

$$
\mathrm{RC} = p\left(2 - q^2\right) + q^2\,(\mathrm{RC}).
$$

Thus, the asymptotic relative cost is

$$
\mathrm{RC} = \frac{p\left(2 - q^2\right)}{1 - q^2} = \frac{2 - q^2}{1 + q}. \tag{2.9}
$$

Case 2. Composites of Size $k = 3$ Are Used. Processing of one composite results in one, two, or three tests. Consider

$$
\begin{aligned}
E[T_m] &= E[T_m|J = 0]\,q^3 + E[T_m|J = 1]\,p + E[T_m|J = 2]\,qp \\
&\quad + E[T_m|J = 3]\,q^2 \\
&= \left\{1 + E\left[T_{m-3}\right]\right\}q^3 + \left\{2 + E\left[T_{m-1}\right]\right\}p \\
&\quad + \left\{3 + E\left[T_{m-2}\right]\right\}qp + \left\{3 + E\left[T_{m-3-1}\right]\right\}q^2p \\
&= 2 + q - 2q^3 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2E\left[T_{m-3}\right], \quad \text{for} \quad m \geq 3.
\end{aligned}
$$

As before, $T_0 = 0$, $T_1 = 1$, and T_2 is the same as in Case 1. Thus,

$$E[T_2] = 3 - q - q^2,$$

and

$$\begin{aligned} E[T_3] &= 2 + q - 2q^3 + p\left(3 - q - q^2\right) + qp = 5 - 2q - q^2 - q^3, \\ E[T_4] &= 2 + q - 2q^3 + p\left(5 - 2q - q^2 - q^3\right) + qp\left(3 - q - q^2\right) + q^2 p \\ &= 7 - 3q - 2q^2 - 2q^3 + 2q^4, \\ E[T_5] &= 9 - 4q - 3q^2 - 2q^3 + 3q^4 - q^5. \end{aligned}$$

Recall

$$\begin{aligned} E[T_m] &= 2 + q - 2q^3 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 E\left[T_{m-3}\right], \\ qE\left[T_{m-1}\right] &= 2q + q^2 - 2q^4 + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] + q^3 E\left[T_{m-4}\right]. \end{aligned}$$

Subtracting the latter from the former and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(2 + q - 2q^3\right) + q^3\left\{E\left[T_{m-3}\right] - E\left[T_{m-4}\right]\right\}.$$

For large m, this difference is essentially a constant equal to the relative cost. Thus,

$$\mathrm{RC} = \frac{2 + q - 2q^3}{1 + q + q^2}. \tag{2.10}$$

Case 3. Composites of Size $k = 4$ Are Used. Again, processing of a composite results in one or three tests. Thus,

$$\begin{aligned} E[T_m] &= \left\{1 + E\left[T_{m-4}\right]\right\} q^4 + \left\{3 + E\left[T_{m-1}\right]\right\} p \\ &\quad + \left\{3 + E\left[T_{m-2}\right]\right\} qp + \left\{3 + E\left[T_{m-3}\right]\right\} q^2 p + \left\{3 + E\left[T_{m-4}\right]\right\} q^3 p \\ &= 3 - 2q^4 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] + q^3 E\left[T_{m-4}\right]. \end{aligned}$$

Subtracting $qE\left[T_{m-1}\right]$ and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(3 - 2q^4\right) + q^4\left\{E\left[T_{m-4}\right] - E\left[T_{m-5}\right]\right\}.$$

For large m, the relative cost is

$$\mathrm{RC} = \frac{3 - 2q^4}{1 + q + q^2 + q^3}. \tag{2.11}$$

Case 4. Composites of Size $k = 5$ Are Used. In this case, one, three, or four tests are required for processing a composite. Thus,

$$\begin{aligned} E[T_m] &= \left\{1 + E\left[T_{m-5}\right]\right\} q^5 + \left\{3 + E\left[T_{m-1}\right]\right\} p + \left\{3 + E\left[T_{m-2}\right]\right\} qp \\ &\quad + \left\{3 + E\left[T_{m-3}\right]\right\} q^2 p + \left\{4 + E\left[T_{m-4}\right]\right\} q^3 p + \left\{4 + E\left[T_{m-5}\right]\right\} q^4 p \\ &= 3 + q^3 - 3q^5 + pE\left[T_{m-1}\right] + qpE\left[T_{m-2}\right] + q^2 pE\left[T_{m-3}\right] \\ &\quad + q^3 pE\left[T_{m-4}\right] + q^4 E\left[T_{m-5}\right]. \end{aligned}$$

Subtracting $qE\left[T_{m-1}\right]$ and simplifying gives

$$E[T_m] - E\left[T_{m-1}\right] = p\left(3 + q^3 - 3q^5\right) + q^5\left\{E\left[T_{m-5}\right] - E\left[T_{m-6}\right]\right\}.$$

For large m this is the relative cost resulting in

$$\mathrm{RC} = \frac{3 + q^3 - 3q^5}{1 + q + q^2 + q^3 + q^4}. \tag{2.12}$$

Note that the asymptotic relative cost with a composite sample size k is the ratio of two polynomials in q, the denominator being a geometric series $1 + q + q^2 + \cdots + q^k$. The numerators of the respective relative costs corresponding to composite sample sizes of $k = 6, 7, 8, 9$, and 10 are tabulated below:

k	Numerator of the relative cost
6	$3 + q - q^3 + q^4 - 3q^6$
7	$3 + q - 3q^7$
8	$4 - 3q^8$
9	$4 + q^7 - 4q^9$
10	$4 + q^3 - q^5 + q^8 - 4q^{10}$

The optimal composite sample size depends on the number of individual samples to be classified and on the prevalence p of individual samples possessing the trait, and hence testing positive. Table 2.7 gives the optimal composite sample size as a function of p for values of $m = 2$–12. See also Fig. 2.14 for the relative costs corresponding to these composite sample sizes.

Table 2.7 can be used interactively by choosing or estimating the optimal composite sample size to use initially. After the first composite sample has been processed, the optimal composite sample size is determined by entering the table with the number of remaining individual samples to be classified. Also, it is possible to update the estimate of q after some initial samples have been classified.

In general, one would like to optimize the entropy-based procedure by permitting k in Step 1 and (k_1, k_2) in Step 2 to vary depending upon the current size of the unclassified pool. On the basis of exhaustive computer searches, Snyder and

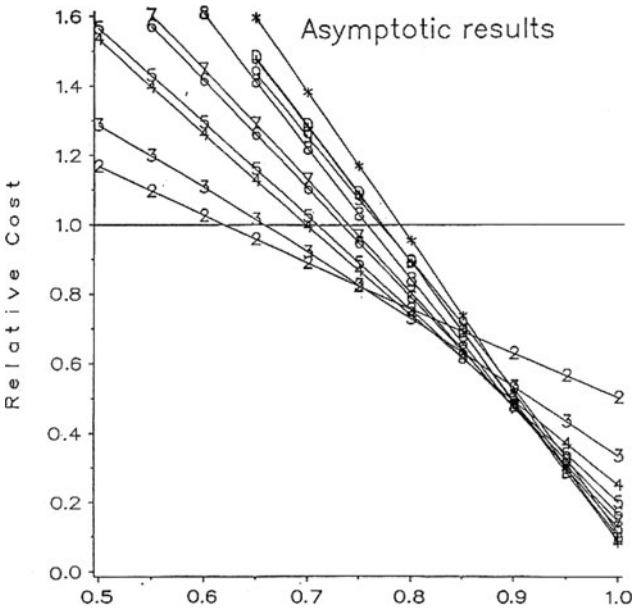

Fig. 2.14 Relative cost for composite sample sizes of $k = 2$–12

Larson (1969) have published such optimized algorithms for $p = 0.01(0.01)0.1$ and where the initial size of the unclassified pool can be as large as $m = 50$.

2.2.8 Exhaustive Retesting in the Presence of Classification Errors

The problem is to classify every individual sample as polluted or not polluted using presence/absence measurements. Suppose that there is a positive probability of misclassifying any sample, either individual or composite, and assume that the probability of misclassification depends only on whether or not the sample is polluted. In particular, composite samples and individual samples have the same misclassification rates. Let r_n be the probability of a false-negative classification. That is,

$$r_n = \Pr[\text{negative test result} \mid \text{sample is polluted}].$$

Similarly, let r_p be the probability of a false-positive classification. That is,

$$r_p = \Pr\left[\text{positive test result} \mid \text{sample is not polluted}\right].$$

Now consider using the exhaustive retesting procedure with composites of size k. Let

$$d_n = \Pr\left[\text{negative classification} \mid \text{individual sample is polluted}\right]$$

Table 2.7 Optimal composite sample size k over the tabulated ranges of p for $m = 1$–16, 25, 50, and ∞

k	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
1	0.0–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.0– 0.62
2	–	0.0–0.38	0.29–0.38	0.24–0.38	0.27–0.38	0.25–0.38	0.25–0.38	0.25– 0.38
3	–	–	0.0–0.29	–	0.19–0.27	0.17–0.25	0.21–0.25	0.19–0.25
4	–	–	–	0.0–0.24	–	0.16–0.17	0.14–0.21	0.12–0.19
5	–	–	–	–	0.0–0.19	–	–	–
6	–	–	–	–	–	0.0–0.16	–	–
7	–	–	–	–	–	–	0.0–0.14	–
8	–	–	–	–	–	–	–	0.0–0.12
k	$m=9$	$m=10$	$m=11$	$m=12$	$m=13$	$m=14$	$m=15$	$m=16$
1	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0	0.38–1.0
2	0.25–0.38	0.22–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.25–0.38	0.24–0.38
3	0.18–0.25	0.19–0.22	0.19–0.25	0.19–0.25	0.18–0.25	0.18–0.38	0.18–0.25	0.19–0.24
4	–	0.17–0.19	0.13–0.19	0.14–0.19	0.16–0.18	0.16–0.18	0.15–0.18	0.13–0.19
5	0.11–0.18	0.10–0.17	–	0.13–0.14	0.12–0.16	0.12–0.16	0.11–0.15	–
6	–	–	0.12–0.13	–	–	–	–	–
7	–	–	0.09–0.12	0.08–0.13	0.08–0.12	0.07–0.12	–	0.11–0.13
8	–	–	–	–	–	–	–	0.10–0.11
9	0.89–1.0	–	–	–	–	–	0.07–0.11	–
10	–	0.0–0.10	–	–	–	–	–	–
11	–	–	0.0–0.09	–	–	–	–	0.07–0.10
12	–	–	–	0.0–0.08	–	–	–	–
13	–	–	–	–	0.0–0.08	–	–	–
14	–	–	–	–	–	0.0– 0.07	0.0–0.07	–
15	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	0.0–0.07

$m = 25$

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0.38–1.0	0.25–0.38	0.18–0.25	0.15–0.18	0.11–0.15
$k=7$	$k=9$	$k=10$	$k=16$	$k=25$
0.09–0.11	0.08–0.09	0.06–0.08	0.03–0.06	0.0–0.03

$m = 50$

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=7$
0.38–1.0	0.25–0.38	0.18–0.25	0.15–0.18	0.11–0.15	0.09–0.11
$k=8$	$k=9$	$k=15$	$k=17$	$k=24$	$k=50$
0.08–0.09	0.06–0.08	0.05–0.06	0.03–0.05	?–0.03	0.0–?

∞ (asymptotic case)

$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0.38–1.0	0.25–0.38	0.18–0.25	0.14–0.18	0.11–0.14
$k=7$	$k=8$	$k=9$	$k=16$	$k=32$
0.09–0.11	0.08–0.09	0.06–0.08	0.03–0.06	?–0.03

The question mark (?) indicates the detection limit (which is a small positive number).

and

$$d_p = \Pr\left[\text{positive classification} \mid \text{individual sample is not polluted}\right].$$

Consider the computation of d_n. There are two ways in which an individual sample can be misclassified as negative. First, the composite is misclassified as negative, so that every individual sample is automatically (but incorrectly) classified as negative. The probability of this happening is r_n. Second, the composite is correctly

classified as positive (which occurs with probability $1 - r_n$), but the individual sample incorrectly tests negative (the probability of this happening is r_n). Thus

$$d_n = r_n + (1 - r_n)r_n = 2r_n - r_n^2.$$

If the misclassification rate r_n is small, then the negative misclassification rate with exhaustive retesting is approximately twice that for individual testing.

Similarly, the probability of a false-positive classification can be shown to be

$$d_p = r_p \left\{ r_p + \left(1 - q^{k-1}\right)\left(1 - r_n - r_p\right) \right\}.$$

For example, suppose $r_n = r_p = 0.2$ and suppose 10% of the individual samples are polluted. Using a composite sample size of $k = 4$, we have $d_n = 0.36$ and $d_p = 0.07$. Notice that d_n is approximately twice r_n while the false-positive rate under compositing, d_p, is substantially less than for individual testing.

Misclassification also affects the relative cost of a retesting procedure. The exhaustive retesting procedure results in either one test or $k + 1$ tests to process k individual samples. If the tests were free of error, then a single test is required when all k items are unpolluted. In the presence of testing error, a single test can also occur when the composite incorrectly tests as negative. Examining the two cases, we see that

$$\begin{aligned} \Pr[\text{one test}] &= r_n \left(1 - q^k\right) + \left(1 - r_p\right) q^k \\ &= r_n + q^k \left(1 - r_n - r_p\right). \end{aligned}$$

Also

$$\begin{aligned} \Pr[k + 1 \text{ tests}] &= 1 - \Pr[\text{one test}] \\ &= 1 - r_n - q^k \left(1 - r_n - r_p\right). \end{aligned}$$

The relative cost of classification becomes

$$\begin{aligned} \text{RC} &= \frac{1}{k} + \Pr[k + 1 \text{ tests}] \\ &= 1 + \frac{1}{k} - r_n - q^k \left(1 - r_n - r_p\right). \end{aligned}$$

Note that this expression reduces to the relative cost given in (2.2) when $r_n = r_p = 0$.

2.2.9 Other Costs

Laboratory procedures become more costly and error-prone when numerous steps must be performed in sequence. From this point of view, the exhaustive retesting

method of Dorfman is certainly the easiest composite strategy to implement since no more that two steps are required. By contrast, the sequential retesting method may require $k + 1$ steps.

But what, precisely, is meant by the number of steps? Conceptually, let us suppose that each test requires one unit of time and, further, that all tests are performed as soon as possible. That is, a test is deferred only if its execution requires the results of other, earlier, tests. Then a composite classification design is said to be *R-step* if its performance to completion could require as many as R units of time. We also refer to R as the *maximum duration* of the design. Clearly, $R = 2$ for exhaustive retesting.

Retesting strategies require that aliquots or duplicates be maintained for each separate item. The maintenance of these duplicates can be a significant portion of total cost and a relevant consideration in deciding which design to select. In addition, another source of error is introduced if true duplicates are difficult to achieve – and the impact of this error source will grow as the needed number of duplicates grows. We define the *maximum aliquot count* (MAC) to be the number of duplicates of each item that must be available to complete the procedure under all possible circumstances. For simple hierarchical designs the MAC is the same as the maximum duration of the design. For feedback designs like the entropy-based procedure, the MAC and the maximum duration can be different – and each can be unreasonably large. These issues are explored further in the Exercises.

2.3 Continuous Response Variables

The previous sections have examined retesting strategies for presence/absence measurements, i.e., where the response variable is binary. The rationale for the different strategies is found in the following two properties.

Property *N*. If the composite is negative, then *every* item in the composite is negative.

Property *P*. If the composite is positive, then *at least one* item in the composite is positive.

Property N is the fundamental premise of compositing for classification and accounts for the method's efficiency since it allows an entire group of items to be classified on the basis of a single measurement. Property P is the justification for curtailment and can lead to improved efficiencies, but is not fundamental to compositing for classification.

We now want to turn our attention to response variables X that are nonnegative but continuously distributed. An individual item is classified as positive if $X \geq c$ for that item, where c is a specified criterion level. The proportion of positive items, i.e., the *prevalence*, is given as $p = \Pr[X \geq c]$. Consider a composite whose k items have individual values $X_1, X_2, \ldots, X_k$. Barring measurement error, the measured

response Y on the composite is the average of $X_1, X_2, \ldots, X_k$ so that

$$kY = X_1 + X_2 + \cdots + X_k.$$

We cannot classify the composite as negative whenever $Y < c$, for then Property N would fail. But, since X is nonnegative, it is certainly the case that $X_i < c$ for all i whenever $X_1 + X_2 + \cdots + X_k < c$. Thus, Property N will be true provided a composite is classified as negative whenever $kY < c$. The probability of a negative composite is then

$$\begin{aligned} q_k &= \Pr[\text{composite is negative}] \\ &= \Pr[X_1 + X_2 + \cdots + X_k < c]. \end{aligned} \tag{2.13}$$

When the individual items can be regarded as statistically independent, then the distribution of $X_1 + X_2 + \cdots + X_k$ in (2.13) is that of the k-fold convolution of X. Throughout we suppose that the individual items can in fact be treated as independent.

The exhaustive retesting procedure is exactly the same as in the case of a binary response: Measurement is made on a composite of size k. All individual items are classified as negative, if the composite is negative; otherwise, measurement is made individually on each item. As in the case of a binary response, the relative saving is

$$\begin{aligned} \text{RS} &= \Pr[\text{composite is negative}] - 1/k \\ &= q_k - 1/k. \end{aligned} \tag{2.14}$$

Recall that $q_k = q^k = (1 - p)^k$ for a binary response variable. Notice that RS depends upon k and also upon p through the criterion level c. But RS also depends upon the distribution of X because of the k-fold convolution occurring in (2.13). This latter dependence is the principal distinction between the continuous and binary response scenarios. Unfortunately, the relative savings is very sensitive to the underlying distribution of X, as we shall see below.

We have computed the relative savings as a function of k and p for the gamma distribution with index parameter a and for the lognormal distribution with logarithmic variance σ^2. The results, for $k = 2, 3, 4$, are shown in Fig. 2.15. Since the relative savings, considered as a function of p, does not depend upon the scaling of X, it was convenient to index the distributions in Fig. 2.15 by their coefficients of variation CV. The corresponding parameter values are

CV	a	σ^2
1.5	0.44	1.09
1.0	1.00	0.83
0.5	4.00	0.47

For comparison purposes, Fig. 2.15 also shows the relative savings for a binary response variable (Bernoulli distribution).

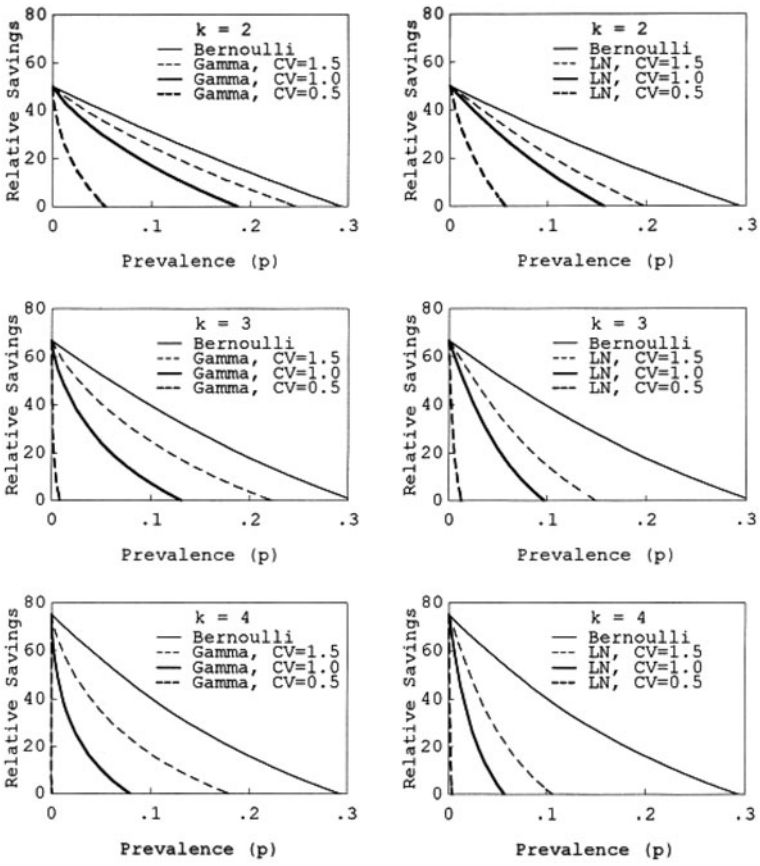

Fig. 2.15 Relative savings of exhaustive retesting when the response variable is binary (Bernoulli) or continuously distributed with a gamma distribution or a lognormal distribution (LN). Relative savings is expressed as a percent and the coefficient of variation is denoted by CV

Several conclusions emerge from an examination of Fig. 2.15:

1. For fixed k and p, the relative savings is greater for the binary response variable than for any continuous response variable.
2. Within each family of distributions (gamma and lognormal) the relative savings *increases* with the skewness of the distribution and, conversely, in the other direction the relative savings decreases as the distribution approaches normality.
3. For fixed p, the performance is quite sensitive to k and to the underlying distribution of X. The relative savings decreases rapidly with increasing k and with decreasing variability of X.

In view of the sensitivity described in item 3, there would be little purpose in attempting to optimize with respect to k. In fact, it appears unlikely that one would want to consider composite sizes much larger than $k = 2$ or $k = 3$.

We now examine items 1 and 2 to determine the extent to which they do or do not hold generally. Let p be fixed and determine the criterion level c by the requirement that $\Pr[X \geq c] = p$. We define a random variable $X(p)$ and its corresponding distribution by truncating X to the interval $[0, c)$ and then rescaling to the interval $[0, 1)$ so that

$$X(p) = \frac{1}{c} X|_{x<c}. \tag{2.15}$$

Now, the event $\{X_1 + \cdots + X_k < c\}$ is contained in the event $\{X_1 < c, \ldots, X_k < c\}$ which implies that q_k is given by

$$\begin{aligned} \Pr[X_1 + \cdots + X_k < c] &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_k < c] \\ &\quad \times \Pr[X_1 < c_1, \ldots, X_k < c] \\ &= \Pr[X_1(p) + \cdots + X_k(p) < 1] q^k, \end{aligned} \tag{2.16}$$

where $q = 1 - p$ and $X_1(p), \ldots, X_k(p)$ are independent realizations of $X(p)$. Equation (2.16) implies that

$$q_k \leq q^k, \tag{2.17}$$

so that the relative savings is always greater for a binary response variable than for a continuous response variable.

A distribution function F is stochastically smaller than a distribution function G provided $F(x) \geq G(x)$ uniformly in x. Given two response variables X and $\tilde{X}$, we shall say that X is more *zero-concentrated* than $\tilde{X}$ if the distribution of $X(p)$ is stochastically smaller than that of $\tilde{X}(p)$ for all p. In this case,

$$\Pr[X_1(p) + \cdots + X_k(p) < 1] \geq \Pr[\tilde{X}_1(p) + \cdots + \tilde{X}_k(p) < 1].$$

Comparing with (2.16), we see for all k and p that the relative savings of exhaustive retesting is greater for the more zero-concentrated response variable. This result is related to item 2 above. We generally associate skewness of a nonnegative random variable with a heavy right-hand tail. But for typical families of distributions, the right tail and the left tail are linked in a way that makes the left tail more zero-concentrated as the right tail becomes more elongated. Thus, typically but not always, a large skewness goes hand in hand with a large relative savings.

2.3.1 Quantitatively Curtailed Exhaustive Retesting

It would be possible to curtail the exhaustive retesting procedure along the lines used for a binary response variable. But the quantitative nature of the measurements allows for a more sophisticated form of curtailment. Suppose the measurement Y on a composite indicates that retesting is needed, i.e., $kY \geq c$. After the first j items have been measured, the individual values $X_1, \ldots, X_j$ are available and, consequently, the total for the remaining items can be calculated as

$$X_{j+1} + \cdots + X_k = kY - (X_1 + \cdots + X_j).$$

If this residual total is less than c, then these $k - j$ remaining items can be classified as negative without further testing. If the residual total exceeds c *and* if $j < k - 1$ then measurement is made on item $j + 1$ and the procedure iterates. If $j = k - 1$, then the value X_k is known so that item k can be classified without measurement on that item.

The number of retests R_k can take the values $0, 1, 2, \ldots, k - 1$ and the total number of measurements, $T_k = 1 + R_k$, ranges from 1 to k. Now $R_k \geq j$ means that item j must be tested, which is equivalent to saying that $X_j + \cdots + X_k \geq c$. Thus,

$$\begin{aligned}
&\Pr[R_k \geq 0] = 1 \\
&\Pr[R_k \geq 1] = \Pr[X_1 + \cdots + X_k \geq c] = 1 - q_k \\
&\Pr[R_k \geq 2] = \Pr[X_2 + \cdots + X_k \geq c] = 1 - q_{k-1}.
\end{aligned}$$

This pattern continues until

$$\Pr[R_k \geq k - 1] = \Pr[X_{k-1} + X_k \geq c] = 1 - q_2.$$

Now

$$\begin{aligned}
E[T_k] &= \Pr[T_k > 0] + \Pr[T_k > 1] + \cdots + \Pr[T_k > k - 1] \\
&= \Pr[R_k \geq 0] + \Pr[R_k \geq 1] + \cdots + \Pr[R_k \geq k - 1] \\
&= k - (q_2 + q_3 + \cdots + q_k).
\end{aligned}$$

From this, we obtain that the relative cost is

$$\mathrm{RC} = (q_2 + q_3 + \cdots + q_k)/k,$$

and the relative savings becomes

$$\mathrm{RS} = 1 - (q_2 + q_3 + \cdots + q_k)/k.$$

As the exercises indicate, quantitative curtailment improves the performance of exhaustive retesting rather markedly. But keep in mind that the derivations of this

section do not account for the effects of measurement error, imperfect mixing, or imperfect duplicates.

2.3.2 Binary Split Retesting

As in the case of presence/absence measurement, any positively testing composite sample is divided into two groups as nearly equal in size as possible. In the presence/absence case, both the composite samples are tested. For continuous measurements, only one composite sample need be formed and tested. The value of the second composite sample can be calculated. Let T_k be the number of tests necessary to classify all k samples starting with a composite sample of size k. Let $U_j = X_1 + \cdots + X_j$, and let $V_j = X_{j+1} + \cdots + X_k = U_k - U_j$. Here we assume that the X_i's are independent and identically distributed Bernoulli random variables. Consider the five mutually exclusive cases defined in terms of values of J_k given below.

Let $\left[\frac{k}{2}\right]$ be the largest integer less than or equal to $k/2$. Then $V_{\left[\frac{k}{2}\right]}$ has the same distribution as $U_{k-\left[\frac{k}{2}\right]}$. Let

$$J_k = \begin{cases} 0 & \text{if} \quad U_k < c \\ 1 & \text{if} \quad U_k \geq c, U_{\left[\frac{k}{2}\right]} < c, V_{\left[\frac{k}{2}\right]} < c \\ 2 & \text{if} \quad U_{\left[\frac{k}{2}\right]} \geq c, V_{\left[\frac{k}{2}\right]} < c \\ 3 & \text{if} \quad U_{\left[\frac{k}{2}\right]} < c, V_{\left[\frac{k}{2}\right]} \geq c \\ 4 & \text{if} \quad U_{\left[\frac{k}{2}\right]} \geq c, V_{\left[\frac{k}{2}\right]} \geq c. \end{cases}$$

Then

$$\begin{aligned} \Pr[J_k = 0] &= q_k, \\ \Pr[J_k = 2] &= \left(1 - q_{\left[\frac{k}{2}\right]}\right) q_{k-\left[\frac{k}{2}\right]}, \\ \Pr[J_k = 3] &= q_{\left[\frac{k}{2}\right]} \left(1 - q_{k-\left[\frac{k}{2}\right]}\right), \\ \Pr[J_k = 4] &= \left(1 - q_{\left[\frac{k}{2}\right]}\right) \left(1 - q_{k-\left[\frac{k}{2}\right]}\right). \end{aligned}$$

Thus

$$\begin{aligned} \Pr[J_k = 1] &= 1 - \Pr[J_k = 0] - \Pr[J_k = 2] - \Pr[J_k = 3] - \Pr\left[j_k = 4\right] \\ &= q_{\left[\frac{k}{2}\right]} q_{k-\left[\frac{k}{2}\right]} - q_k. \end{aligned}$$

So

$$
\begin{aligned}
E[T_k] &= E[E(T_k|J_k)] = \sum_{j=0}^{4} E(T_k|J_k = j)\Pr\left[J_k = j\right] \\
&= 1 \cdot \Pr[J_k = 0] + 2\Pr[J_k = 1] + \left(1 + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\right)\Pr[J_k = 2] \\
&\quad + \left(1 + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\right)\Pr[J_k = 3] \\
&\quad + \left(2 + \left\{E\left(T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right) - 1\right\} + \left\{E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil}\right] - 1\right\}\right)\Pr[J_k = 4] \\
&= \Pr[J_k = 0] + 2\Pr[J_k = 1] + \Pr[J_k = 2] + \Pr[J_k = 3] \\
&\quad + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\{\Pr[J_k = 2] + \Pr[J_k = 4]\} \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\{\Pr[J_k = 3] + \Pr[J_k = 4]\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
E[T_k] &= 1 + \Pr[J_k = 1] - \Pr[J_k = 4] + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{\lceil \frac{k}{2} \rceil}\right) \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{k-\lceil \frac{k}{2} \rceil}\right) \\
&= q_{\lceil \frac{k}{2} \rceil} + q_{k-\lceil \frac{k}{2} \rceil} - q_k \\
&\quad + E\left[T_{\lceil \frac{k}{2} \rceil}|U_{\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{\lceil \frac{k}{2} \rceil}\right) \\
&\quad + E\left[T_{k-\lceil \frac{k}{2} \rceil}|U_{k-\lceil \frac{k}{2} \rceil} \geq c\right]\left(1 - q_{k-\lceil \frac{k}{2} \rceil}\right).
\end{aligned}
$$

Now

$$
\begin{aligned}
E\left[T_j\right] &= E\left[T_j|U_j < c\right]\Pr\left[U_j < c\right] + E\left[T_j|U_j \geq c\right]\Pr\left[U_j \geq c\right] \\
&= q_j + E\left[T_j|U_j \geq c\right](1 - q_j).
\end{aligned}
$$

So

$$
E\left[T_j|U_j \geq c\right] = \frac{E\left[T_j\right] - q_j}{1 - q_j}.
$$

Therefore,

$$E[T_k] = q_{\lceil \frac{k}{2} \rceil} + q_{k-\lceil \frac{k}{2} \rceil} - q_k + E\left[T_{\lceil \frac{k}{2} \rceil}\right] - q_{\lceil \frac{k}{2} \rceil} + E\left[T_{k-\lceil \frac{k}{2} \rceil}\right] - q_{k-\lceil \frac{k}{2} \rceil}$$
$$= E\left[T_{\lceil \frac{k}{2} \rceil}\right] + E\left[T_{k-\lceil \frac{k}{2} \rceil}\right] - q_k, \quad k = 2, 3, \ldots.$$

Now $E[T_1] = 1$ so, $E[T_2] = 2 - q_2$, $E[T_3] = 3 - q_2 - q_3$, $E[T_4] = 2(2 - q_2) - q_4$, etc.

This recurrence (difference) equation can be solved iteratively for composite samples of arbitrary size k. If the composite sample size is a power of 2, the recurrence formula can be solved, giving

$$E(T_{2^r}) = 2^r - q_{2^r} - 2q_{2^{r-1}} - \cdots - 2^{r-1}q_2, \quad r = 2, 3, \ldots. \tag{2.18}$$

The following examples may be useful:

$$\begin{aligned}
E[T_5] &= 5 - 2q_2 - q_3 - q_5, \\
E[T_6] &= 6 - 2q_2 - 2q_3 - q_6, \\
E[T_7] &= 7 - 3q_2 - q_3 - q_4 - q_7, \\
E[T_9] &= 9 - 4q_2 - q_3 - q_4 - q_5 - q_9, \\
E[T_{10}] &= 10 - 4q_2 - 2q_3 - 2q_5 - q_{10}, \\
E[T_{11}] &= 11 - 4q_2 - 3q_3 - q_5 - q_6 - q_{11}, \\
E[T_{12}] &= 12 - 4q_2 - 4q_3 - 2q_6 - q_{12}, \\
E[T_{13}] &= 13 - 5q_2 - 3q_3 - q_4 - q_6 - q_{13}, \\
E[T_{14}] &= 14 - 6q_2 - 2q_3 - 2q_4 - 2q_7 - q_{14}, \\
E[T_{15}] &= 15 - 7q_2 - q_3 - 3q_4 - q_7 - q_8 - q_{15}.
\end{aligned}$$

It is interesting to note that the coefficients in the expressions for $E(T_k)$ add up to 1. The relative cost can be found by dividing $E[T_k]$ by the composite sample size, i.e., $\mathrm{RC}_k = E[T_k]/k$.

The optimal binary split retesting procedure starts with the largest possible composite sample size. To understand why this is so, consider combining two composite samples to form a larger composite sample. If the larger composite sample tests negative, then one test is saved. If the larger composite sample tests positive, then one of the original (smaller) composite samples is tested and the value of the other composite sample is calculated. This results in a total of two tests, namely one on the larger composite sample and the other on one of the smaller composite samples. That is to say, it takes the same number of tests to reach this point if the two composite samples were tested separately. In practice, the composite sample size will be limited by the number of subsamples that can be mixed or ratio of the detection limit to the action level.

2.3.3 Entropy-Based Retesting

This method appears to be inappropriate for continuous response variables. Recall that a positively testing composite results in the formation of a subcomposite of size about half of that of the original composite. If this subcomposite tests positive, then the remaining items in the original composite are returned to the pool of unclassified items. However, the total for these remaining items can be calculated, and it seems unreasonable not to use this information. If these items are not returned to the unclassified pool, then the resulting procedure reduces to the curtailed binary split retesting procedure.

2.4 Cost Analysis of Composite Sampling for Classification

2.4.1 Introduction

Sampling plans for environmental and public health monitoring often involve expensive laboratory methods for quantifying observations on individual sample units. United States Environmental Protection Agency (US EPA) and the regulated community spend an estimated $5 billion every year on collecting data for research, regulatory decision making, and regulatory compliance (US EPA, 1994). Johnson and Patil (2001) have carried out a cost analysis of composite sampling for classification. The cases of presence/absence and continuous measurements are considered. The general cost expression is using probability theory and the relative cost of composite sampling is derived in comparison with the conventional method of using individual sample measurements.

Cost analysis of composite sampling is not as easy as determining the expected number of measurements to be made. Composite sampling involves costs that do not arise in the conventional method of making measurements on individual sample units. For instance, forming a composite sample of several soil samples requires careful laboratory procedure of cleaning the containers and rinsing the solvent before forming every composite. Similarly, when composite sampling is used for sampling of fluids or gases, appropriate procedures have to be carefully implemented while forming composite samples. Also, archiving aliquots of individual samples for possible retesting involves storage and retrieval costs. Finally, any extra labor costs must be taken into account before concluding as to whether or not composite sampling will be truly cost-effective.

2.4.2 General Cost Expression

A general cost expression is derived by taking into consideration various cost components and different retesting strategies. The following notation is used in the derivation of the cost expression and relative cost of composite sampling:

m = number of individual sample units to be classified,
n = number of composite sample units,
k = number of individual sample units contributing to a single composite sample,
C_s = cost of acquisition of an individual sample unit,
C_a = cost of archiving an individual sample unit,
C_c = cost of forming a composite sample,
C_t = cost of testing a sample unit, either individual or composite,
Y_k = number of composites to be formed, each of size k,
T_k = number of tests for classifying k individual sample units in a composite.

Then the relative cost of composite sampling is given by

$$\mathrm{RC} = \{(C_s + C_a) + C_c E[Y_k] + C_t E[T_k]\}/(C_s + C_t).$$

The procedure to be followed for classifying m individual sample units by composite sampling with selective retesting is as follows:

1. Obtain m individual sample units.
2. Obtain an aliquot from every sample unit and archive the remaining material for possible retesting.
3. Form n composite samples, each consisting of aliquots from l individual sample units.
4. Analyze the n composite samples.
5. Classify a composite sample as "clean" or "contaminated." If a composite sample is "clean," then all individual samples contributing aliquots to that composite are classified as "clean." Otherwise, retesting must be undertaken to classify individual samples as "clean" or "contaminated."
6. Retest the archived sample units based on the result of Step 5. All the m individual sample units are classified.

2.4.3 *Effect of False Positives and False Negatives on Composite Sample Classification*

Testing mechanisms are subject to some degree of error. In the case of binary classification, an error is either a false-positive or a false-negative test. Let r_{n} and r_{p} denote the rates of these two errors, respectively. These rates are defined by the following probability statements:

$r_{\mathrm{p}} = \Pr(\text{positive test result} \mid \text{clean sample})$,
$r_{\mathrm{n}} = \Pr(\text{negative test result} \mid \text{contaminated sample})$.

Composite sampling with retesting reduces the overall false-positive error rate with a trade-off that the false-negative error rate can be magnified due to

compositing. The false-negative error rate may be controlled by retesting some of the negative testing composites, though this would increase the expected number of tests.

The interest is in $E[Y_k]$ and $E[T_k]$ for the specified values of C_s, C_a, C_c, C_t, and k. In situations where r_p and r_n are not negligible, the overall design false-positive error rate, d_p, *and* false-negative error rate, d_n, are also derived.

2.4.4 Presence/Absence Measurements

The measurement of a sample returns a binary response indicating presence or absence of the trait of interest in the tested sample. The Bernoulli model with parameter p is appropriate for this situation under the assumption that all m individual measurements are independent and identically distributed with probability p of testing positive.

2.4.4.1 Exhaustive Retesting

Exhaustive retesting, as proposed by Dorfman (1943), does not result in re-formation of composites and hence $E[Y_k] = 1$. The number of tests will be 1 (when the composite tests negative) or $k + 1$ (when the composite tests positive). Writing $q = 1 - p$, where p denotes the probability that the trait is present in an individual sample, the expected number of tests is given by

$$\begin{aligned} E[T_k] &= 1 \cdot q^k + (k+1) \cdot (1 - q^k) \\ &= 1 + k(1 - q^k). \end{aligned}$$

The overall design false-negative error rate is given by

$$\begin{aligned} d_\mathrm{n} &= r_\mathrm{n} + (1 - r_\mathrm{n}) r_\mathrm{n} \\ &= 2r_\mathrm{n} - r_\mathrm{n}^2. \end{aligned}$$

The overall design false-positive error rate is given by

$$\begin{aligned} d_\mathrm{p} &= r_\mathrm{p} \cdot [\Pr(\text{retest} | \text{at least one of } k-1 \text{ is positive}) \cdot (1 - q^k) \\ &\qquad + \Pr(\text{retest} | \text{all } k-1 \text{ are negative})\, q^{k-1}] \\ &= r_\mathrm{p} \cdot [(1 - r_\mathrm{n})(1 - q^{k-1} + r_\mathrm{p} \cdot q^{k-1}] \\ &= r_\mathrm{p} \cdot [1 - r_\mathrm{n} - q^{k-1}(1 - r_\mathrm{n} - r_\mathrm{p})]. \end{aligned}$$

$$\begin{aligned} \Pr(T_k = 1) &= r_\mathrm{n}(1 - q^k) + (1 - r_\mathrm{p})\, q^k \\ &= r_\mathrm{n} + q^k(1 - r_\mathrm{n} - r_\mathrm{p}), \end{aligned}$$

$$\begin{aligned}\Pr(T_k = k+1) &= 1 - \Pr(T_k = 1)\\ &= 1 - r_{\mathrm{n}} - q^k(1 - r_{\mathrm{n}} - r_{\mathrm{p}}).\end{aligned}$$

Therefore,

$$E[T_k] = k + 1 - k[r_{\mathrm{n}} + q^k(1 - r_{\mathrm{n}} - r_{\mathrm{p}})].$$

2.4.4.2 Sequential Retesting

Sterret (1957) proposed sequential retesting method as an improvement in the exhaustive retesting method. In this method,

$$\begin{aligned}E[Y_k] &= 1 + (k-2)p,\\ E[T_k] &= 2k - (k-3)q - q^2 - (1 - q^{k+1})/(1-q), \quad k = 2, 3, \ldots.\end{aligned}$$

If r_{n} and r_{p} are not negligible, then

$$\begin{aligned}E[Y_k] &= 1 + (k-2)[r_{\mathrm{p}}q + (1 - r_{\mathrm{n}})p],\\ E[T_k] &= 2k - (k-3)[r_{\mathrm{n}}p + (1 - r_{\mathrm{p}})q] - [r_{\mathrm{n}} + (1 - r_{\mathrm{p}})q]^2\\ &\quad - \{1 - [r_{\mathrm{n}}p + (1 - r_{\mathrm{p}})q]^{k+1}\}/\{r_{\mathrm{p}}\, q + (1 - r_{\mathrm{n}})p\}, \quad k = 3, 4, \ldots.\end{aligned}$$

2.4.4.3 Binary Split Retesting

The binary split retesting method of Gill and Gottlieb (1974) entails a recurrence relation for the expected number of composites to be formed and for expected number of tests to be carried out.

$$\begin{aligned}E[Y_k] &= E[Y_{k1}] + E[Y_{k2}] + 1 - 2q^k, \quad k = 4, 5, \ldots,\\ E[T_k] &= E[T_{k1}] + E[T_{k2}] + 1 - 2q^k, \quad k = 2, 3, \ldots,\end{aligned}$$

where $k_1 = k_2 = k/2$ if k is even and $k_1 = (k-1)/2$ and $k_2 = (k+1)/2$ if k is odd.

The design false-positive rate is given by

$$\begin{aligned}d_{\mathrm{p}} =&[(1 - r_{\mathrm{n}})(1 - q^k) + r_{\mathrm{p}}q^k] \times [(1 - r_{\mathrm{n}})(1 - q^{k1}) + r_{\mathrm{p}}q^{k1}]\\ &\times [(1 - r_{\mathrm{n}})(1 - q^{k11}) + r_{\mathrm{p}}q^{k11}] \times \cdots \times r_{\mathrm{p}}.\end{aligned}$$

Here, k_1 and k_2 denote sizes of subcomposites of the composite of size k. These are in turn split into subcomposites of sizes k_{11} and k_{12}, k_{21} and k_{22}, respectively, and so on. The design false-negative error rate d_{n} for an initial composite of size k is given by

$$\begin{aligned} d_{\mathrm{n}}(k) &= r_{\mathrm{n}} - r_{\mathrm{n}}^2 + d_{\mathrm{n}}(k/2) \quad \text{if k is even,} \\ &= r_{\mathrm{n}} - r_{\mathrm{n}}^2 + d_{\mathrm{n}}((k+1)/2) \quad \text{if k is odd.} \end{aligned}$$

Here, the argument of d_{n} indicates the composite sample size.

The expected number of composites and tests is, respectively, given by

$$\begin{aligned} E[Y_k] &= E[Y_{k1}] + E[Y_{k2}] + 1 - 2[(1 - r_{\mathrm{p}})q + r_{\mathrm{n}}p]^k, \quad k = 4, 5, \ldots, \\ E[T_k] &= E[T_{k1}] + E[T_{k2}] + 1 - 2[(1 - r_{\mathrm{p}})q + r_{\mathrm{n}}p]^k, \quad k = 2, 3, \ldots. \end{aligned}$$

These equations can be solved recursively for the appropriate value of k.

2.4.5 Continuous Measurements

Measurement on a continuous random variable results in classifying a sample unit as negative if its measured value is less than some numerical criterion. However, a size k composite cannot be classified as negative using the same criterion because it would not imply that every individual sample unit contributing to the composite is negative. For this purpose, if c denotes the criterion for an individual sample, the numerical criterion for a composite sample of size k is c/k.

If $X_1, X_2, \ldots, X_k$ denote individual sample values and Y denotes the composite value, then the probability that a composite sample tests negative is

$$\begin{aligned} q_k &= \Pr[X_1 + X_2 + \cdots + X_k < c] \\ &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_{\mathrm{k}} < \mathrm{c}] \times \Pr[X_1 < c, \ldots, X_k < c] \\ &= \Pr[X_1 + \cdots + X_k < c | X_1 < c, \ldots, X_k < c] \times q^k, \end{aligned}$$

where $q = \Pr[X_1 < c]$ and $X_1, \ldots, X_k$ are independent and identically distributed.

Since $q_k < q^k$, the relative cost for measuring a continuous variable has a lower bound representing the relative cost for measuring a presence/absence variable. The expected number of composites and tests heavily depends on the probability distribution of the individual sample values. Approximations can be obtained through expressions in case of presence/absence measurement. However, such an approximation can have a dual impact on the result.

On the one hand, this leads to over-optimistic results due to an upper bound on the probability of a composite sample testing negative. On the other hand, the presence/absence expressions are based on independence among individual sample measurements achieved through random formation of composite samples. When individual sample values are autocorrelated, better strategies of forming composites can be developed to improve the performance of composite sampling by reducing the relative cost. In this case, there is no comparison with the case of presence/absence measurements.

Chapter 3
Identifying Extremely Large Observations

3.1 Introduction

It is a common experience in several environmental problems of site characterization, cleanup evaluation, and compliance monitoring that the interest is in both the average and the extremely high contamination levels. Even though compositing, at least under idealistic conditions, incurs no loss of information for estimating the population mean, there is a loss of information on individual sample values, particularly extreme values. This has been a limitation on application of composite sampling techniques to environmental problems. There is very little literature about the detection of large individual sample values. Casey et al. (1985) give a method to predict the maximum sample value using composite sample measurements. Gore and Patil (1993) have developed a statistical method to identify the largest individual sample value without exhaustively measuring all the individual samples.

Under the assumption of continuous measurements, only one sample is likely to have the largest value. Identifying this sample is then comparable to detecting the defective item in a set known to contain exactly one defective. Can we use a classification procedure for identifying the largest individual sample value? The following observation is relevant to this question.

The classification problem for a continuous measurement has a criterion value c, and it is desired to identify every individual sample having a value that exceeds or equals c. Note that every retesting procedure involves at least as many measurements as there are individual samples satisfying this criterion. As a consequence, if the value of c is small then a large proportion of samples may satisfy the criterion, and hence the relative cost of classification will be high. In an extreme situation the criterion value can be so small that all individual samples are measured exhaustively. On the other hand, if the criterion value c is large, then not too many samples will satisfy the criterion, and hence the relative cost of classification will be small. Again, in an extreme situation the criterion value can be so large that no individual sample satisfies it, and hence the only information available at the end is on composite sample measurements. As a third possible scenario, consider the situation where the criterion value is exactly between the largest and the second largest sample values. In this case, the classification will result in identifying the sample maximum at a

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_3, © Springer Science+Business Media, LLC 2011

very small relative cost. However, since the sample maximum is both variable and unknown, it is unlikely to encounter this situation. However, this concept suggests a procedure that may involve a succession of classification problems, where the criterion value changes progressively.

Suppose measurements on n composites, each of size k, are available. Select a composite at random and exhaustively test all the constituent samples. Treat the largest of these individual sample values as the criterion value and initiate a classification process. As soon as an individual sample value exceeds the criterion value, the criterion value is changed to this individual sample value. In this way, the successive criterion values will be ascending in magnitude like record values, and finally we will have the sample maximum identified. How many measurements are required to identify the sample maximum by this method? It is a variable that depends on the choice of the composite that is selected to initiate the process. It is then obvious to optimize the choice of this composite so as to minimize the total number of measurements leading to the identification of the sample maximum. There are two choices available at this point of time. First, fix the number of measurements that can be made on individual samples and infer on the sample maximum using these measurements optimally. Second, select the initial composite sample with an objective of minimizing the total number of measurements required to identify the sample maximum.

The methods presented in this chapter for identifying the sample maximum are based on the two choices described above. The first, discussed in Section 3.2, was initially developed by Casey et al. (1985), where constituent samples of only the largest measuring composite are subjected to further measurement. The second, discussed in Section 3.3, is presented in Gore and Patil (1993), where the number of measurements required to identify the sample maximum is a variable, but is minimized by careful successive selection of composites for measurements on constituent individual samples.

3.2 Prediction of the Sample Maximum

Casey et al. (1985) developed a method to predict the maximum individual sample value by testing individual samples comprising the composite with the largest measurement. Since there is no certainty that this method will always identify the true maximum, they evaluated the probability that this method identifies the sample maximum by simulation.

Pollution control standards frequently specify both time-dependent arithmetic pollutant means and instantaneous maximum permitted values. Because of their potential toxicological significance, knowledge about extreme values is particularly important for the purpose of enforcing water quality standards. The maximum pollutant concentration should be an observed value or be estimated from other measurements.

The principal barrier to the detection of violations is the cost of extensive and comprehensive monitoring. Due to the random component of the data, no method exists that will find the maximum concentration with certainty unless continuous monitoring is used. Any sampling method adopted must be able to identify the maximum pollutant concentration a large proportion of the time. Failing this identification, the method must be able to signal the existence of excessive pollutant levels by finding some "large" value. A method is developed which can predict the maximum from a finite set of sequential samples without testing all samples. The following assumptions are used:

1. The process of collecting samples is distinct from their measurement
2. The cost of sample measurement is high relative to that of collection
3. The sample measurements have high positive autocorrelation

There are many situations in water, air, and industrial process monitoring where the collection and testing of samples are distinct. An incremental change in the number of tests performed is important because the cost of testing is typically much larger than that of sampling if laboratory analysis is required. The assumption of high positive autocorrelation is fundamental to the method developed. It permits the estimation of information about some of the pollutant samples which are collected but not tested.

Composite sampling, a common practice in water pollution monitoring, involves the physical pooling of a set of sequential samples prior to measurement. The result of this process is an arithmetic average of the samples that were composited. Assume there are m samples aggregated into n sequential groups. Then, within each group, a fixed portion of each sample is pooled to form a total of $[m/n]^*$ composite samples which are subsequently measured. In the presence of high positive autocorrelation, the maximum concentration among the m samples will tend to be surrounded by samples with high values. As a consequence, the composite sample that contains the maximum sample value will also tend to have a relatively high measurement. This suggests that the search for the maximum can be concentrated among the individual observations that formed the composites with the highest measured levels. Only a portion of each individual sample may be pooled when forming the composites, since the remaining portion must be retained for later analysis once the maximum among the composites is identified.

Primary first-order compositing (PFOC) consists of several steps. Initially, the composites are formed and measured, and the composite with the maximum level is identified. Then, all the samples that formed this composite are measured. The maximum of these sample measurements is the estimate of the maximum for all samples.

Several alternative composite methods should be considered. Improved performance in detecting extreme values would be guaranteed if sample examination were not confined to the composite with the highest measurement. The logical extension would include an analysis of the samples from the second highest composite, and the approach could be incrementally extended to other composites with lower values.

The resulting increase in performance would, of course, entail higher laboratory and related testing costs.

3.3 The Sweep-Out Method to Identify the Sample Maximum

Consider a composite sample of size k. Let $x_1, x_2, \ldots, x_k$ be the k individual sample values, and let y be the composite sample measurement. Further, let $x_{k:k}$ denote the maximum of the k individual sample values. That is,

$$x_{k:k} = \max\{x_1, x_2, \ldots, x_k\}.$$

Observe that

$$y \leq x_{k:k} \leq ky.$$

This inequality implies that the measurement on every composite sample gives a lower bound as well as an upper bound for the largest value among its constituent individual samples. It is then interesting to make the following observation.

Consider two composite samples. Let the composite sample sizes be k_1 and k_2, the composite measurements y_1 and y_2, and the maximum individual sample values $x_{k_1:k_1}$ and $x_{k_2:k_2}$. Without loss of generality assume that $y_1 < y_2$. In general, this does not imply $x_{k_1:k_1} \leq x_{k_2:k_2}$. However, if $k_1 y_1 \leq y_2$, then it can be inferred that $x_{k_1:k_1} \leq x_{k_2:k_2}$, and hence the first composite sample cannot contain the individual sample with the largest value. It is thus clear that there is no need to consider the first composite sample when searching for the individual sample with the largest value. In this way, we may eliminate a significant number of composite samples as not containing the individual samples having large values. This elimination process may finally leave us with a very few composite samples as possibly containing individual samples having large values. All the individual samples constituting these composite samples may then be subjected to measurement in order to identify the individual samples having large values.

Using the above reasoning, we obtain a sweep-out method as follows:

Identify the composite sample with the largest measurement. Denote the size of this composite by $k_{\max}$ and the measurement by $y_{\max}$. Clearly, the largest value, $x_{\max}$, say, of a constituent individual sample satisfies

$$y_{\max} \leq x_{\max} \leq k_{\max} y_{\max}.$$

Any other composite of size k and measurement y cannot contain an individual sample having a value that exceeds $x_{\max}$ if $ky \leq x_{\max}$. However, if $x_{\max} \leq ky$, then measure every individual sample that constitutes the composite having the measurement y. If an individual sample value exceeds $x_{\max}$, then $x_{\max}$ is redefined and assigned this value. If there is no composite sample that satisfies $x_{\max} \leq ky$, then the search for the largest individual sample value is complete.

The following algorithm describes the steps of the sweep-out method for identification of extremely large individual sample:

Step 1. Locate the composite sample with the largest measurement.
Step 2. Exhaustively measure every individual sample that constitutes this composite.
Step 3. Identify the largest observation M among the k individual measurements thus obtained.
Step 4. Check if any composite sample has a measurement Y that exceeds M/k, where k is the composite sample size.
Step 5. If there is any composite sample satisfying the condition in step 4, then exhaustively measure all of its constituent samples, identify the individual sample measurements that are larger than M, then replace M by the largest of these measurements, and repeat step 4. If there are two or more samples satisfying the condition in step 4, then start with the composite sample that has the largest measurement among all such composites.
Step 6. If no more composite samples satisfy the condition in step 4, then the search for the largest measurement is complete.
Step 7. The same method with the second largest measurement will identify the second largest measurement, etc.

3.4 Extensive Search of Extreme Values

The sweep-out method described in the previous section needs to be examined further so that its cost effectiveness in identifying extreme values can be evaluated. Note that, since exhaustive testing of all individual samples (without compositing) identifies all individual values, identification of extreme values is achieved simply by arranging the individual sample values in a descending order of magnitude. Thus the method of exhaustive testing involves as many measurements as the number of individual samples.

Another point worth noting at this stage is that we have assumed some spatial autocorrelation between individual sample values. In case of positive autocorrelation, samples from neighboring grid points are more likely to have similar values than samples from randomly selected grid points. In the extreme case of a perfect spatial autocorrelation, it is possible that the four individual samples with the largest measurements are composited in a single composite. In this case, measuring the four individual samples from a single composite would lead to the identification of four extremely large individual sample values. If this phenomenon is repeated in all the composites, then we will obtain a perfectly linear relationship between the number of extreme values identified and the number of measurements required.

The observations made above on the sweep-out method establish the need for an investigation of the statistical properties of the sweep-out method. The

performance of the sweep-out method will be partly determined by the spatial autocorrelation structure among the individual sample values, partly by the statistical distribution of these individual sample values and partly also by the compositing plan as well as the composite sample size. While there is a need to examine the relevant statistical properties and issues, the results of this chapter seem to be encouraging enough to recommend the use of composite sampling techniques in situations where the interest is not restricted to estimating the population mean, but it is also desired to identify extremely large individual sample values.

As a consequence of identifying individual samples having large values, the sweep-out method can provide estimates for the upper percentiles of the distribution of individual sample values. In several environmental problems of compliance monitoring, it is common to encounter the following two situations:

1. Compliance limits are specified in terms of upper percentiles of the statistical distribution of individual sample measurements. For considerations of cost, time, and effort, compliance is to be verified with composite sample measurements.
2. Compliance limits are specified in terms of estimated upper percentiles (such as 99th or 95th) of the statistical distribution of composite measurements. However, permit authorities may prefer to work with individual samples instead in order to monitor for permit compliance purposes (Kahn, 1991, personal communication).

In the former situation, the proposed sweep-out method does the job. The sweep-out method needs to be implemented on the composites formed for compliance monitoring until the desired percentage of upper extreme values of individual samples is identified. In the latter situation, the permit authority needs a guidance on how to make an adjustment in the compliance limits so that individual samples could be used for the purpose of monitoring compliance. Here, the sweep-out method may not be applicable if the permit authority cannot possibly measure individual samples. In such a situation, a separate statistical method of predicting extreme individual sample values may be necessary. Such a method may assume a statistical distribution and an autocorrelation model for the individual sample values. However, if the permit authority has access to the original individual samples, and if additional measurements on these individual samples are possible, then the sweep-out method of this chapter can be implemented to express compliance limits in terms of individual sample values rather than composite sample measurements.

3.5 Application

The sweep-out method is illustrated by applying it to data on polychlorinated biphenyl (PCB) concentration in surface soil samples at the Armagh compressor station along the gas pipeline of the Texas Eastern Gas Pipeline Company in Pennsylvania.

The Armagh compressor station is located in West Wheatfield Township, Indiana County, Pennsylvania. The site includes one compressor building along with several other buildings on 79 acres. There are two known liquid pits. The surrounding area contains 64 residences within 1 mile of the station. Some of these houses have private wells; however, a public water supply line was recently installed. None of the private wells has been contaminated as a result of the Texas Eastern operations. There is one wetland situated within one-half mile of the site. Richard Run, which flows to the south of the site, is classified as a cold water fishery. There are no public recreational facilities near the station (see Texas Eastern Gas Pipeline Company, 1989a, b).

Onsite soils are defined as being within the confines of the station site fencing and are accessible only by Texas Station personnel and authorized site visitors. The detection limit for PCB concentration in a surface soil sample is 1 part per million (ppm). The cleanup criteria for onsite soils are specified by an average overall PCB concentration of 5 ppm in soils between 0 and 6 in. depth. The objective of the onsite surface soil sampling is to characterize the presence of PCBs at the Armagh site in terms of the average PCB concentration in onsite surface soils. Potential sources of PCB have been identified, and a rectangular grid is laid out around each such source. Four different onsite grids are identified by the alphabetic codes A through D. Grid points are identified by a two-digit row number and an alphabetic column code. Sampling of the surface soil is done at selected grid points in two distinct phases. The second phase is undertaken to fill in locations not covered during phase I. Grid D is not sampled during phase I, but is sampled during phase II. Phase II locations are generally farther away from the potential PCB source, and the measured PCB concentrations tend to be lower during this phase. A total of 130 onsite surface soil samples are collected during phase I and 228 during phase II as follows:

	Phase I	**Phase II**
Grid A	78 samples	106 samples
Grid B	16 samples	16 samples
Grid C	36 samples	32 samples
Grid D		74 samples

The distance between consecutive rows as well as between consecutive columns is 25 ft. For computerization of the data and to facilitate statistical analysis of the same using statistical computer packages, the alphabetic column codes are converted into numeric codes, with A converted into 1, B into 2, and so on. Row and column codes of the sampling locations in grids B, C, and D are shifted so as to synchronize them with the codes of the grid A. This synchronization enables plotting of all the sampled grid points on the same graph, as in Fig. 3.1a.

All surface soil samples are collected from the 0 to 6 in. depth using a bucket auger, trowel, or scoop. Vegetation, rocks, and other debris that interfered with sample collection are removed. Soil samples for PCB analysis are placed in a stainless steel bucket/bowl and mixed with a trowel to obtain a homogeneous sample. All

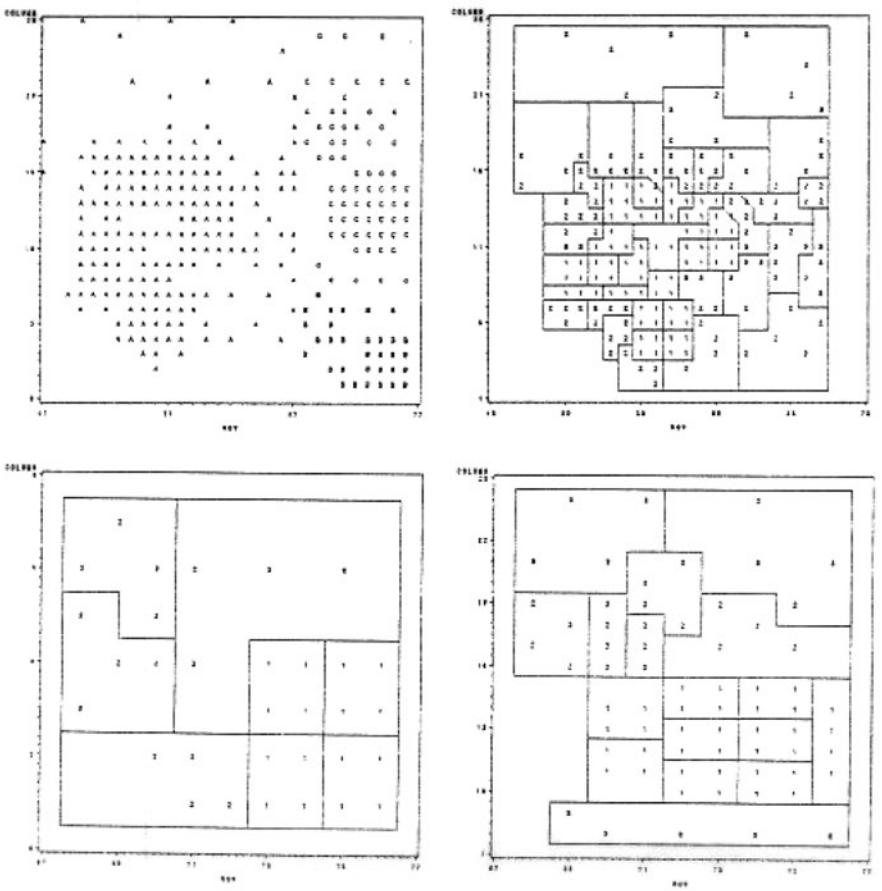

Fig. 3.1 **(a)** Sampling locations on grids A, B, and C. A: grid A, B: grid B, C: grid C; **(b)** sampling locations and compositing scheme for grid A, 1: phase I, 2: phase II; **(c)** sampling locations and compositing scheme for grid B; 1: phase I, 2: phase II; and **(d)** sampling locations and compositing scheme for grid C, 1: phase I, 2: phase II

rocks, twigs, etc., are removed from the sample. The sample is then placed into a jar using the stainless steel trowel.

Choice of composite sample size is dictated by two considerations. On the one hand, the relative savings in measurement costs increases with the composite sample size, and, therefore, the larger the composite sample size, the more cost-effective it is. On the other hand, compositing has a dilution effect. That is, if an individual sample having a large value is combined with other individual samples with relatively small values in a single composite, then the large individual sample value is diluted as the result of compositing. If the composite sample size is too large, then the dilution can lead to non-detection of a large individual sample value. If, for instance, the detection limit is d and if it is important to detect every individual

sample having a value of c or more (the criterion value), then the composite sample size should never exceed c/d so that possible dilution will not result in non-detection of any individual sample with a value of c or more. In any particular problem, it is therefore necessary to specify the detection limit and the criterion value so that the optimal composite sample size may be determined.

In view of the dilution problem, a criterion value for the PCB concentration of 5 ppm and a detection limit of 1 ppm imply that the composite sample size should not exceed 5. Moreover, it is usually the case that analytical variability is, in a relative sense, high at low concentrations, particularly in the region of detection levels. Analytical variability can lead to misclassification and thus can diminish the cost-effectiveness of composite sampling techniques. This additional consideration of analytical variability further limits the choice of the composite sample size. It was therefore decided to choose the composite sample size to be 4. Only in a few cases, where the spatial arrangement makes it impossible or impracticable to identify exactly four neighboring sample locations, composite samples of sizes 3 or 5 are formed.

Boswell and Patil (1990) have investigated strategies for composite sample formation when samples are positively spatially autocorrelated. The purpose of the analysis is to classify every individual sample as exceeding or not exceeding a specified criterion value. After comparing four different choices of compositing strategies for classification of individual samples, Boswell and Patil conclude that, when there is positive spatial dependence among the individual sampling locations, compositing of samples from neighboring points, as nearly in a square region as possible, increases the cost-efficiency of composite sampling. Due to positive spatial dependence, these samples are likely to exhibit greater homogeneity within themselves than randomly selected samples.

When samples of solids, such as soil, are composited, it is not easy to achieve homogeneous composites, and therefore the mean of the composite sample measurements can be more variable than the mean of the corresponding individual sample values. However, if care is taken to composite only homogeneous individual samples, then the variability of the mean of composite sample measurements resulting from imperfect mixing will be minimal. It is therefore recommended to form composites in such a way that the individual samples within every composite are more homogeneous than those in different composites.

In order to maximize the within-composite homogeneity, it is decided that individual samples collected from contiguous locations be composited. Considering the fact that the four grids represent different sources of PCBs and considering the temporal distance between the two sampling phases, it is also decided that all composites be formed within a sampling phase and also within a grid. This would also be attractive from the management and operational point of view. These considerations lead to the decision of compositing only individual samples collected from contiguous locations belonging to the same grid and sampled during the same sampling phase.

After the composite sample size and the composite sample formation strategy are determined, it is necessary to identify the sampling locations to be composited.

There is a considerable subjectivity involved at this stage since not all the grid points are included in the sampling plan, and the sampled grid is therefore not exactly rectangular. However, enough precaution is taken to avoid selection bias in the composite sample formation. First, even though measurements on PCB concentrations at sampled locations are available, formation of composite samples is based only on the geographical positions of sampling locations. Second, a few other choices for composite sample formation are implemented for comparison with that used for the analysis reported here. As the estimate of the mean PCB concentration does not depend on the particular choice of the compositing method, the estimate of the population variance is used as the criterion for this comparison. Since there is no significant difference between the various choices of composite sample formation, only one set of composites is used in this chapter. Sampling locations and formation of composites are shown in the schematic plots in Fig. 3.1. Table 3.1 shows the individual sample values and simulated composite sample measurements.

To illustrate the sweep-out method in the case of the Armagh site, we note that the highest PCB concentration in a composite sample (composite number 25 in Table 3.1) is 4897.5 ppm. Since the size of this composite is 4, the highest PCB concentration in an individual sample cannot exceed 19,590 ppm. Exhaustive testing of the constituent samples results in identifying the highest PCB concentration in an individual sample, which is 10,000. Note that there is a composite sample (composite number 5 in Table 3.1) with a PCB concentration of 3999.5 ppm and hence may contain an individual sample with PCB concentration exceeding 10,000 ppm. Upon measuring every individual sample in this composite, it is indeed found to be the case, as there is an individual sample with PCB concentration of 10,700 ppm. This implies that no composite sample with a PCB concentration of 2675 ppm or less can contain an individual sample with PCB concentration exceeding 10,700 ppm. Since there is no other composite sample with a measurement exceeding 2675 ppm, the sampling location with the largest PCB concentration has been identified. Note that this requires only 8 measurements in addition to the 90 composite sample measurements.

Figure 3.2 shows a scatterplot of individual sample values plotted against the simulated composite sample measurements. The two rays from the origin indicate the upper and the lower bounds on the largest individual sample value for every composite sample measurement. Thus, corresponding to the composite sample with a measurement of 4897.5 ppm, the upper bound for the maximum individual sample value in this composite is 19,590 ppm, while the lower bound is 4897.5 ppm, which is the same as the composite sample measurement. Since 4897.5 ppm is the largest composite sample measurement, individual samples in this composite are measured separately, and an individual sample with a PCB concentration of 10,000 ppm is identified. A horizontal line through the point identifying this individual sample indicates that there is only one composite (composite number 5 in Table 3.1) which can possibly contain an individual sample with a PCB concentration of more than 10,000 ppm. Making measurements on all the individual samples constituting this composite, one locates an individual sample with a PCB concentration of 10,700 ppm. There is no other composite that can contain an individual

Table 3.1 Individual sample values and simulated composite sample measurements

Composite sample	Phase	Individual sample value	Composite sample measurement	Composite sample	Phase	Individual sample value	Composite sample measurement
01	I	2.9, 3.1, 22, 22	12.5	46	II	1.9, 1.6, 82, 390	118.9
02	I	21, 298, 18, 1880	554.3	47	II	1.4, 1, 530, 320	213.1
03	I	9.4, 51, 319, 1.0	95.1	48	II	160, 180, 19, 320	169.8
04	I	105, 30, 22, 67	56.0	49	II	5.4, 1.7, 0.0, 15	5.5
05	I	18, 2320, 10700, 2960	3999.5	50	II	7.7, 6.9, 310, 19	85.9
06	I	38, 2.5, 13, 154	51.9	51	II	27, 23, 21, 5	19
07	I	1.1, 12, 55, 8.7	19.2	52	II	7.5, 2.2, 55, 80	36.2
08	I	13, 1.9, 2.9, 22	10.0	53	II	7.7, 4.3, 24, 250	71.5
09	I	129, 12, 44, 22	51.8	54	II	4.3, 6.4, 20, 33	15.9
10	I	1.6, 1070, 1.0, 64	284.2	55	II	436, 9.5, 120. 21, 58	128.9
11	I	13, 3.8, 3, 6.8	6.9	56	II	1.5, 160, 180, 1000	335.4
12	I	13, 3.8, 2.8, 6.9	6.1	57	II	2.9, 15, 150, 12, 11	38.2
13	I	34, 28, 745, 3850	1164.3	58	II	2.9, 26, 1.2, 1.3	7.9
14	I	50, 18, 17, 34	29.8	59	II	24, 2.6, 3.5, 18	12.0
15	I	4.6, 22, 1.0, 42	17.4	60	II	3.9, 27, 5.4, 12	12.1
16	I	14, 3.3, 1.5, 2.6	5.4	61	II	72, 38, 7.1, 35	38.0
17	I	2.4, 1390, 3, 672	516.9	62	II	52, 37, 66, 38	48.3
18	I	8.9, 661, 20, 18	177.0	63	II	1.3, 2.1, 15, 4.4	5.7
19	I	18, 24, 26	22.7	64	II	60, 79, 8.7, 150	74.4
20	I	3.5, 16, 20	13.2	65	II	16, 24, 18, 160	54.5
21	I	97, 70, 14, 150	82.8	66	II	150, 210, 18, 13	97.8
22	I	37, 72, 40, 33	45.5	67	II	26, 7.8, 43, 49	31.5
23	I	38, 44, 83, 30	48.8	68	II	46, 24, 18, 12	25
24	I	38, 100, 140, 47	81.3	69	II	38, 12, 140, 60	62.5
25	I	590, 7100, 10000, 1900	4897.5	70	II	26, 14, 190, 61,33	64.8
26	I	670, 940, 240, 290	535	71	II	340, 190, 10	180
27	I	74, 200, 120, 220	153.5	72	II	0.0, 0.0, 0.0, 0.0	0.0

Table 3.1 (continued)

Composite sample	Phase	Individual sample value	Composite sample measurement	Composite sample	Phase	Individual sample value	Composite sample measurement
28	I	280, 260, 10, 250	200	73	II	0.0, 0.0, 0.0, 1.1	0.3
29	I	44, 110, 660, 230	261	74	II	1.1, 2.8, 4.2, 6.6	3.7
30	I	580, 1100, 1300, 4900	1970	75	II	6.9, 16, 7, 13	10.8
31	I	110, 80, 210, 12	103	76	II	11, 13, 6.4, 8	9.6
32	I	75, 890, 170, 550	421.3	77	II	0, 236, 7.2, 2.4	61.4
33	I	2300, 420, 520, 1300	1135	78	II	5.8, 535, 1.1, 0.0	135.5
34	II	0.0, 1.2, 1.67	2.5	79	II	0.0, 1.4, 4.9, 0.0	1.6
35	II	5.7, 17, 4.3, 36	15.8	80	II	0.0, 0.0, 5.1, 6.3	2.9
36	II	28, 170, 10, 62	67.5	81	II	7.9, 14, 20, 31	18.2
37	II	300, 6.4, 53	119.8	82	II	52, 1, 500, 46	162.3
38	II	16, 18, 150, 27	52.8	83	II	16, 5, 36, 64	30.3
39	II	6.2, 7.1, 31, 38	20.6	84	II	40, 38, 68, 7.5	38.4
40	II	16, 66, 61, 340, 1500	396.6	85	II	40, 33, 36, 17	31.5
41	II	1.3, 3.5, 2.1, 8.8	3.9	86	II	35, 4, 170	52.3
42	II	7.5, 2.7, 1.6, 11	5.7	87	II	110, 200, 4.2	104.7
43	II	0.0, 0.0, 17, 2.8	5.0	88	II	7.4, 3.3, 21, 2.3	8.5
44	II	1.1, 5.9, 350, 17	93.5	89	II	3.8, 35, 20, 17	19.0
45	II	3.2, 5, 11, 5.1	6.1	90	II	23, 17, 3, 6.8	12.5

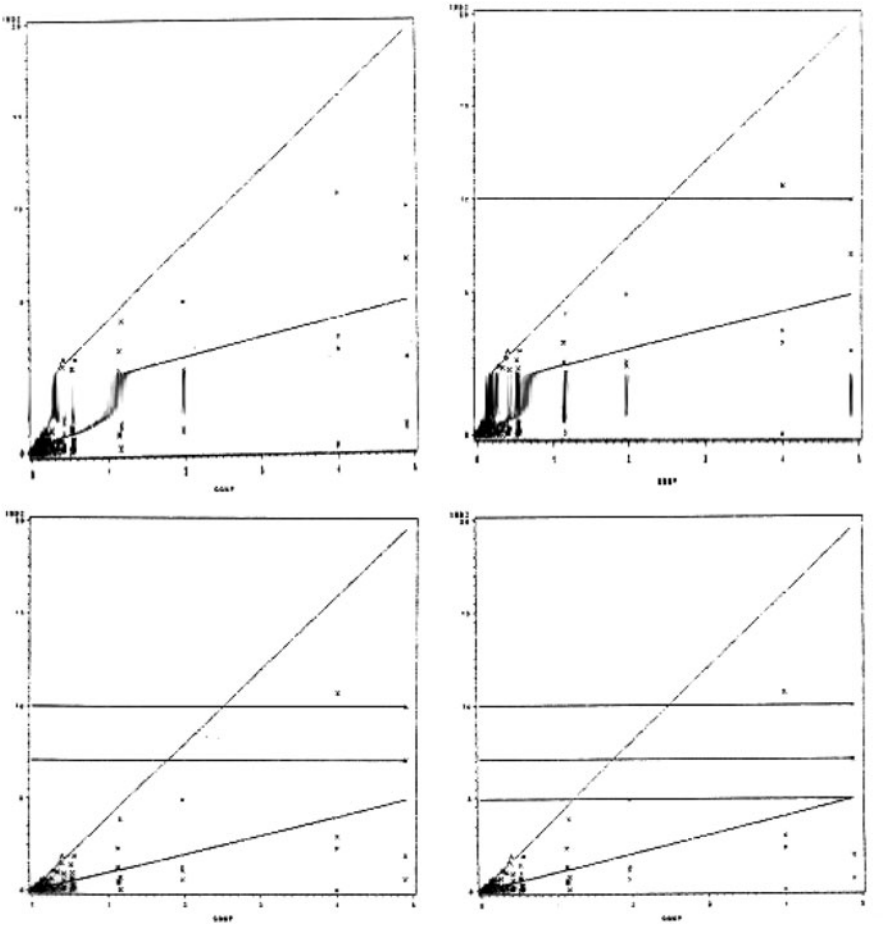

Fig. 3.2 Illustration of the sweep-out method. Individual sample values (*Y-axis*) vs. composite sample measurements (*X-axis*) in 1000 ppm. **(a)** The upper and lower bounds for the largest individual values. **(b)** Measurements on individual samples from only two composites identify two largest individual values. **(c)** Measurements on individual samples from only two composites identify the three largest individual values. **(d)** Measurements on individual samples from three composites identify the four largest individual values

sample with a PCB concentration exceeding 10,700 ppm, as is evident from Fig. 3.2b. Making measurements on the eight individual samples constituting two composites has thus identified the individual sample with the largest PCB concentration. This search can easily be extended to identify more individual samples with high PCB concentration. Figure 3.2c, d shows how additional measurements on 12 individual samples constituting only 3 composites help identify the individual samples with the 4 largest PCB concentrations. In other words, with only 12 measurements in addition to the 90 measurements on the simulated composite

samples, we are able to identify the 4 individual samples with the highest PCB concentrations.

In order to investigate the relationship between the number of extreme values identified and the number of composites retested, we extend the sweep-out method for the Armagh site to all the 90 composites. Figure 3.3 gives a graphical summary of these results. It is interesting to note the concavity of the curve in the graph. This implies that identification of every additional extreme value initially requires relatively more measurements. Another way of interpreting the graph is illustrated by the following statement: When about 20% largest individual sample values are identified, which in turn implies that individual samples from the 50 composites are measured, we already have information on 200 individual sample values. It is with this information that identification of further extreme values appears to require a relatively smaller number of measurements as the sweep-out method progresses.

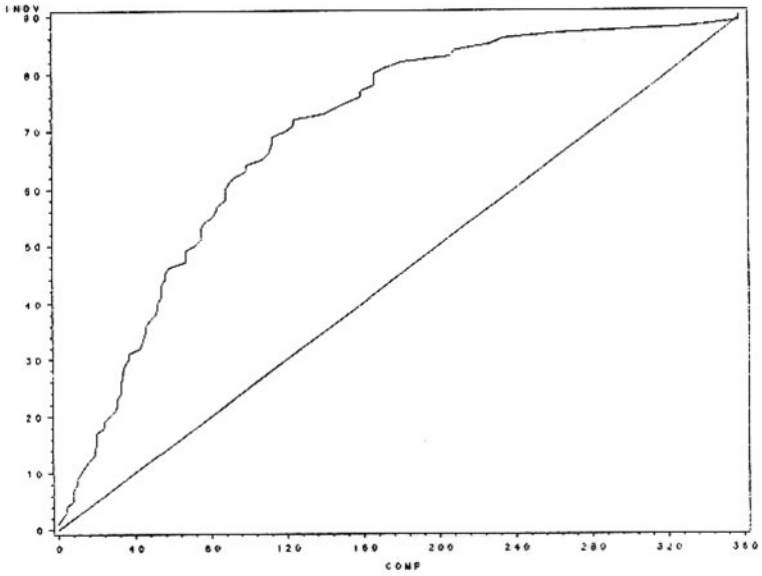

Fig. 3.3 Number of composites retested (*Y-axis*) vs. number of extreme values identified (*X-axis*). The *diagonal line* represents the optimal case in which exactly four extreme values are identified for every composite

3.6 Two-Way Composite Sampling Design

A two-way composite sampling design forms a rectangular or square array of sampling units. A row-composite is formed from individual sample units in a row. A column-composite is similarly formed from individual sample units in a column. Every individual sample unit thus contributes to a row-composite and a

column-composite. Measurements are made on all row-composites and all column-composites.

Let X_{ij} denote the value of the individual sample unit that contributes to the ith row-composite and jth column-composite, $i = 1, \ldots, r, j = 1, \ldots, c$. Let $Y_{i\cdot}$ denote the value of the ith row-composite and $Y_{\cdot j}$ denote the value of the jth column-composite, $i = 1, \ldots, r, j = 1, \ldots, c$.

Now rearrange the individual samples in the array so that the values of row-composites and column-composites are in descending order. Let $Y_{[i]\cdot}$ denote the value of the ith ordered row-composite, $i = 1, \ldots, r$, and let $Y_{\cdot[j]}$ denote the value of the jth ordered column-composite, $j = 1, \ldots, c$. Let $X_{[i][j]}, i = 1, \ldots, r, j = 1, \ldots, c$, denote the individual sample value corresponding to the ith ordered row-composite and jth ordered column-composite. Let $W_{[i]\cdot}$ denote the composite sample total for the ith ordered row-composite, $i = 1, \ldots, r$, and $W_{\cdot[j]}$ denote the composite sample total for the jth ordered column-composite, $j = 1, \ldots, c$. Table 3.2 illustrates the arrangement described so far.

Table 3.2 Arrangement for two-way sweep-out method

	Column					
Row	1	2	–	c	Value	Total
1	$X_{[1][1]}$	$X_{[1][2]}$	–	$X_{[1][c]}$	$Y_{[1]\cdot}$	$W_{[1]\cdot}$
2	$X_{[2][1]}$	$X_{[2][2]}$	–	$X_{[2][c]}$	$Y_{[2]\cdot}$	$W_{[2]\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	$X_{[r][1]}$	$X_{[r][2]}$	–	$X_{[r][c]}$	$Y_{[r]\cdot}$	$W_{[r]\cdot}$
Value	$Y_{\cdot[1]}$	$Y_{\cdot[2]}$	–	$Y_{\cdot[c]}$		
Total	$W_{\cdot[1]}$	$W_{\cdot[2]}$	–	$W_{\cdot[c]}$		

The two-way sweep-out method is described in the following algorithm.

1. Input values of $\{Y_{[i]\cdot}, i = 1, \ldots, r\}$ and $\{Y_{\cdot[j]}, j = 1, \ldots, c\}$.
2. Define $R = 0$, $K = 0$, $k_{[i]\cdot} = c, i = 1, \ldots, r$, and $k_{\cdot[j]} = r, j = 1, \ldots, c$. Compute values of $W_{[i]\cdot} = Y_{[i]\cdot} \times k_{[i]\cdot}, i = 1, \ldots, r$, and $W_{\cdot[j]} = Y_{\cdot[j]} \times k_{\cdot[j]}, j = 1, \ldots, c$.
3. Arrange the $r \times c$ individual samples so that row-composites and column-composites are in descending order.
4. Define $U_{\mathrm{R}} = \{\mathrm{g} : 1 \leq g \leq r, W_{[g]\cdot} > R\}$,
   $U_{\mathrm{C}} = \{\mathrm{h} : 1 \leq h \leq c, W_{\cdot[h]} > R\}$.
5. If U_{R} or U_{C} is empty, then stop the search and follow step 17.
6. Define $i = \min\{g : g \,\varepsilon\, U_{\mathrm{R}}\}$.
7. Define $j = \min\{h : h \,\varepsilon\, U_{\mathrm{C}}\}$.
8. If the individual sample representing the cell (i, j) has not been measured so far, follow step 11.
9. If $Y_{[i]\cdot} < Y_{\cdot[j]}$, include i in U_{R} and follow step 8.
10. If $Y_{[i]\cdot} \geq Y_{\cdot[j]}$, include j in U_{C} and follow step 8.

11. If $k_{[i]\cdot} > 1$ and $k_{\cdot[j]} > 1$, follow step 14.
12. If $k_{[i]\cdot} = 1$, define $X_{[i][j]} = Y_{[i]\cdot}$ and follow step 15.
13. If $k_{\cdot[j]} = 1$, define $X_{[i][j]} = Y_{\cdot[j]}$ and follow step 15.
14. Make a measurement on the individual sample representing the (i, j)th cell and denote it by $X_{[i][j]}$. Increment K by 1.
15. Update $R \leftarrow \max\{R, X_{[i][j]}\}$.
16. Update the ith row-composite:

$$\begin{aligned} W_{[i]\cdot} &\leftarrow W_{[i]\cdot} - X_{[i][j]}, \\ k_{[i]\cdot} &\leftarrow k_{[i]\cdot} - 1, \\ Y_{[i]\cdot} &\leftarrow W_{[i]\cdot}/k_{[i]\cdot}. \end{aligned}$$

Similarly, update the jth column-composite:

$$\begin{aligned} W_{\cdot[j]} &\leftarrow W_{\cdot[j]} - X_{\text{[i][j]}}, \\ k_{\cdot[j]} &\leftarrow k_{\cdot[j]} - 1, \\ Y_{\cdot[j]} &\leftarrow W_{\cdot[j]}/k_{\cdot[j]}. \end{aligned}$$

Write 0 in the (i, j)th cell and follow step 3.

17. The value of R is the largest individual sample value and K is the number of measurements made on individual sample units.

3.7 Illustrative Example

Application of the two-way composite sampling design for finding the largest individual sample value is illustrated with an artificial data set.

Consider a two-way composite sampling design for 16 individual samples arranged in a square of 4 rows and 4 columns. The individual sample values are as follows.

Arrange the 16 individual sample values in a square of 4 rows and 4 columns. Calculate row totals, column totals, row averages, and column averages.

	Column					
Row	1	2	3	4	Value	Total
1	46	58	160	41	76.25	305
2	177	45	43	49	78.50	314
3	62	109	68	60	74.75	299
4	77	167	56	37	84.25	337
Value	90.5	94.75	81.75	44.25		
Total	362	379	327	177		

Step 2. Define $R = 0$ and $K = 0$.

Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.

	Column					
Row	1	2	3	4	Value	Total
1	167	77	56	37	84.25	337
2	45	177	43	49	78.50	314
3	58	46	160	41	76.25	305
4	109	62	68	60	74.75	299
Value	94.75	90.5	81.75	44.25		
Total	379	362	327	177		

Step 4. $U_{\mathrm{R}} = \{1, 2, 3, 4\}$,
$U_{\mathrm{C}} = \{1, 2, 3, 4\}$.

Step 5. Both U_{R} and U_{C} are non-empty and hence proceed to step 6.

Step 6. Define $i = 1$.

Step 7. Define $j = 1$.

Step 8. Since the individual sample representing the cell (1,1) has not been measured so far, proceed to step 11.

Step 11. Since $k_{[1]\cdot} > 1$ and $k_{\cdot[1]} > 1$, proceed to step 14.

Step 14. Make a measurement on the individual sample representing the cell (1,1) and obtain the value 167. Increment K by 1 so that $K = 1$.

Step 15. Update $R = 167$.

Step 16. Update the first row-composite: $W[1]\cdot = 170, k[1]\cdot = 3, Y[1]\cdot = 56.67$.

Similarly, update the first column-composite: $W\cdot[1] = 212, k\cdot[1] = 3, Y\cdot[1] = 70.67$.

Write 0 in the cell (1,1) to obtain the following table:

	Column					
Row	1	2	3	4	Value	Total
1	0	77	56	37	56.67	170
2	45	177	43	49	78.50	314
3	58	46	160	41	76.25	305
4	109	62	68	60	74.75	299
Value	70.67	90.5	81.75	44.25		
Total	212	362	327	177		

Proceed to step 3.

Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.

Step 4. $U_{\mathrm{R}} = \{1, 2, 3, 4\}$,
$U_{\mathrm{C}} = \{1, 2, 3, 4\}$.

	Column					
Row	1	2	3	4	Value	Total
1	177	43	45	49	78.50	314
2	46	160	58	41	76.25	305
3	62	68	109	60	74.75	299
4	77	56	0	37	56.67	170
Value	90.5	81.75	70.67	44.25		
Total	362	327	212	177		

Step 5. Both U_R and U_C are non-empty and hence proceed to step 5.
Step 6. Define $i = 1$.
Step 7. Define $j = 1$.
Step 8. Since the individual sample representing the cell (1,1) has not been measured so far, proceed to step 11.
Step 11. Since $k_{[1]\cdot} > 1$ and $k_{\cdot[1]} > 1$, proceed to step 14.
Step 14. Make a measurement on the individual sample representing the cell (1,1) and obtain the value 177. Increment K by 1 so that $K = 2$.
Step 15. Update $R = 177$.
Step 16. Update the first row-composite: $W[1]\cdot = 137, k[1]\cdot = 3, Y[1]\cdot = 45.67$.
Similarly, update the first column-composite: $W\cdot[1] = 185, k\cdot[1] = 3, Y\cdot[1] = 61.67$.
Write 0 in the cell (1,1) to obtain the following table:

	Column					
Row	1	2	3	4	Value	Total
1	0	43	45	49	45.67	137
2	46	160	58	41	76.25	305
3	62	68	109	60	74.75	299
4	77	56	0	37	56.67	170
Value	61.67	81.75	70.67	44.25		
Total	185	327	212	177		

Proceed to step 3.
Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.
Step 4. $U_R = \{1, 2\}$,
$U_C = \{1, 2, 3\}$.
Step 5. Both U_R and U_C are non-empty and hence proceed to step 5.
Step 6. Define $i = 1$.
Step 7. Define $j = 1$.

	Column					
Row	1	2	3	4	Value	Total
1	160	58	46	41	76.25	305
2	68	109	62	60	74.75	299
3	56	0	77	37	56.67	170
4	43	45	0	49	45.67	137
Value	81.75	70.67	61.67	44.25		
Total	327	212	185	177		

Step 8. Since the individual sample representing the cell (1,1) has not been measured so far, proceed to step 11.

Step 11. Since $k_{[1]\cdot} > 1$ and $k_{\cdot[1]} > 1$, proceed to step 14.

Step 14. Make a measurement on the individual sample representing the cell (1,1) and obtain the value 160. Increment K by 1 so that $K = 3$.

Step 15. R is unchanged at 177.

Step 16. Update the first row-composite: $W[1]\cdot = 145, k[1]\cdot = 3, Y[1]\cdot = 48.33$.

Similarly, update the first column-composite: $W\cdot[1] = 167, k\cdot[1] = 3, Y\cdot[1] = 55.67$.

Write 0 in the cell (1,1) to obtain the following table:

	Column					
Row	1	2	3	4	Value	Total
1	0	58	46	41	48.33	145
2	68	109	62	60	74.75	299
3	56	0	77	37	56.67	170
4	43	45	0	49	45.67	137
Value	55.67	70.67	61.67	44.25		
Total	167	212	185	177		

Proceed to step 3.

Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.

	Column					
Row	1	2	3	4	Value	Total
1	109	62	68	60	74.75	299
2	0	77	56	37	56.67	170
3	58	46	0	41	48.33	145
4	45	0	43	49	45.67	137
Value	70.67	61.67	55.67	44.25		
Total	212	185	167	177		

Step 4. $U_R = \{1\}$,
$U_C = \{1, 2\}$.
Step 5. Both U_R and U_C are non-empty and hence proceed to step 5.
Step 6. Define $i = 1$.
Step 7. Define $j = 1$.
Step 8. Since the individual sample representing the cell (1,1) has not been measured so far, proceed to step 11.
Step 10. Since $k_{[1]\cdot} > 1$ and $k_{\cdot[1]} > 1$, proceed to step 14.
Step 14. Make a measurement on the individual sample representing the cell (1,1) and obtain the value 109. Increment K by 1 so that $K = 4$.
Step 15. R is unchanged at 177.
Step 16. Update the first row-composite: $W[1]\cdot = 190, k[1]\cdot = 3, Y[1]\cdot = 63.33$.
Similarly, update the first column-composite: $W\cdot[1] = 103, k\cdot[1] = 2, Y\cdot[1] = 51.50$.
Write 0 in the cell (1,1) to obtain the following table:

	Column					
Row	1	2	3	4	Value	Total
1	0	62	68	60	63.33	190
2	0	77	56	37	56.67	170
3	58	46	0	41	48.33	145
4	45	0	43	49	45.67	137
Value	51.5	61.67	55.67	44.25		
Total	103	185	167	177		

Proceed to step 3.
Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.

	Column					
Row	1	2	3	4	Value	Total
1	62	68	0	60	63.33	190
2	77	56	0	37	56.67	170
3	46	0	58	41	48.33	145
4	0	43	45	49	45.67	137
Value	61.67	55.67	51.5	44.25		
Total	185	167	103	177		

Step 4. $U_R = \{1\}$,
$U_C = \{1\}$.
Step 5. Both U_R and U_C are non-empty and hence proceed to step 5.
Step 6. Define $i = 1$.
Step 7. Define $j = 1$.

Step 8. Since the individual sample representing the cell (1,1) has not been measured so far, proceed to step 11.

Step 11. Since $k_{[1]\cdot} > 1$ and $k_{\cdot[1]} > 1$, proceed to step 14.

Step 14. Make a measurement on the individual sample representing the cell (1,1) and obtain the value 62. Increment K by 1 so that $K = 5$.

Step 15. R is unchanged at 177.

Step 16. Update the first row-composite: $W[1]\cdot = 128$, $k[1]\cdot = 2$, $Y[1]\cdot = 64$.

Similarly, update the first column-composite: $W\cdot[1] = 123$, $k\cdot[1] = 2$, $Y\cdot[1] = 61.50$.

Write 0 in the cell (1,1) to obtain the following table:

	Column					
Row	1	2	3	4	Value	Total
1	0	68	0	60	64.00	128
2	77	56	0	37	56.67	170
3	46	0	58	41	48.33	145
4	0	43	45	49	45.67	137
Value	61.50	55.67	51.5	44.25		
Total	123	167	103	177		

Proceed to step 3.

Step 3. Arrange the individual samples to obtain the following table. The row averages and column averages in this table are in descending order.

	Column					
Row	1	2	3	4	Value	Total
1	0	68	0	60	64.00	128
2	77	56	0	37	56.67	170
3	46	0	58	41	48.33	145
4	0	43	45	49	45.67	137
Value	61.50	55.67	51.5	44.25		
Total	123	167	103	177		

Step 4. $U_{\mathrm{R}} = \{\}$,
$U_{\mathrm{C}} = \{\}$.

Step 5. Since U_{R} and U_{C} are both empty, stop the search and declare $R = 177$ as the largest individual sample value.

Note that $K = 5$ is the number of measurements made on individual samples in addition to the eight measurements made on composite samples (four row-composites and four column-composites). In this way, the largest individual sample value is obtained by making a total of 13 measurements instead of 16 as would be required by the conventional method of making measurement for every individual sample.

3.8 Analysis of Composite Sampling Data Using the Principle of Maximum Entropy

J. H. Carson, Jr. (2001), has proposed a new tool for hot spot detection with no (or minimal) retesting based on the principle of maximum entropy. The methodology is easy to implement and can accommodate multiple criteria for evaluating site remediation. Very simple decision rules are provided by the new methodology that complements use of composite sampling for controlling residual mean concentrations.

3.8.1 Introduction

A risk-based cleanup standard C_s for a COC represents a bound for average residual concentration of contaminant. This bound corresponds to the bound on the probability of an adverse effect. After a removal activity, the mean will be tested to verify if the average residual concentration is less than C_s. Since a population having the mean less than C_s has values greater as well as smaller than C_s, compliance in terms of the mean does not imply compliance in terms of individual values. It is interesting to note that most of the regulators insist that every individual value be less than C_s.

The heavy right tail of the distribution of environmental contaminant prompts an auxiliary criterion for controlling extreme concentrations. For instance, the following can make a valid set of criteria for remediation:

- A separate standard for an upper percentile.
- A stringent standard for every individual sample value (possibly a small multiple of C_s).
- A limit on the proportion of sample values that exceed C_s.
- A small probability of an undetected hot spot.

Here a "sample" is a physical sample of material as opposed to an ensemble of observations. Similarly, the processes of collecting samples and testing them must be distinguished. A "hot spot" is defined as a small area where concentration of contaminant exceeds the "hot spot threshold" with high probability.

3.8.1.1 Hot Spot Detection Based on Composite Sample Values

Composite sampling involves physical blending of individual samples or their chemical extracts for measuring some physical or chemical properties. Compositing results in averaging of characteristics and hence composite samples carry more information about means than do individual samples. Composite sampling therefore provides better estimates and more powerful tests concerning the mean. Even though compositing reduces the variance of the mean, it does not erase the information about extreme values completely. Many regulators think that compositing is used to dilute or mask hot spots. Although compositing has the potential of achieving these

goals, it is possible to judiciously use compositing and detect hot spots when composite sample values are available. Due to the observational economy of composite sampling, the sample coverage can be improved so as to increase the probability of hitting the hot spot. It is then necessary to develop decision rules to indicate when a component sample in a composite represents a hot spot.

The threshold C_s for individual sample value is not appropriate as hot spot threshold. Carson (2001) proposes $3C_s$ as the hot spot threshold for the following reasons:

- A factor of 3 is within the minimum uncertainty factor used in noncarcinogenic risk estimates (USEPA, 1986, pp. 6–15).
- Carcinogenic risk estimates are more conservative.
- The undetected hot spots should allow minimal exposure probability.

When composite sampling is implemented, a decision rule must be developed to indicate that a component sample represents a hot spot. This requires a decision rule similar to the following:

- Define a hot spot threshold (H).
- Define a "clean" rule R_c for deciding that a composite does not include a sample from a hot spot.
- Define a "hot" rule R_h for deciding that a composite includes a sample from a hot spot.

For composites indicating the presence of a sample from a hot spot, a possible action is retesting or remediation. Both of these actions being expensive, Carson (2001) suggests a new method that estimates the probability that none of the component samples is from a hot spot without requiring any retesting.

3.8.2 Modeling Composite Sampling Using the Principle of Maximum Entropy

It is possible to estimate the probability of at least one of the component sample values exceeding the hot spot threshold using a simple probability model for unobserved values of component samples in composites. The model assumes the following:

- Component samples contribute equal material to a composite.
- Every composite sample is mixed thoroughly.
- Composite sample value is an unbiased estimator of the mean of component sample values.
- All combinations of component sample values that result in the same composite sample value are equally likely.

The assumption of "equally likely" is same as putting the maximum entropy probability distribution on the simplex.

3.8.2.1 What Is the Simplex?

For a positive integer k, let $\mathbf{U} = (U_1, \ldots, U_k, U_{k+1})$ be a $(k+1)$-dimensional vector. The k-simplex is the set

$$S_k = \{\mathbf{U} | U_1 + U_2 + \cdots + U_k + U_{k+1} = 1 \text{ and } U_i \geq 0, i = 1, \ldots, k+1\}.$$

Individual values of component samples in a composite of size $k+1$ can be modeled as random variables distributed on the k-simplex.

3.8.2.2 Maximum Entropy Applied to the k-Simplex

Note that S_k is a k-dimensional set embedded in a $(k+1)$-dimensional space. This is due to the fact that the sum of the $k+1$ coordinates is constrained to be 1. Under the assumptions that every component sample contributes an equal quantity to the composite and that the composite has been mixed well before making the measurement, the $k+1$ individual sample values have a uniform distribution on S_k.

Let $C_D^{k+1} = (\times)_{i=1}^{k+1}[0, D]$ denote the $(k+1)$-dimensional cube with opposite vertices at the origin and the point $(D, \ldots, D)$. The symbol $(\times)$ denotes the Cartesian product of sets. The intersection of S_k and C_D^{k+1} is the set points in the simplex that have all coordinates less than or equal to D. The probability of this event is the ratio of the area of this intersection to the area of the k-simplex S_k. Now, consider the probability that all component sample values of a composite of size k are less than A when composite sample value is C. This is calculated as indicated above with $D = A/kC$.

For higher dimensions ($k > 3$), this probability can be calculated directly, but the calculation becomes too complicated. In this case, Monte Carlo simulation is an easy and effective solution.

3.8.3 When Is the Maximum Entropy Model Reasonable in Practice?

Consider a situation in which the analytical results of composites show a flat variogram, no obvious spatial pattern, and a right-skewed distribution. Approximate statistical independence of composite (and of component) samples would be a reasonable assumption due to flat variogram, lack of pattern or trend, and independence of measurement errors. Gamma distribution with shape parameter α and scale parameter β is a candidate distribution for the distribution of composite sample data due to the right skew in the distribution. Even the exponential distribution is a candidate distribution if the right tail is not very thick.

Further, it is a common practice to form composite samples from component samples that are mutually related, either by spatial proximity or by some other common feature. In such a case, component samples within a composite are not independent. Such complications make it impossible to draw accurate quantitative

inferences about variography and skewness of component samples based on composite sample data. Nonetheless, quantitative inference about the variogram and skewness of the distribution of component samples is possible under assumptions like that of spatial stationarity, nonoverlapping composite formation, and so on. The principle of maximum entropy seems to be reasonable in many situations of site characterization.

Chapter 4
Estimating Prevalence of a Trait

4.1 Introduction

When presence/absence measurements are made on individual sampling units, selected from a population, it is possible to either classify every sampling unit as testing positive/negative or compute the proportion of sampling units that test positive. While the former is desired when the interest is in identifying those sampling units that possess a certain trait (and hence test positive), the latter is useful when the interest is not in the status of individual sampling units, but only in characterizing the population by the proportion of sampling units that possess the trait. The problem, then, is to estimate the proportion of sampling units in the population that possess the trait, using presence/absence measurements. Here, the values on the individual sampling units in the population are assumed to be independent and identically distributed random variables. The usual procedure (without forming composite samples) is to select and measure m individual sampling units and to estimate the prevalence p by the sample proportion $\bar{X}_m$ of individual sampling units that possess the trait. This provides an unbiased estimator of p, the population proportion of sampling units that possess the trait, since it is easy to note that

$$E[\bar{X}_m] = p \text{ and } \text{Var}[\bar{X}_m] = \frac{p(1-p)}{m}.$$

Instead of selecting and measuring m individual sampling units, it is possible to consider the following composite sampling procedure. Select $nk > m$ individual samples, form and measure $n < m$ composite samples of size k each. When the cost of making measurement (testing) is large compared to that of collecting samples, the procedure that uses composite samples may have a much smaller cost. Further, an estimator based on the composite sample measurements may even have a smaller mean square error.

The practice of estimating p with measurements on composite samples has been widely used in estimating the proportion of infective vectors of disease-causing agents among biological populations. An early use of composite sampling for estimating the prevalence of a trait is reported by scientists studying virus transmission rates of insects causing plant diseases (see Watson, 1936, for example).

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_4, © Springer Science+Business Media, LLC 2011

A number of reported studies have discussed statistical aspects of composite sampling for estimating the prevalence (see, for instance, Gibbs and Gower, 1960; Chiang and Reeves, 1962; Thompson, 1962; Kerr, 1971; Griffiths, 1972; Sobel and Elashoff, 1975; Loyer, 1983; Swallow, 1985, 1987; Burrows, 1987).

In the composite sampling procedure, a total of n composite samples, each consisting of k individual sampling units, are tested for the occurrence of a trait. The trait is such that a composite sample is judged to test positive for the trait if at least one of its constituent individual sampling units possesses the trait. For example, a plant is judged as diseased after one or more of k vectors feeding on the plant transmit the disease-causing agent. The identity of the individual sampling units possessing the trait is not of interest in this case. If the probability of occurrence of the trait in a single individual sampling unit is p and if individual sampling units are independent, then the probability that a composite sample possesses the trait is $p_y = 1 - (1 - p)^k$.

Composite sampling for making a comparison between disease transmission rates has been used to compare resistance of genetic sources (see Swallow, 1985). Garner et al. (1988) suggest that composite sampling can also protect the confidentiality of humans when, for example, a subject's knowledge of the presence of a disease might dissuade him/her from volunteering for individual testing for the occurrence of the disease. Thompson (1960) and Swallow (1985) draw parallels between estimating p by composite sampling and estimating densities of bacteria on dilution plates (Fisher, 1921) or estimating the densities of plants in a quadrat (Bartlett, 1935) by their presence or absence. The solution to the density estimation problem is often modeled by the Poisson distribution, and the Poisson and binomial distributions are very similar when p is small and n is large. In the problem of estimating the prevalence of a trait with composite sample measurements, most of the reported studies assume a binomial distribution for the number of individual samples that possess the trait.

4.2 The Maximum Likelihood Estimator

Let $y_1, y_2, \ldots, y_n$ represent the measurements on n composite samples of size k each. Each of these measurements is 1 if the trait is present in the composite sample, and it is 0 otherwise. The $\{y_i\}$ will then be observations on n independent Bernoulli random variables with parameter $p_y = 1 - q^k$ where $q = 1 - p$. The likelihood function of q is given by $L(q) = \left(1 - q^k\right)^{n\bar{y}_n} + \left(q^k\right)^{n\left(1-q^k\right)}$. Therefore, the log-likelihood function of q is $\ell(q) = \ln(L(q)) = n\bar{y}_n \ln\left(1 - q^k\right) + n\left(1 - \bar{y}_n\right) \ln(q^k)$. The maximum likelihood estimator of p is

$$\hat{p} = 1 - (1 - \bar{y}_n)^{1/k}. \tag{4.1}$$

Although $1 - \bar{y}_n$, the proportion of composite samples that test negative, is an unbiased estimator of $(1 - p)^k$, a bias is introduced in $\hat{p}$ by taking the kth root of

$1 - \bar{y}_n$. In fact, $\hat{p}$ is positively biased whenever $k > 1$ as is shown below using properties of the binomial distribution and Jensen's inequality for the expectation of a convex function of a random variable:

$$\begin{aligned} E[\hat{p}] &= 1 - E\left[(1 - \bar{y}_n)^{1/k}\right] \geq 1 - (1 - E[\bar{y}_n])^{1/k} \\ &= 1 - \left(1 - \left[1 - (1 - p)^k\right]\right)^{1/k} = p. \end{aligned}$$

The mean and the variance of $\hat{p}$ are, respectively, given by

$$E(\hat{p}) = 1 - \sum_{i=0}^{n} \left(\frac{i}{n}\right)^{1/k} \binom{n}{i} \left[(1 - p)^k\right]^i \left[1 - (1 - p)^k\right]^{n-i}, \tag{4.2}$$

$$\mathrm{Var}[\hat{p}] = \sum_{i=0}^{n} \left(\frac{i}{n}\right)^{2/k} \binom{n}{i} \left[(1 - p)^k\right]^i \left[1 - (1 - p)^k\right]^{n-i} - \left[1 - E(\hat{p})\right]^2. \tag{4.3}$$

The maximum likelihood estimator $\hat{p}$ is expanded into a Taylor series around the true parameter p so as to eliminate the leading term in the bias. This gives the following expansion for the maximum likelihood estimator $\hat{p}$:

$$\hat{p} = p + (\hat{p} - p)\frac{d\hat{p}}{dp} + \frac{1}{2!}(\hat{p} - p)^2 \frac{d^2\hat{p}}{dp^2} + \cdots.$$

Taking term by term expectation of the Taylor series expansions of $\hat{p}$ and $(\hat{p} - p)^2$, we obtain

$$E[\hat{p}] = p + p_y(1 - p)\left(\frac{k - 1}{2k^2}\right)\left[\frac{1}{n(1 - p_y)} + \frac{(1 - 2p_y)(2k - 1)}{3kn^2(1 - p_y)^2} + 0(p_y^{-3})\right] \tag{4.4}$$

and

$$\begin{aligned} \mathrm{MSE}[\hat{p}] = \frac{p_y(1 - p)^2}{k^2}\Bigg[&\frac{1}{n(1 - p_y)} + \frac{1}{n^2(1 - p_y)^2}\left(3\left(\frac{k - 1}{k}\right)^2 p_y\right. \\ &\left. + 2\left(\frac{k - 1}{k}\right)(1 - p_y)\right) + 0(p_y^{-3})\Bigg]. \end{aligned} \tag{4.5}$$

The maximum likelihood estimator $\hat{p}$ is consistent and has the asymptotic variance

$$\text{a Var}[\hat{p}] = \frac{1 - (1 - p)^k}{nk^2(1 - p)^{k-2}}. \tag{4.6}$$

Several researchers have investigated the relationship between the bias in $\hat{p}$ and p, k, and n (see, for instance, Gibbs and Gower, 1960; Kerr, 1971; Griffiths, 1972;

Loyer, 1983; Swallow, 1985, 1987; Boswell and Patil, 1987). Optimal composite sample sizes have also been determined for specified values of k (Loyer, 1983; Swallow, 1985), for specified values of $m = k \cdot n$ (Swallow, 1985), and for a range of the cost of measuring composite samples relative to that of measuring individual samples (Swallow, 1987).

Burrows (1987) proposed an alternative estimator which is a better estimator of p in that it has smaller bias and mean square error than the maximum likelihood estimator. Further it is just as easy to calculate as the maximum likelihood estimator. The maximum likelihood estimator may not therefore be recommended at all. See Section 4.4 for a comparison between the two estimators.

4.3 Alternative Estimators

The maximum likelihood estimator $\hat{p}$ given above is biased because it is derived from an unbiased estimator of a nonlinear function, $(1-p)^k$, of p. One approach to the estimation of nonlinear functions, according to Burrows (1987), is found in Haldane (1955) and Anscombe (1956). Burrows uses an estimator of the form $1 - \left[\frac{n\bar{Y}_n + a}{n+b}\right]^{1/k}$, where a and b are arbitrary constants chosen so as to eliminate the dominant term in the Taylor series expansion of the bias. The resulting estimator is surprisingly simple, namely,

$$\tilde{p} = 1 - \left[1 - \alpha \bar{y}_n\right]^{1/k}, \quad \text{where } \alpha = 2kn/(2kn + k - 1).$$

Note that $\tilde{p}$ has a smaller bias and a smaller MSE than those of $\hat{p}$ for the values of $k = 2, 50$. The bias in $\tilde{p}$ is obtained from that in $\hat{p}$ after removing the first term in the Taylor series expansion, namely, $\frac{1}{n(1-p_y)}$. Thus, $\tilde{p}$ has a uniformly smaller bias than $\hat{p}$. Similarly, the MSE of $\tilde{p}$ is uniformly smaller than that of $\hat{p}$. The optimal composite sample size corresponding to a specified value of $\tilde{p}$ is different from the optimal composite sample size corresponding to the same value of $\hat{p}$. For combinations of p, k, and n examined by Burrows, the bias of $\tilde{p}$ attained a maximum of 5.2% of the bias of $\hat{p}$, though higher optimal n were required to minimize the MSE for the estimator $\tilde{p}$ than were required for the estimator $\hat{p}$. Burrows gives a table listing the optimal composite sample sizes for both the estimators and the ratio $\frac{\text{MSE}[\tilde{p}]}{\text{MSE}[\hat{p}]}$ (in percentage form) for various values of n and p using the appropriate optimal composite sample sizes.

A simple moment-type estimator of p is proposed by Gastwirth and Hammick (1989). Under the assumptions made above, this estimator reduces to

$$\bar{p} = \frac{1}{nk} \sum_{j=1}^{n} Y_j,$$

with

$$\text{bias}\left[\bar{p}\right] = \frac{1 - kp - (1-p)^k}{k} \tag{4.7}$$

and

$$\text{MSE}\left[\bar{p}\right] = \frac{1}{k^2 n} \times$$
$$\left[(n-1)(1-p)^{2k} + (2n - 2knp + 1)(1-p)^k - k^2 np\left(\frac{2}{k} - p\right) + n\right]. \tag{4.8}$$

Gastwirth and Hammick observe that the bias and the MSE of this simple estimator are satisfactory only for small values of $p(\leq 0.02)$ and small composite sample sizes ($k \leq 10$).

4.4 Comparison Between $\hat{p}$ and $\tilde{p}$

Since the individual sample values are independent Bernoulli random variables with a common parameter p, the average of n composite sample measurements, $n\bar{Y}_n$, has a binomial distribution with parameters n and $\left(1 - q^k\right)$. This distribution, for given values of m, k, and p, can be evaluated on a computer to obtain the means, the variances, and the MSEs of both $\hat{p}$ and $\tilde{p}$. For example,

$$E[\tilde{p}] = \sum_{i=0}^{n}\left[1 - \left(1 - \frac{ri}{n}\right)^{\frac{1}{k}}\right] b\left(i; m, \left(1 - q^k\right)\right). \tag{4.9}$$

Table 4.1 gives the result of some calculations carried out in order to illustrate how the two estimators compare in their biases as well as in their mean squared errors (MSEs).

4.5 Estimation of Prevalence in Presence of Measurement Error

If measurement errors are likely to occur, then the estimators of prevalence need to be adjusted accordingly. In particular, if there is a probability α of a false-positive test result and a probability β of a false-negative test result, then the probability of a positive test response, p_+, is given by

$$p_+ = (1-p)\alpha + p(1-\beta).$$

The prevalence p can then be expressed in terms of p_+, α, and β as follows:

$$p = \frac{p_+ - \alpha}{1 - \alpha - \beta},$$

Table 4.1 Bias and mean squared error of the maximum likelihood estimate $\hat{p}$ and the alternative estimate $\tilde{p}$ for selected composite sample sizes (k) when the individual prevalence rate is $p = 0.1$ and the sample size is $n = 25$

k	Bias($\hat{p}$)	MSE($\hat{p}$)	Bias($\tilde{p}$)	MSE($\tilde{p}$)
7	0.002555	0.000797	0.000031	0.000735
8	0.002795	0.000755	0.000036	0.000686
9	0.003053	0.000732[a]	0.000042	0.000652
10	0.003340	0.000733	0.000050	0.000628
11	0.003684	0.000777	0.000760	0.000614
12	0.004137	0.000915	0.000773	0.000605
13	0.004801	0.001248	0.000788	0.000603[b]
14	0.005847	0.001947	0.000105	0.000605
15	0.007527	0.003266	0.000123	0.000611
20	0.038920	0.032273	0.000071	0.000654
21	0.052627	0.045217	−0.000021	0.000656[c]
22	0.069465	0.061118	−0.000162	0.000653
35	0.469203	0.429973	−0.008624	0.000374
36	0.500732	0.458432	−0.009731	0.000365
37	0.530902	0.485585	−0.010875	0.000362[d]
38	0.559597	0.511337	−0.012049	0.000365
39	0.586741	0.535630	−0.013247	0.000370

[a] Minimum MSE($\hat{p}$) at $k = 9$
[b] Local minimum MSE($\hat{p}$) at $k = 13$
[c] Local maximum MSE($\hat{p}$) at $k = 21$
[d] Local minimum MSE($\hat{p}$) at $k = 37$

and the maximum likelihood estimator of p is accordingly adjusted and is written as

$$\hat{p} = 1 - \left(1 - p^*\right)^{1/k},$$

where

$$p^* = \frac{\bar{y} - \alpha}{1 - \alpha - \beta}.$$

(Note that $\bar{y}$ is the maximum likelihood estimator of p_+.) For more details, see Garner et al. (1990).

Chapter 5
A Bayesian Approach to the Classification Problem

5.1 Introduction

When the measurements are present/absent, the optimal composite sample size can be determined for any specified retesting procedure if the prevalence p of individual samples possessing the trait is known. Chapter 2 contains the discussion and derivations of the necessary results for this purpose. Clearly, then, the optimal composite sample size depends on this prevalence. However, since the prevalence is usually not known prior to sampling, it is not always possible to accurately determine the optimal composite sample size.

Suppose p is the prevalence of polluted samples. Assuming that the pollution affects every sample independently, p is the probability that any individual sample is polluted. Let k denote the composite sample size. Determination of the optimal composite sample size without knowing p can lead to one of the two types of errors: a strategic error and a tactical error. *A strategic error* is the error of adapting a wrong strategy; that is, either we employ composite sampling when it would result in a relatively higher cost than the cost of exhaustive testing of individual samples or we do not employ composite sampling when it would result in a lower relative cost than that of exhaustive testing. *A tactical error* is the error of choosing a nonoptimal composite sample size which may result in a relative cost smaller than unity but still higher than what could have been achieved with the optimal composite sample size.

For example, consider a situation with the true prevalence rate of $p = 0.5$. In this case exhaustive testing has a smaller relative cost compared to that of composite sampling with any composite sample size $k \geq 3$. The graph in Fig. 5.1a shows the relative cost that results from a strategic error of using composite samples with values of k from 3 to 50 when the prevalence rate is $p = 0.5$. Next, consider a situation with the true prevalence rate of $p = 0.05$. Suppose a strategic error of not employing composite sample techniques is made, resulting in a relative cost of unity. Had we used exhaustive retesting with any composite sample size k from 3 to 50, we would have achieved a relative cost lower than unity, as shown in the graph in Fig. 5.1b. Finally, consider a situation with the prevalence rate of $p = 0.01$. The optimal composite sample size in this case is $k = 11$, with a relative cost of 0.1956. Incorrect prediction of p, resulting in a tactical error of using a nonoptimal value

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_5, © Springer Science+Business Media, LLC 2011

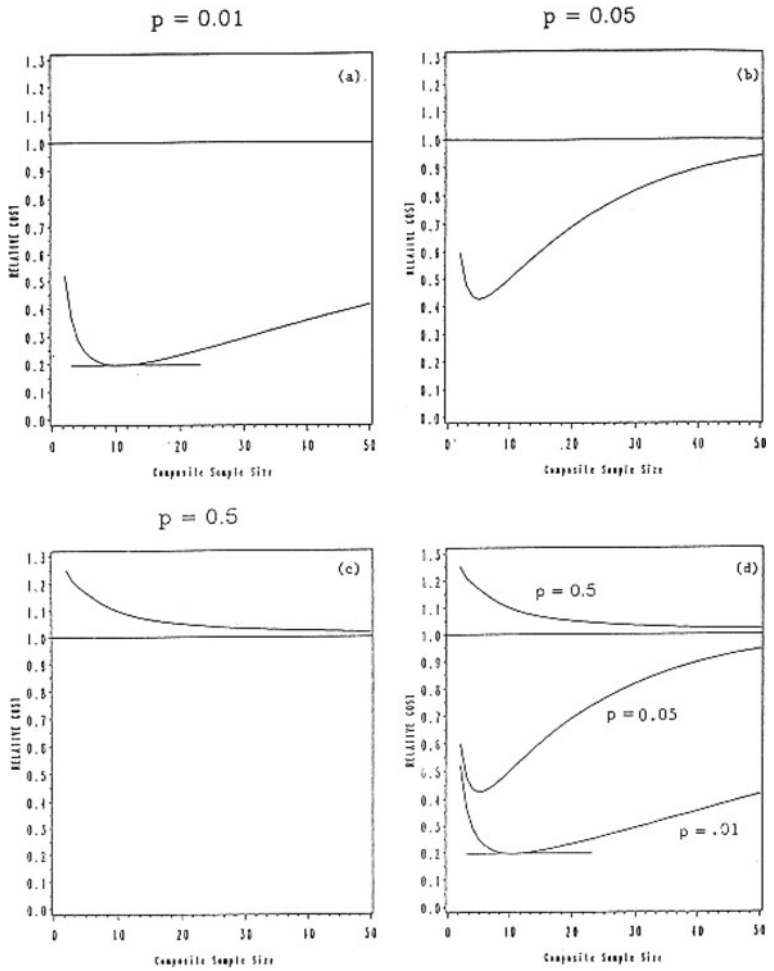

Fig. 5.1 **(a)** Effect of a tactical error. The *small horizontal line* represents the relative cost if the optimal composite sample size is used. The *curve* represents the actual relative cost of composite sampling as a function of the composite sample size. **(b)** and **(c)** Effects of strategic errors. The *horizontal line* at relative cost = 1 represents the relative cost of exhaustive testing. The *curve* represents the actual relative cost of composite sampling as a function of the composite sample size. **(d)** Shows **(a)**, **(b)**, and **(c)** on a single graph to facilitate comparison

of k, will incur a relative cost higher than 0.1956. However, the relative cost is still smaller than unity, the relative cost of exhaustive testing. The graph in Fig. 5.1c depicts this situation. Fig. 5.1d combines the graphs for all the three cases described above.

A Bayesian formulation allows us to approach this problem more realistically. Usually, there is some prior information, some local knowledge, or some expert opinion available on the prevalence of samples that possess the trait. Using some

or all of these forms of information, it is possible to specify a prior probability distribution for the prevalence and thereby predict a value of the prevalence. This predicted value of the prevalence is then used to determine the composite sample size. After a batch of samples are classified using any of the classification procedures described earlier (see Chapter 2), the information on the prevalence can be updated by evaluating the posterior distribution of the prevalence, given the number of samples classified as possessing the trait. In monitoring situations, this posterior can be used as the prior at the subsequent monitoring stage. In this way, the prior and the empirical information together may result in a composite sample size that converges to the optimal value, as the monitoring progresses.

We now formalize the notation used in the remainder of this chapter. Samples are to be classified in successive stages of sampling. Let p denote the true but unknown prevalence of polluted samples. We assume that the pollution process acts independently and identically on individual samples. Then p is also the probability that a randomly selected sample is polluted. That is, the probability, p, that a given sample is polluted is the same for every individual sample and does not depend on which, if any, of the other samples are polluted. For $i = 1, 2, \ldots$, suppose that n_i samples are classified during the ith sampling stage. This may be done using composite samples or by exhaustively testing all the n_i individual samples, because the outcome of the classification does not depend on the particular method used for classifying samples. This is so because we consider only those procedures that accurately classify the samples. To distinguish the true prevalence p from its predicted value, we use p_i to denote the predicted value of p for the ith sampling stage. This predicted value, p_i, is used to determine the composite sample size k_i for use at the ith sampling stage. As the result of classification of n_i samples, there will be a random number X_i of samples classified as polluted. By the assumptions made above, X_i has a binomial distribution with parameters n_i and p. When a prior distribution of p is used, the parameters of the prior distribution will also have a subscript indicating the sampling stage. For instance, imposing a conjugate prior on our belief about p implies a beta distribution with parameters α and β, say. For $i = 1, 2, \cdots$, let α_i and β_i indicate the parameters of the prior distribution used at the ith sampling stage. The predicted value of p for the ith sampling stage is then given by $p_i = \alpha_i/(\alpha_i + \beta_i)$. Note that p_i is never a parameter in the usual sense; it is a predicted value and thus is neither the true value nor an unknown.

Under the usual method of exhaustively testing every individual sample, the total number of measurements is the same as the number of individual samples to be classified. That is, there is one measurement per individual sample and hence, the relative cost is 1. The problem is to decide whether to form composite samples for making measurements or to carry out exhaustive testing. For the purpose of illustrating the difference between the two alternatives, we consider only exhaustive retesting as the composite sampling procedure of classification.

In order to implement exhaustive retesting, it is necessary to choose the composite sample size k ($k = 1$ implies exhaustive testing). The optimal choice of k minimizes the relative cost for a specified value of p. One possible method of

choosing k is to predict the value of the prevalence p based on the available prior information, and then one can determine the corresponding optimal composite sample size for exhaustive retesting. After k individual samples in a single composite sample are classified, the information on the prevalence can be updated using the number of samples classified as polluted, and this procedure can be repeated at every monitoring stage.

5.2 Bayesian Updating of p

As described in the preceding section, samples are to be classified in successive stages. Suppose that the available prior information, local knowledge, and expert opinion on the prevalence of polluted samples are expressed in terms of a beta distribution with parameters α and β. The prior pdf of p is then given by

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad 0 < p < 1;$$

where $\alpha > 0$, $\beta > 0$, and

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

where Γ is the gamma function.

Suppose n samples are classified and X of them are classified as polluted. Then, for a given value of p, the random variable X follows the binomial distribution with parameters n and p. That is, the probability mass function of X is given by

$$\Pr[X = x|p] = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n; \quad 0 < p < 1.$$

The conditional pdf of p given that $X = x$ is then derived as follows.

The conditional pdf $f(p|X = x)$ of p given that $X = x$ is defined by

$$\begin{aligned} f(p|X = x) &= \frac{f(p)\Pr[X = x|p]}{\Pr[X = x]} \\ &= \frac{f(p)\Pr[X = x|p]}{\int_0^1 \Pr[X = x|p] f(p) dp}. \end{aligned}$$

Now,

$$\int_0^1 \Pr[X = x|p] f(p) dp = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \cdot \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} dp$$

$$= \frac{\binom{n}{x}}{B(\alpha, \beta)} \int_0^1 p^{\alpha+x-1} (1-p)^{n+\beta-x-1} dp$$

$$= \binom{n}{x} \cdot \frac{B(\alpha + x, n + \beta - x)}{B(\alpha, \beta)}.$$

We therefore have

$$f(p|X = x) = \frac{\frac{\binom{n}{x}}{B(\alpha,\beta)} p^{\alpha+x-1} (1-p)^{n+b-x-1}}{\frac{\binom{n}{x} B(\alpha+x, n+\beta-x)}{B(\alpha,\beta)}}$$

$$= \frac{1}{B(\alpha + x, n + \beta - x)} p^{\alpha+x-1} (1-p)^{n+\beta-x-1}, \quad 0 < p1.$$

Note that this is the pdf of a beta distribution with parameters $\alpha + x$ and $n + \beta - x$.

We now establish the updating formula for the predicted value of the prevalence of polluted samples at successive sampling stages.

Suppose that the parameters of the initial prior distribution of p are denoted by α_1 and β_1. Note that this prior distribution of p is to be used for predicting the value of p at the first sampling stage. As noted earlier, the prior expectation of p is to be used as the predicted value of p. Now the expectation of the beta distribution with parameters α_1 and β_1 is given by

$$p_1 = \frac{\alpha_1}{\alpha_1 + \beta_1}.$$

Let k_1 denote the composite sample size that minimizes the relative cost $1 + \frac{1}{k} - (1 - p_1)^k$. Let n_1 individuals be classified during the first sampling stage, and further let X_1 denote the number of samples classified as polluted. Then the posterior distribution of p given X_1 is a beta distribution with parameters $\alpha_2 = \alpha_1 + X_1$ and $\beta_2 = n_1 + \beta_1 - X_1$. This posterior distribution of p is used as its prior distribution for the second sampling stage. That is, for the second sampling stage,

$$p_2 = \frac{\alpha_1 + X_1}{n_1 + \beta_1 - X_1}$$

is the predicted value of the prevalence of polluted samples, and k_2 is the corresponding recommended composite sample size.

Proceeding in the same manner from one sampling stage to another, we obtain a sequence of the predicted values of the prevalence of polluted samples at successive sampling stages. Thus,

$$p_i = \frac{\alpha_i}{\alpha_i + \beta_i}, \quad i = 1, 2, \ldots;$$

where

$$\alpha_i = \alpha_{i-1} + X_{i-1};$$
$$\beta_i = n_{i-1} + \beta_{i-1} - X_{i-1}; \quad i = 1, 2, \ldots.$$

That is, for positive integers $r \geq 2$, we have

$$\alpha_r = \alpha_1 + \sum_{i=1}^{r-1} X_i,$$
$$\beta_r = \beta_1 + \sum_{i=1}^{r-1} n_i - \sum_{i=1}^{r-1} X_i.$$

The predicted value $p_r = \frac{\alpha_r}{\alpha_r + \beta_r}$ is used to determine the composite sample size k_r at the rth sampling stage, $r = 1, 2, \ldots$.

Since closed-form formulas are not available for the relative cost at successive stages, computer calculations are used to calculate the binomial probabilities for all possible cases and to evaluate the relative cost for the first 10 stages of monitoring. The number of cases in the computation increases geometrically with the number of sampling stages. The updating is done after each composite sample has been processed. That is, n_i is taken as k_i, $i = 1, 2, \ldots, 10$.

Since composite sample techniques are cost-efficient only for relatively small prevalences, it was decided to use J-shaped beta prior distributions and to restrict the prevalence to be smaller than 0.5. If the mean p is restricted to be smaller than 0.5 then the J-shaped beta distribution with mean μ and the largest possible variance has $\beta = 1$ and $\alpha = \mu/(1 - \mu)$, that is, $\lambda = 1.0$ below. If the mean is $\mu = 0.5$, then the distribution is the uniform distribution. The J-shaped distribution with mean $\mu < 0.5$ and the smallest possible variance has $\alpha = 1$ and $\beta = (1 - \mu)/\mu$, that is, $\lambda = 0.0$ below. Then J-shaped curves with intermediate variances are obtained by taking the following linear combinations:

$$\alpha = \lambda \left(\frac{\mu}{1 - \mu} \right) + (1 - \lambda) \cdot 1,$$
$$\beta = \lambda \cdot 1 + (1 - \lambda) \left(\frac{1 - \mu}{\mu} \right),$$

with $\lambda = 0.0, 0.5, 0.6, 0.7, 0.8, 0.9$, and 1.0. The variances of these seven distributions are shown in Fig. 5.2a, b. Large variances cause large changes when updating. When the initial predicted prevalence is far from the true prevalence, large changes are desirable.

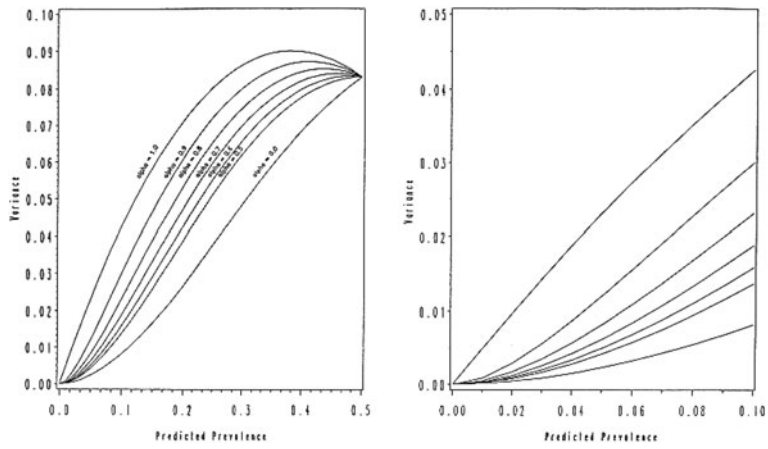

Fig. 5.2 Variances for J-shaped beta distributions: $\alpha = \lambda\left(\frac{\mu}{1-\mu}\right) + (1-\lambda)\cdot 1$, $\beta = \lambda \cdot 1 + (1-\lambda)\left(\frac{1-\mu}{\mu}\right)$; $\lambda = 1.0$ gives maximum variance; $\lambda = 0.0$ gives minimum variance

However, when the initial predicted prevalence is close to the true prevalence, large changes cause the procedure to deviate from the optimal procedure, and it may take several sampling stages for it to approach the optimal procedure.

Although the emphasis is on the relative cost, the average composite sample sizes are also calculated. This chapter assumes that a composite sample of any size can be formed. Small predicted prevalences result in large composite sample sizes. When $\alpha = 1.0$, the composite sample size shoots up dramatically for small predicted prevalences and small true prevalences and takes several sampling stages to come back down. The use of larger predicted prevalence and/or the use of smaller values of α may avoid this problem.

5.3 Minimization of the Expected Relative Cost

Recall that the relative cost for the exhaustive retesting procedure, given by $\mathrm{RC} = 1 + 1/k - q^k$, is a function of the composite sample size k and the prevalence $p = 1-q$ of the trait under study. In decision-theoretic notation, since the composite sample size is a decision of the statistician and the prevalence of the trait is beyond the statistician's control, the relative cost is a loss function, which could be written as

$$L(k, q) = 1 + 1/k - q^k.$$

The approach in the foregoing discussion imposed a conjugate prior on the prevalence p, giving the expected prevalence $\hat{p}$, say, and then the composite sample size, $\hat{k}$, say, that minimizes $L(k, \hat{q})$ was determined. That is, $\hat{k}$ satisfies

$$L\left(\hat{k}, \hat{q}\right) = \min_{k} \; L(k, \hat{q}).$$

In this section, we consider a decision-theoretical approach, minimizing the expected loss rather than the loss function at the expected prevalence. Thus, we use the prior distribution with a density f, say, to compute the risk function

$$R(k, f) = E[L(k, q)] = \int_0^1 L(k, q) f(p) dp$$

and then determine the composite sample size k that minimizes $R(k, f)$. The conjugate prior distribution of the prevalence, p, is a beta distribution, which has two parameters, α and β. Thus, the risk function can be expressed as

$$\begin{aligned} R(k, \theta) &= \int_0^1 L(k, q) \frac{1}{B(\alpha, \beta)} p^{\alpha-1} q^{\beta-1} dp \\ &= 1 + \frac{1}{k} - \frac{B(\alpha, \beta + k)}{B(\alpha, \beta)}, \end{aligned}$$

where $\theta = (\alpha, \beta)$ and $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. The composite sample size k that minimizes $R(k, \theta)$ will be the optimal composite sample size since it minimizes the expected relative cost under the beta prior distribution.

The composite sample size k determined by this method is a step function of α and β, which are the parameters of the beta prior distribution. Note that the composite sample size determined by minimizing the relative cost for the predicted prevalence is a function of the mean of the prior distribution, $\alpha/(\alpha + \beta)$. Thus, the method of the preceding section would return the same composite sample size for all pairs (α, β) which have $\alpha/(\alpha+\beta)$ constant. To get an idea as to how the values of k determined by minimizing $R(k, \theta)$ change with α and β, some numerical results were obtained. These are shown in Fig. 5.3a, b. Note that the contour lines in these figures are not straight lines.

The sequential updating of the prior information can be described as follows. Let the initial beta prior on p have the parameters α_1 and β_1. For the Dorfman retesting scheme, the composite sample size for the first stage k_1 is the value of k which minimizes $R(k, \theta_1)$, where $\theta_1 = (\alpha_1, \beta_1)$. Then composite samples of size k_1 are formed and tested. Suppose X_1 individual samples out of n_1 are classified as polluted in the first stage.

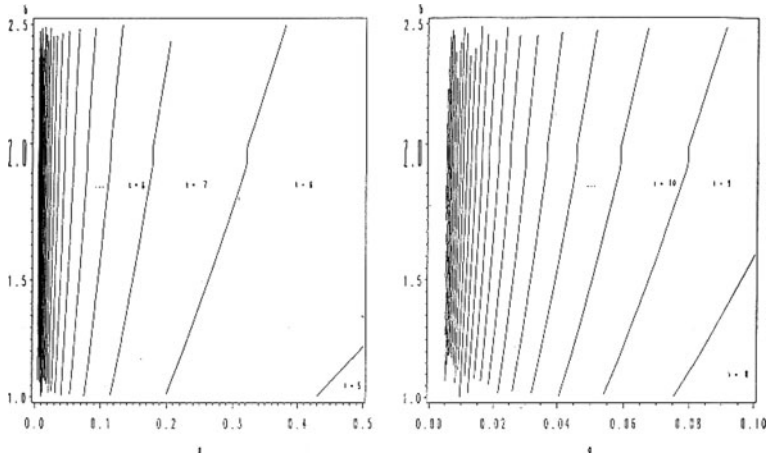

Fig. 5.3 Regions of optimal composite sample size k as a function of α and β

The posterior distribution of p given X_1 is then beta with parameters $\alpha_2 = \alpha_1 + X_1$ and $\beta_2 = \beta_1 + n_1 - X_1$. This distribution is used as the prior distribution at the second stage. For Dorfman retesting scheme, the composite sample size, k_2, corresponds to the risk function $R(k, \theta_2)$, where $\theta_2 = (\alpha_2, \beta_2)$. Composite samples of size k_2 are then formed and tested, yielding X_2 individual samples that test positive for pollution. In general, after r sampling stages, as in the preceding section, the posterior distribution of p is the beta distribution with parameters

$$\alpha_{r+1} = \alpha_1 + \sum_{i=1}^{r} X_i \quad \text{and}$$

$$\beta_{r+1} = \beta_1 + \sum_{i=1}^{r} n_i - \sum_{i=1}^{r} X_i .$$

This distribution is then used as the prior distribution of p for the $(r+1)$th stage.

5.4 Discussion

Even though composite sample techniques are known to be cost-efficient, it is not always possible to arrive at the optimal composite sample size. The potential risk of not being able to arrive at the optimal composite sample size is negotiated by the sequential empirical Bayesian approach described in this chapter. Since closed-form expressions are not available, numerical results are worked out for graphical presentation in this chapter. An alternative to the computations could be a simulation study, but it would result only in approximate results. It was therefore preferred to present

the results of computations along with the algorithms. Due to limitations on the computing time, only the first 10 sampling stages were included in the computations reported here. The algorithms, however, are very general and can be implemented to include any desired number of sampling stages for the purpose of computing the relative costs as well as the composite sample sizes at all the sampling stages.

This chapter deals with a situation where all the individual samples have the same probability p of being polluted, independently of one another. A more general situation could have the probability that a sample is polluted changing from sample to sample. Another possibility is regarding the type of measurement. This chapter assumes a presence/absence measurement, and therefore every sample is simply classified as testing positive or negative. If the measurement is continuous, then the classification of samples will involve a criterion value for a composite sample, which depends on the composite sample size and the criterion value for an individual sample. Thus, if c is the criterion value for an individual sample and if k is the composite sample size, then the criterion value for a composite sample (of size k) is c/k. If the method of measurement is subject to a detection limit d, then one must have $c/k \geq d$ in order to prevent any false-negative tests. This requirement restricts the composite sample size $k \leq c/d$.

Chapter 6
Inference on Mean and Variance

6.1 Introduction

A common purpose of compositing sampling is to draw statistical inference on the population mean, and possibly on the population variance. As noted earlier (see Chapter 1), sampling units may be selected from a finite population or from a bulk population. In the former case, the sampling units are defined and exist before sampling while in the latter case, the sampling units are created by the sampling process. In either case, however, the sampling units as selected from the population are called individual samples so that the remainder of this chapter applies equally to both a finite population and a bulk or integrated population. Every individual sample has a unique value for the variable of interest, and this value is called the individual sample value. Individual sample values are denoted by X with a possible subscript identifying an individual sample. Thus, the individual sample values for m individual samples would be denoted by $X_1, X_2, \ldots, X_m$. Usually the individual sample values are assumed to be independent and identically distributed with mean $E[X_i] = \mu_x$ and $\text{Var}[X_i] = \sigma_x^2$. If an individual sample is subjected to measurement, then its measured value, also called the individual sample measurement, need not coincide with the corresponding individual sample value due to a possible measurement error. It is a common practice to derive an aliquot, also called an increment, from an individual sample for making measurement. In such a case, the individual sample is said to be homogeneous if every aliquot derived from this individual sample has the same value as that of the individual sample. Under the assumption of homogeneity of all individual samples, making measurements on either individual samples or aliquots derived therefrom would yield identical results, except for a possible measurement error.

A composite sample of size k is formed by selecting k individual samples and then pooling them together. It is also possible that a composite sample is formed by pooling k aliquots derived from the k individual samples to be composited. It is necessary to homogenize the composite sample in order to make it a homogeneous primary sampling unit. Otherwise, the nonhomogeneity of the composite sample should be accommodated in the statistical treatment of composite sample measurements. Also, if the aliquots derived from the individual samples that contribute to a composite sample are unequal in volume, then their contributions to the composite

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_6, © Springer Science+Business Media, LLC 2011

sample value will also be unequal. In this case, the composite sample value is a weighted average of the individual sample values. In addition, it is not always possible to determine or control the volumes of the individual samples that contribute to a composite sample. As a consequence, the relative proportions of these volumes have to be treated as random variables. This case is called the case of composite sampling with random weights. The foregoing discussion makes it clear that drawing statistical inference on the population mean using composite sample data involves a variety of complexities through homogeneity or nonhomogeneity of the composite samples, possible randomness of weights, and possible measurement error.

In this chapter, we relate the population parameters to the parameters of the distribution of composite sample data. It is important to note that the population is modeled only through its first two moments, and no distributional assumption is required on the population. As the result of compositing several individual samples, the composite sample value is an average of the constituent individual sample values. A consequence of this averaging is that the probability distribution of composite sample values is closer to the normal distribution than that of individual sample values. The normality of the composite sample values is more pronounced if the population distribution is moderately non-normal. That is, if the population distribution is not highly skewed, then the composite sample values may be assumed to be approximately normally distributed. This observation permits not only estimation of population parameters using composite sample data but also construction of confidence intervals and tests of hypotheses involving these parameters. In this chapter, we develop the necessary statistical theory for estimation of population parameters, construction of a confidence interval for the population mean, and construction of the test for a hypothesis in the population mean. For a discussion on tests of hypotheses with composite sample data, see also Mack and Robinson (1985), Messner et al. (1990), Neptune et al. (1990), and Edland and van Belle (1994).

6.2 Notation and Basic Results

6.2.1 Notation

Let X denote the individual sample values for the characteristic of interest (observed or conceptual), let Y (and occasionally Z) denote the composite sample value, and let W denote the weights of the individual samples within a composite. We shall use uppercase letters to denote random variables and lowercase letters for constants. Moreover, lowercase boldface letters will denote random or constant vectors and uppercase boldface letters will denote random or constant matrices. Let ϵ denote the measurement error that may occur while making measurement on a sample. Subscripts will be used to distinguish between different samples – individual, composite, as well as subsamples.

Let n be the number of composite samples; k be the composite sample size, that is, the number of individual samples contributing to a single composite

sample; and s be the number of subsamples drawn from a single composite sample. Define $\boldsymbol{\mu}_x = E[\boldsymbol{x}]$, the expected value of $\boldsymbol{x}$; $\boldsymbol{\Sigma}_x = \text{Var}[\boldsymbol{x}]$, the variance/covariance matrix of $\boldsymbol{x}$; and $\boldsymbol{\Gamma}_{x,z} = \text{cov}[\boldsymbol{x}, \boldsymbol{z}]$, the covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{z}$. Let $\boldsymbol{x}_{ji} = \left[X_{ji1}, X_{ji2}, \ldots, X_{jik}\right]'$ be the vector of values taken up by the k individual samples forming the ith subsample of the jth composite sample, $i = 1, \ldots, s;\ j = 1, \ldots, n$; $\boldsymbol{w}_{ji} = \left[W_{ji1}, W_{ji2}, \ldots, W_{jik}\right]'$ be the vector of weights (compositing proportions) with which the k individual samples contribute to the ith subsample of the jth composite sample, $i = 1, \ldots, s;\ j = 1, \ldots, n$; and $Y_{ji} = \sum_{\ell=1}^{k} W_{ji\ell} X_{ji\ell} = \boldsymbol{w}'_{ji} \boldsymbol{x}_{ji}$ be the measurement on the ith subsample from the jth composite sample, $i = 1, \ldots, s;\ j = 1, \ldots, n$.

6.2.2 Basic Results

Let $\boldsymbol{x}$ be a random k-vector with expectation $\boldsymbol{\mu}_x$ and variance/covariance matrix $\boldsymbol{\Sigma}_x$ and let $\boldsymbol{c}$ be a constant k-vector. We have the following results for the expectation and the variance of $\boldsymbol{c}'\boldsymbol{x}$.

Lemma 6.2.1 *Expectation of* $\boldsymbol{c}'\boldsymbol{x}$:

$$E\left[\boldsymbol{c}'\boldsymbol{x}\right] = \boldsymbol{c}'\boldsymbol{\mu}_x.$$

Proof Note that

$$\boldsymbol{c}'\boldsymbol{x} = \sum_{\ell=1}^{k} c_\ell X_\ell$$

and therefore

$$\begin{aligned} E\left[\boldsymbol{c}'\boldsymbol{x}\right] &= E\left[\sum_{\ell=1}^{k} c_\ell X_\ell\right] \\ &= \sum_{\ell=1}^{k} c_\ell E\left(X_\ell\right) \text{ by linearity of expectation} \\ &= \boldsymbol{c}'\boldsymbol{\mu}_x. \end{aligned}$$

Lemma 6.2.2 *Variance of* $\boldsymbol{c}'\boldsymbol{x}$:

$$\text{Var}\left[\boldsymbol{c}'\boldsymbol{x}\right] = \boldsymbol{c}'\boldsymbol{\Sigma}_x\boldsymbol{c}.$$

Proof Note that

$$\boldsymbol{c}'\boldsymbol{x} = \sum_{\ell=1}^{k} c_\ell X_\ell$$

and therefore

$$\begin{aligned}\mathrm{Var}\left[\boldsymbol{c}'\boldsymbol{x}\right] &= \mathrm{Var}\left[\sum_{\ell=1}^{k} c_\ell X_\ell\right] \\ &= \sum_{\ell=1}^{k} c_\ell^2 \mathrm{Var}(X_\ell) \\ &+ \sum_{\ell \neq \ell'} c_\ell c_{\ell'} \mathrm{cov}\left(X_\ell, X_{\ell'}\right) \\ &= \boldsymbol{c}' \boldsymbol{\Sigma}_x \boldsymbol{c}.\end{aligned}$$

Corollary 6.2.1

$$E\left[\boldsymbol{x}\boldsymbol{x}'\right] = \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x'.$$

Proof Note that

$$\begin{aligned}\boldsymbol{\Sigma}_x &= E\left[\left(\boldsymbol{x} - \boldsymbol{\mu}_x\right)\left(\boldsymbol{x} - \boldsymbol{\mu}_x\right)'\right] \\ &= E\left[\boldsymbol{x}\boldsymbol{x}'\right] - \boldsymbol{\mu}_x \boldsymbol{\mu}_x'\end{aligned}$$

and therefore

$$E\left[\boldsymbol{x}\boldsymbol{x}'\right] = \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x \boldsymbol{\mu}_x',$$

and hence the corollary is proved.

Lemma 6.2.3 *Let* $\boldsymbol{x}$ *be a random k-vector, and let* $\boldsymbol{c}$ *and* $\boldsymbol{d}$ *be constant k-vectors. Then*

$$\mathrm{cov}\left[\boldsymbol{c}'\boldsymbol{x}, \boldsymbol{d}'\boldsymbol{x}\right] = \boldsymbol{c}' \boldsymbol{\Sigma}_x \boldsymbol{d}.$$

Proof Consider

$$\mathrm{cov}\left[\boldsymbol{c}'\boldsymbol{x}, \boldsymbol{d}'\boldsymbol{x}\right] = \mathrm{cov}\left(\sum_{\ell=1}^{k} c_\ell X_\ell, \sum_{\ell'=1}^{k} d_{\ell'} X_{\ell'}\right)$$

$$\begin{aligned}&= \left[\left(\sum_{\ell=1}^{k} c_\ell X_\ell\right)\left(\sum_{\ell'=1}^{k} d_{\ell'} X_{\ell'}\right) - \left(\sum_{\ell=1}^{k} c_\ell \mu_\ell\right)\left(\sum_{\ell'=1}^{k} d_{\ell'} \mu_{\ell'}\right)\right] \\ &= E\left[\sum_{\ell=1}^{k}\sum_{\ell'=1}^{k} c_\ell d_{\ell'} X_\ell X_{\ell'}\right] - \left(\boldsymbol{c}'\boldsymbol{\mu}_x\right)\left(\boldsymbol{d}'\boldsymbol{\mu}_x\right)\end{aligned}$$

$$
\begin{aligned}
&= E\left[\boldsymbol{c}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{d}\right] - \boldsymbol{c}'\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\boldsymbol{d} \\
&= \boldsymbol{c}'E\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{d} - \boldsymbol{c}'\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\boldsymbol{d} \\
&= \boldsymbol{c}'\left[\boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right]\boldsymbol{d} - \boldsymbol{c}'\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\boldsymbol{d} \text{ by Corollary 6.2.1} \\
&= \boldsymbol{c}'\boldsymbol{\Sigma}_x\boldsymbol{d},
\end{aligned}
$$

and hence the lemma is proved.

Let $\boldsymbol{x}$ be a random k-vector and let $\boldsymbol{C}$ be a $k \times k$ constant matrix. We then have the following result.

Lemma 6.2.4 *Expectation of* $\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x}$*:*

$$E\left[\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x}\right] = \boldsymbol{\mu}_x'\boldsymbol{C}\boldsymbol{\mu}_x + \operatorname{tr}\left[\boldsymbol{C}\boldsymbol{\Sigma}_x\right].$$

Proof Note that

$$\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x} = \operatorname{tr}\left[\boldsymbol{C}\boldsymbol{x}\boldsymbol{x}'\right]$$

and therefore

$$E\left[\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x}\right] = E\left[\operatorname{tr}\boldsymbol{C}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right] = \operatorname{tr}\left[\boldsymbol{C}E\left(\boldsymbol{x}\boldsymbol{x}'\right)\right]$$

by linearity of expectation in Lemma 6.2.3:

$$
\begin{aligned}
&= \operatorname{tr}\left[\boldsymbol{C}\left(\boldsymbol{\mu}_x\boldsymbol{\mu}_x' + \boldsymbol{\Sigma}_x\right)\right] \\
&= \operatorname{tr}\left[\boldsymbol{C}\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right] + \operatorname{tr}\left[\boldsymbol{C}\boldsymbol{\Sigma}_x\right] \\
&= \boldsymbol{\mu}_x'\boldsymbol{C}\boldsymbol{\mu}_x + \operatorname{tr}\left[\boldsymbol{C}\boldsymbol{\Sigma}_x\right].
\end{aligned}
$$

6.3 Estimation Without Measurement Error

Suppose n composite samples are formed using known volumes of k individual (usually secondary) sampling units each. In case all the primary sampling units are homogeneous, composite sample values are simply obtained as weighted averages of values of the respective constituent primary sampling units. Suppose $m = kn$. Suppose n composites are formed from k primary sampling units each and are measured. Let $Y_1, \ldots, Y_n$ denote the n composite sample measurements. If $V_{j1}, V_{j2}, \ldots, V_{jk}$ are volumes of the k secondary sampling units that form the jth composite sample and if $X_{j1}, X_{j2}, \ldots, X_{jk}$ are the corresponding primary sampling unit values, then

$$Y_j = W_{j1}X_{j1} + W_{j2}X_{j2} + \cdots + W_{jk}X_{jk},$$

where

$$W_{ji} = \frac{V_{ji}}{V_{j1} + \cdots + V_{jk}}.$$

Now

$$E\big[Y_j\big] = \sum_{i=1}^{k} W_{ji} E\left[X_{ji}\right] = \mu_x,$$

since

$$\sum_{i=1}^{k} W_{ji} = 1$$

and

$$\boldsymbol{\mu}_x = \mu_x \mathbf{1}_k.$$

Further, assuming $\boldsymbol{\Sigma}_x = \sigma_x^2 \boldsymbol{I}_m$, we have

$$\begin{aligned} \mathrm{Var}\big[Y_j\big] &= \sum_{i=1}^{k} W_{ji}^2 \mathrm{Var}\left[X_{ji}\right] \\ &= \sigma_x^2 \left(\sum_{i=1}^{k} W_{ji}^2 \right), \end{aligned}$$

provided $X_{j1}, \ldots, X_{jk}$ are independent.

If the k secondary sampling units that form a composite have the same volume, then $W_{ji} \equiv \frac{1}{k}$. In this case, $E\big[Y_j\big] = \mu_x$ and $\mathrm{Var}\big[Y_j\big] = \sigma_x^2/k$. Further, if $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ is the average of the n composite sample measurements, then $E\big[\bar{Y}_n\big] = \mu_x$ and $\mathrm{Var}\left[\bar{Y}_n\right] = \left[\sigma_x^2/k\right]/n = \sigma_x^2/m$, since $m = nk$. Recall that the average of $m = nk$ individual sample measurements yield $E\big[\bar{X}_m\big] = \mu_x$ and $\mathrm{Var}\big[\bar{X}_m\big] = \sigma_x^2/m$. In other words, making $m = nk$ measurements on individual sampling units yield an unbiased estimator of the population mean μ_x with a variance of σ_x^2/m. On the other hand, only n measurements on composite sample also yield an unbiased estimator of μ_x with the same variance σ_x^2/m. It is then obvious that at least under the idealistic assumptions of independence of primary sampling units and compositing of equal volumes of secondary sampling units, composite sampling achieves substantial savings in the cost of measurement by reducing the total number of measurement from $m = nk$ to n.

6.4 Estimation in the Presence of Measurement Error

The measurements, either on individual or on composite samples, are not always exact. Measurements are often made with error. In the presence of measurement errors, the measured value of a sampling unit or a sample is different from its true value. For instance, the measurement on an individual sample, T_i, may be related to the true value, X_i, by the following relationship:

$$T_i = X_i + \epsilon_i, \quad i = 1, \ldots, m.$$

Assuming $\{\epsilon_i\}$ to be independent and identically distributed with $E[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma_\epsilon^2$, we have

$$E[T_i] = \mu_x \quad \text{and} \quad \text{Var}[T_i] = \sigma_x^2 + \sigma_\epsilon^2.$$

If a sample of m primary sampling units is selected at random and measured, then the sample mean $\bar{T}_m$ is an unbiased estimator of the population mean μ_x. Further, assuming that measurement errors are independent of sample values, we also have

$$\text{Var}\left[\bar{T}_m\right] = \left(\sigma_x^2 + \sigma_\epsilon^2\right)/m.$$

Suppose n composites are formed from k primary sampling units each. If measurement error has the same behavior when measuring composite samples as when measuring individual samples, then the composite sample measurement Y_j will be given by

$$Y_j = \sum_{i=1}^{k} W_{ji} X_{ji} + \epsilon_j,$$

where $W_{ji} = V_{ji}/(V_{j1} + \cdots + V_{jk})$, and ϵ_j is the measurement error with $E\left[\epsilon_j\right] = 0$, $\text{Var}\left[\epsilon_j\right] = \sigma_\epsilon^2$. Therefore, $E\left[Y_j\right] = \sum_{i=1}^{k} W_{ji} E\left[X_{ji}\right] = \mu_x$, and assuming that the measurement error is independent of measured values, $\text{Var}\left[Y_j\right] = \sigma_x^2 \left(\sum_{i=1}^{k} W_{ji}^2\right) + \sigma_\epsilon^2$.

If equal volumes of all the k individual samples are used to form composite samples, then $W_{ji} \equiv \frac{1}{k}$ and hence $E\left[Y_j\right] = \mu_x$, $\text{Var}\left[Y_j\right] = \sigma_x^2/k + \sigma_\epsilon^2$, $j = 1, 2, \ldots, n$.

If $\bar{Y}_n$ is the average of the n composite sample measurements, then assuming independence between composite samples, we obtain

$$\begin{aligned} E\left[\bar{Y}_n\right] &= \mu_x, \\ \text{Var}\left[\bar{Y}_n\right] &= \left(\sigma_x^2/k + \sigma_\epsilon^2\right)/n \\ &= \sigma_x^2/m + \sigma_\epsilon^2/n. \end{aligned}$$

Compare the variance of the composite sample mean $\bar{Y}_n$ with that of the individual sample mean $\bar{X}_m$, which is given by

$$\mathrm{Var}\left[\bar{X}_m\right] = \sigma_x^2/m + \sigma_\epsilon^2/m.$$

Here, the individual sample mean has a smaller variance, but the composite sample mean has a smaller cost of measurement. Composite sample mean is based on only n measurements as opposed to $m = nk$ measurements required to obtain the individual sample mean. If the cost of measurement is relatively high, then composite sampling may result in substantial savings in the cost of measurement, although it entails a reduction in the precision as well.

6.5 Maintaining Precision While Reducing Cost

If a certain level of precision is desired about the mean, then composite sampling may be able to maintain that precision level, while reducing the overall cost of sampling and of measurement relative to measuring each individual sample (Paasivirta and Paukku, 1989a). Given the same model assumptions as before, the variance of the mean of m individual samples measured with error is

$$\mathrm{Var}\left[\bar{X}_m\right] = \left(\sigma_x^2 + \sigma_\epsilon^2\right)/m,$$

while the variance of the mean of n composite samples, each of size k, is

$$\mathrm{Var}\left[\bar{Y}_n\right] = (\sigma_x^2/k + \sigma_\epsilon^2)/n.$$

Maintaining precision in the mean between estimators based on m individual samples and on n composite samples can be accomplished by restricting m

$$m \geq n(\lambda + (1-\lambda)/k),$$

where $\lambda = \sigma_\epsilon^2 \left(\sigma_x^2 + \sigma_\epsilon^2\right)$ is the ratio of measurement error variability relative to the sum of measurement error and sampling variability.

If C_s and C_a represent the cost of sampling and the cost of analysis or measurement, respectively, then the total sampling and analytical costs of m individual samples and of n composite samples, each of size k, are $m(C_\mathrm{s}+C_\mathrm{a})$ and $n(kC_\mathrm{s}+C_\mathrm{a})$, respectively. Here the additional cost of forming the composite samples, after the individual samples have been taken, is negligible.

Substantial cost savings can be realized by composite sampling, as compared to measuring each individual samples, while maintaining precision about the mean. This is especially true when the analytical measurements are precise, but costly, relative to sampling.

For example, if the cost of measurement is 10 times the cost of taking a sample, then it is only 24% as costly to composite 4 sets of 10 individual samples as it is to measure each of 30 individual samples, yet precision in the mean is maintained when $\lambda \geq 0.05$.

In practice, there may be a number of attributes to measure, each with a different λ. Therefore, the optimal choice of k and n may differ for each attribute, but only one choice can be made. Possible design strategies might include a compromise choice, utilizing the importance of the attributes to help determine the design or preserving precision about all attributes.

6.6 Estimation of σ_x^2 and σ_ϵ^2

Using composite samples as described in Section 6.4, the components of variance, σ_x^2 and σ_ϵ^2, are confounded and cannot be decomposed. However, the mean and these two variance components can be separately estimated by taking multiple composite samples of different sizes (Cameron, 1951).

Let $\bar{Y}_1$ and $\bar{Y}_2$ be the averages of n_1 and n_2 composite sample measurements of sizes k_1 and k_2, respectively, and let S_1^2 and S_2^2 be the corresponding sample mean squares of the composite sample measurements. Now S_i^2 is an unbiased estimator of

$$\mathrm{Var}\left[\bar{Y}_i\right] = \left(\frac{\sigma_x^2}{k_i} + \sigma_\epsilon^2\right)/n_i$$

for $i = 1, 2$. Plugging in S_i^2 for $\mathrm{Var}\left[\bar{Y}_i\right]$ and solving these two equations we obtain the moment estimators of the two variance components as given below:

$$\tilde{\sigma}_x^2 = \left(\frac{1}{k_1} - \frac{1}{k_2}\right)^{-1}\left(n_1 S_1^2 - n_2 S_2^2\right),$$
$$\tilde{\sigma}_\epsilon^2 = (k_1 - k_2)^{-1}\left(k_1 n_1 S_1^2 - k_2 n_2 S_2^2\right).$$

Further, since both $\bar{Y}_1$ and $\bar{Y}_2$ are unbiased estimators of μ_x, we have for any real number a,

$$E\left[a\bar{Y}_1 + (1-a)\bar{Y}_2\right] = \mu_x$$

and

$$\begin{aligned}\mathrm{Var}\left[a\bar{Y}_1 + (1-a)\bar{Y}_2\right] &= a^2\mathrm{Var}\left[\bar{Y}_1\right] + (1-a)^2\mathrm{Var}\left[\bar{Y}_2\right] \\ &= \frac{a^2}{n_1}\left[\frac{\sigma_x^2}{k_1} + \sigma_\epsilon^2\right] + \frac{(1-a)^2}{n_2}\left[\frac{\sigma_x^2}{k_2} + \sigma_\epsilon^2\right].\end{aligned}$$

The value of a that minimizes this variance is

$$a^* = n_1\sigma_2^2/(n_1\sigma_2^2 + n_2\sigma_1^2),$$

where

$$\sigma_1^2 = \frac{\sigma_x^2}{k_1} + \sigma_\epsilon^2,$$
$$\sigma_2^2 = \frac{\sigma_x^2}{k_2} + \sigma_\epsilon^2.$$

This value can be estimated by replacing σ_i^2 by S_i^2, giving

$$\tilde{a} = n_1 S_2^2/(n_1 S_2^2 + n_2 S_1^2).$$

Therefore

$$\tilde{\mu}_x = \tilde{a}\bar{Y}_1 + (1 - \tilde{a})\bar{Y}_2$$

is an estimator of μ_x.

In general, $\tilde{\mu}_x$ is a biased estimator of μ_x since $\tilde{a}$ and $\bar{Y}_1$, $\bar{Y}_2$ are not stochastically independent. However, if the distribution of the individual sample values is the normal distribution, then $\tilde{\mu}_x$ is an unbiased estimator of μ_x. This is true since in this case $\bar{Y}_1$, $\bar{Y}_2$, S_1^2, and S_2^2 are all stochastically independent.

6.7 Estimation of Population Variance

Loss of information on individual sample values has been a major limitation of composite sampling procedures. When measurements are obtained on composite samples, information on individual sample values is lost. As a consequence, information on sample-to-sample variation within composites is also lost. If estimation of population variance is desired, then the sample-to-sample variation within composites may have to be compared with the variation within composites and may have to be compared with the variation between composite sample measurements. Using the notation of the preceding sections, let $X_{ji}, i = 1, \ldots, k_i;\ j = 1, \ldots, n$, denote the individual sample values and let Y_j, $j = 1, \ldots, n$, denote the composite sample measurement. For simplicity, we assume that all composite samples are formed from k individual samples so that $m = nk$. We also assume that $W_{ji} \equiv \frac{1}{k}$ so that $Y_j = \frac{1}{k}\sum_{i=1}^{k} X_{ji}$, $j = 1, 2, \ldots, n$.

The individual sample mean $\bar{X}_m$ is an unbiased estimator of the population mean μ_x, and the variance of $\bar{X}_m$ is given by σ_x^2/m. An unbiased estimator of σ_x^2 is given by

$$S_x^2 = \frac{1}{m-1}\sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ji} - \bar{X}_m)^2.$$

Every composite sample measurement Y_j has expectation μ_x and variance σ_x^2/k. The composite sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ is an unbiased estimator of the population mean μ_x, and the variance of $\bar{Y}_n$ is given by σ_x^2/m. An unbiased estimator of σ_x^2 using composite sample measurements is given by

$$S_y^2 = \frac{k}{m-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2.$$

The following identity is helpful in understanding how composite sampling incurs a loss of information on sample-to-sample variation within composites. The total variation among the nk individual sample values can be expressed as the sum of two components, one corresponding to the variation between individual sample values within composite samples and the other corresponding to the variation between composite sample measurements:

$$\sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ji} - \bar{X}_m)^2 = \sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ji} - Y_j)^2 + k\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2,$$

where $\bar{X}_m \equiv \bar{Y}_n$ by the assumptions that $W_{ji} \equiv \frac{1}{k}$, and there is no measurement error. If composite samples are formed by randomly grouping m individual samples into n groups of size k each, then the above identity can theoretically give three unbiased estimators of σ_x^2, namely,

$$S_x^2 = \frac{1}{m-1}\sum_{j=1}^{n}\sum_{i=1}^{k}\left[X_{ji} - \bar{X}_m\right]^2,$$

$$S_y^2 = \frac{k}{m-1}\sum_{j=1}^{n}\left[Y_j - \bar{Y}_n\right]^2,$$

and

$$S_w^2 = \frac{1}{n(k-1)}\sum_{j=1}^{n}\sum_{i=1}^{k}\left[X_{ji} - Y_j\right]^2.$$

Note that S_x^2 uses only individual sample values including their mean, while S_y^2 is based solely on composite sample measurements, again including their mean. The third estimator will not be available in any application of composite sampling, since it involves individual sample values as well as composite sample measurements. If composite samples are formed and measured, then no measurements will be taken on individual samples. On the other hand, if measurements are made on individual samples, then composite samples will not be formed, and hence no measurements will be available on composite samples. It is interesting to note that it gives a measure of sample-to-sample variation within composites. As a consequence of nonavailability of this value, there is a loss of information of the sample-to-sample variation within composite samples. As for the unbiasedness of these estimators, we make the following observations:

- If composite samples are formed by randomly selecting k individual samples each, then all the three estimators of σ_x^2 given above are unbiased.
- If composites are formed with an objective of increasing heterogeneity of individual samples within composites, then

$$E\left[S_w^2\right] > \sigma_x^2,$$

and hence

$$E\left[S_y^2\right] < \sigma_x^2.$$

If only estimation of the population mean is of interest, then formation of internally heterogeneous composites results in an unbiased estimator $\bar{Y}_n$ of the population mean μ_x with a precision that is higher than that of the individual sample estimator $\bar{X}_m$.

- Formation of internally homogeneous composites will imply that

$$E\left[S_w^2\right] < \sigma_x^2,$$

and hence

$$E\left[S_y^2\right] > \sigma_x^2.$$

This case is considered more conservative than the case of random formation of composites, since the confidence interval for the population mean based on composite sample measurements will be wider than that based on individual sample values. As a consequence, the probability that the confidence interval derived from composite sample measurements straddles the population mean is at least as high as the corresponding probability for the confidence interval derived from individual sample values.

The above discussion may be summarized in an analysis of variance as in Table 6.1.

Table 6.1 Analysis of Variance

Source of variation	Degree of freedom	Sum of squares	Expected mean square[a]
Between composites	$n-1$	$k\sum_{j=1}^{n}(Y_j-\bar{Y}_n)^2$	σ_x^2
Within composites	$m-n=n(k-1)$	$\sum_{j=1}^{n}\sum_{i=1}^{k}\left[X_{ji}-Y_j\right]^2$	σ_x^2
Between individual samples	$m-1$	$\sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ji}-\bar{X}_m)^2$	σ_x^2

[a] Expectation is taken under the assumption of random formation of composites

6.8 Confidence Interval for the Population Mean

The composite sample mean provides an unbiased estimator of the population mean. The composite sample mean square provides the standard error of this estimator. Using these two sample quantities, it is possible to construct a confidence interval for the population mean under certain assumptions. Since compositing of several individual samples causes the composite sample value to be an average of the individual sample values, there is a physical realization of the central limit theorem. That is, the composite sample values are approximately normally distributed if the population is moderately skewed. A confidence for the population mean can then be constructed due to the approximate normality of the composite sample values. Thus, if $Y_1, \ldots, Y_n$ are the n composite sample values, then the composite sample mean is given by $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Similarly, the composite sample mean square is $S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)^2$, and an unbiased estimator of the population variance σ^2 is given by kS_y^2, where k is the composite sample size.

Assuming a normal distribution for the composite sample values $Y_1, \ldots, Y_n$, we obtain the following confidence interval for the population mean μ:

$$\bar{Y}_n \pm t_{n-1,1-\alpha/2} \cdot S_y/\sqrt{kn},$$

where α is the probability that the confidence interval does not straddle the true population mean. The confidence interval based on n individual sample values, $X_1, \ldots, X_n$, would have been given by

$$\bar{X}_n \pm t_{n-1,1-\alpha/2} S_x/\sqrt{n},$$

which is the standard form of a confidence interval for the population mean. Note that the confidence interval based on composite sample values is always at least as narrow as that based on individual sample values. As a matter of fact, the confidence

interval based on composite samples of size 2 is about 29% shorter than the one based on individual sample values.

6.9 Tests of Hypotheses in the Population Mean

Site characterization of a Superfund Site or a hazardous waste site often involves comparing the field data with established safety standards. Cleanup evaluation demands a comparison between pre- and post-remediation data. Compliance monitoring requires routine checks to verify that contamination is under compliance norms. All these situations call for carrying out statistical tests of hypotheses for the population means. In some cases, the test is a single-sample procedure while in some others, it is a two-sample problem. In either, it is necessary to assume that the population values follow some probability distribution, with hypothesized mean μ_x and variance σ_x^2. If this distribution is moderately skewed, then the composite sample values are approximately normally distributed with the same mean μ_x but with a smaller variance σ_x^2/k, where k is the composite sample size. We discuss the one-sample problem and the two-sample problem separately in the following sections.

6.9.1 One-Sample Tests

When composite sample data are collected to evaluate the status of a population in terms of its mean, there are two possibilities. First, there is no prior knowledge or standard as to what the mean is anticipated or supposed to be. Second, there is a statutory standard for comparison, and the field data are to be compared against this standard value of the mean. For instance, it may be stipulated that the concentration of lead in drinking water should not exceed a certain level for the water to be considered safe. In the latter situation, data are collected on n composite samples of size k each, and the composite sample mean is used to test the hypothesis that the population mean does not exceed the standard.

Suppose the n composite sample values are denoted by $Y_1, \ldots, Y_n$, and the composite sample mean and mean square are $\bar{Y}_n$ and S_y^2, respectively. Suppose the standard for the population mean is expressed in terms of a value μ_0 so that the following hypothesis is of interest:

$$\mathrm{H}_0 : \mu_x \leq \mu_0$$

against the alternative

$$\mathrm{H}_1 : \mu_x > \mu_0.$$

The most powerful test for this hypothesis is obtained under the assumption that the composite sample values follow a normal distribution with the mean μ_x and the variance $\sigma_y^2 = \sigma_x^2/k$. The test statistic is given by

$$t = \frac{\bar{Y}_n - \mu_0}{S_y/\sqrt{n}}.$$

The test statistic t follows the Student's t distribution with $n-1$ degrees of freedom. The null hypothesis is accepted if

$$t \leq t_{n-1,1-\alpha},$$

where α is the significance level and $t_{n-1,1-\alpha}$ is the critical value for the Student's t distribution with $n-1$ degrees of freedom at the significance level α.

6.9.2 Two-Sample Tests

When two sets of composite samples are available and it is desired to compare the means of the populations these sets represent, then we begin by computing the sample statistics $\bar{Y}_1$, $\bar{Y}_2$, S_1^2, and S_2^2. Let n_1 and n_2 be the respective number of composite samples in two sets. Suppose the hypothesis to test is

$$\mathrm{H}_0 : \mu_1 - \mu_2 \leq 0$$

against the alternative

$$\mathrm{H}_0 : \mu_1 - \mu_2 > 0.$$

The test statistic corresponding to the most powerful test for the null hypothesis is given by

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma}_{\text{pooled}}/\sqrt{1/n_1 + 1/n_2}},$$

where t follows the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom,

$$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\,\hat{\sigma}_1^2 + (n_2 - 1)\,\hat{\sigma}_2^2}{n_1 + n_2 - 2}},$$

$\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ being the variance estimates obtained from the samples of respective sizes n_1 and n_2.

The test is carried out by comparing the test statistic against the appropriate critical value $t_{n_1+n_2-2,1-\alpha}$ of the Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Often the null hypothesis may be

$$\mathrm{H}_0 : \mu_1 - \mu_2 \leq c_0$$

against the alternative

$$H_0 : \mu_1 - \mu_2 > c_0.$$

In this case, the test statistic is also adjusted for c_0 and is given by

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - c_0}{\sigma_{\text{pooled}}/\sqrt{1/n_1 + 1/n_2}}.$$

The composited value of the test statistic is compared with the critical value of the Student's t distribution, and the null hypothesis is accepted or rejected accordingly. Tests for one-sided hypotheses can be easily developed by replacing $1 - \alpha$ by $1 - \alpha/2$ in the defining equations for the critical values of the null distributions of the respective test statistics.

6.10 Applications

6.10.1 Comparison of Arithmetic Averages of Soil pH Values with the pH Values of Composite Samples

Peech (1965) stated that the relationship between soil pH and base saturation (BS) level is linear in the acid range of New York soils. The BS level of a composite sample made up of cores with variable BS levels can be expected to equal the average of the BS levels of the cores. Thus the observed pH of the composite sample should be similar to the arithmetic average of the pH levels of the individual cores if the relationship between pH and BS is indeed linear. Baker et al. (1981) made comparisons between arithmetically averaged soil pH values and pH values of composited soil samples (Table 6.2).

6.10.2 Comparison of Random and Composite Sampling Methods for the Estimation of Fat Contents of Bulk Milk Supplies

Milk supplies for manufacturing in Ireland are paid for generally on the basis of fat content. The fat content is determined on a composite sample. In a report on milk sampling in the Irish Dairy Industry it was recommended that the optimum sampling and testing frequency should be determined to establish supply quality with an acceptable precision. Sampling and testing schemes of milk supplies vary over countries, but little information is available on their precision. Connolly and O'Connor (1981) report the results of an experiment where 61 herd milk supplies in 3 different creamery locations were sampled during the trials. Samples were analyzed for fat, and in addition each one was used in the formation of a composite. Fat content was determined by the Milk Tester Automatic. Fat percentage can be

Table 6.2 Comparisons of pH values of composited samples (consisting of equal weights of four replicates) with arithmetic averages of individual pH values

			Relative lime rate			
Soil type	Subgroup	Sample	0	1	2	4
Sultan sil	Aquic	Average	5.44	5.63	5.66	5.77
	Xerofluvent	Composite	5.43	5.65	5.67	5.75
Puyallup sal	Fluventic	Average	5.04	5.36	5.52	6.03
	Haploxeroll	Composite	5.06	5.37	5.51	6.1
Kitsap ℓ	Dystric	Average	5.36	5.65	5.91	6.16
	Xerofluvent	Composite	5.36	5.67	5.89	6.17
Nisqually	Pachic	Average	4.76	5.40	5.63	6.19
ℓsa	Xerofluvent	Composite	4.81	5.43	5.64	6.26
Alderwood ℓ	Dystric	Average	5.18	5.51	5.76	6.03
	Entic Durochrept	Composite	5.20	5.53	5.71	6.00
Norma sil	Flaguentic	Average	4.72	4.97	5.34	5.67
	Humaquept	Composite	4.72	4.96	5.34	5.70
Buckley ℓ	Typic	Average	5.36	5.60	5.86	6.50
	Humaquept	Composite	4.72	4.96	5.34	5.70

Source: Baker et al. (1981)

Table 6.3 Comparison of composite and yield-weighted estimates of fat percentage for three locations

	Location		
	A	B	C
Composite fat (%)	3.729	3.700	3.576
Weighted fat (%)	3.788	3.681	3.576

Source: Connolly and O'Connor (1981)

estimated by composite sampling method or, if information is available on every collection, by a weighted mean of the fat percentage of each collection, weighted by the milk yield in the collection.

In the experimental data, the fat percentage was estimated by the weighted average and compared with the composite estimate. The results are summarized in Table 6.3.

6.10.3 Optimization of Sampling for the Determination of Mean Radium-226 Concentration in Surface Soil

The US Environmental Protection Agency (EPA) has certain standards for contamination of soil with ^{226}Ra. These standards (US EPA, 1983) require that the average ^{226}Ra concentration over a 100 m^2 area not exceed 5 p Ci/g above background in the top 15 cm of soil. The normal background level of ^{226}Ra is typically 1–2 p Ci/g (see Myrick et al., 1983). Determination of compliance at a reasonable cost requires

extremely efficient surveying techniques, since there are thousands of potentially contaminated sites to be examined.

Williams et al. (1989) analyzed data for five sites. In this study, soil samples are collected from sites by three different methods: 10-composite sampling, 10-composite sampling, and individual or post-hole sampling. A 10-composite sample consists of 10 aliquots of soil weighing ∼50 g each and taken at approximately uniformly spaced points over the site. A 20-composite sample consists of 20 aliquots weighing ∼25 g each, collected at approximately uniformly spaced points over the entire site. An individual sample is collected with a post-hole digger from an area of about 500 cm^2, but only a small portion of the collected and mixed soil (roughly 500 g) is taken for the sample. As nearly as practical, all samples were taken uniformly at depths from 0 to 15 cm. Results are summarized in Table 6.4.

Table 6.4 Summary of soil sample data

Site	Area (m^2)	Type of sample	Mean (p Ci/g)	VEM[a]
A	15	Individual	14.2	14.2
		20-Composite	14.2	
		10-Composite	13.7	
B	130	Individual	7.3	9.0
		20-Composite	9.6	
		10-Composite	10.2	
C	270	Individual	19.3	17.3
		20-Composite	15.2	
		10-Composite	10.1	
D	30	Individual	7.9	7.8
		20-Composite	7.7	
		10-Composite	7.8	
E	200	Individual	51.3	57.9
		20-Composite	64.1	
		10-Composite	58.3	

[a] Unbiased estimate of the mean ^{226}Ra concentration
Source: Williams et al. (1989)

Chapter 7
Composite Sampling with Random Weights

7.1 Introduction

Composite samples are formed by physically mixing aliquots of individual samples. If aliquots of equal volumes are used, then the composite sample values are simply the arithmetic averages of individual sample values. If aliquots are not equal in volume, then statistical techniques need to be adjusted to account for unequal volumes. In this case, the composite sample values are weighted averages of individual sample values. The weights associated with individual sample values are proportional to the volumes of aliquots of the respective individual samples. If volumes of the aliquots are known, so that the weights also are known, then they can be treated as constants. The statistical properties of the composite sample values follow rather easily from those of the individual sample values. However, it can be the case that the volumes of the aliquots are unknown because they have resulted from a random process. In such a case, the statistical properties of the composite sample values depend not only on those of the individual sample values but also on those of the volumes (or, equivalently, of the weights) and are affected by the interrelationships between the individual sample values and the volumes of the aliquots.

The following examples illustrate how random weights can result from random processes beyond the experimenter's control:

A. The weights are generated by a random mechanism.

The effluents of several plants in an industrial zone are to be sampled to make a measurement on the concentration of a particular contaminant. These plants discharge their effluents directly into a stream. Samples are collected from the stream, and then analytical measurements are made on these samples.

Since the samples are not collected separately from individual plants, the experimenter has no control over the volumes of effluents from different individual plants that contribute to the samples collected from the stream. Moreover, these volumes depend on the production processes of the plants; they may vary from time to time, resulting in volumes varying randomly from sample to sample. It is therefore necessary to treat the weights as random variables when analyzing the composite sample measurements.

B. The weights are fixed by the experimenter, but the compositing apparatus fails to homogenize the composite sample.

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_7, © Springer Science+Business Media, LLC 2011

At a hazardous waste site, an estimate of the average concentration level of a certain contaminant is required in order to decide as to whether remediation is called for. Four cores of soil, all equal in size and hence also in volume, are extracted from four random locations on the site. These four cores are then blended to form a composite sample. The instrument used to measure the concentration of the contaminant can only test an amount of soil which is significantly smaller than the amount of soil that the compositing apparatus can composite. It is therefore necessary to subsample the composite sample for making an analytical measurement.

Suppose that blending is inaccurate; that is, blending may leave some clumps of soil unbroken or it may not mix the soil thoroughly. Then the volumes of aliquots that come from different individual samples in one subsample of the composite will be different from those in another subsample. It is then necessary to treat the weights as random variables (though perhaps not observable) when making measurements on subsamples of the heterogeneous composite sample.

C. The weights are generated by the sampling protocol.

It is required to estimate the population density of a particular aquatic community in a large body of water. A net is towed through water at different locations for a fixed amount of time at a constant speed. The amount of water filtered during every sampling episode depends on many random conditions such as the wind velocity, wave height, and water currents. The fraction of each individual sample in the composite sample is then a random quantity, thus giving rise to random weights.

7.2 Expected Value, Variance, and Covariance of Bilinear Random Forms

Lemma 7.2.1 *(Rohde, 1976) Let* $\boldsymbol{x}$, $\boldsymbol{w}$, *and* $\boldsymbol{u}$ *be random* k*-vectors and let* $\boldsymbol{x}$ *be stochastically independent of both* $\boldsymbol{w}$ *and* $\boldsymbol{u}$. *Then*

$$E\left[\boldsymbol{w}'\boldsymbol{x}\right] = \boldsymbol{\mu}_w'\boldsymbol{\mu}_x, \tag{7.1}$$

$$\mathrm{Var}\left[\boldsymbol{w}'\boldsymbol{x}\right] = \boldsymbol{\mu}_x'\boldsymbol{\Sigma}_w\boldsymbol{\mu}_x + \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w + \mathrm{tr}\left[\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_x\right], \tag{7.2}$$

and

$$\mathrm{cov}\left[\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}\right] = \boldsymbol{\mu}_x'\boldsymbol{\Gamma}_{w,u}\boldsymbol{\mu}_x + \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_x\boldsymbol{\mu}_u + \mathrm{tr}\left[\boldsymbol{\Gamma}_{w,u}\boldsymbol{\Sigma}_x\right]. \tag{7.3}$$

Proof The proof is a simple application of conditional expectation. Note that

$$E\left[\boldsymbol{w}'\boldsymbol{x}\right] = E\left[E\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{w}\right)\right]$$

and therefore

$E\left[\boldsymbol{w}'\boldsymbol{x}\right] = E\left[\boldsymbol{\mu}_x'\boldsymbol{w}\right]$ by Lemma 6.2.1 and since $\boldsymbol{x}$ is stochastically independent of $\boldsymbol{w}$

$= \boldsymbol{\mu}_x' \boldsymbol{\mu}_w'$ again by Lemma 6.2.1
$= \boldsymbol{\mu}_w' \boldsymbol{\mu}_x'$ and hence (7.1) is proved.
Next note that

$$\text{Var}\left[\boldsymbol{w}'\boldsymbol{x}\right] = \text{Var}\left[E\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right)\right] + E\left[\text{Var}\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right)\right].$$

First consider

$$E\left[\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right] = \boldsymbol{\mu}_w'\boldsymbol{x} \text{ by Lemma 6.2.1}$$

and also

$$\text{Var}\left[\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right] = \boldsymbol{x}'\boldsymbol{\Sigma}_w\boldsymbol{x} \quad \text{by Lemma 6.2.2.}$$

Therefore

$$\text{Var}\left[E\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right)\right] = \text{Var}\left(\boldsymbol{\mu}_w'\boldsymbol{x}\right) = \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_x\boldsymbol{\mu}_w$$

by Lemma 6.2.2 and also

$$E\left[\text{Var}\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right)\right] = E\left[\boldsymbol{x}'\boldsymbol{\Sigma}_w\boldsymbol{x}\right] = \boldsymbol{\mu}_x'\boldsymbol{\Sigma}_w\boldsymbol{\mu}_x + \text{tr}\left[\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_x\right]$$

by Lemma 6.2.4.
This proves (7.2).
Finally, note that

$$\begin{aligned}\text{cov}\left[\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}\right] &= \text{cov}\left[E\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right), E\left(\boldsymbol{u}'\boldsymbol{x}|\boldsymbol{x}\right)\right] \\ &\quad + E\left[\text{cov}\left(\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}|\boldsymbol{x}\right)\right].\end{aligned}$$

By Lemma 6.2.1, we have

$$E\left[\left(\boldsymbol{w}'\boldsymbol{x}|\boldsymbol{x}\right)\right] = \boldsymbol{\mu}_w'\boldsymbol{x}$$

and

$$E\left[\left(\boldsymbol{u}'\boldsymbol{x}|\boldsymbol{x}\right)\right] = \boldsymbol{\mu}_u'\boldsymbol{x}$$

and by Lemma 6.2.3,

$$\text{cov}\left(\boldsymbol{\mu}_w'\boldsymbol{x}, \boldsymbol{\mu}_u'\boldsymbol{x}\right) = \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_x\boldsymbol{\mu}_u'.$$

Also,

$$\text{cov}\left(\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}|\boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\Gamma}_{w,u}\boldsymbol{x}$$

and hence

$$E\left[\operatorname{cov}\left(\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}|\boldsymbol{x}\right)\right] = E\left[\boldsymbol{x}'\boldsymbol{\Gamma}_{w,u}\boldsymbol{x}\right]$$
$$= \boldsymbol{\mu}_x'\boldsymbol{\Gamma}_{w,u}\boldsymbol{\mu}_x + \operatorname{tr}\left[\boldsymbol{\Gamma}_{w,u}\boldsymbol{\Sigma}_x\right] \quad \text{by Lemma 6.2.4.}$$

Thus

$$\operatorname{cov}\left[\boldsymbol{w}'\boldsymbol{x}, \boldsymbol{u}'\boldsymbol{x}\right] = \boldsymbol{\mu}_w'\boldsymbol{\Sigma}_x\boldsymbol{\mu}_u + \boldsymbol{\mu}_x\boldsymbol{\Gamma}_{w,u}\boldsymbol{\mu}_x + \operatorname{tr}(\boldsymbol{\Gamma}_{w,u}\boldsymbol{\Sigma}_x),$$

and therefore (7.3) is proved. This completes the proof of the lemma.

Elder et al. (1980) relax the assumption that $\boldsymbol{x}$ is stochastically independent of both $\boldsymbol{w}$ and $\boldsymbol{u}$. They make the following assumptions:

$E[\boldsymbol{w}|\boldsymbol{x}] = \boldsymbol{\mu}_w$ for all $\boldsymbol{x}$ (expectation independence);

$\operatorname{Var}[\boldsymbol{w}|\boldsymbol{x}] = \boldsymbol{\Sigma}_w$ for all $\boldsymbol{x}$ (variance independence);

and

$\operatorname{cov}[\boldsymbol{w}, \boldsymbol{u}|\boldsymbol{x}] = \boldsymbol{\Gamma}_{w,u}$ for all $\boldsymbol{x}$ (covariance independence).

Under these weaker assumptions, they prove that Lemmas 7.2.1, 7.2.2, and 7.2.3 hold.

Since the weights add up to unity, we have $\boldsymbol{\imath}'\boldsymbol{w} = 1$, where $\boldsymbol{\imath} = (1, 1, \ldots, 1)'$ is a k-vector. Taking expectation and noting that the expectation of a constant (i.e., a degenerate random variable) is same as its only possible value, we have $E\left[\boldsymbol{\imath}'\boldsymbol{w}\right] = 1$. Also, noting that the variance of a degenerate random variable is zero, we obtain $\operatorname{Var}\left[\boldsymbol{\imath}'\boldsymbol{w}\right] = \boldsymbol{\imath}'\boldsymbol{\Sigma}_w\boldsymbol{\imath} = 0$ (by Lemma 6.2.2).

Elder (1977) points out that the variances of the weights are bounded. Clearly $0 \leq W_i \leq 1 (i = 1, \ldots, k)$ and hence $0 \leq W_i^2 \leq W_i \leq 1$ and so $E(W_i^2) \leq E(W_i)$. Therefore

$$\begin{aligned}\operatorname{Var}[W_i] &= E\left[W_i^2\right] - [E(W_i)]^2 \\ &\leq E(W_i) - [E(W_i)]^2 \\ &= E[W_i][1 - E(W_i)], \quad i = 1, \ldots, k.\end{aligned}$$

7.3 Models for the Weights

The mean and the variance of the composite sample estimator of the population mean depend on the distributions of the weights and of the individual sample values. Two statistical distributions for the weights are studied. Rohde (1976) argues for the Dirichlet distribution, and Elder (1977) formulates a model for the weights which gives the multivariate hypergeometric distribution. This model is the same as that originally used by Brown and Fisher (1972). Elder points out that both this distribution and the Dirichlet distribution converge to a singular multivariate normal distribution under suitable conditions. He suggests this as a reasonable approximation

in many cases because of the physical averaging that occurs due to blending of composite samples.

The basic idea in the literature on random weights is that if reasonable models are set up for the weights in the form of assumptions on their first two moments $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_w$, then, through the expectation and variance of each composite sample value and hence of the composite sample estimator of the population mean can be found by Lemma 6.3.1.

7.3.1 Assumptions on the First Two Moments

If the k individual samples contributing to a single composite sample have the same size, then it is reasonable to assume, by symmetry of the compositing proportions, that such proportions have a common expectation, a common variance, and that every pair of proportions is equally correlated. Thus,

$$\boldsymbol{\mu}_w = \frac{1}{k}\mathbf{1}_k, \tag{7.4}$$

$$\boldsymbol{\Sigma}_w = \sigma_w^2 \left[\frac{k}{k-1}\boldsymbol{I}_k - \frac{1}{1-k}\boldsymbol{J}_k \right], \tag{7.5}$$

where $\boldsymbol{I}_k$ is the identity matrix of order k and $\boldsymbol{J}_k$ is the square matrix of order k with all elements equal to 1.

These two assumptions are widely used in the literature. Elder et al. (1980) refer to them as characterizing an "unbiased" compositing/subsampling procedure.

7.3.2 Distributional Assumptions

It is possible to go further and assume a probability distribution for the weights. As noted earlier, the weights, being proportions, must satisfy

$$\boldsymbol{\mu}_w' \mathbf{1}_k = 1$$

and

$$\boldsymbol{\Sigma}_w \mathbf{1}_k = \mathbf{0}.$$

That is, the distribution of weights must be singular.

There are a number of distributions which may be appropriate for the weights. Rohde (1976) advocates the use of the Dirichlet distribution. In particular, he recommends the one-parameter Dirichlet distribution

$$W_{ji} \sim D(\lambda \mathbf{1}),$$

which implies (7.4) and (7.5). Rohde points out that Dirichlet is just one of the distributions which satisfy (7.4) and (7.5), but he also gives a physical motivation for the choice of the Dirichlet distribution using a theorem by Fabius (1973).

Theorem *(Fabius, 1973) Let*

$$(W_1, \ldots, W_k) \sim D(\lambda_1, \ldots, \lambda_k),$$

and let

$$U_1 = \frac{W_1}{\sum_{i=1}^{k-1} W_i}, \ldots, U_{k-1} = \frac{W_{k-1}}{\sum_{i=1}^{k-1} W_i}.$$

If $(U_1, \ldots, U_{k-1})$ *is independent of* W_k*, then*

$$(U_1, \ldots, U_{k-1}) \sim D(\lambda_1^*, \ldots, \lambda_{k-1}^*),$$

where $\lambda_1^*, \ldots, \lambda_{k-1}^*$ *are suitably defined.*

The physical interpretation of this statistical property is given by Rohde (1976) as follows. Let $W_1, \ldots, W_k$ represent the proportions of the k individual samples in a composite sample. If the kth individual sample were deleted from the composite, then the new proportions would be

$$U_1 = \frac{W_1}{1 - W_k}, \ldots, U_{k-1} = \frac{W_{k-1}}{1 - W_k}.$$

Note that

$$1 - W_k = \sum_{i=1}^{k-1} W_i.$$

In order for the model for proportions not to be influenced unduly by the addition of another sample to the composite, it seems reasonable to assume that $U_1, \ldots, U_{k-1}$ are independent of W_k. Note that $U_i / U_j = W_i / W_j$; that is, the relative proportions between the Us are the same as between the W's. As Rohde points out, this property characterizes Dirichlet distribution. A failure of the Dirichlet distribution to fit implies a dependence between the proportions of individual samples in a composite sample and another sample to be added to the composite sample.

Brown and Fisher (1972) propose the multivariate hypergeometric distribution, based on a physical model for discrete materials or pelletized products, such as grains, pebbles, bales of wool. Elder (1977) shows that both the Dirichlet and the multivariate hypergeometric distributions converge to a singular multivariate normal distribution under suitable asymptotic conditions.

7.4 The Model for Composite Sample Measurements

Let the composite sample measurement be written as

$$Y = W_1 X_1 + \cdots + W_k X_k + \epsilon, \tag{7.6}$$

where $\{W_i\}$ and $\{X_i\}$ are defined in Chapter 6 (see Section 6.2.1) and ϵ is the measurement error. Assume that measurement error is negligible and models the composite sample value (or measurement) by

$$Y = W_1 X_1 + \cdots + W_k X_k = \boldsymbol{w}'\boldsymbol{x}.$$

Assume that the weights $\{W_i\}$ are stochastically independent of the individual sample values $\{X_i\}$. Then by Lemma 6.2.1, we have

$$E[Y] = \boldsymbol{\mu}_w' \boldsymbol{\mu}_x,$$
$$\text{Var}[Y] = \boldsymbol{\mu}_w' \boldsymbol{\Sigma}_x \boldsymbol{\mu}_w + \boldsymbol{\mu}_x' \boldsymbol{\Sigma}_w \boldsymbol{\mu}_x + \text{tr}[\boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_x].$$

We now consider this model in different situations. For example, we consider several composite samples, subsampling of composite samples, and also some probability distributions for the weights.

7.4.1 Subsampling a Composite Sample

Rohde (1976) develops the theory for a single composite sample formed from k individual samples. The proportions of the individual samples used to make up the composite sample are either fixed (known) or random. Elder (1977) (see also Elder et al., 1980) generalizes the theory to c composite samples, each formed from k aliquots and s subsamples taken from each composite sample. Further, t analyses or tests may be carried out on each subsample. This generalization takes into account the variability due to dividing the original individual samples into increments before compositing, the variability due to imperfect mixing of the composite samples and also due to the selection of subsamples, and also the variability of the test procedure itself, causing variation between the results of repeated measurements on the same subsample.

If a composite sample is formed by taking subsamples, also called aliquots or increments, from the k individual samples, then the fractions of the original sample values represented in the composite sample value must have the properties described earlier. If

$$Z = \boldsymbol{w}'\boldsymbol{x} \quad \text{and} \quad Y = \boldsymbol{u}'\boldsymbol{x}$$

are values of two composite samples formed from aliquots of the same k individual samples, then

$$\operatorname{cov}[Z, Y] = \boldsymbol{\mu}_x' \boldsymbol{\Gamma}_{w,u} \boldsymbol{\mu}_x + \boldsymbol{\mu}_w' \boldsymbol{\Sigma}_x \boldsymbol{\mu}_u' + \operatorname{tr}\left[\boldsymbol{\Gamma}_{w,u} \boldsymbol{\Sigma}_x\right].$$

The sum of the weights being unity, $\boldsymbol{\mu}_w' \mathbf{1} = \mathbf{1}$ and $\mathbf{1}' \boldsymbol{\Sigma}_w \mathbf{1} = 0$. In many cases, it is reasonable to assume that the random weights are exchangeable random variables, then $\boldsymbol{\mu}_w = \frac{1}{k}\mathbf{1}$ and

$$\boldsymbol{\Sigma}_w = \sigma_w^2 \left[(1-\rho)\, \boldsymbol{I}_k + \rho \boldsymbol{J}_k\right]. \tag{7.7}$$

Furthermore, since $\mathbf{1}' \boldsymbol{\Sigma}_w \mathbf{1} = 0$, we have $k\,[1 + (k-1)\rho]\,\sigma_w^2 = 0$ or $\rho = \frac{-1}{k-1}$. As Rohde (1976) points out, the symmetric Dirichlet distribution has this property.

Now suppose that the financial restrictions allow only s measurements to be made. We then construct s composite samples and compare the results with those of s measurements on individual samples. As before, let $X_1, \ldots, X_k$ be the values associated with the k individual samples. By subsampling s times from each sample we form s composite samples by combining one subsample from each individual sample. Then $Y_i = \boldsymbol{w}_i' \boldsymbol{x}$ is the value of the ith composite sample, $i = 1, 2, \ldots, s$. We compare $\bar{Y}_s = \frac{1}{s}\sum_{i=1}^{s} Y_i$ with the average $\bar{X}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$ of s randomly selected individual samples from among k. Let

$$\bar{Y}_s = \frac{1}{s}\left(\boldsymbol{w}_1' + \boldsymbol{w}_2' + \cdots + \boldsymbol{w}_s'\right)\boldsymbol{x} = \boldsymbol{v}'\boldsymbol{x}.$$

Clearly, $\bar{Y}_s$ represents a composite sample measurement, and the formulas that were derived earlier apply. Both $\bar{X}_s$ and $\bar{Y}_s$ are unbiased estimators of μ_x. Moreover,

$$\operatorname{Var}\left[\bar{X}_s\right] = \frac{1}{s^2}\mathbf{1}' \boldsymbol{\Sigma}_x \mathbf{1}.$$

To calculate the variance of $\bar{Y}_s$, we need $\boldsymbol{\Sigma}_v = \operatorname{Var}(\boldsymbol{v})$ where $\boldsymbol{v} = \frac{1}{s}\sum_{i=1}^{s} \boldsymbol{w}_i$:

$$\begin{aligned}\boldsymbol{\Sigma}_v = \operatorname{Var}(\boldsymbol{v}) &= \operatorname{Var}\left(\frac{1}{s}\sum_{i=1}^{s} \boldsymbol{w}_i\right) \\ &= \frac{1}{s^2}\sum_{i=1}^{s} \boldsymbol{\Sigma}_w = \frac{1}{s}\boldsymbol{\Sigma}_w,\end{aligned}$$

assuming independence of weights from one subsample to another. Then

$$\begin{aligned}\operatorname{Var}\left[\bar{Y}_s\right] &= \boldsymbol{\mu}_v' \boldsymbol{\Sigma}_x \boldsymbol{\mu}_v + \boldsymbol{\mu}_x' \boldsymbol{\Sigma}_v \boldsymbol{\mu}_x + \operatorname{tr}\left[\boldsymbol{\Sigma}_v \boldsymbol{\Sigma}_x\right] \\ &= \boldsymbol{\mu}_w' \boldsymbol{\Sigma}_x \boldsymbol{\mu}_w + \frac{1}{s}\boldsymbol{\mu}_x' \boldsymbol{\Sigma}_w \boldsymbol{\mu}_x + \frac{1}{s}\operatorname{tr}\left[\boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_x\right].\end{aligned}$$

Now assume $X_1, \ldots, X_k$ to be independent and identically distributed with mean μ_x and variance σ_x^2. Then $\boldsymbol{\mu}_x = \mu_x \mathbf{1}$ and $\boldsymbol{\Sigma}_x = \sigma_x^2 \boldsymbol{I}$ so that $\boldsymbol{\mu}_x' \boldsymbol{\Sigma}_w \boldsymbol{\mu}_x = 0$. It then

follows that

$$\mathrm{Var}\left[\bar{Y}_s\right] = \frac{\sigma_x^2}{k} + \frac{k\sigma_w^2\sigma_x^2}{s} = \frac{\sigma_x^2}{k}\left[1 + \frac{k^2\sigma_w^2}{s}\right]. \tag{7.8}$$

Assume $\boldsymbol{w}$ to have a Dirichlet distribution with parameter $\mathbf{1}$; then

$$\mathrm{Var}\left[\bar{Y}_s\right] = \frac{\sigma_x^2}{k}\left[1 + \frac{k-1}{s(k+1)}\right].$$

Also, then,

$$\mathrm{Var}\left[\bar{X}_s\right] = \frac{\sigma_x^2}{s}.$$

The ratio of the two variances is

$$\frac{\mathrm{Var}\left[\bar{Y}_s\right]}{\mathrm{Var}\left[\bar{X}_s\right]} = \frac{s}{k}\left[1 + \frac{k-1}{s(k+1)}\right].$$

If k is sufficiently large compared to s, then the variance of $\bar{Y}_s$ is much smaller than the variance of $\bar{X}_s$.

It seems that choosing k large in comparison to s would result in the best compositing situation. However, when a composite sample is formed, the physical mixing is often imperfect, resulting in a higher variability, affecting the optimality of the composite sample size. The optimal choice may also depend on the cost of sampling and the cost of analysis of a sample, individual, or composite.

Rohde (1976) points out that the sample variance of the composite sample mean is a biased estimator of σ_x^2. In fact

$$E\left[S_y^2\right] = E\left[\frac{1}{s-1}\sum_{i=1}^{s}(Y_i - \bar{Y}_s)^2\right] = k\sigma_w^2\sigma_x^2.$$

Rohde suggests two approaches to get an unbiased estimator of σ_x^2 from the sample variance of the composite sample measurements. If an independent estimator of σ_w^2 is available, it can be used to estimate σ_w^2 so that S_y^2 can be used, with the necessary adjustment, to obtain an unbiased estimator of σ_x^2. The other approach is to assume some model, such as the Dirichlet distribution, which reduces the number of parameters that need to be estimated.

Elder (1977) points out that Rohde makes the assumption of independence of the weights W_i in different composite samples. When subsamples are repeatedly drawn from the same composite sample for testing, then the resulting composite samples may not have independent weights. In this case, the symmetric Dirichlet distribution

is not appropriate. However, as the number of subsamples s increases, the weights become independent.

Alternatively, using the upper bound for σ_w^2 given by Elder et al. (1980),

$$0 \le \sigma_w^2 \le \frac{k-1}{k^2},$$

we see that

$$\frac{\sigma_x^2}{k} \le \mathrm{Var}\left[\bar{Y}_s\right] \le \frac{\sigma_x^2}{k}\left[1 + \frac{k-1}{s}\right].$$

These bounds on $\mathrm{Var}\left[\bar{Y}_s\right]$ lend themselves to some interesting conclusions about the variability of the composited estimator in the presence of random weights:

(a) Neglecting the randomness of the weights may be misleading in assessing the variance of $\bar{Y}_s$ which is never smaller than the corresponding variance in the "fixed weights" case, namely $\frac{\sigma_x^2}{k}$.
(b) The effect, on σ_w^2, of increasing k within any finite interval is contradictory. On the one hand, since a composite is generally harder to homogenize the larger it is, σ_w^2 may increase with an increasing k; on the other hand, the larger the composite, the smaller each proportion and hence the smaller σ_w^2 will be (in fact, as $k \to \infty$, $\sigma_w^2 \to 0$). Therefore, in the random weights case, it is difficult to predict how $\mathrm{Var}\left[\bar{Y}_s\right]$ varies with the composite sample size k.

7.4.2 Several Composite Samples

If there are $m = nk$ individual samples, then n composite samples, each of size k, can be formed by randomly forming n subsets of size k from the collection of nk samples. Under the assumption that the individual samples are stochastically independent, the composite samples are also stochastically independent, since they comprise disjoint sets of individual samples. Thus a composite sample value is

$$Y_j = \boldsymbol{w}_j' \boldsymbol{x}_j, \quad j = 1, \ldots, n.$$

Assuming that the weights are independent, we find that the composite sample estimator $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ is an unbiased estimator of $\boldsymbol{\mu}_w' \boldsymbol{\mu}_x$. If we further assume that the $\{X_{ji}\}_{i=1}^{k}$ are independent and identically distributed, we have

$$\boldsymbol{\mu}_x = \mu_x \mathbf{1}_m \quad \text{and} \quad \boldsymbol{\Sigma}_x = \sigma_x^2 \boldsymbol{I}_m.$$

Under this assumption, it is easy to see that

$$
\begin{aligned}
E\left[\bar{Y}_n\right] &= \mu_x, \\
\mathrm{Var}\left[\bar{Y}_n\right] &= \frac{1}{nk}\sigma_x^2 + \frac{k}{n}\sigma_x^2\sigma_w^2 \\
&= \frac{\sigma_x^2}{nk}[1 + k^2\sigma_w^2].
\end{aligned}
$$

Again, using the upper bound on σ_w^2 given by Elder et al. (1980), namely,

$$
0 \le \sigma_w^2 \le \frac{k-1}{k^2},
$$

we find that

$$
\frac{\sigma_x^2}{nk} \le \mathrm{Var}\left[\bar{Y}_n\right] \le \frac{\sigma_x^2}{n}.
$$

The inequality on the left implies that the composite sample estimator of μ_x, due to the randomness of the weights, cannot have a variance smaller than that of the average of the nk individual sample values. On the other hand, since composite sampling involves exactly n measurements, the inequality on the right implies that the composite sampling, in spite of random weights, achieves a smaller variance for the estimator of μ_x compared to the average of n individual sample values, i.e., one individual sample from every set of k.

7.4.3 Subsampling of Several Composite Samples

Now we consider a combination of the two preceding methods. Here, with a collection of $m = nk$ individual samples, we form n subsets of k individual samples each and then take s subsamples from each composite sample. The foregoing results can be combined to obtain the following. The composite sample mean, $\bar{Y}$, is an unbiased estimator of the population mean μ_x with the variance

$$
\mathrm{Var}\left[\bar{Y}\right] = \frac{\sigma_x^2}{nk}\left[1 + \frac{k^2}{s}\sigma_w^2\right].
$$

Using the upper bound given by Elder et al. (1980), we see that

$$
\frac{\sigma_x^2}{nk} \le \mathrm{Var}\left[\bar{Y}\right] \le \frac{\sigma_x^2}{nk}\left[1 + \frac{k-1}{s}\right].
$$

While the inequality on the left yields the same interpretation as given in the preceding section, writing the upper bound on $\mathrm{Var}\left[\bar{Y}\right]$ as

$$\mathrm{Var}\big[\bar{Y}\big] \le \frac{\sigma^2}{ns}\left[1 + \frac{s-1}{k}\right]$$

it can be seen that, in the presence of random weights, the composite sample estimator cannot be worse, in terms of the variance, than the comparable "uncomposited" estimator $\bar{X} = \frac{1}{ns}\sum_{i=1}^{ns} X_i$ as long as $s = 1$, i.e., one subsample is selected from each composite. If $s > 1$, the composited estimator may have a larger variance than the uncomposited one, and some knowledge of σ_w^2 is necessary to assess the effect of compositing.

7.4.4 Measurement Error

In addition to the assumptions of model discussed in this chapter, Elder et al. (1980) make the assumption that there may be error in measurement. Suppose n composites, each of size k, are formed as earlier. From each composite sample, select s subsamples and run t analyses on each subsample. Let the result of the ℓth test on the jth subsample from the ith composite sample be denoted by $Y_{ij\ell}$. Then

$$Y_{ij\ell} = \boldsymbol{w}_{ij}'\boldsymbol{x}_{ij} + \epsilon_{ij\ell},$$
$$i = 1, \ldots, n; \quad j = 1, \ldots, s; \quad \ell = 1, \ldots, t;$$

where $\boldsymbol{x}_{ij}$ depends on the composite sample index i. This model incorporates the variability generated by dividing a sample into s aliquots; this is called within-increment variability. The variability of the results of repeated testing on the same sample (subsample) results in the additional term $\epsilon_{ij\ell}$. The total of nst measurements are assumed to have independent and identically distributed errors with

$$E\big[\epsilon_{ij\ell}\big] = 0, \quad \mathrm{Var}\big[\epsilon_{ij\ell}\big] = \sigma_t^2,$$
$$i = 1, \ldots, n; \; j = 1, \ldots, s; \; \ell = 1, \ldots, t.$$

The technique used to find the mean and the variance of the composite sample estimator $\bar{Y}$ is similar to that used earlier. The structure on the weights requires

$$E\big[\boldsymbol{w}_{ij}\big] = E\big[(W_{ij1}, \ldots, W_{ijk})\big]' = \frac{1}{k}\mathbf{1};$$
$$\mathrm{Var}\big[\boldsymbol{w}_{ij}\big] = \sigma_w^2 \boldsymbol{I}.$$

It is then easy to show that

$$\bar{Y} = \frac{1}{nst}\sum_{i=1}^{n}\sum_{j=1}^{s}\sum_{\ell=1}^{t} Y_{ij\ell}$$

is an unbiased estimator of μ_x, with variance

$$\mathrm{Var}\left[\bar{Y}\right] = \frac{\sigma_x^2}{nk} + \frac{k\sigma_w^2\sigma_x^2}{ns} + \frac{\sigma_t^2}{nst}.$$

If the aliquots used to form the composites are either made up of discrete (equal-sized) pieces or are divided into equal-sized pieces, then a hypergeometric distribution can be used to model the number of pieces from different increments that make up the composite samples that are further broken down into subsamples (Brown and Fisher, 1972; see also Elder, 1977). Assume each increment to be composed of g equal-sized pieces and that each subsample is formed from a random selection of G pieces. Then each subsample is a composite sample with weights $W_{j\ell} = g_{j\ell}/G$, where j is the index of the subsample, ℓ is the index of the increment, and $g_{j\ell}$ is the number of pieces from increment ℓ in the subsample j. From the hypergeometric model,

$$E[g_{j\ell}] = G/k$$

and

$$\mathrm{Var}\left[g_{j\ell}\right] = G(kg - G)(k-1)/\left[k^2(kg-1)\right].$$

Then

$$\sigma_w^2 = \frac{(k-1)(kg-G)}{k^2G(kg-1)}.$$

Taking limit as both g and G increase to infinity, while keeping g/G constant, results in the distribution of the weights approaching a singular normal distribution; see Elder (1977).

The optimal values of k, n, and t depend on many factors. Elder (1977) discusses this problem, pointing out that closed-form solutions do not exist. Generally, if the cost of sampling is relatively small, then k can be much larger than n. If the within-increment variability is small, then s is small.

We conclude the discussion of random weights with an observation that, under suitable assumptions, the problem of random weights in composite sampling can be handled with appropriate statistical methods, and valid conclusions can be drawn regarding the unbiasedness of the composite sample estimator of the population mean. The precision of this estimator can also be estimated and compared with an uncomposited estimator.

7.5 Applications

The literature on composite sample techniques for estimation of the population parameters of a continuous measurement includes a number of applications of composite sampling with random weights to a variety of environmental problems. Some important applications are discussed here.

7.5.1 Sampling Frequency and Comparison of Grab and Composite Sampling Programs for Effluents

The impetus behind the development of the theory of grab and composite sampling was to enable the Illinois Environmental Protection Agency to evaluate alternative sampling requirements for dischargers to the state's waterways, as well as to optimize its own sampling programs. Janardan and Schaeffer mention that the law required that dischargers report the results of 24-h composite samples for the monthly maximum and mean. The agency was proposing changes in the averaging requirement in order to provide a more flexible standard. Before any changes were made, the influence of different sampling procedures on the demonstration of compliance with the regulations had to be evaluated. To make this assessment, grab and composite samples were generated by simulation of distributions observed in actual data.

Some results of the simulation for the normal distribution are given in Table 7.1. These data clearly show that the information content of a series of grab samples is better than that of the same number of composites. The opposite is true when only single samples of each type are compared. Further, the results show that the sample variance of the composites underestimates the true variance in accordance with theory.

Table 7.1 Summary statistics for grab and 24-h composite samples generated from normal distribution ($\mu = 18.17, \sigma = 8.47$)

Statistics	Grab samples	Composite samples	Monthly means (G)	Monthly means (C)
Sample mean	18.52	18.19	18.52	18.18
Sample variance	70.83	4.34	3.06	0.1645
Expected variance	71.74	4.22	3.59	3.0509

7.5.2 Theoretical Comparison of Grab and Composite Sampling Programs

Schaeffer and Janardan develop the theoretical results for grab and composite sampling programs. They use the data from "Waste water sampling methodologies and

flow measurement techniques" by Harris and Keffer (1974) USEPA Region VII Field Investigation Study. Table 7.2 gives grab sampling statistics while Table 7.3 gives composite sampling statistics.

Table 7.2 Grab sampling statistics

	5-Day biochemical oxygen demand	Chemical oxygen demand	Solids
Sample size (m)	32	32	32
Sample mean $\bar{X}$	207.38	233.65	207.38
Sample variance $\hat{\sigma}_x^2$	1326.56	9020.04	21,136.52

Data are taken from the study "Waste water sampling methodologies and flow measurements techniques" by Harris and Keffer (1974) (table XVI)

Table 7.3 Composite samples based on (sample) volume proportional to total flow since last sample[a]

	5-Day biochemical oxygen demand	Chemical oxygen demand	Solids
Sample mean	208.97	240.47	238.52
Sample variance	1186.11	5518.54	5047.82
Estimate of σ^2	9488.88	44,148.32	40,382.56

This table is constructed from data of table XVI of "Waste water sampling methodologies and flow measurement techniques" by Harris and Keffer (1974)
[a] The data on proportions W_i in column 2 of this table are by Huibregtse and Moser (1976) (table 2.1, p. 5)

7.5.3 Grab vs. Composite Sampling: A Primer for the Manager and Engineer

Schaeffer, Kerster, and Janardan report the result of a simulation study comparing grab and composite samples. As they observed, in no case did the simulator results differ significantly from the theoretical expectations. All composite runs returned about the same means as the grab samples. The composites show the expected loss of information relative to grabs. The results are summarized in Table 7.4.

7.5.4 Composite Samples Overestimate Waste Loads

Wastewater treatment plant performance is monitored by the collection and analysis of samples from the process stream for physical, chemical, and microbiological constituents. Samples may be broadly classified as "grab" or "composite." Grab samples represent the composition of the flow at a given instant in time, irrespective of the flow volume. Composite samples represent an average composition in the flow over time (usually 24 h) and may or may not be proportional to the flow. Flow

Table 7.4 Summary statistics for grab and composite samples

Sample		Obtained mean[a]	Sample variance[b]	Corrected variance[c]	Sw^2	Distribution shape (C/G)
A.	BOD (mg/l)					
	Grab	18.52	70.83	71.74		Normal
	Composite	18.19	4.34	6.13	7.18 E-4	
	Monthly mean grab	18.52	3.06	3.59		
	Monthly mean composite	18.18	0.16	4.42		
B.	BOD (mg/l)					
	Grab	2.30	1.29	1.22		Lognormal
	Composite	2.33	7.77 E-2	1.01E-1	5.19 E-4	
	Monthly mean grab	2.30	5.73 E-2	6.08 E-2		
	Monthly mean composite	2.33	3.31 E-3	7.89 E-2		
C.	Industrial BOD (mg/l)					
	Grab	74.35	3315.61	2570.49		Gamma
	Composite	77.52	167.71	218.76	5.29 E-4	
	Monthly mean grab	74.01	69.33	128.53		
	Monthly mean composite	77.52	8.07	170.26		
D.	Lagoon BOD (mg/l)					
	Grab	28.89	822.14	834.50		Gamma
	Composite	27.36	37.59	52.05	6.68 E-4	
	Monthly mean grab	28.49	28.59	41.73		
	Monthly mean composite	27.36	1.51	38.31		
E.	Lagoon suspended solids (mg/l)					
	Grab	33.92	627.55	603.23		Gamma
	Composite	35.62	38.68	62.83	1.08 E-3	
	Monthly mean grab	33.92	39.53	30.16		
	Monthly mean composite	35.62	3.09	39.89		
F.	Industrial zinc (mg/l)					
	Grab	0.22	1.73 E-2	1.73 E-2		Beta
	Composite	0.22	0.98 E-3	1.22 E-3	4.30	
	Monthly mean grab	0.22	1.00 E-3	8.65 E-4		
	Monthly mean composite	0.22	5.01 E-5	9.92 E-4		

[a] For daily sample means $n = 720$; for monthly means $n = 36$

[b] Sample variance was obtained from the individual observations using $\mathrm{Var}[X] = \left[k\sum X_i^2 - \left(\sum X_i\right)^2\right]/\left(k^2 - k\right)$

[c] Corrected variances were obtained using equations (10) and (11) of Schaeffer et al. (1983)

proportional (FP) sampling is one of two ways: fixed time with sample volume proportional to flow (VP) or fixed volume with time proportional to flow (TP). Non-flow proportional composites (NFP) are usually taken as a fixed volume at fixed times.

Composite samples are generally believed to be more representative than grab samples of process stream average performance. Work has shown that if flows and concentrations are uncorrelated and lack autocorrelation, the same applies to the mean concentration of grabs and composites. However, the true variance of FP composites is larger than that of grab samples or TP composites. Furthermore, loads computed from FP composites are biased because of the volume weighting of subsamples during FP compositing.

When flows and concentrations are correlated, then VP and TP composites produce biased estimates of the mean and variance of concentrations, as well as of loads. In addition, the data in this study show that the bias arising from the correlation between flow and concentration substantially exceeds that arising from flow proportioning. As a result, regulatory monitoring data obtained from composite samples must be viewed with suspicion.

In the study reported by Schaeffer et al. (1983), samples from two treatment plants (Freeport and St. Charles, IL) were analyzed for the total suspended solids (TSS) by "Standard Methods" procedure 20913 (dried at 180°C). At St. Charles, the grab samples were also analyzed by drying at 105°C (procedure 209A). Ammonia (NH_3) analyses at both plants were by ion-selective electrode (procedure 417E). Flows (m^3/s) were monitored continuously at both facilities.

Table 7.5 summarizes the data for Freeport, and Table 7.6 summarizes the data for St. Charles. The tables give the number of observations, mean, standard deviation, skewness, kurtosis, minimum, and maximum.

Table 7.5 Freeport effluent concentrations and loads. The standard deviations (SD) are computed directly from sample data; variance corrections for compositing and for autocorrelation are not included

	Concentrations (ppm)			Loads (ppm × m^3/s)		
Parameter	Maximum	Mean	Skew	Maximum	Mean	Skew
N	Minimum	SD	Kurtosis	Minimum	SD	Kurtosis
Hourly grabs						
NH_3	18.0	12.8	0.3	4.3	2.5	0.0
167	9.4	1.8	3.1	1.2	0.8	2.0
TSS	1480.0	971.2	0.3	390.0	190.0	0.4
167	704.0	125.3	4.9	86.0	64.0	3.1
Daily time proportioned composites						
NH_3	24.3	11.4	0.2	4.9	2.3	0.2
100	2.8	3.5	4.6	0.4	0.7	5.2
TSS	1158.0	862.3	0.1	316.5	172.1	0.2
99	658.0	102.9	2.8	81.5	40.2	3.9
Daily volume proportioned composites						
NH_3	22.5	11.6	0.0	4.6	2.3	0.1
100	0.6	3.4	4.7	0.1	0.7	4.5
TSS	1222.0	884.8	0.0	314.6	176.6	0.1
99	674.0	116.2	2.9	81.2	41.6	3.5

Source: Schaeffer et al. (1983)

Table 7.6 St. Charles effluent concentrations and loads. The standard deviations (SD) are computed directly from sample data; variance corrections for compositing and for autocorrelation are not included

	Concentrations (ppm)			Loads (ppm × m^3/s)		
Parameter	Maximum	Mean	Skew	Maximum	Mean	Skew
N	Minimum	SD	Kurtosis	Minimum	SD	Kurtosis
Hourly grabs						
NH_3	15.5	7.3	0.1	2.9	1.3	0.1
168	2.0	4.0	2.1	0.3	0.7	1.8
TSS-105	122.0	28.0	1.2	23.0	4.9	1.9
191	0.0	24.6	3.8	0.0	4.7	4.4
TSS-180	32.0	8.0	0.9	6.3	1.4	2.0
163	0.0	7.1	3.3	0.0	1.3	5.0
Daily time proportioned composites						
NH_3	17.5	9.6	0.1	3.3	2.2	0.1
50	5.0	2.8	2.8	1.1	0.5	2.8
TSS-105	13.0	39.0	2.2	30.1	8.7	1.9
65						
Daily volume proportioned composites						
NH_3	17.5	12.6	0.0	4.3	2.8	0.8
63	6.0	3.0	2.3	1.7	0.5	4.3
TSS-105	99.0	42.0	0.6	24.1	9.5	0.4
66	7.0	22.5	3.2	1.1	5.1	3.3

Source: Schaeffer et al. (1983)

7.5.5 Composite Samples for Foliar Analysis

Foliar analysis has frequently been used in forestry to detect nutrient deficiencies, predict fertilizer requirements, and monitor uptake and recycling of nutrients. Substantial savings in cost can often result if laboratory analyses are performed on a composite of the field samples rather than on the individual samples. The composite sample consists of a thorough mixture of a number of field samples considered adequate to represent the population in question. The practice is based on an assumption that a valid estimate of the mean of some characteristic of the population may be obtained by analysis of the single composite sample.

Several different methods of combining individual samples to produce a composite sample have been used in forestry. The methods differ in the weights of individual samples which are combined. For instance, in grab sampling, a handful of needles is obtained from each of a number of trees and combined, i.e., the weight of the individual sample is unknown. Alternatively, equal weights of needles from each tree may be combined, or equal numbers of needles from each tree may be combined. In the latter case the samples are effectively combined in proportion to the average weight of individual needles. Samples may also be combined in proportion to some measure of tree size such as basal area.

Although the statistical properties and requirements of composite samples of bulk materials such as wool (Brown and Fisher, 1972), water (Rohde, 1976, 1979),

soil (Peterson and Calvin, 1965; Cameron et al., 1971), and other substances (Kratochvil and Taylor, 1981) have been published, there is no corresponding account of different methods for compositing samples of tree foliage. In this study several methods of creating compositing samples for plots within a factorial experiment were examined. The effects of the methods of estimates of plots and treatment means and their implications for sampling intensity were assessed. Tables 7.7, 7.8, and 7.9 summarize the results of this study.

Table 7.7 Effect of site preparation, fertilizer, and weed control on tree growth, needle weight, and foliar concentration of nitrogen and phosphorus, within-plot variation in these, and the correlations between them (*Pinus radiata*, age 3 years, Belanglo State Forest, NSW)

Variable	Nil	Weed control		Fertilizer weed control	Significant effects[a]
a. Treatment means for nutrient concentrations and growth variables					
Nitrogen, N(%)	1.84	1.98	1.99	1.87	Pr*
Phosphorus, P(%)	0.114	0.125	0.169	0.160	F***
Fascicle weight, W(mg)	14.7	28.4	21.7	42.1	F**, W***
Height, H(m)	0.76	1.14	1.43	2.28	F.W*
Basal area, B(cm^2)	1.33	6.56	6.63	31.16	F.W**
b. Within-plot standard deviations of nutrient concentrations and coefficients of variation for growth variables					
Nitrogen	0.200	0.274	0.226	0.287	Pr*
Phosphorus	0.0278	0.1209	0.0218	0.0320	F.W*
Fascicle weight	0.383	0.354	0.408	0.319	ns
Height	0.252	0.287	0.270	0.227	F.P*
Basal area	0.720	0.674	0.689	0.466	ns
c. Correlations between nutrient concentrations and growth variables					
N:W	0.143	–0.156	0.150	0.084	ns
N:H	0.152	–0.198	0.148	0.252	F.W**, Pr.W*
N:B	0.243	–0.188	0.161	0.162	ns
P:W	0.397	0.228	0.051	0.101	F*
P:H	0.356	0.211	–0.025	0.423	F.W*
P:B	0.344	0.187	–0.065	0.283	F.W*

[a] Pr, site preparation; F, fertilizer; W, weed control; ns, not significant *, **, ***, at 5, 1, and 0.1%, respectively

7.5.6 Lateral Variability of Forest Floor Properties Under Second-Growth Douglas-Fir Stands and the Usefulness of Composite Sampling Techniques

One of the purposes of the study conducted by Carter and Lowe is to evaluate the accuracy of composite sample data in relation to the mean of individual samples used to create the composite sample.

The mass of each nutrient contributed to the composite sample from an individual forest floor subsample is the product of the nutrient concentration in the subsample with the mass of the subsample used. Assuming that weighing and analytical errors

Table 7.8 Effect of site preparation, fertilizer, and weed control on the differences between treatment means obtained by compositing needle samples on an equal weight basis and those obtained using three compositing methods using unequal sample weights

Compositing	Nil	Weed control	Fertilizer	Fertilizer and weed control	Significant effects[a]
Nitrogen (%)					
Height	0.0065	–0.0227	0.0104	0.0186	F.W*
Fascicle weight	0.0119	–0.0156	0.0111	0.0114	Pr.F.W*
Basal area	0.0325	–0.0623	0.0241	0.0243	F.W*, Pr.W*
Phosphorus (%)					
Height	0.0024	0.0013	0.0000	0.0035	F.W*
Fascicle weight	0.0042	–0.0013	0.0003	0.0014	ns
Basal area	0.0068	0.0030	–0.0011	0.0052	F.W*

[a] See footnote to Table 7.7

Table 7.9 Within-plot variation of nutrient mass obtained by using different compositing methods and the number of randomly chosen sample trees required from a 25-tree plot to estimate the weighted mean to within 10 and 20% of the true value

			Number of samples			
	Within-plot standard deviation		±10%		±20%	
Compositing factor	N (%)	P (%)	N (%)	P (%)	N (%)	P (%)
Equal weight	0.244	0.025	7	10	4	5
Height	0.574	0.052	16	l8	9	10
Fascicle weight	0.749	0.063	19	20	11	13
Basal area	1.250	0.096	22	23	17	17

are negligible, a nutrient's concentration in a weighted composite sample will be equal to the weighted composite sample and to the weighted arithmetic mean of the nutrient concentration in the subsamples. Carter and Lowe (1986) study this assumption, comparing analytical variables determined from weighted composite samples with the same variable calculated as the weighted arithmetic mean of individual subsamples. The composite samples were tested to determine whether they were significantly different from the mean of the individual samples.

Chapter 8
A Linear Model for Estimation with Composite Sample Data

8.1 Introduction

The foregoing chapters covered various issues involving composite sample data with their statistical treatment. The problems vary from classification of individual samples to drawing inference on population parameters, especially the population mean. The measurements can be either presence/absence or continuous. A common theme in all these chapters is the need to establish a relationship between individual sample values and composite sample values. In Chapter 1, for instance, we observe that $Y = 1 - \prod_{i=1}^{k} (1 - X_i)$, where each X_i is a binary variable, taking a value 0 or 1, $i = 1, \ldots, k$, and hence Y is also binary. Here, Y is the composite sample value and $X_i,\ i = 1, \ldots, k$, are individual sample values. In Chapter 6, we noted that the composite sample value Y_j is a weighted average of individual sample values and is expressed as $Y_j = W_{j1}X_{j1} + \cdots + W_{jk}X_{jk}$. In Chapter 7, we considered the case where the weights of individual sample values in the above expression are random. In this chapter, we discuss a unified approach to express composite sample values in terms of individual sample values when the measurements are continuous. Due to the physical averaging of sample values upon compositing, it is proposed that a linear model best represents the functional relationship between individual sample values and composite sample values.

As for a statistical treatment of composite sample data, it is necessary to have a functional relationship between the two sets of values, individual sample values and composite sample values. With a linear model to express this relationship, most of the procedures used in linear statistical inference can be extended easily to composite sample data. For instance, if the individual sample values are generated by a process that can be visualized as a linear model with a factor at several levels, then composite sample data can be viewed as either a nested or a cross-over design where the composite samples are treated very much like blocking of individual samples by a factor at several levels. Thus, if n composite samples are formed, then we have divided all the available individual samples into n blocks, each block representing a composite sample. However, it must be noted here that, unlike in case of block designs, the variation among individual samples within each composite sample is never observed, and hence any inference from composite sample data has to be

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_8, © Springer Science+Business Media, LLC 2011

based only on variation between composite samples, as opposed to the variation between individual samples if the individual samples themselves were subjected to measurement.

8.2 Motivation for a Unified Model

Compositing is used in many different areas of application, with different materials and procedures. As we saw in the earlier chapters, it may be preceded by sampling schemes and followed by subsampling and laboratory measurement phases, which are as varying as desired due to the wide range of situations involved.

The methodology for analyzing data from composite samples has been developed mostly in an attempt to cope with problems which arise in specific areas of application. As a consequence, there has often been a failure to recognize both the common statistical structure of problems in different areas and the usefulness of methods beyond the particular problem for which they have been proposed. Setting up of a model, as general as possible, for the analysis of continuous measurements from composite samples may facilitate a conceptualization of all the aspects of composite sampling procedures and of the objectives of data analysis, including the unification of terminology and notation and the development of methods for new applications.

The purpose of the model introduced in this chapter is to offer a flexible tool for handling simultaneously both the basic features mentioned in the Introduction and the complications presented in the two preceding chapters. In particular, features of the compositing procedures which should be included in this general model are as follows:

1. The presence of different average levels of the response variable in the population under study, especially in meaningful subgroups in the population corresponding to combinations of potentially explanatory variables (as, for instance, different depths in a water body, different age–sex–race groups in a sample survey on a human population).
2. The nature of the population or lot or physical medium under study. In particular, the presence of natural segments (as, for instance, bags or bins where material to be sampled is stored, sites in a spatially allocated population) and heterogeneity of the material at various scales (between and within natural segments, within individual units in a segment, etc.).
3. The physical, chemical, or biological process which takes place in the compositing procedure. In particular, whether or not this process is such that the analyst controls the proportions of original material entering into each composite sample (case of fixed weights) or these proportions result from some random mechanism (case of random weights), and whether or not this process eliminates the heterogeneity (if present) of the original sampled material.
4. The characteristics of the measurement phase, including presence and magnitude of measurement errors.

8.3 The Model

Several models used in the literature to analyze measurements on a continuous variable from composite samples may be regarded as special cases of the following general model.

Suppose $\boldsymbol{x}$ is the $m \times 1$ vector of values assumed by the individual samples; $\boldsymbol{F}$ is an $m \times p$ known matrix; $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of fixed-effect parameters; $\boldsymbol{R}$ is an $m \times q$ known matrix; $\boldsymbol{\gamma}$ is a $q \times 1$ unknown vector of random-effect parameters; $\boldsymbol{\epsilon}$ is an $m \times 1$ vector of random disturbances; $\boldsymbol{y}$ is the $n \times 1$ vector of observations on the composite samples; $\boldsymbol{U}$ is an $n \times m$ random matrix of compositing weights; and $\boldsymbol{\eta}$ is an $n \times 1$ vector of measurement errors.

With the above notation, the general model can be stated in the following form:

$$\boldsymbol{x} = \boldsymbol{F\beta} + \boldsymbol{R\gamma} + \boldsymbol{\epsilon}, \tag{8.1}$$

$$\boldsymbol{y} = \boldsymbol{Ux} + \boldsymbol{\eta}. \tag{8.2}$$

The following assumptions are made concerning the random components of submodel (8.1):

$$E(\boldsymbol{\gamma}) = \boldsymbol{0}_q, \tag{8.3}$$

$$\text{Var}[\boldsymbol{\gamma}] = \boldsymbol{\Sigma}_\gamma, \tag{8.4}$$

$$E[\boldsymbol{\epsilon}] = \boldsymbol{0}_m, \tag{8.5}$$

$$\text{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}_\epsilon, \tag{8.6}$$

$$\text{cov}[\boldsymbol{\gamma}, \boldsymbol{\epsilon}] = \boldsymbol{0}_{q \times m}. \tag{8.7}$$

The matrix $\boldsymbol{U}$ in submodel (8.2) has the generic element:

$$U_{ji} = \begin{cases} W_{ji} & \text{if the } i\text{th individual} \\ & \text{sample contributes to the} \\ & j\text{th composite sample;} \\ 0 & \text{otherwise.} \end{cases}$$

It is convenient, as exemplified in the preceding chapter, to have the non-zero weights arranged to be contiguous, i.e., to form a sub-(row)vector of each row of $\boldsymbol{U}$. This can be achieved by rearranging the elements of $\boldsymbol{x}$ accordingly so that individual samples entering the same composite sample are themselves contiguous. This is accomplished by permuting the elements of $\boldsymbol{x}$ by means of an appropriate permutation matrix $\boldsymbol{P}$ of order m. Hence, $\boldsymbol{U}$ can be written in the form

$$\boldsymbol{U} = \boldsymbol{WP}.$$

Since the statistical characterization of the non-zero weights in $\boldsymbol{U}$ is invariant under permutation, we will concentrate on the matrix $\boldsymbol{W}$ when dealing with the statistical properties of the compositing weights. The matrix $\boldsymbol{W}$ is considered as a sample data matrix, whose jth row is a realization of a random vector $\boldsymbol{w}_j$ with moments:

$$E\big[\boldsymbol{w}_j\big] = \boldsymbol{\mu}_{w_j}, \tag{8.8}$$

$$\mathrm{Var}\big[\boldsymbol{w}_j\big] = \boldsymbol{\Sigma}_{w_j}, \tag{8.9}$$

and cross-covariance matrix between any pair of rows:

$$\mathrm{cov}\big[\boldsymbol{w}_j, \boldsymbol{w}_{j'}\big] = \boldsymbol{\Gamma}_{w_j, w_{j'}}, \qquad j \neq j'. \tag{8.10}$$

In this, note that the compositing weights are always in the form of proportions, i.e., they are divided by the total weight (volume, number, quantity, etc.) of individual samples in each composite sample, and therefore

$$\boldsymbol{w}_j' \mathbf{1}_m = 1. \tag{8.11}$$

As a consequence,

$$\boldsymbol{\mu}_{w_j}' \mathbf{1}_m = 1 \qquad \forall j, \tag{8.12}$$

$$\mathbf{1}_m' \boldsymbol{\Sigma}_{w_j} \mathbf{1}_m = 0 \qquad \forall j, \tag{8.13}$$

for $\mathbf{1}_m' \boldsymbol{\Sigma}_{w_j} \mathbf{1}_m$ is the variance of the degenerate random variable $\boldsymbol{w}_j' \mathbf{1}_m$

$$\text{and} \qquad \mathbf{1}_m' \boldsymbol{\Gamma}_{w_j, w_{j'}} \mathbf{1}_m = 0, \qquad j \neq j', \tag{8.14}$$

since $\mathbf{1}_m' \boldsymbol{\Gamma}_{w_j, w_{j'}} \mathbf{1}_m$ is the covariance between two degenerate random variables.

Sometimes we shall denote the $n \times m$ matrix of expectations of $\boldsymbol{W}$ by $\boldsymbol{M}_{\boldsymbol{W}}$ and the $nm \times nm$ variance/covariance matrix of $\boldsymbol{W}$ by $\boldsymbol{\Sigma}_{\boldsymbol{W}}$, that is,

$$E[\boldsymbol{W}] = \boldsymbol{M}_{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{\mu}_{w_1}' \\ \vdots \\ \boldsymbol{\mu}_{w_n}' \end{bmatrix}, \tag{8.15}$$

$$\mathrm{Var}(\boldsymbol{W}) = \boldsymbol{\Sigma}_{\boldsymbol{W}} = \begin{bmatrix} \boldsymbol{\Sigma}_{w_1} & \boldsymbol{\Gamma}_{w_1, w_2} & \cdots & \boldsymbol{\Gamma}_{w_1, w_n} \\ \boldsymbol{\Gamma}_{w_2, w_1} & \boldsymbol{\Sigma}_{w_2} & \cdots & \boldsymbol{\Gamma}_{w_2, w_1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Gamma}_{w_n, w_1} & \boldsymbol{\Gamma}_{w_n, w_2} & \cdots & \boldsymbol{\Sigma}_{w_n} \end{bmatrix}. \tag{8.16}$$

Again, by (8.12)

$$\boldsymbol{M}_{\boldsymbol{W}} \mathbf{1}_n = \mathbf{1}_n. \tag{8.17}$$

By using (8.13) and (8.14), we see that

$$\mathbf{1}'_{nm}\boldsymbol{\Sigma}_{\boldsymbol{W}}\mathbf{1}_{nm} = 0, \tag{8.18}$$

i.e., $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ is singular.

Finally, for the vector of measurement errors

$$E(\boldsymbol{\eta}) = \mathbf{o}_n, \tag{8.19}$$
$$\text{Var}[\boldsymbol{\eta}] = \boldsymbol{\Sigma}_{\eta}, \tag{8.20}$$
$$\text{cov}[\boldsymbol{\gamma}, \boldsymbol{\eta}] = \boldsymbol{O}_{q\times n}, \tag{8.21}$$
$$\text{cov}[\boldsymbol{\epsilon}, \boldsymbol{\eta}] = \boldsymbol{O}_{m\times n}, \tag{8.22}$$
$$\text{cov}[\boldsymbol{w}_j, \boldsymbol{\eta}] = \boldsymbol{O}_{m\times n} \qquad \forall j. \tag{8.23}$$

8.4 Discussion of the Assumptions

The basic idea which underlies model (8.1) and (8.2) is that the data from composite sampling arise in two stages: first, the mechanism generating $\boldsymbol{x}$, which we would investigate if we decided to carry out a traditional data analysis without compositing, and second, the mechanism generating $\boldsymbol{y}$, the measurements we actually analyze after compositing. From the point of view of statistical inference, we are in the following situation: we (usually) want to make inference on the mechanism generating $\boldsymbol{x}$, but we must do this by analyzing $\boldsymbol{y}$ and taking into consideration any additional disturbances due to its generating mechanism. Some relevant questions then are

i. Which features of the process generating $\boldsymbol{x}$ are preserved in $\boldsymbol{y}$?
ii. What restrictions should we put on the mechanism generating $\boldsymbol{y}$ in order to make inference on those aspects of the mechanism generating $\boldsymbol{x}$ that are of particular interest?

8.4.1 The Structural/Sampling Submodel

The first mechanism is modeled by (8.1), which we shall sometimes call the *structural/sampling submodel*, in that it accounts for structural effects of explanatory variables on $\boldsymbol{x}$ through $\boldsymbol{F\beta}$ and for sources of variability due to the sampling scheme through $\boldsymbol{R\gamma}$. In the terminology of theory of linear models, this is a *mixed model*, since the structure is modeled via a set of fixed effects (and hence the letter $\boldsymbol{F}$ for the coefficient matrix) while the sampling design affects $\boldsymbol{x}$ in terms of variance/covariance components and hence via a set of random effects (whereby the letter $\boldsymbol{R}$ for the coefficient matrix). In general, the columns of $\boldsymbol{F}$ will refer to variables, either quantitative or qualitative or mixed, that do not require the same testing

procedure as required for measuring $\boldsymbol{x}$ (and $\boldsymbol{y}$), since the desire to avoid the cost and effort of carrying out such a procedure on all original samples is precisely one of the main motivations for compositing. For example, in the report of an analysis of data obtained through the National Human Adipose Tissue Survey of the US Environmental Protection Agency presented by Orban et al., (1990), $\boldsymbol{F}$ is the incidence matrix of four qualitative variables (namely, census region, age group, race, sex) fixed by sampling design and observed on donors of tissue specimens, which were the individual samples in that application.

The structure of $\boldsymbol{\Sigma_\gamma}$ reflects the effects of the sampling design on the variance/covariance structure of $\boldsymbol{x}$. The parent population may be finite or infinite; appropriate choices of $\boldsymbol{R}$ and of $\boldsymbol{\Sigma_\gamma}$ will account for complexities of the sampling scheme such as stratification, multiple stages, varying probability sampling.

The variance/covariance matrix $\boldsymbol{\Sigma}_\epsilon$ of the error component $\boldsymbol{\epsilon}$ has, in most situations, the form $\boldsymbol{\Sigma}_\epsilon = \sigma^2 \boldsymbol{I}_m$, but more complicated structures are possible. In particular, a non-diagonal $\boldsymbol{\Sigma}_\epsilon$ may account for correlated errors which may result from spatial autocorrelation of individual sample values on samples selected from neighboring sites or from temporal autocorrelation of individual samples taken sequentially over time.

8.4.2 The Compositing/Subsampling Submodel

The second mechanism, which we call as the *compositing/subsampling submodel*, generating the observation $\boldsymbol{y}$ that we actually analyze, is modeled by (8.2). Again, there are two aspects of this submodel: the compositing process, whose effect on the compositing proportions is accounted for, after permutation $\boldsymbol{P}$, through $\boldsymbol{W}$, and the subsequent laboratory measurement procedure, whose features are reflected in $\boldsymbol{\eta}$.

As for $\boldsymbol{\eta}$, the assumption $E(\boldsymbol{\eta}) = \mathbf{o}_n$ simply means that the laboratory measurement is *valid*, that is, it does not give systematically biased measurements. The variance/covariance matrix of $\boldsymbol{\eta}$ usually has the form $\boldsymbol{\Sigma}_\eta = \sigma_\eta^2 \boldsymbol{I}_n$, because it is uncommon for carefully planned laboratory testing procedures to yield measurements with reliability varying from test to test or with measurement errors correlated over pairs of tests. However, in principle it is possible to take nonstandard features of the measurement phase into account by complicating the assumed structure of $\boldsymbol{\Sigma}_\eta$.

The matrix of weights, $\boldsymbol{W}$, is the only component of model (8.1) and (8.2) which is specific to the compositing process itself. As such, this component of submodel (8.2) deserves some detailed illustration, which will be given in the discussion to follow.

8.4.3 The Structure of the Matrices $\boldsymbol{W}$, $\boldsymbol{M_W}$, and $\boldsymbol{\Sigma_W}$

Various features of the compositing design, such as the number of composite sample size k, the number of composite samples n, subsampling of composite samples, and the choice of individual samples for inclusion in every composite sample, impose

specific structures on $\boldsymbol{W}$. A stochastic characterization of $\boldsymbol{W}$, especially regarding it being random or not, affects the structures of $\boldsymbol{M}_{\boldsymbol{W}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{W}}$. It is rather difficult to give a general treatment of such structures without complicating the notation and derivations. Instead, we begin with a fairly general case in which s (>1) subsamples are drawn from each of the c (>1) composite samples of size k. Clearly, the case $c = 1, s > 1$ reported by Brown and Fisher (1972) and by Rohde (1976) and the case $c > 1, s = 1$ are just special cases of this more general case. Here $m = ck$ is the total number of individual samples and $n = cs$ is the total number of measurements. In order to avoid additional subscripts and also without loss of generality, the number of individual samples k_j and the number of subsamples s_j are assumed to be the same for every composite sample, $j = 1, \cdots, c$. That is, $k_j = k$, and $s_j = s\ \forall j$. Furthermore, it is assumed that the compositing procedure is *exclusive*, i.e., that each individual sample contributes to one and only one composite sample.

In this general setting, the $cs \times ck$ matrix $\boldsymbol{W}$ is patterned as follows:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{O}_{s\times k} & \dots & \boldsymbol{O}_{s\times k} \\ \boldsymbol{O}_{s\times k} & \boldsymbol{W}_2 & \dots & \boldsymbol{O}_{s\times k} \\ \dots & \dots & \boldsymbol{W}_j & \vdots \\ \boldsymbol{O}_{s\times k} & \boldsymbol{O}_{s\times k} & \dots & \boldsymbol{W}_c \end{bmatrix}, \tag{8.24}$$

where

$$\boldsymbol{W}_j = \begin{bmatrix} \boldsymbol{w}'_{j1} \\ \boldsymbol{w}'_{j2} \\ \vdots \\ \boldsymbol{w}'_{ji} \\ \vdots \\ \boldsymbol{w}'_{js} \end{bmatrix} \tag{8.25}$$

is the $s \times k$ sub-matrix of non-zero proportions with which the k individual samples contribute to the s subsamples from the jth composite sample. The block diagonal pattern of $\boldsymbol{W}$ is due to the assumption of exclusiveness, since if a compositing procedure is exclusive then two rows $\boldsymbol{w}'_{ji}$, $\boldsymbol{w}'_{j'i'}$ of $\boldsymbol{W}$, pertaining to two different composite samples, $j \neq j'$, cannot overlap, i.e., they cannot have non-zero weights in the same column. Clearly, the (ck)-vector $\boldsymbol{x}$ is also patterned accordingly:

$$\boldsymbol{x} = [\boldsymbol{x}'_1\ \boldsymbol{x}'_2 \dots \boldsymbol{x}'_j \dots \boldsymbol{x}'_c]', \tag{8.26}$$

where $\boldsymbol{x}_j = [x_{j1}, x_{j2} \dots x_{j\ell} \dots x_{jk}]'$ is the k-vector of individual sample values associated with the k individual samples that form the jth composite sample and

hence, under the assumption of homogeneity of individual samples, with the sk aliquots contributing to the s-subsamples from the c composite samples.

Grouping of the rows of $\boldsymbol{W}$ into subsets corresponding to different composite samples imposes a particular structure on the moments of these rows, since it is clear that within each composite sample the weights are generated by the same random vector, whereas weights in different composite samples may be generated by different random vectors. In other words, there are c random vectors, generating the s non-zero weights in $\boldsymbol{w}_j$, $j = 1, \ldots, c$, with expected values $\boldsymbol{\mu}_{w_j}$, variance/covariance matrices $\boldsymbol{\Sigma}_{w_j}$, and cross-covariance matrices $\boldsymbol{\Gamma}_{w_j}$. Moreover, we shall denote by $\boldsymbol{\Gamma}_{w_j,w_\ell}$ the cross-covariance matrix between the random vectors generating the non-zero weights in the jth and the ℓth composite samples, $j, \ell = 1, \ldots, c, \quad j \neq \ell$.

At the first sight it may seem an unnecessary complication to consider $\boldsymbol{W}_j$ as a sample data sub-matrix, whose non-zero elements in the ith row, $i = 1, \ldots, s$, are a realization of a random vector $\boldsymbol{w}_j \sim \left(\boldsymbol{\mu}_{w_j}, \boldsymbol{\Sigma}_{w_j}, \boldsymbol{\Gamma}_{w_j}\right)$, and of $\boldsymbol{W}$ as a sample data matrix having the $\boldsymbol{W}_j$s as diagonal blocks. However, this permits translating all the prior information about the compositing process in terms of the statistical properties of the compositing weights. In the following discussion, we shall try to illustrate this translation in case of some common compositing situations.

8.4.3.1 Fixed Weights

If the proportions of the aliquots of individual samples which make up the composite samples are fixed, i.e., they are constant over all the composite samples, then each of the rows $\boldsymbol{w}'_{ji}$ in (8.25) may be thought of as a realization from a degenerate multivariate random variable $\boldsymbol{w}_j$ with

$$\begin{aligned} \boldsymbol{\mu}_{w_j} &= \boldsymbol{w}_j, \\ \boldsymbol{\Sigma}_{w_j} &= \boldsymbol{O}_k, \\ \Gamma_{w_j} &= \boldsymbol{O}_k, \\ \text{and } \boldsymbol{\Gamma}_{w_j,w_\ell} &= \boldsymbol{O}_k, \quad j \neq \ell. \end{aligned}$$

These can be written in a compact form as follows:

$$\boldsymbol{M}_{\boldsymbol{W}} = \boldsymbol{W}, \tag{8.27}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \boldsymbol{O}_{nm} \tag{8.28}$$

so that $\Pr\{\boldsymbol{W} = \boldsymbol{M}_{\boldsymbol{W}}\} = 1$.

The two most common situations in which the weights may be assumed to be fixed are as follows:

(a) The weights (volumes or number) of individual samples are fixed by the analyst and no subsampling is made, i.e., measurements are made on the composite samples. This is, the case, for example, with the National Human Adipose

Tissue Survey, where the analysts in the US EPA Office of Toxic Substance decide as to which tissue specimens would enter each composite sample and then analyze all the assembled tissue in each composite, since the primary reason for compositing is the need for more tissue mass per analysis sample.

(b) Subsampling is made and physical mixing is (or may be assumed to be) perfect. In this case, the weights will not only be fixed but also be known if the analyst can predetermine them by design or observe them before compositing. Otherwise, they are fixed but unknown. The latter situation may occur, for example, if grab samples of water are taken with volumes proportional to flow and are directly poured into a container and perfectly amalgamated, in order to estimate waste loads. Then the volumes of the contributions of each grab sample to each subsample from the container are unknown, but they are the same in different subsamples. As a consequence, they are not realizations of a random vector or, to be consistent with the approach followed in this chapter, they are realizations of a degenerate random vector $\boldsymbol{w}$ with an unknown mean $\boldsymbol{\mu}_w$ and a null variance/covariance matrix.

8.4.3.2 Random Weights: The Case of *c* Composite Samples and *s* Subsamples from Each

In Section 7.1 we illustrated, with the help of two examples, the two mechanisms which may generate random weights. To give a more formal treatment of the topic, let us consider the expected value and the variance/covariance matrix of $\boldsymbol{W}$ in the case considered here, i.e., that of $s > 1$ subsamples from each of the $c > 1$ composite samples of size k. Since these matrices are highly patterned, their manipulation is facilitated by the use of some results on the Kronecker product and on elementary matrices and vectors, which are given in Section 8.7. For further details on the use of these mathematical tools, see Graham (1981).

Owing to the presence of non-random zeros, the $\boldsymbol{w}_{ji}$ vectors have the following pattern:

$$\boldsymbol{w}_{ji} = [\mathbf{o}'_k \; \mathbf{o}'_k \; \ldots \boldsymbol{w}'_{ji} \ldots \mathbf{o}'_k \; \mathbf{o}'_k]'. \tag{8.29}$$

As a consequence, $\boldsymbol{M}_{\boldsymbol{W}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ are themselves characterized by blocks of patterned zeros. Suppose $\boldsymbol{E}^{\text{bs}}_{ii'}$ is an $s \times s$ elementary matrix, having a 1 in the (i, i')th position and 0 elsewhere, which identifies the pair of subsamples (i, i'); the superscript *bs* reminds that this is a *b*etween *s*ubsamples operator; $\boldsymbol{E}^{\text{bc}}_{j\ell}$ is a $c \times c$ elementary matrix, having a 1 in the (j, ℓ)th position and 0 elsewhere, which identifies the pair of composite samples (j, ℓ); the superscript *bc* reminds that this is a *b*etween *c*omposite samples operator. Using formula (8.58) and the above notation, we may write $\boldsymbol{M}_{\boldsymbol{W}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ in a compact way as follows:

$$\boldsymbol{M}_{\boldsymbol{W}} = \sum_{j=1}^{c} \left(\boldsymbol{E}^{\text{bc}}_{jj} \otimes \mathbf{1}_s \otimes \boldsymbol{\mu}'_{w_j} \right), \tag{8.30}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sum_{j=1}^{c} \Bigg[\sum_{i=1}^{s} (\boldsymbol{E}_{jj}^{\mathrm{bc}} \otimes \boldsymbol{E}_{ii}^{\mathrm{bs}}) \otimes \boldsymbol{E}_{jj}^{\mathrm{bc}} \otimes \boldsymbol{\Sigma}_{w_j} + \sum_{\substack{i=1 \\ i \neq i'}}^{s} \sum_{i'=1}^{s} (\boldsymbol{E}_{jj}^{\mathrm{bc}} \otimes \boldsymbol{E}_{ii'}^{\mathrm{bs}}) \otimes \boldsymbol{E}_{jj}^{\mathrm{bc}} \otimes \boldsymbol{\Gamma}_{w_j} \Bigg] + \sum_{\substack{j=1 \\ j \neq \ell}}^{c} \sum_{\ell=1}^{c} \sum_{i=1}^{s} \sum_{i'=1}^{s} (\boldsymbol{E}_{j\ell}^{\mathrm{bc}} \otimes \boldsymbol{E}_{ii'}^{\mathrm{bs}}) \otimes \boldsymbol{E}_{j\ell}^{\mathrm{bc}} \otimes \boldsymbol{\Gamma}_{w'_j, w_\ell}. \tag{8.31}$$

In the literature, it is usually assumed that $\boldsymbol{\Gamma}_{w_j} = \boldsymbol{O}_k$, i.e., that weights in different subsamples from the same composite are uncorrelated. Rohde (1976, p. 277) further assumes they are independent, pointing out that this is "reasonable so long as s is small relative to k."

It should be noted here that, although the assumption $\boldsymbol{\Gamma}_{w_j} = \boldsymbol{O}_k$ is often satisfied, it is not necessarily so. For example, if the number of subsamples which can be drawn from each composite is finite and a moderate to large number of them is actually selected, then the weights in different subsamples are bound to be negatively correlated (see, for example, Elder et al., 1980).

As for the cross-covariance matrix between vectors of weights in different composite samples, $\boldsymbol{\Gamma}_{w_j, w_\ell}$, the values in these covariance matrices depend essentially on whether or not compositing is exclusive. Recall that by exclusiveness we mean that all the aliquots taken from a particular individual sampling unit enter a single composite sample. Clearly, if compositing is exclusive, then the weights in any one composite sample cannot affect the weights in any other composite sample, and therefore $\boldsymbol{\Gamma}_{w_j, w_\ell} = \boldsymbol{O}_k \quad \forall j, \ell;\ j \neq \ell$. Another way of interpreting this is to consider a whole exclusive compositing procedure as the "union" of c independent compositing procedures.

Although exclusiveness is assumed in nearly all reported investigations in the literature, it is conceptually possible to have non-exclusive compositing schemes. For example, if sampling costs are not negligible, it may be justified to select k individual sampling units to draw c aliquots from each of them and then to form c composite samples using one aliquot from each individual sample. In this case, there would be a "perfect non-exclusiveness." That is, each individual sampling unit would contribute to *all* the composite samples.

Clearly, in the non-exclusive case, the matrices $\boldsymbol{\Gamma}_{w_j, w_\ell}$ can still be $\boldsymbol{O}_k$ if the random mechanism generating the weights in one composite does not interact with that generating the weights in any other composite. However, at least conceptually, the possibility of correlated weights in different composited samples cannot be ruled out in non-exclusive compositing procedures, and we have allowed for this possibility in (8.30) to retain the maximum possible generality in this theoretical presentation.

8.4.3.3 Unbiased Compositing/Subsampling Procedures

A compositing process and the subsequent subsampling step are said to constitute an *unbiased compositing/subsampling procedure* if "on the average subsamples consist of equal proportions of material from all increments in the composite" (see Elder et al., 1980). For unbiased procedures, we have

$$\boldsymbol{\mu}_{w_j} = \mu \mathbf{1}_k \qquad \forall j. \tag{8.32}$$

Constraint (8.12) implies that $\mu = \frac{1}{k}$. Hence, for the matrix of expected values $\boldsymbol{M}_{\boldsymbol{W}}$ we get

$$\boldsymbol{M}_{\boldsymbol{W}} = \boldsymbol{I}_c \otimes \mathbf{1}_s \otimes \frac{1}{k}\mathbf{1}_k'. \tag{8.33}$$

The following assumptions on the second moments of the weights are widely used in the literature:

(a) All the weights have the same variance σ_w^2
(b) All pairs of weights in the same subsample have the same covariance
(c) Weights in different subsamples from the same composite as well as in different composites are uncorrelated

These assumptions, along with constraints (8.13) and (8.14), yield

$$\begin{aligned}
\boldsymbol{\Sigma}_{w_j} &= \sigma_w^2 \left[\left(\frac{k}{k-1} \right) \boldsymbol{I}_k - \left(\frac{1}{k-1} \right) \boldsymbol{J}_k \right], \\
\boldsymbol{\Gamma}_{w_j} &= \boldsymbol{O}_k, \\
\boldsymbol{\Gamma}_{w_j, w_\ell} &= \boldsymbol{O}_k.
\end{aligned}$$

Hence, for the variance/covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ we obtain

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sum_{j=1}^{c} \sum_{i=1}^{s} \left(\boldsymbol{E}_{jj}^{\mathrm{bc}} \otimes \boldsymbol{E}_{ii}^{\mathrm{bs}} \right) \otimes \boldsymbol{E}_{jj}^{bc} \otimes \sigma_w^2 \left[\left(\frac{k}{k-1} \right) \boldsymbol{I}_k - \left(\frac{1}{k-1} \right) \boldsymbol{J}_k \right]. \tag{8.34}$$

Assumptions (8.33) and (8.34) may be considered as an extended definition of unbiasedness, insofar as their practical meaning is, with the insightful words used by Elder et al. (1980), that "all increments (compositing objects in our terminology) receive the same treatment in the compositing/subsampling procedure." In the remainder of this chapter, we shall refer to (8.33) and (8.34) as describing an "unbiased composite/subsample procedure."

8.5 Moments of x and y

Suppose the random vector $\boldsymbol{x}$ and the random matrix $\boldsymbol{W}$ are expectation, variance, and covariance independent. That is,

$$E[\boldsymbol{W}|\boldsymbol{x}] = E[\boldsymbol{W}] \qquad \forall \boldsymbol{x}, \tag{8.35}$$

$$\text{Var}[\boldsymbol{W}|\boldsymbol{x}] = \text{Var}[\boldsymbol{W}] \qquad \forall \boldsymbol{x}. \tag{8.36}$$

These conditions, which represent a matrix version of those first introduced by Bohrnstedt and Goldberger (1969) and later used by Elder (1977), basically imply that the random mechanism producing the weights does not interact with the random mechanism generating the X-values in the population.

Under this assumption and using the specifications given in Section 8.3 for the structural/sampling and the compositing/subsampling submodels, the expected values and variance/covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ may be derived:

$$E[\boldsymbol{x}] = \boldsymbol{F}\boldsymbol{\beta}, \tag{8.37}$$

$$\text{Var}[\boldsymbol{x}] = E\left[(\boldsymbol{R}\boldsymbol{\gamma} + \boldsymbol{\epsilon})(\boldsymbol{R}\boldsymbol{\gamma} + \boldsymbol{\epsilon})\right] = \boldsymbol{R}\boldsymbol{\Sigma}_{\gamma}\boldsymbol{R}' + \boldsymbol{\Sigma}_{\epsilon}, \tag{8.38}$$

$$E[\boldsymbol{y}] = E(\boldsymbol{W}\boldsymbol{x}) = E(\boldsymbol{W})E(\boldsymbol{x}) = \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{F}\boldsymbol{\beta}, \tag{8.39}$$

$$\begin{aligned}\text{Var}[\boldsymbol{y}] &= \text{Var}[\boldsymbol{W}\boldsymbol{x}] + \text{Var}[\boldsymbol{\eta}] \\ &= \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{R}\boldsymbol{\Sigma}_{\gamma}\boldsymbol{R}'\boldsymbol{M}'_{\boldsymbol{W}} + \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_{\epsilon}\boldsymbol{M}'_{\boldsymbol{W}} \\ &\quad + \left[(\boldsymbol{I}_n \otimes \boldsymbol{\beta}'\boldsymbol{F}')\boldsymbol{\Sigma}_{\boldsymbol{W}}(\boldsymbol{I}_n \otimes \boldsymbol{F}\boldsymbol{\beta})\right] \\ &\quad + \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \text{tr}\left[\left(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right)\boldsymbol{R}\boldsymbol{\Sigma}_{\gamma}\boldsymbol{R}'\right]\boldsymbol{e}'_{j'} \\ &\quad + \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \text{tr}\left[\left(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\epsilon}\right]\boldsymbol{e}'_{j'} + \boldsymbol{\Sigma}_{\eta}.\end{aligned} \tag{8.40}$$

For the derivation of (8.39) and (8.40), see Section 8.8.

8.6 Complex Sampling Schemes Before Compositing

In many applications of composite sampling, the population of interest is structured and/or practical reasons make simple random sampling inadvisable. In all these cases, a complex sampling scheme, i.e., a sampling design with two or more sampling stages, with stratification and/or selection of clusters (or segments) at some of such stages, must be employed for drawing the final individual sampling units to be composited. The features of such complex sampling scheme must be taken into account in the analysis of the final measurement on the composite samples,

since they usually inflate the variance of each final measurement and sometimes introduce correlation between final measurements.

One powerful and flexible approach for taking the complex sampling scheme into consideration is to account for the effects of the various complications of the sampling design through random-effect parameters in the model for X_i, the value characterizing the ith individual sampling unit. There is a vast literature on this topic, especially from researchers in the area of analytical use of sample survey data from finite populations; a good review is in Skinner et al. (1989).

8.6.1 Segmented Populations

A simple and very common instance of structured population is provided by situations in which the population of interest is naturally divided into *segments*, i.e. into clusters of spatially, temporally, or physically aggregated objects. Examples of such populations include

(i) bales of wool (Cameron, 1951; Brown and Fisher, 1972);
(ii) sites in a water body (Rohde, 1979);
(iii) fertilizer in bags; and
(iv) materials in bins.

Let us illustrate the situation by a simple example (see Fig. 8.1). Some hazardous waste material is stored in 100 barrels. In order to take appropriate action, an estimate of the average concentration level of a highly toxic contaminant X is needed. The measurement process is very costly, and therefore the analysts resort to compositing. Four barrels (say nos. 2, 27, 55, and 65) are randomly chosen and two individual samples are taken from each barrel at different depths.

The two individual samples from barrel 2 and the two from barrel 27 are composited; a similar procedure is carried out for the two individual samples from barrel 55 and barrel 65. Therefore, two composite samples of size 4 are finally measured at the laboratory to determine the average concentration level in the batch of 100 barrels. Again, to keep things easier, we assume that the measurement error is negligible and may therefore be ignored.

8.6.2 Estimating the Mean in Segmented Populations

To introduce the basic concepts associated with the problem of estimating the mean of X in a segmented population, let us generalize the features of the previous example.

Let us first suppose that the population is composed of a very large number, $B \to \infty$, of segments. cb segments are randomly selected and grouped into c subsets. From each segment in each subset, a constant number a of compositing objects

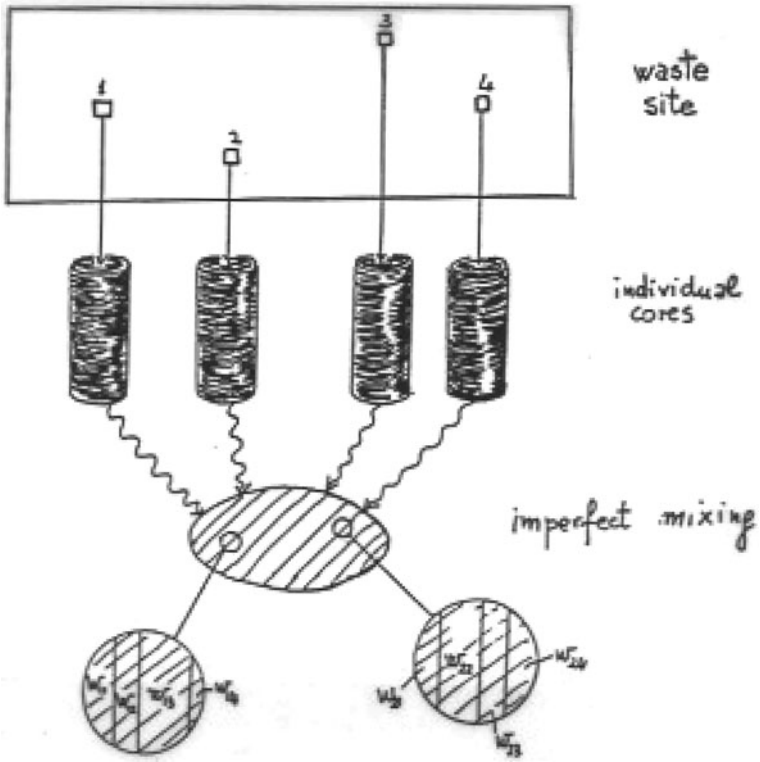

Fig. 8.1 Hazardous waste material in barrels example

is drawn randomly. The $k = ab$ compositing objects are finally composited; therefore, there are c composite samples of size $k = ab$, and each is measured without error. A graphical representation of such a compositing procedure is given in Fig. 8.2.

This layout may be modeled as follows. Consider the random-effects linear model expressing the X-value in each individual sample as an additive function of the general mean μ_x (the parameter of interest), an effect due to the random selection of barrels and an error component. This model may be written as follows:

$$
\begin{aligned}
X_{jhi} = \mu_x + \gamma_{jh} + \epsilon_{jhi}, \quad & j = 1, \ldots, c \\
& h = 1, \ldots, b \\
& i = 1, \ldots, a \\
\text{with} \qquad \gamma_{jh} \ \text{i.i.d.}(0, \sigma_\gamma^2), \quad & j = 1, \ldots, c \\
& h = 1, \ldots, b \\
\epsilon_{jhi} \ \text{i.i.d.}(0, \sigma_\epsilon^2), \quad & j = 1, \ldots, c \\
& h = 1, \ldots, b \\
& i = 1, \ldots, a \, .
\end{aligned}
$$

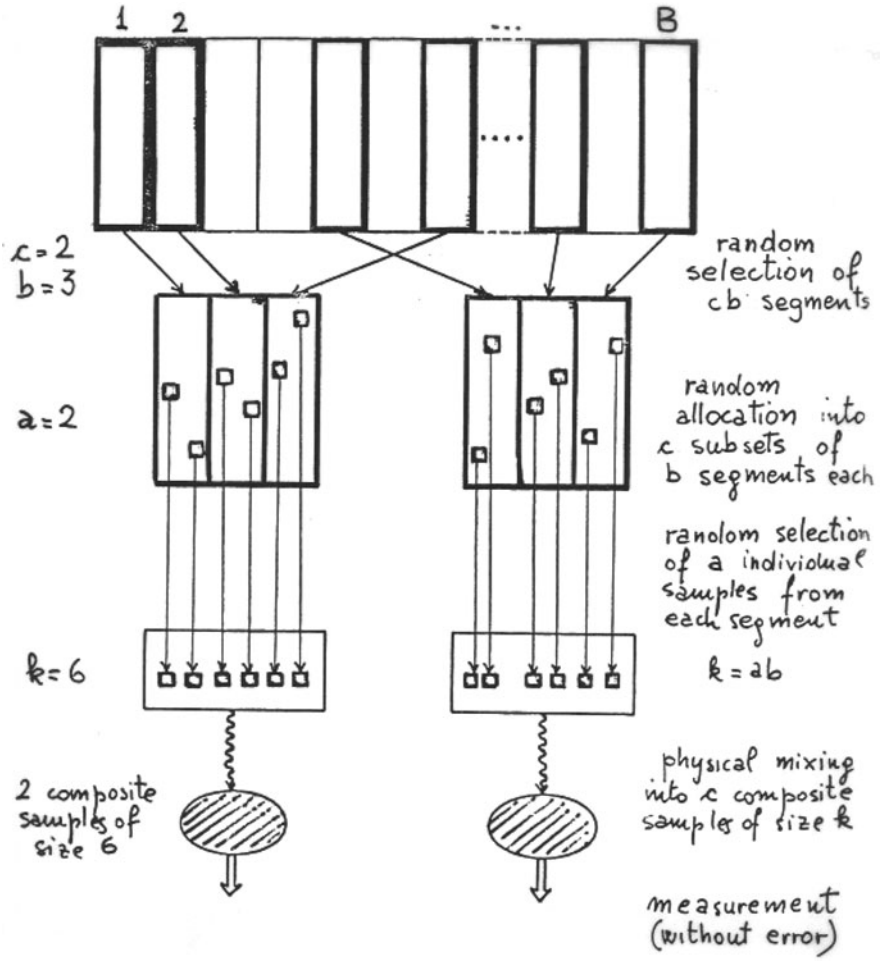

Fig. 8.2 Sampling/compositing scheme for segmented populations (with one composite per subset)

Hence

$$\begin{aligned}
\mathrm{Var}\big[X_{jhi}\big] &= \sigma_\gamma^2 + \sigma_\epsilon^2, \\
\mathrm{cov}\big[X_{jhi}, X_{jhi'}\big] &= \sigma_\gamma^2, \quad i \neq i', \\
\mathrm{cov}\big[X_{jhi}, X_{jh'i'}\big] &= 0, \quad h \neq h'; \quad \forall i, i', \\
\mathrm{cov}\big[X_{jhi}, X_{j'h'i'}\big] &= 0, \quad j \neq j'; \quad \forall h, h'; \quad \forall i, i'.
\end{aligned}$$

Denote the final measurement on the jth composite sample by

$$Y_j = \sum_{h=1}^{b} \sum_{i=1}^{a} W_{jhi} X_{jhi}, \quad j = 1, \ldots, c.$$

If $W_{jhi} = \frac{1}{k}\ \forall j, h, i$, then

$$\begin{aligned} E\big[Y_j\big] &= \mu_x && \forall j, \\ \mathrm{Var}\big[Y_j\big] &= \frac{\sigma_\gamma^2}{b} + \frac{\sigma_\epsilon^2}{k} && \forall j, \\ \mathrm{cov}\big[Y_j, Y_{j'}\big] &= 0, && j \neq j', \end{aligned}$$

and the composited mean $\bar{Y} = \frac{\sum_{j=1}^{c} Y_j}{c}$ has expected value:

$$E\left[\bar{Y}\right] = \mu_x$$

and variance:

$$\mathrm{Var}\big[\bar{Y}\big] = \frac{\sigma_\gamma^2}{cb} + \frac{\sigma_\epsilon^2}{ck}.$$

That is, the composited mean is an unbiased estimator of μ_x, but the variance of $\bar{Y}$ is inflated, when compared to that of a composited mean in unsegmented populations, by a quantity $\frac{\sigma_\gamma^2}{cb}$. Notice that it is still possible to obtain an unbiased estimator of $\mathrm{Var}\big[\bar{Y}\big]$, since

$$E\left[\frac{\sum_{j=1}^{c}(Y_j - \bar{Y})^2}{c(c-1)}\right] = \frac{\sigma_\gamma^2}{cb} + \frac{\sigma_\epsilon^2}{ck} = \mathrm{Var}\big[\bar{Y}\big],$$

but σ_γ^2 and σ_ϵ^2 are not separately estimable.

8.6.3 Estimating Variance Components in Segmented Populations

If the variability of X in the parent population is of direct interest, then careful design of the compositing procedure makes it possible to get an estimate of σ_ϵ^2.

To show this, let us now suppose that gb segments are randomly selected and randomly grouped into g groups of b segments each. From each segment in each

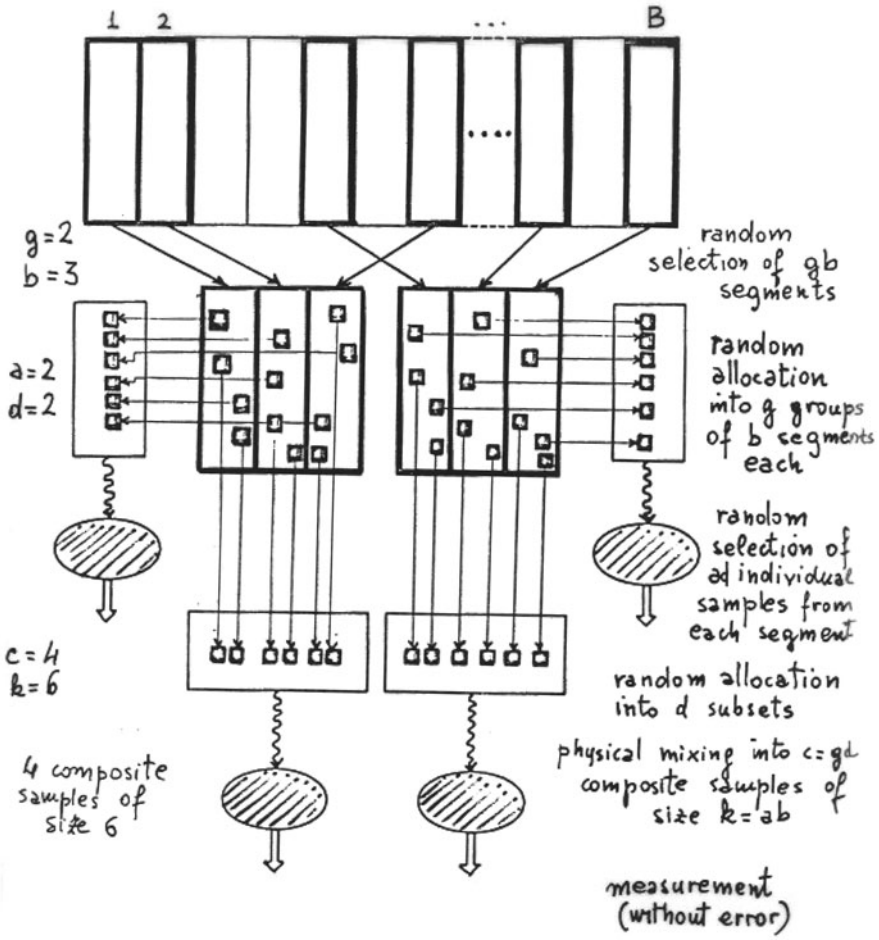

Fig. 8.3 Sampling/compositing scheme for segmented populations (with d composites per subset)

group, a compositing objects are randomly drawn and composited; the procedure is repeated until d composites are formed within each group. So, there are a total of $c = gd$ composites of size $k = ab$ each. This compositing procedure is depicted in Fig. 8.3.

Let $X_{\ell jhi}$ be the X-value for the ith individual sample from the segment h of the ℓth group, which enters the jth composite sample. This value can then be modeled as follows:

$$X_{\ell jhi} = \mu_x + \gamma_{\ell h} + \epsilon_{\ell hi},$$

with

$$\begin{aligned}
&\gamma_{\ell h} \text{ i.i.d. } (0, \sigma_\gamma^2), \quad \ell = 1, \ldots, g;\ h = 1, \ldots, b \\
&\epsilon_{\ell h} \text{ i.i.d. } (0, \sigma_\epsilon^2), \quad \ell = 1, \ldots, g;\ h = 1, \ldots, b;\ \ i = 1, \ldots, a.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\operatorname{Var}\left[X_{\ell jhi}\right] &= \sigma_\gamma^2 + \sigma_\epsilon^2, \\
\operatorname{cov}\left[X_{\ell jhi}, X_{\ell jhi'}\right] &= \sigma_\gamma^2, \quad i \neq i', \\
\operatorname{cov}\left[X_{\ell jhi}, X_{\ell jh'i'}\right] &= 0, \quad h \neq h'\ \forall i, i', \\
\operatorname{cov}\left[X_{\ell jhi}, X_{\ell j'hi'}\right] &= \sigma_\gamma^2, \quad j \neq j'\ \forall i, i', \\
\operatorname{cov}\left[X_{\ell jhi}, X_{\ell j'h'i'}\right] &= 0, \quad h \neq h', \\
\operatorname{cov}\left[X_{\ell jhi}, X_{\ell j'h'i'}\right] &= 0, \quad \ell \neq \ell'.
\end{aligned}$$

Let $Y_{\ell j} = \sum_{h=1}^{b} \sum_{i=1}^{a} W_{\ell jhi} X_{\ell jhi}$ be the final measurement, without error, on the jth composite in the ℓth group. Then,

$$\begin{aligned}
\operatorname{Var}\left[Y_{\ell j}\right] &= \frac{\sigma_\gamma^2}{b} + \frac{\sigma_\epsilon^2}{k} \qquad \forall \ell, j, \\
\operatorname{cov}\left[Y_{\ell j}, Y_{\ell j'}\right] &= \frac{\sigma_\gamma^2}{b}, \qquad j \neq j', \\
\operatorname{cov}\left[Y_{\ell j}, Y_{\ell' j'}\right] &= 0, \qquad \ell \neq \ell';\ \forall j, j'.
\end{aligned}$$

The composite sample mean is now

$$\bar{Y} = \sum_{\ell=1}^{g} \sum_{j=1}^{d} Y_{\ell j}/gd,$$

and it is also possible to define the group means

$$\bar{Y}_\ell = \sum_{j=1}^{d} Y_{\ell j}/d, \qquad \ell = 1, \ldots, g.$$

The total sum of squares for composite sample measurements may be decomposed into two components as follows:

$$\sum_{\ell=1}^{g}\sum_{j=1}^{d}(Y_{\ell j}-\bar{Y})^2=\sum_{\ell=1}^{g}\sum_{j=1}^{d}(Y_{\ell j}-\bar{Y}_{\ell})^2+\sum_{\ell=1}^{g}d(\bar{Y}_{\ell}-\bar{Y})^2,$$

and it is possible to show that

$$E\left[\sum_{\ell=1}^{g}\sum_{j=1}^{d}(Y_{\ell j}-\bar{Y}_{\ell})^2\right]=\sigma_{\epsilon}^2\left(\frac{d(g-1)}{k}\right).$$

That is, $\frac{k}{d(g-1)}\sum_{\ell=1}^{g}\sum_{j=1}^{d}(Y_{\ell j}-\bar{Y}_{\ell})^2$ is an unbiased estimator of σ_{ϵ}^2.

8.7 Estimating the Effect of a Binary Factor

For illustration purposes, we shall introduce the basic problems encountered in the estimation of fixed-effects parameters in linear models with composited data in the context of the simplest possible example, that of a binary factor.

As an example, let us suppose that a county is divided into two areas: one in which the prevailing economic activity is industrial and one in which the prevailing economic activity is agricultural. A monitoring program is set up in order to control the quality of underground water in the county. The purpose of the program is twofold: to obtain accurate estimates of the average concentration level of some contaminant X and to evaluate the differential impact of industrial and agricultural activities on such concentration. In other words, the questions to be addressed are how much contaminant X is present, on average, in the country underground water? Given the contaminant X is a side product of both industry and farming, which of the two activities is more responsible for its presence and concentration?

Two wells are randomly chosen in the industrial area and two in the agricultural area (see Fig. 8.4). Due to cost and time constraints, the monitoring program staff decides to use some form of compositing on the individual grab samples of water taken from the wells. Two composite samples are formed mixing two of the individual sampling units in each composite. The composite samples are then completely tested, without subsampling, at the laboratory to determine the concentration of X. The measurements are taken with an error which is so negligible, compared with natural variability of X, that they may be considered to be free of errors.

This situation may be modeled as follows. Let

$$X_{hi}=\mu_x+\beta_h+\epsilon_{hi},\qquad \begin{matrix}h=1,2,\\ i=1,2\end{matrix}$$

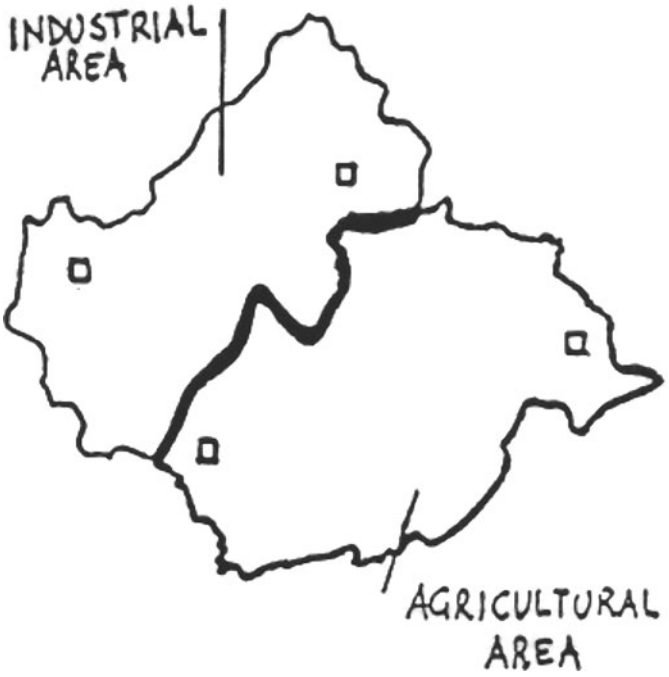

Fig. 8.4 Industrial vs. agricultural water quality example

with β_1, β_2 unknown fixed parameters,

$$\sum_{h=1}^{2} \beta_h = 0 \Rightarrow \beta_2 = -\beta_1$$
$$\epsilon_{hi} \quad \text{i.i.d.} \sim (0, \sigma_\epsilon^2)$$

be the model expressing the effect of the two levels ($h = 1$: industrial area; $h = 2$: agricultural area) of the binary factor F on the variable X of interest in the individual samples. Notice that the two individual samples in each area may be thought of as replicates and that, since the number of replicates in the areas is constant, the individual sampling design is balanced.

Let

$$Y_j = \sum_{h=1}^{2} \sum_{i=1}^{2} U_{jhi} X_{hi}, \qquad j = 1, 2,$$

where

$$U_{jhi} = \begin{cases} W_{jhi} & \text{if the } i\text{th individual sampling} \\ & \text{unit at the factor-level } h \\ & \text{contributes to the } j\text{th composite;} \\ 0 & \text{otherwise} \end{cases}$$

be the compositing model, which describes the way the four individual samples are composited into two composites of size 2. We assume that the final measurement on each composite is made without error.

In matrix form,

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{with } E[\boldsymbol{\epsilon}] = \mathbf{0}, \text{Var}[\boldsymbol{\epsilon}] = \sigma_\epsilon^2 \boldsymbol{I}_m, \tag{8.41}$$

$$\boldsymbol{y} = \boldsymbol{U}\boldsymbol{x}, \tag{8.42}$$

where

$$\boldsymbol{x} = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{bmatrix}, \quad \boldsymbol{F} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_x \\ \beta_1 \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} u_{111} & u_{112} & u_{121} & u_{122} \\ u_{211} & u_{212} & u_{221} & u_{222} \end{bmatrix}.$$

Note that all the design and model features concerning the individual samples are summarized in $\boldsymbol{F}$, while $\boldsymbol{U}$ summarizes all the features of the compositing design (assignment of individual samples to the composites and the composite sample size). These features are fairly easy to understand if we consider their basic components.

The matrix $\boldsymbol{F}$ can be thought of as a product of two matrices:

$$\boldsymbol{F} = \boldsymbol{G}\boldsymbol{L},$$

where the matrix $\boldsymbol{G}$ describes the number of replicates for each combination of levels of the explanatory factors (in our example, the number of wells drawn in each area) and $\boldsymbol{L}$ gives a matrix representation of the model we are studying.

In our example

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

$$\boldsymbol{L} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

The matrix $\boldsymbol{U}$ can also be thought of as a product of two matrices:

$$\boldsymbol{U} = \boldsymbol{W}P.$$

The matrix $\boldsymbol{W}$ contains the weights reordered in such a way that in each row the non-zero weights $0 < w_{jhi} \leq 1$ are contiguous. When, as in our example, each individual sampling unit contributes to just one composite, then we say that the compositing procedure is *exclusive* and the matrix $\boldsymbol{W}$ has the additional feature that there is only one non-zero weight in each column. In our example:

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & 0 & 0 \\ 0 & 0 & w_{21} & w_{22} \end{bmatrix},$$

where w_{ji} is the proportion with which the ith individual sampling unit in the jth composite sample contributes to that composite.

In order to achieve this convenient form for $\boldsymbol{W}$, individual sampling units must be reordered so that units entering the same composite are themselves contiguous in vector $\boldsymbol{x}$. This is accomplished by the permutation matrix $\boldsymbol{P}$. For example, if the first and fourth individual sampling units (x_{11}, x_{22}) are to be mixed to form the first composite and the second and third (x_{12}, x_{21}) to form the second composite, then it is convenient to have the vector $\boldsymbol{x}$ reordered as $\boldsymbol{x}^* = [x_{11}, x_{22}, x_{12}, x_{21}]'$. This is accomplished by a permutation matrix $\boldsymbol{P}$ that depends on the particular compositing scheme. Hence, given the number of non-zero weights in each row of $\boldsymbol{W}$, the permutation matrix $\boldsymbol{P}$ plays the role of an assignment matrix, which identifies the individual sampling units that are allocated to every composite sample.

To see how the estimation of $\boldsymbol{\beta}$ can be handled in this context, observe that (8.41) and (8.42) may be rewritten as a single model as follows:

$$\boldsymbol{y} = \boldsymbol{U}F\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\epsilon} = \boldsymbol{D}\boldsymbol{\beta} + \boldsymbol{\zeta}. \tag{8.43}$$

Since $\mathrm{Var}[\boldsymbol{y}] = \boldsymbol{W}\mathrm{Var}[\boldsymbol{\epsilon}]\boldsymbol{W}' = \sigma_\epsilon^2 \boldsymbol{W}\boldsymbol{W}'$ depends on the structure of $\boldsymbol{W}$ and may therefore display heteroscedasticity even if $\mathrm{Var}[\boldsymbol{\epsilon}]$ does not, we must use a generalized least squares approach to estimate $\boldsymbol{\beta}$. The normal equations are

$$(\boldsymbol{D}'[\boldsymbol{W}\boldsymbol{W}']^{-1}\boldsymbol{D})\hat{\boldsymbol{\beta}} = \boldsymbol{D}'[\boldsymbol{W}\boldsymbol{W}']^{-1}\boldsymbol{y}. \tag{8.44}$$

The basic problem in the estimation of fixed linear models with composited data can now be addressed more clearly: are $\boldsymbol{D}$ and $\boldsymbol{\zeta}$ such that $\boldsymbol{\beta}$ is still identifiable? We shall show by means of two somewhat extreme cases that the answer to this question depends on the choice of $\boldsymbol{W}$, i.e., upon the compositing design.

8.7.1 Fully Segregated Composites

Suppose that, in the industrial vs. agricultural area example, the analyst may afford to make two measurements a day. One reasonable way of forming two composites out of the four available wells is to take two wells in each area and composite the water grab samples from them. This compositing design is depicted in Fig. 8.5. With this choice, each composite is formed only by water samples from one of the two areas, either the industrial or the agricultural. This is what we call "full segregation" of the composites.

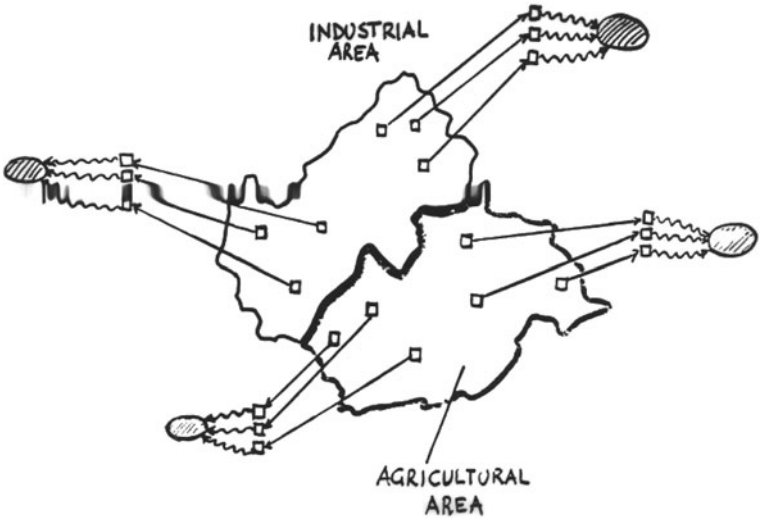

Fig. 8.5 Industrial vs. agricultural water quality example: the "fully segregated composite samples" case

We have

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & 0 & 0 \\ 0 & 0 & w_{21} & w_{22} \end{bmatrix},$$

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \boldsymbol{I}_4,$$

so that

$$\boldsymbol{D} = [\boldsymbol{WPGL}] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Clearly, rank$(\boldsymbol{D}) = 2$ so that segregation has preserved in $\boldsymbol{D}$ the full rank characterization of $\boldsymbol{F}$, and therefore μ_x and $\boldsymbol{\beta}_1$ are both estimable with the composited measurements. We get

$$[\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{D}] = [\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{P}'\boldsymbol{W}'(\boldsymbol{WW}')^{-1}\boldsymbol{WPGL}]$$

$$= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{w_{11}^2+w_{12}^2} & 0 \\ 0 & \frac{1}{w_{21}^2+w_{22}^2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{w_{11}^2+w_{12}^2} + \frac{1}{w_{21}^2+w_{22}^2} & \frac{1}{w_{11}^2+w_{12}^2} - \frac{1}{w_{21}^2+w_{22}^2} \\ \frac{1}{w_{11}^2+w_{12}^2} - \frac{1}{w_{21}^2+w_{22}^2} & \frac{1}{w_{11}^2+w_{12}^2} + \frac{1}{w_{21}^2+w_{22}^2} \end{bmatrix}.$$

Hence

$$|\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{D}| = \frac{4}{\left(w_{11}^2 + w_{12}^2\right)\left(w_{21}^2 + w_{22}^2\right)}$$

and

$$[\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{D}]^{-1}$$

$$= \begin{bmatrix} \frac{(w_{11}^2+w_{12}^2)+(w_{21}^2+w_{22}^2)}{4} & \frac{(w_{11}^2+w_{12}^2)-(w_{21}^2+w_{22}^2)}{4} \\ \frac{(w_{11}^2+w_{12}^2)-(w_{21}^2+w_{22}^2)}{4} & \frac{(w_{11}^2+w_{12}^2)+(w_{21}^2+w_{22}^2)}{4} \end{bmatrix}. \tag{8.45}$$

We also have

$$\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{y} = \begin{bmatrix} \frac{(w_{21}^2+w_{22}^2)y_1+(w_{11}^2+w_{12}^2)y_2}{(w_{11}^2+w_{12}^2)(w_{21}^2+w_{22}^2)} \\ \frac{(w_{21}^2+w_{22}^2)y_1-(w_{11}^2+w_{12}^2)y_2}{(w_{11}^2+w_{12}^2)y_2(w_{21}^2+w_{22}^2)} \end{bmatrix}. \tag{8.46}$$

Using (8.45) and (8.46) in (8.44), we obtain

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{y_1+y_2}{2} \\ \\ \frac{y_1-y_2}{2} \end{bmatrix}, \tag{8.47}$$

with

$$\text{Var}\left[\hat{\boldsymbol{\beta}}\right] = \sigma_\epsilon^2 \left[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}\right]^{-1}.$$

Therefore, with fully segregated composite samples, we find that

(a) μ_x and β_1 are estimable and their estimates are the "natural" ones: the composited mean for μ_x and the semi-difference between the measurements of the composite sample characterized by level 1 and the measurement on the composite samples characterized by level 2;
(b) inspection of (8.45) shows that the effect of compositing on the variances of $\hat{\mu}_x$ and $\hat{\beta}_1$ and on their covariance depends on the weights:

- If the weights are chosen to be different for the two composites, $w_{11} \neq w_{22}$, then the two estimates in $\hat{\boldsymbol{\beta}}$ are correlated (positively if $w_{11} > w_{21}$, negatively if $w_{11} < w_{21}$).
- If the weights are equal, the two estimates are uncorrelated and their variances attain the minimum when the weights are chosen to be uniform ($w_{11} = w_{12} = w_{21} = w_{22} = \frac{1}{2}$).

The latter result may be further formalized to show that indeed the uniform weights choice in the fully segregated case gives a variance/covariance matrix $\text{Var}\left[\hat{\boldsymbol{\beta}}\right]$ which satisfies the important property of being D-optimal, i.e., of having minimum determinant.

Recall that the variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{Var}\left[\hat{\boldsymbol{\beta}}\right] = \sigma_\epsilon^2 \left[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}\right]^{-1} = \sigma_\epsilon^2 \left[\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{P}'\boldsymbol{W}'\left(\boldsymbol{W}\boldsymbol{W}'\right)^{-1}\boldsymbol{W}\boldsymbol{P}\boldsymbol{G}\boldsymbol{L}\right]^{-1}.$$

The estimate $\hat{\boldsymbol{\beta}}$ is optimal if it has a "small" variance in some sense. The most common criterion for judging if the variance is reasonably small is to compute some scalar function of it. In the case of a variance/covariance matrix, the most widely used function is its determinant, which corresponds to the generalized variance

$$|\text{Var}\left[\hat{\boldsymbol{\beta}}\right]| = \sigma_\epsilon^2 |\boldsymbol{D}'\left(\boldsymbol{W}\boldsymbol{W}'\right)^{-1}\boldsymbol{D}|^{-1}.$$

Clearly, larger values of $|\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}|$ correspond to smaller values of $|\text{Var}\left[\hat{\boldsymbol{\beta}}\right]|$. The variance/covariance matrix $\text{Var}\left[\hat{\boldsymbol{\beta}}\right]$ with the minimum determinant, and hence with the minimum generalized variance, is said to be "D-optimal".

Recall that

$$|\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}| = \frac{4}{\left(w_{11}^2 + w_{12}^2\right)\left(w_{21}^2 + w_{22}^2\right)}.$$

If we define the quantity Q by

$$Q = Q(\mathbf{W}) = |\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{P}'\boldsymbol{W}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{W}\boldsymbol{P}\boldsymbol{G}\boldsymbol{L}|\,, \tag{8.48}$$

then, by imposing the constraints

$$w_{11} + w_{12} = 1 \quad \text{and} \quad w_{21} + w_{22} = 1\,,$$

we can write Q as the following function of w_{11} and w_{21} only

$$Q(\mathbf{W}) = \frac{4}{\left(2w_{11}^2 - 2w_{12}^2 + 1\right)\left(2w_{21}^2 - 2w_{22}^2 + 1\right)}.$$

It can be seen, either analytically or by inspection of the graph in Fig. 8.6, that

$$\max_{\mathbf{W}} Q(\mathbf{W}) = 16,$$

which is achieved when

$$w_{11} = w_{12} = 1/2 \quad \text{and} \quad w_{21} = w_{22} = 1/2\,,$$

which is the case of uniform weights.

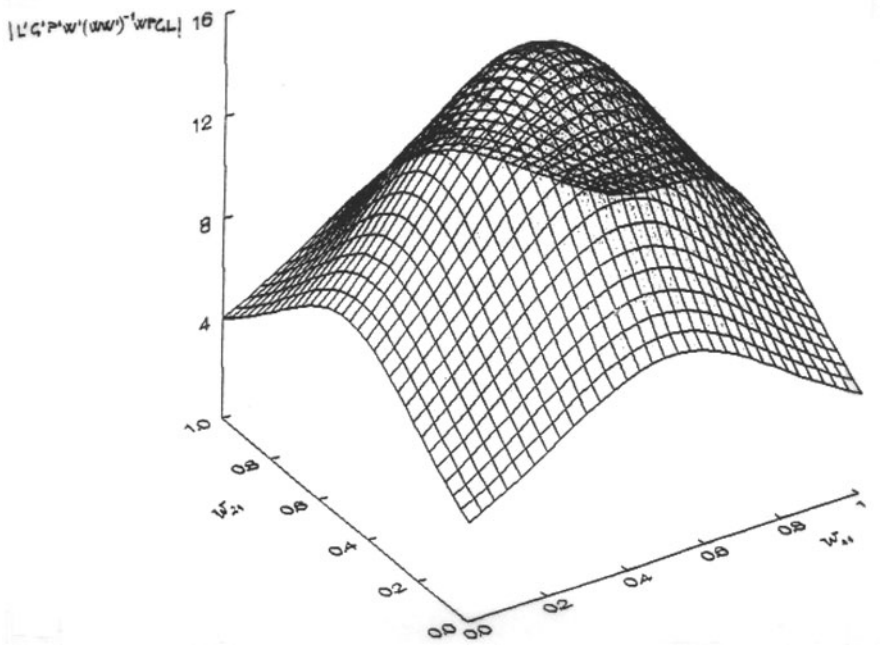

Fig. 8.6 Graph of $Q(\mathbf{W})$ defined in (8.48) vs. w_{11} and w_{21} for the "fully segregated composite samples" case

8.7.2 Fully Confounded Composites

Suppose now that the monitoring staff of the industrial vs. agricultural area example may afford only one measurement a day. To give a "fair" estimate of the average concentration level of the contaminant X, the analysts decide they must have samples from both areas in the composite everyday. Therefore, they form the composite taking everyday a different pair of wells, one in the industrial area and one in the agricultural area (Fig. 8.7). Clearly, with this choice, all the composites contain water from both areas. This is what we call "full confounding" of the composites.

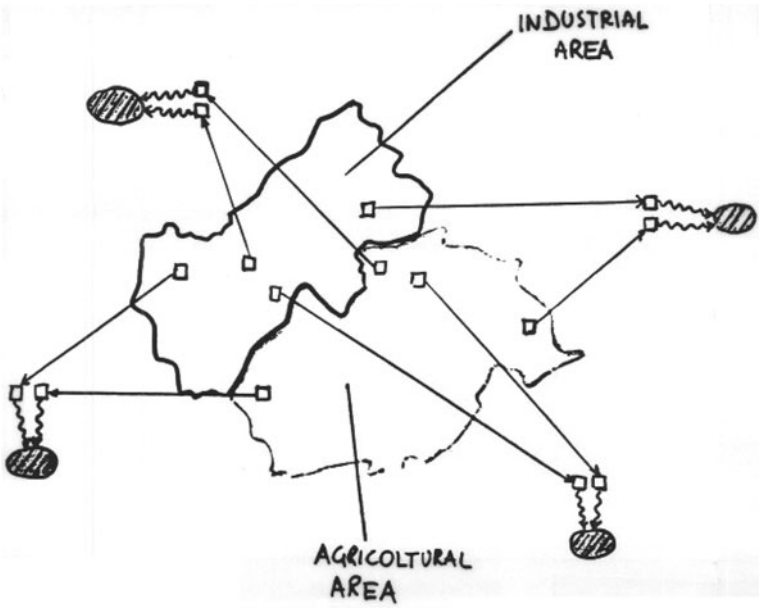

Fig. 8.7 Industrial vs. agricultural water quality example: the "fully confounded composite samples" case

Under this design, we have

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & 0 & 0 \\ 0 & 0 & w_{21} & w_{22} \end{bmatrix},$$

$$\boldsymbol{P} = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,1\,0\,0 \\ 0\,0\,0\,1 \end{bmatrix}.$$

Notice that the only difference from the "fully segregated composites" case is in matrix $\boldsymbol{P}$, which assigns the third individual sampling unit x_{21} to the first composite and the second sampling unit x_{12} to the second.

We obtain

$$\boldsymbol{D} = [\boldsymbol{WPGL}] = \begin{bmatrix} 1 & w_{11} - w_{12} \\ 1 & w_{21} - w_{22} \end{bmatrix}.$$

Clearly, unlike in the context of equation 2.1.1, the rank of $\boldsymbol{D}$ depends now on the choice of the weights:

(a) if the weights of the individual samples characterized by level 1 of $\boldsymbol{F}$ are equal over the two pairs, $w_{11} = w_{21} = w_1$, then the differences $(w_{11} - w_{12})$ and $(w_{21} - w_{22})$ are equal, and the second column of $\boldsymbol{D}$ is proportional to the first, i.e., rank $(\boldsymbol{D}) = 1$; in this instance, only the function $\mu_x + (1 - 2w_1)\beta_1$ is estimable. In particular if $w_{11} = w_{12} = \frac{1}{2}$, $j = 1, 2$, that is, if the two individual samples in each composited pair are taken of the same size (volume, mass, amount, etc.), then the second column of $\boldsymbol{D}$ is identically 0, μ_x is estimable, but no estimation of β_1 can be undertaken;

(b) estimation of both μ_x and β_1 is still possible if the differences of the weights $w_{11} - w_{12}$ are purposely chosen to be different from pair to pair, for then rank $(\boldsymbol{D}) = 2$:

$$[\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{D}] = [\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{P}'\boldsymbol{W}'(\boldsymbol{WW}')^{-1}\boldsymbol{WPGL}]$$

$$= \begin{bmatrix} 1 & 1 \\ w_{11} - w_{12} & w_{21} - w_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{w_{11}^2 + w_{12}^2} & 0 \\ 0 & \frac{1}{w_{21}^2 + w_{22}^2} \end{bmatrix} \begin{bmatrix} 1 & w_{11} - w_{12} \\ 1 & w_{21} - w_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{w_{11}^2 + w_{12}^2} + \frac{1}{w_{21}^2 + w_{22}^2} & \frac{w_{11} + w_{12}}{w_{11}^2 + w_{12}^2} + \frac{w_{21} + w_{22}}{w_{21}^2 + w_{22}^2} \\ \frac{w_{11} + w_{12}}{w_{11}^2 + w_{12}^2} + \frac{w_{21} + w_{22}}{w_{21}^2 + w_{22}^2} & \frac{(w_{11} + w_{12})^2}{w_{11}^2 + w_{12}^2} + \frac{(w_{21} + w_{22})^2}{w_{21}^2 + w_{22}^2} \end{bmatrix}.$$

Hence,

$$|\boldsymbol{D}'(\boldsymbol{WW}')^{-1}\boldsymbol{D}| = |\boldsymbol{L}'\boldsymbol{G}'\boldsymbol{P}'\boldsymbol{W}'(\boldsymbol{WW}')^{-1}\boldsymbol{WPGL}|$$

$$= \frac{[(w_{11} - w_{12}) - (w_{21} - w_{22})]^2}{(w_{11}^2 + w_{12}^2)(w_{21}^2 + w_{22}^2)}.$$

Inspection of $|\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}|$ gives another way of handling the estimability problem. If $w_{11} = w_{21}$, then $|\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}| = 0$, and the normal equations $[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}]\boldsymbol{\beta} = \boldsymbol{D}(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{y}$ cannot be uniquely solved by inverting $[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}]$ to give the GLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Now suppose the weights are chosen so as to give rank $(D) = 2$, then the matrix $[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}]$ is invertible and the inverse has the following form:

$$[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}]^{-1} =$$
$$\begin{bmatrix} \frac{(w_{11}-w_{12})^2(w_{21}^2+w_{22}^2)+(w_{21}-w_{22})^2(w_{11}^2+w_{12}^2)}{[(w_{11}-w_{12})(w_{21}-w_{22})]^2} & \frac{-(w_{11}-w_{12})(w_{21}^2+w_{22}^2)-(w_{21}-w_{22})(w_{11}^2+w_{12}^2)}{[(w_{11}-w_{12})(w_{21}-w_{22})]^2} \\ \frac{-(w_{11}-w_{12})(w_{21}^2+w_{22}^2)-(w_{21}-w_{22})(w_{11}^2+w_{12}^2)}{[(w_{11}-w_{12})(w_{21}-w_{22})]^2} & \frac{(w_{21}^2+w_{22}^2)+(w_{11}^2+w_{12}^2)}{[(w_{11}-w_{12})(w_{21}-w_{22})]^2} \end{bmatrix}.$$

We also find

$$\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{y} = \begin{bmatrix} \frac{(w_{21}^2+w_{22}^2)y_1+(w_{11}^2+w_{12}^2)y_2}{(w_{11}^2+w_{12}^2)(w_{21}^2+w_{22}^2)} \\ \frac{(w_{11}-w_{12})(w_{21}^2+w_{22}^2)y_1+(w_{21}-w_{22})(w_{11}^2+w_{12}^2)y_2}{(w_{11}^2+w_{12}^2)(w_{21}^2+w_{22}^2)} \end{bmatrix}, \tag{8.49}$$

and hence

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{(w_{21}^2-w_{22}^2)y_2-(w_{11}^2-w_{12}^2)y_1}{[(w_{11}-w_{12})(w_{21}-w_{22})]} \\ \frac{y_1-y_2}{[(w_{11}-w_{12})(w_{21}-w_{22})]} \end{bmatrix}, \tag{8.50}$$

with $\text{Var}\left[\hat{\boldsymbol{\beta}}\right] = \sigma_\epsilon^2[\boldsymbol{D}'(\boldsymbol{W}\boldsymbol{W}')^{-1}\boldsymbol{D}]^{-1}$.

From (8.49) and (8.50) it is seen that in the case of fully confounded composites:

(a) both μ_x and β_1 are still estimable; their estimates approach the fully segregated ones in (8.50) when $w_{11} \to 1$ and $w_{22} \to 1$, that is, when the weights are chosen to be as different as possible; at the extreme, when $(w_{11} = 1, w_{12} = 0)$ and $(w_{21} = 0, w_{22} = 1)$, segregation is again achieved by using only one individual sampling unit in each composite and the estimates in (8.50) are equal to those in (8.47).

(b) inspection of (8.50) shows that the more similar the two vectors of weights in the two composites, the larger the variances of $\hat{\mu}$ and $\hat{\beta}_1$ and their covariance; furthermore, covariance will be positive if both w_{11} and $w_{21} < \frac{1}{2}$; negative if both w_{11} and $w_{21} > \frac{1}{2}$.

Again, this latter result may be better summarized by studying the determinant of $V(\hat{\boldsymbol{\beta}})$, i.e., the conditions under which such matrix attains D-optimality.

With $Q(\mathbf{W})$ defined as in (8.48), we obtain

$$Q(\mathbf{W}) = \frac{4(w_{11} - w_{21})^2}{(2w_{11}^2 - 2w_{11} + 1)(2w_{21}^2 - 2w_{21} + 1)}.$$

It can be seen, either analytically or by inspection of the graph in Fig. 8.8, that

$$\max_{\mathbf{W}} Q(\mathbf{W}) = 4\,,$$

which is attained when

$$w_{11} = 1\,, \quad w_{12} = 0\,, \quad w_{21} = 0\,, \quad w_{22} = 1\,,$$

i.e., when full segregation is again achieved by using only one individual sample in each composite sample.

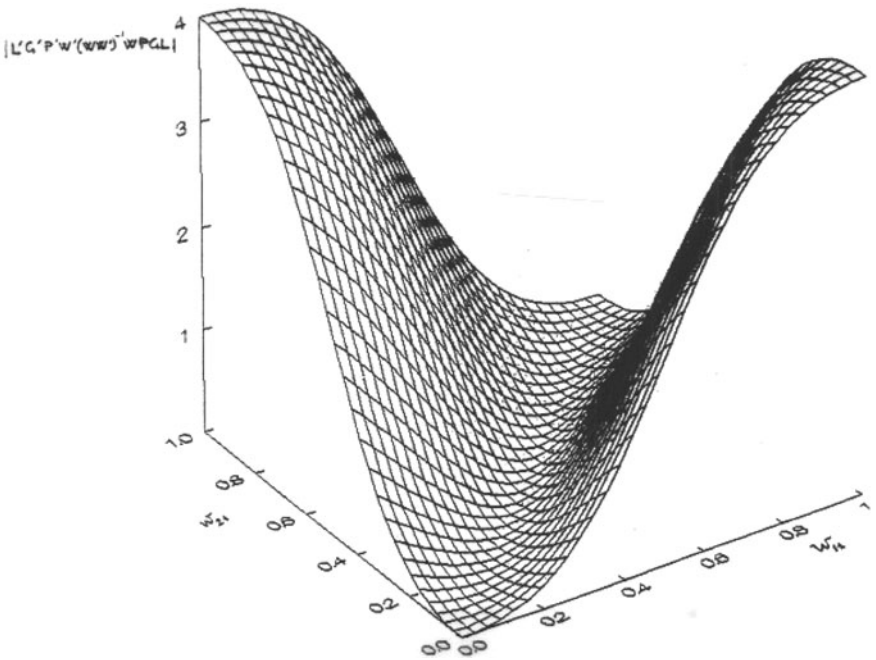

Fig. 8.8 Graph of $Q(\mathbf{W})$ defined in (8.48) vs. w_{11} and w_{21} for the "fully confounded composite samples" case

8.8 Elementary Matrices and Kronecker Products

An $n \times m$ matrix having a 1 in position i, j and 0 elsewhere is denoted by $\boldsymbol{E}_{ij}$ and is termed an elementary matrix (of order $n \times m$). The reference to the dimensions may

be dropped, unless it is necessary to avoid confusion, in which case the notation $\boldsymbol{E}_{ij}$ $(n \times m)$ may be used.

An n-vector having a 1 in position i and 0 elsewhere is denoted by $\boldsymbol{e}_i$ and is termed an elementary vector (of order n). Again, if the dimension is not obvious from the context, the notation $\boldsymbol{e}_i(n)$ should be used.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be any two matrices of order $n \times m$ and $p \times q$, respectively. Then their Kronecker product is defined as follows:

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1m}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2m}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \dots & a_{nm}\boldsymbol{B} \end{bmatrix}.$$

8.8.1 Decomposition of Block Matrices

Any $(cm \times cm)$ block matrix $\boldsymbol{A}$ of the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \dots & \boldsymbol{A}_{1c} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \dots & \boldsymbol{A}_{2c} \\ \vdots & \vdots & \boldsymbol{A}_{j\ell} & \vdots \\ \boldsymbol{A}_{c1} & \boldsymbol{A}_{c2} & \dots & \boldsymbol{A}_{cc} \end{bmatrix}, \tag{8.51}$$

where $\boldsymbol{A}$ is an $(m \times m)$ submatrix, may be represented as a function of its blocks as follows:

$$\begin{aligned} \boldsymbol{A} = & \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \\ \vdots & \vdots & \boldsymbol{O}_m & \vdots \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \end{bmatrix} + \begin{bmatrix} \boldsymbol{O}_m & \boldsymbol{A}_{12} & \dots & \boldsymbol{O}_m \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \\ \vdots & \vdots & \boldsymbol{O}_m & \vdots \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \end{bmatrix} \\ & + \cdots + \begin{bmatrix} \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{O}_m \\ \vdots & \vdots & \boldsymbol{O}_m & \vdots \\ \boldsymbol{O}_m & \boldsymbol{O}_m & \dots & \boldsymbol{A}_{cc} \end{bmatrix} \\ = & \sum_{j=1}^{c} \sum_{\ell=1}^{c} (\boldsymbol{E}_{j\ell} \otimes \boldsymbol{A}_{j\ell}). \end{aligned} \tag{8.52}$$

Notice that, conversely,

$$\boldsymbol{A}_{j\ell} = [\boldsymbol{O}_m \dots \boldsymbol{I}_m \dots \boldsymbol{O}_m]\,\boldsymbol{A}\begin{bmatrix}\boldsymbol{O}_m\\ \vdots\\ \boldsymbol{I}_m\\ \vdots\\ \boldsymbol{O}_m\end{bmatrix} = (\boldsymbol{e}'_j \otimes \boldsymbol{I}_m)\boldsymbol{A}(\boldsymbol{e}_\ell \otimes \boldsymbol{I}_m).$$

Hence,

$$\boldsymbol{A} = \sum_{j=1}^{c}\sum_{\ell=1}^{c}(\boldsymbol{E}_{j\ell} \otimes [(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m)\boldsymbol{A}(\boldsymbol{e}_\ell \otimes \boldsymbol{I}_m)]).$$

This may be written in a more symmetrical way using properties of the Kronecker product:

$$\begin{aligned}\boldsymbol{A} &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}(\boldsymbol{e}_j\boldsymbol{e}'_\ell) \otimes [(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m)\boldsymbol{A}(\boldsymbol{e}_\ell \otimes \boldsymbol{I}_m)]\\ &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}\{\boldsymbol{e}_j \otimes [(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m)\boldsymbol{A}]\}\{\boldsymbol{e}'_\ell \otimes (\boldsymbol{e}_\ell \otimes \boldsymbol{I}_m)\}\\ &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}(\boldsymbol{e}_j \otimes \boldsymbol{e}'_j \otimes \boldsymbol{I}_m)\boldsymbol{A}(\boldsymbol{e}'_\ell \otimes \boldsymbol{e}_\ell \otimes \boldsymbol{I}_m)\\ &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}(\boldsymbol{e}_j \otimes \boldsymbol{e}'_j) \otimes \boldsymbol{I}_m]\boldsymbol{A}[(\boldsymbol{e}'_\ell \otimes \boldsymbol{e}_\ell) \otimes \boldsymbol{I}_m]\end{aligned}$$

and, since for elementary vectors $\boldsymbol{e}_i \otimes \boldsymbol{e}'_k = \boldsymbol{e}_i\boldsymbol{e}'_k = \boldsymbol{E}_{ik}$:

$$= \sum_{j=1}^{c}\sum_{\ell=1}^{c}[\boldsymbol{E}_{jj} \otimes \boldsymbol{I}_m]\boldsymbol{A}[\boldsymbol{E}_{\ell\ell} \otimes \boldsymbol{I}_m]. \tag{8.53}$$

The representations of block matrices in (8.52) and (8.53) may be repeatedly used for matrices with more than one level of blocking, i.e., matrices whose blocks are block-submatrices, whose blocks are block sub-submatrices, and so on. For example, if each $\boldsymbol{A}_{j\ell}$ matrix in (8.51) is an $(sk \times sk)$ block matrix:

$$\boldsymbol{A}_{j\ell} = \begin{bmatrix} {}_{j\ell}\boldsymbol{B}_{11} & {}_{j\ell}\boldsymbol{B}_{12} & \cdots & {}_{j\ell}\boldsymbol{B}_{1s} \\ {}_{j\ell}\boldsymbol{B}_{21} & {}_{j\ell}\boldsymbol{B}_{22} & \cdots & {}_{j\ell}\boldsymbol{B}_{2s} \\ \vdots & \vdots & {}_{j\ell}\boldsymbol{B}_{ih} & \vdots \\ {}_{j\ell}\boldsymbol{B}_{s1} & {}_{j\ell}\boldsymbol{B}_{s2} & \cdots & {}_{j\ell}\boldsymbol{B}_{1s} \end{bmatrix} = \sum_{i=1}^{s}\sum_{h=1}^{s}(\boldsymbol{E}_{ih} \otimes_{j\ell} \boldsymbol{B}_{ih}), \tag{8.54}$$

where ${}_{j\ell}\boldsymbol{B}_{ih}$ is a $(k \times k)$ sub-submatrix, then

$$\begin{aligned} \boldsymbol{A} &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}\left[\boldsymbol{E}_{j\ell} \otimes \sum_{i=1}^{s}\sum_{h=1}^{s}(\boldsymbol{E}_{ih} \otimes_{j\ell} \boldsymbol{B}_{ih})\right] \\ &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}\left[\sum_{i=1}^{s}\sum_{h=1}^{s}\boldsymbol{E}_{j\ell} \otimes (\boldsymbol{E}_{ih} \otimes_{j\ell} \boldsymbol{B}_{ih})\right] \\ &= \sum_{j=1}^{c}\sum_{\ell=1}^{c}\left[\sum_{i=1}^{s}\sum_{h=1}^{s}\left(\boldsymbol{E}_{j\ell} \otimes \boldsymbol{E}_{ih}\right) \otimes_{j\ell} \boldsymbol{B}_{ih}\right]. \end{aligned} \tag{8.55}$$

This nested decomposition is useful when working with the variance/covariance matrix of the random matrix of weights $\boldsymbol{W}$, $\boldsymbol{\Sigma}_{\boldsymbol{W}}$, in the general case of $c > 1$ composites and $s > 1$ subsamples from each, since in this case there are two levels of blocking: composites and subsamples within composites. Here $m = ck$; we get

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sum_{j=1}^{c}\sum_{\ell=1}^{c}\sum_{i=1}^{s}\sum_{h=1}^{s}\left(\boldsymbol{E}_{j\ell}^{\text{bc}} \otimes \boldsymbol{E}_{ih}^{\text{bs}}\right) \otimes C(\boldsymbol{w}_{ji}, \boldsymbol{w}_{\ell h}), \tag{8.56}$$

where $C(\boldsymbol{w}_{ji}, \boldsymbol{w}_{\ell h})$ is the $ck \times ck$ covariance matrix of the vectors $\boldsymbol{w}_{ji}$ and $\boldsymbol{w}_{\ell h}$ of weights; $\boldsymbol{E}_{j\ell}^{\text{bc}}$ is a $c \times c$ between composites elementary matrix; $\boldsymbol{E}_{ih}^{\text{bs}}$ is an $s \times s$ between subsamples elementary matrix.

Owing to the presence of necessarily zero weights, the generic random vector of weights $\boldsymbol{w}_{ji}$ has the following structure:

$$\begin{matrix} & \text{composite 1} & \text{composite 2} & & \text{composite } i & & \text{composite } c & \\ \boldsymbol{w}_{ji} = \Big(& \mathbf{o}_k' & \mathbf{o}_k' & \cdots & \boldsymbol{w}_{ij}^{*\prime} & \cdots & \mathbf{o}_k' & \Big), \end{matrix}$$

where $\boldsymbol{w}_{ji}^{*}$ is the k-subvector generating the only non-zero weights.

As a consequence, the only non-null covariance and cross-covariance matrices are those between the subvectors of non-zero weights $C(\boldsymbol{w}_{ji}^{*}, \boldsymbol{w}_{\ell h}^{*})$ for all j, ℓ, i, and h. Thus, the covariance matrices in (8.56) are patterned as follows:

$$C(\boldsymbol{w}_{ji}, \boldsymbol{w}_{\ell h}) = E_{j\ell}^{\text{bc}} \otimes C(\boldsymbol{w}_{ji}^{*}, \boldsymbol{w}_{\ell h}^{*}). \tag{8.57}$$

Summarizing, we may write

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sum_{j=1}^{c}\sum_{\ell=1}^{c}\sum_{i=1}^{s}\sum_{h=1}^{s}\left(\boldsymbol{E}_{j\ell}^{\mathrm{bc}} \otimes \boldsymbol{E}_{ih}^{\mathrm{bs}}\right) \otimes \boldsymbol{E}_{j\ell}^{\mathrm{bc}} \otimes C(\boldsymbol{w}_{ji}^{*}, \boldsymbol{w}_{\ell h}^{*}). \tag{8.58}$$

8.9 Expectation and Dispersion Matrix When Both W and x Are Random

8.9.1 The Expectation of Wx

Let $\boldsymbol{x}$ be a random m-vector with

$$\begin{aligned} E[\boldsymbol{x}] &= \boldsymbol{\mu}_x, \\ \mathrm{Var}[\boldsymbol{x}] &= \boldsymbol{\Sigma}_x, \end{aligned}$$

and let $\boldsymbol{W}$ be a random $n \times m$ matrix such that

$$\begin{aligned} E[\boldsymbol{W}] = \boldsymbol{M}_{\boldsymbol{W}} &= \begin{bmatrix} \boldsymbol{\mu}'_{w_1} \\ \vdots \\ \boldsymbol{\mu}'_{w_n} \end{bmatrix}, \\ \mathrm{Var}[\boldsymbol{W}] = \boldsymbol{\Sigma}_{\boldsymbol{W}} &= \begin{bmatrix} \boldsymbol{\Sigma}_{w_1} & \boldsymbol{\Gamma}_{w_1,w_2} & \cdots & \boldsymbol{\Gamma}_{w_1,w_n} \\ \boldsymbol{\Gamma}_{w_2,w_1} & \boldsymbol{\Sigma}_{w_2} & \cdots & \boldsymbol{\Gamma}_{w_2,w_1} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\Gamma}_{w_n,w_1} & \boldsymbol{\Gamma}_{w_n,w_2} & \cdots & \boldsymbol{\Sigma}_{w_n} \end{bmatrix}. \end{aligned} \tag{8.59}$$

It is convenient to think of $\mathrm{Var}[\boldsymbol{W}]$ in terms of the "vec" operator:

$$\mathrm{Var}[\boldsymbol{W}] = E\left\{\left[\mathrm{vec}\left(\boldsymbol{W}' - \boldsymbol{M}'_{\boldsymbol{W}}\right)\right]\left[\mathrm{vec}\left(\boldsymbol{W}' - \boldsymbol{M}'_{\boldsymbol{W}}\right)\right]'\right\}. \tag{8.60}$$

Notice that, since $\mathrm{vec}\left(\boldsymbol{W}' - \boldsymbol{M}'_{\boldsymbol{W}}\right)$ is an $nm \times 1$ vector, $\mathrm{Var}[\boldsymbol{W}]$ is of order $nm \times nm$.

Furthermore, let us assume that $\boldsymbol{W}$ has its first two moments free of $\boldsymbol{x}$, i.e.,

$$E_W[\boldsymbol{W}|\boldsymbol{x}] = E[\boldsymbol{W}] \qquad \forall \boldsymbol{x}, \tag{8.61}$$

$$\mathrm{Var}_W[\boldsymbol{W}|\boldsymbol{x}] = \mathrm{Var}[\boldsymbol{W}] \qquad \forall \boldsymbol{x}, \tag{8.62}$$

where E_W and Var_W denote the expectation and variance/covariance operators with respect to the random matrix $\boldsymbol{W}$. It immediately follows that

$$E[\boldsymbol{W}\boldsymbol{x}] = E_W[(\boldsymbol{W}|\boldsymbol{x})]\, E_x[\boldsymbol{x}] = E_W[\boldsymbol{W}]\, E_x[\boldsymbol{x}] = \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\mu}_x. \tag{8.63}$$

The derivation of $\text{Var}[\boldsymbol{W}\boldsymbol{x}]$ is less straightforward, and we need a few intermediate results to show it.

Result 1 *Let $\boldsymbol{e}_j$ be an elementary n-vector having a* 1 *in position j and* 0 *elsewhere. Then*

$$E_x\left\{\text{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{x}'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{x}\right)\right]\right\} = \left(\boldsymbol{e}_j' \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x'\right) + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\Sigma}_x\right)\right]. \tag{8.64}$$

Proof By cyclic commutativity of the trace operator

$$E_x\left\{\text{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{x}'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{x}\right)\right]\right\} = E_x\{\text{tr}[\boldsymbol{\Sigma}_{\boldsymbol{W}}(\boldsymbol{e}_{j'} \otimes \boldsymbol{x})(\boldsymbol{e}_j' \otimes \boldsymbol{x}').$$

Applying the Kronecker mixed product rule, the left-hand side of (8.64) becomes

$$\begin{aligned}
&= E_x\left\{\text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{x}\boldsymbol{x}'\right)\right]\right\} \\
&= \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes E\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)\right] \\
&= \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \left(\boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)\right)\right],
\end{aligned}$$

and, by distributivity of the Kronecker product, this is

$$= \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\Sigma}_x\right) + \boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)\right],$$

by Kronecker mixed product rule

$$= \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\Sigma}_x\right)\right] + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x\right)\left(\boldsymbol{e}_j' \otimes \boldsymbol{\mu}_x'\right)\right],$$

by cyclic commutativity of the trace operator

$$= \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\Sigma}_x\right)\right] + \text{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x\right)\right],$$

finally, since the argument of the second trace is a scalar, we obtain

$$= \left(\boldsymbol{e}_j' \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x\right) + \text{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'}\boldsymbol{e}_j' \otimes \boldsymbol{\Sigma}_x\right)\right],$$

which completes the proof.

Result 2 *With* $\boldsymbol{e}_j$ *as in Result* 1, *we have*

$$E_x\left\{\left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{I}_n \otimes \boldsymbol{x}\right]\right\} = \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x'\right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x\right)\right] + \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j tr\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{I}_m\right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right) \boldsymbol{\Sigma}_x\right] \boldsymbol{e}_{j'}'. \tag{8.65}$$

Proof On using (8.53), we may emphasize the patterned structure of $\boldsymbol{\Sigma}_{\boldsymbol{W}}$ given in (8.59):

$$\boldsymbol{\Sigma}_{\boldsymbol{W}} = \sum_{j=1}^{n}\sum_{j'=1}^{n} \left(\boldsymbol{E}_{jj} \otimes \boldsymbol{I}_m\right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{E}_{j'j'} \otimes \boldsymbol{I}_m\right).$$

Hence,

$$\begin{aligned} &E_x\left\{\left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{I}_n \otimes \boldsymbol{x}\right]\right\} \\ &= E_x\left\{\left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right]\left[\sum_{j=1}^{n}\sum_{j'=1}^{n} \left(\boldsymbol{E}_{jj} \otimes \boldsymbol{I}_m\right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{E}_{j'j'} \otimes \boldsymbol{I}_m\right)\right]\left[\boldsymbol{I}_n \otimes \boldsymbol{x}\right]\right\} \\ &= E_x\left\{\sum_{j=1}^{n}\sum_{j'=1}^{n} \left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right]\left[\boldsymbol{E}_{jj} \otimes \boldsymbol{I}_m\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{E}_{j'j'} \otimes \boldsymbol{I}_m\right]\left[\boldsymbol{I}_n \otimes \boldsymbol{x}\right]\right\}, \end{aligned}$$

by the Kronecker mixed product rule

$$\begin{aligned} &= E_x\left\{\sum_{j=1}^{n}\sum_{j'=1}^{n} \left[\boldsymbol{E}_{jj} \otimes \boldsymbol{x}'\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{E}_{j'j'} \otimes \boldsymbol{x}\right]\right\} \\ &= \sum_{j=1}^{n}\sum_{j'=1}^{n} E_x\left\{\left[\boldsymbol{E}_{jj} \otimes \boldsymbol{x}'\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{E}_{j'j'} \otimes \boldsymbol{x}\right]\right\} \\ &= \sum_{j=1}^{n}\sum_{j'=1}^{n} E_x\left\{\left[\boldsymbol{e}_j \boldsymbol{e}_j' \otimes \boldsymbol{x}'\right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\boldsymbol{e}_{j'} \boldsymbol{e}_{j'}' \boldsymbol{x}\right]\right\} \end{aligned}$$

using again the Kronecker mixed product rule

$$= \sum_{j=1}^{n}\sum_{j'=1}^{n} E_x\left\{\boldsymbol{e}_j\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{x}'\right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{x}\right)\right] \boldsymbol{e}_{j'}'\right\}$$

since the quantity in square brackets is a scalar

$$= \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j E_x \left\{ \mathrm{tr}\left[\left(\boldsymbol{e}'_j \otimes \boldsymbol{x}' \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{x} \right) \right] \right\} \boldsymbol{e}'_{j'}.$$

Thus, using (8.64),

$$E_x \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{x}' \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{I}_n \otimes \boldsymbol{x} \right) \right] = \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \left(\boldsymbol{e}'_j \otimes \boldsymbol{\mu}'_x \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x \right) \boldsymbol{e}'_{j'}$$
$$+ \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \boldsymbol{e}'_j \otimes \boldsymbol{\Sigma}_x \right) \right] \boldsymbol{e}'_{j'}.$$

Now,

$$\begin{aligned}
&\sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \left(\boldsymbol{e}'_j \otimes \boldsymbol{\mu}'_x \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{\mu}_x \right) \boldsymbol{e}'_{j'} \\
&= \sum_{j=1}^{n}\sum_{j'=1}^{n} \left(\boldsymbol{e}_j \boldsymbol{e}'_j \otimes \boldsymbol{\mu}'_x \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \boldsymbol{e}'_j \otimes \boldsymbol{\mu}_x \right) \\
&= \left[\sum_{j=1}^{n} \left(\boldsymbol{e}_j \boldsymbol{e}'_j \otimes \boldsymbol{\mu}'_x \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \sum_{j'=1}^{n} \left(\boldsymbol{e}'_{j'} \boldsymbol{e}'_{j'} \otimes \boldsymbol{\mu}_x \right) \right] \\
&= \left[\left(\sum_{j=1}^{n} \boldsymbol{E}_{jj} \right) \otimes \boldsymbol{\mu}'_x \right] \boldsymbol{\Sigma}_{\boldsymbol{W}} \left[\left(\sum_{j=1}^{n} \boldsymbol{E}_{j'j'} \right) \otimes \boldsymbol{\mu}_x \right] \\
&= \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}'_x \right) \boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x \right) \right]
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \boldsymbol{e}'_j \otimes \boldsymbol{\Sigma}_x \right) \right] \boldsymbol{e}'_{j'} \\
&= \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} 1 \boldsymbol{e}'_j \right) \otimes \left(\boldsymbol{I}_m \boldsymbol{\Sigma}_x \boldsymbol{I}_m \right) \right] \boldsymbol{e}'_{j'} \\
&= \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{W}} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m \right) \boldsymbol{\Sigma}_x \left(\boldsymbol{e}'_j \otimes \boldsymbol{I}_m \right) \right] \boldsymbol{e}'_{j'}
\end{aligned}$$

by cyclic commutativity of the trace operator

$$= \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_x\right]\boldsymbol{e}_{j'}'$$

and result (8.65) follows. This completes the proof of Result 2.

8.9.2 Variance/Covariance Matrix of $\boldsymbol{Wx}$

We can now derive the following expression for the variance/covariance matrix of $\boldsymbol{Wx}$:

$$\begin{aligned}\mathrm{Var}[\boldsymbol{Wx}] &= \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_x\boldsymbol{M}_{\boldsymbol{W}}' + \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x\right)\right] \\ &\quad + \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_x\right]\boldsymbol{e}_{j'}' . \qquad (8.66)\end{aligned}$$

For the proof, we write $\widetilde{\boldsymbol{W}} = \boldsymbol{W} - \boldsymbol{M}_{\boldsymbol{W}}$ so that

$$\begin{aligned}\mathrm{Var}[\boldsymbol{Wx}] &= E_x\left\{\mathrm{Var}_w[\boldsymbol{Wx}|\boldsymbol{x}]\right\} + \mathrm{Var}_x\left\{E_w\left[\boldsymbol{Wx}|\boldsymbol{x}\right]\right\} \\ &= E_x\left[E_w\left\{\left(\boldsymbol{Wx} - \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{x}\right)\left(\boldsymbol{Wx} - \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{x}\right)'\right\}\right] + \mathrm{Var}_x\left[\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{x}\right] \\ &= E_x\left\{E_w\left[\left(\boldsymbol{W} - \boldsymbol{M}_{\boldsymbol{W}}\right)\left(\boldsymbol{x}\boldsymbol{x}'\right)\left(\boldsymbol{W} - \boldsymbol{M}_{\boldsymbol{W}}\right)'\right]\right\} + \mathrm{Var}_x\left[\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{x}\right] \\ &= E_x\left\{E_w\left[\left(\widetilde{\boldsymbol{W}}\boldsymbol{x}\right)\left(\boldsymbol{x}'\widetilde{\boldsymbol{W}}'\right)\right]\right\} + \boldsymbol{M}_{\boldsymbol{W}}\mathrm{Var}[\boldsymbol{x}]\boldsymbol{M}_{\boldsymbol{W}}' \\ &= E_x\left\{E_w\left[\left(\boldsymbol{I}_n \otimes \boldsymbol{x}'\right)\left(\mathrm{vec}\,\widetilde{\boldsymbol{W}}'\right)\left(\mathrm{vec}\,\widetilde{\boldsymbol{W}}'\right)'\left(\boldsymbol{I}_n \otimes \boldsymbol{x}\right)\right]\right\} \\ &\quad + \boldsymbol{M}_{\boldsymbol{W}}\mathrm{Var}[\boldsymbol{x}]\boldsymbol{M}_{\boldsymbol{W}}' \\ &= E_x\left\{\left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right]E_w\left[\left(\mathrm{vec}\,\boldsymbol{W}'\right)\left(\mathrm{vec}\,\boldsymbol{W}'\right)'\right]\left(\boldsymbol{I}_n\boldsymbol{x}\right)\right\} \\ &\quad + \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_x\boldsymbol{M}_{\boldsymbol{W}}' \\ &= E_x\left\{\left[\boldsymbol{I}_n \otimes \boldsymbol{x}'\right]\mathrm{Var}[\boldsymbol{W}]\left[\boldsymbol{I}_n \otimes \boldsymbol{x}\right]\right\} + \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_x\boldsymbol{M}_{\boldsymbol{W}}'.\end{aligned}$$

On using (8.65), the chain of equalities continues as

$$\begin{aligned}&= \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_x\boldsymbol{M}_{\boldsymbol{W}}' + \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x\right)\right] \\ &\quad + \sum_{j=1}^{n}\sum_{j'=1}^{n} \boldsymbol{e}_j \mathrm{tr}\left[\left(\boldsymbol{e}_j' \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m\right)\boldsymbol{\Sigma}_x\right]\boldsymbol{e}_{j'}' .\end{aligned}$$

This result can also be written as

$$\mathrm{Var}[\boldsymbol{Wx}] = \boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{\Sigma}_x\boldsymbol{M}_{\boldsymbol{W}}' + \left[\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x'\right)\boldsymbol{\Sigma}_{\boldsymbol{W}}\left(\boldsymbol{I}_n \otimes \boldsymbol{\mu}_x\right)\right]$$

$$+ \sum_{j=1}^{n} \sum_{j'=1}^{n} \boldsymbol{e}_j \left[\text{vec} \left(\boldsymbol{e}_j' \otimes \boldsymbol{I}_m \right) \right]' \left(\boldsymbol{\Sigma}_{\boldsymbol{W}} \otimes \boldsymbol{\Sigma}_x \right) \left[\text{vec} \left(\boldsymbol{e}_{j'} \otimes \boldsymbol{I}_m \right)' \right] \boldsymbol{e}_{j'}'.$$

Notice that, in the special case of a linear combination $\boldsymbol{w}'\boldsymbol{x}$ of random vectors (i.e., $n = 1$), (8.66) reduces to

$$\text{Var}\left[\boldsymbol{w}'\boldsymbol{x}\right] = \boldsymbol{\mu}_w' \boldsymbol{\Sigma}_x \boldsymbol{\mu}_w + \boldsymbol{\mu}_x' \boldsymbol{\Sigma}_w \boldsymbol{\mu}_x + \text{tr}\left[\boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_x\right],$$

as already noted by Elder et al. (1980).

Chapter 9
Composite Sampling for Site Characterization and Cleanup Evaluation

9.1 Data Quality Objectives

Environmental data are collected for making and/or defending certain decisions. Using data quality objectives (DQO) to plan new data collection activities helps assure that the right type and quality of information will be collected and that the analysis of the collected information will lead to efficient decisions.

The DQO process is a total quality management tool developed by the US Environmental Protection Agency (EPA) to facilitate the planning of data collection activities. It requires the planners to focus their planning by specifying the application (use of the data), the decision criteria, and the acceptable probability of making an incorrect decision. The process is suitable for sequential consideration of relevant issues.

DQOs are specifications required to design an investigation. These specifications can be put in the form of a process, where each step of the process identifies a specification. The DQO process developed by EPA has the following steps:

1. State the problem
2. Identify the decision to be made
3. Identify inputs to the decision
4. Define the boundaries of the study
5. Develop decision rule
6. Specify limits on uncertainty
7. Optimize design for obtaining data

Data quality objectives enable the investigator to evaluate the potential consequences of uncertainty before data collection is undertaken and to specify the acceptable range for uncertainty in the decision that will be based on the results of the investigation. Most importantly, the DQO process can save resources by identifying efficient and cost-effective data collection protocols.

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_9, © Springer Science+Business Media, LLC 2011

The seven stages of the DQO process are described briefly below.

1. State the problem to be resolved.

A clear and well-stated description of the problem and any resources, time, or other restrictions on the data collection activity are essential in order to optimally utilize the available resources.

State the problem.

- Describe the problem as you currently understand it.
- Consider the importance of social and political considerations to the problem.
- Organize and review relevant information, including preliminary studies, and indicate the source and reliability of the information. Conduct literature searches and explore ongoing studies to ensure that the problem has not previously been resolved.
- If it is a complex problem, then organize your understanding of it by identifying components of the problem, each of which could potentially be addressed by a separate study. Try to prioritize among these components for further planning. Examine whether new data are critical to resolving this problem.

2. Identify the decision.

- Describe initial ideas on approaches to resolving the problem.
- State the range of actions that might be taken based on the outcome of this study. Consider agency policies that may influence these actions (e.g., agency emphasis on pollution prevention rather than source containment or treatment).
- Specify the criteria for taking these actions as specific "If . . . , then . . . " scenarios when possible. If the criteria are not known at this time, then specify how they will be established.
- State the decision as a choice among alternative courses of action that will resolve one or more components of the problem

The decision maker (data user) should be involved in this step and is encouraged to provide general guidance for taking action.

3. Identify the inputs.

Make a list of environmental variables or characteristics to be measured, the criteria for taking action, and any other information needed to make the decision. Confirm that each environmental characteristic in the list can be measured.

4. Define the scope of the study.

It is necessary to have a detailed description of the population for which the decision will be made, including area and time period involved. When the population consists of people and objects, it is important to define space and time boundaries and the other characteristics that determine what belongs in and out of the population.

Alternatively, the population may consist of a continuous medium (air, water, soil). In this case, the portion of this medium that belongs to the population can

usually be defined just by the spatial and temporal boundaries, although there may be other characteristics that help to define it. For example, a survey of the toxic contaminant levels in surface waters of the population could consist of the top 10–12 m of water (the epilimnetic waters). Other characteristics might include meteorological conditions (wind speeds less than 15 mph, for instance).

Sometimes we are not able to sample from the entire population. In this case, we either make inferences only to that portion of the population that can be measured or we make assumptions that allow us to infer to the entire population.

Make sure that the practical considerations (identified in the problem step) are consistent with these boundaries.

5. Develop a Decision Rule.

Prepare a statement that defines how environmental data will be summarized and used to make the decision.

After the data have been collected, they are summarized to form a result, which is compared to the criteria for taking action to make the decision. The purpose of this step is to integrate the outputs from previous steps into a single statement specifying how environmental data will be summarized and used to make the decision, including quantitative criteria for determining what action to take.

It is important that someone with statistical expertise be involved in this step to be certain that the decision rule is stated in a manner that leads to an efficient design.

Develop a decision rule as an "If . . ., then . . . " statement that incorporates the study result, the criteria for taking action, and the actions(s) that will be taken under various possible scenarios.

Confirm that you will need all the data you will be collecting. If not, then define a more narrowly focused set of input variables.

6. Specify limits on uncertainty.

There is always some error in environmental data. As a result, some degree of uncertainty will exist in any decision based on environmental data. The limits on uncertainty should be based on careful consideration of the consequences of incorrect conclusions. There are two types of decision errors for all studies that will support a decision on whether or not to take action: false-positives and false-negatives. The definition of what constitutes false-positive and false-negative errors depends on how the decision is defined. In this step, limits on uncertainty are established and stated as acceptable probabilities of making incorrect decisions; i.e., acceptable false-positive and false-negative error rates for the decision.

Define false-positive (f(+)) errors for the decision and describe scenarios in which each type of error might take place.

Determine if false-positive or false-negative errors are of greater concern.

Establish an acceptable probability for the occurrence of each of these errors. Also, specify a region of indifference, the area in which you choose not to control the probability of an incorrect outcome because, under the stated conditions, either decision is acceptable. This region may be narrow or broad and must be acceptable to the decision maker.

Combine the probability statements into a formal statement of the levels of uncertainty that can be tolerated in the results.

7. Optimize the design.

In this step, statistical techniques are used to develop and evaluate various designs for the study that meet the specifications from the DQO process. The data collected using these designs should enable the decision to be made subject to error rates no greater than those specified in the limits on uncertainty (given that the assumptions on which the design was developed hold true).

Obtain the information needed to develop alternative designs: the limits on uncertainty from the proceeding step; any budget or time constraints; any practical considerations; cost estimates for all study activities; estimates of the inherent variability of variables or environmental characteristics to be measured; and estimates of the variability that will be introduced by the sampling and measurement process.

Select the most cost-efficient design that has acceptable performance and meets all other needs of the decision maker including political and social concerns.

Confirm that the design will yield useful results even when conditions are more adverse than those expected or assumed.

If it appears that there is no design that will meet both the limits on uncertainty and the budget constraints, then determine whether to compromise by relaxing the limits on uncertainty or other practical constraints or by finding additional funding to achieve the desired limits on uncertainty within the specified boundaries for the study.

9.2 Optimal Composite Designs

The concept of optimality is very important in environmental sampling in general and composite sampling in particular. Unlike in hard sciences, several factors affect the sampling units resulting in high variability of the environmental characteristic being observed. On the other hand, due to high costs of sampling and measurements, the sample size can at most be moderate, making it more difficult to keep the errors within stipulated bounds. It is therefore very important to optimize both sampling and compositing designs.

Suppose the variability of composite sample measurements is modeled, as in the preceding chapter, in terms of a linear model. Then the total variation among composite sample measurements will have several variance components: σ_t^2, the variance due to measurement error; σ_w^2, the variance component that accounts for randomness of compositing weights; σ^2, the sampling variance or the population variance; and σ_c^2, the variance component due to variability among individual sample values within a single composite sample. Each of these variance components is estimable, and their respective estimators are as follows: σ_t^2 is estimated by the variability among measurements on the s subsamples drawn from the same individual or composite sample and σ^2 is estimated by the sample variance for individual samples.

9.2.1 Cost of a Sampling Program

Suppose the composite sample size is denoted by k, the number of composite samples by n, and the number of individual samples by m. Due to the possibility of a measurement error, measurements are made on s subsamples of every sample, either individual or composite. While the total sampling effort is measured by the number of individual samples selected (m), the total analytical effort is measured by the total number of measurements (ms when individual samples are used to make measurements and ns when composites samples are used to make measurements). Further, suppose c_s is the cost of sampling per sampling unit (either individual or composite), c_t is the cost of measurement (testing) per sample, again either individual or composite, and c_k the cost of collecting and processing an individual sample before compositing. With these components of the cost of sampling and the cost of measurement, a cost function can be defined that gives the cost of an analysis, which covers both sampling and analytical tests. If the total cost is denoted by C, then we have, for a situation where individual samples are subjected to measurements:

$$C = mc_s + snc_t.$$

Since a composite sample is made up of k individual sampling units, we can replace c_s with kc_k to yield the following equation:

$$C = n\left(kc_k + sc_t\right).$$

For specified values of s and k, this equation can be inverted to yield

$$n = \frac{C}{kc_k + sc_t}.$$

This enables one to compute the number of composite samples that one can afford.

9.2.2 Optimal Allocation of Resources

For sampling of individual samples, the optimum number of analytical tests to make on each individual sample can be computed directly using the following formula:

$$\hat{s} = \sqrt{\frac{c_s S_t^2}{c_t S^2}},$$

where S_t^2 is an estimate of σ_t^2 and S^2 is an estimate of σ^2. The value of $\hat{s}$ is truncated to an integer and constrained to be greater than or equal to unity.

Analogous formulas can be derived for s and k in composite sampling by simultaneously minimizing both the variance of a sample mean and the cost of a sampling

plan. This is done by computing the derivative (with respect to s and k) of the product of cost and variance, setting these derivatives equal to zero, and then solving for the optimal values of s and k (Kendall and Stuart, 1966). The number of composite samples, n, is determined so that the plan is affordable and the desired statistical test has sufficient statistical power. The number of composite samples must also be greater than 1 so that an estimate of the sampling variance can be obtained.

The results depend on the model for the compositing weights. If the compositing weights are fixed and all equal to $\frac{1}{k}$, then explicit formulas cannot be obtained for s and k except for their ratio

$$\frac{\hat{s}}{\hat{k}} = \sqrt{\frac{S_t^2 c_k}{S^2 c_t}}.$$

Thus an iterative or trial and error approach is required as described below. Given a pair of s and k values (in the above ratio), n can be computed so as to achieve a desired total cost using the following formula:

$$\hat{n} = \frac{C}{kc_k + sc_t}.$$

Alternatively, one can solve for a value of n that yields a specified standard error of a sample mean using the following formula:

$$\hat{s} = \frac{\frac{S_t^2}{s} + \frac{S^2}{k}}{S_{\bar{Y}}^2},$$

where $S_{\bar{Y}}^2$ is an estimate of the composite sample variance. If $\hat{s} < 1$, then a single replicate should be used. For this model k should be as large as one can afford.

The value for m depends on the limitation of funds and the desired statistical power one wishes to have for testing differences between subpopulations. The formula for m to stay within a fixed cost is as given above. The formula to achieve a desired standard error for this model is

$$\hat{m} = \frac{\frac{1}{k^2}\frac{S_t^2}{s} + \frac{S^2}{k}}{S_{\bar{Y}}^2}.$$

9.2.3 Power of a Test and Determination of Sample Size

A sampling program needs to have sufficient sample sizes so that if there are important differences between subpopulations, they are likely to be detected statistically. Sokal and Rohlf (1981, p. 263.) give the following relationship to determine the required sample size for comparing two means:

$$n = 2\left(\frac{\sigma}{\delta}\right)^2 \left(t_{\alpha[v]} + t_{2(1-p)[v]}\right)^2,$$

where δ is the smallest true difference that it is desired to detect, σ is the true standard deviation for a sample, v is the degrees of freedom of the estimate of σ, and α is the significance level one plans to use. The probability p is the desired probability that an observed difference, as small as δ, will be found to be statistically significant. It is 1 minus the probability of a type II error. This formula is based on the assumption that the difference between means follows the normal distribution.

In the case of composite samples, this formula can be expressed as

$$\hat{n}\hat{s} = 2\frac{\text{EMS}}{\delta^2}(t_{\alpha[v]} + t_{2(1-k)[v]})^2,$$

where EMS is the expected value for the mean square for differences among sample values and v is the degrees of freedom that one will have for the mean square in a statistical analysis based on the planned sampling program (usually $v = m - 1$). The other symbols are as defined above. This equation must be used as part of an iterative cycle since the degrees of freedom, v, depend on m and the expected mean square depends on both n and k. The solution of this equation is further complicated by the fact that for small sample sizes one cannot ignore the fact that n, k, and m are integers. Unique solutions are not always possible.

9.2.4 Algorithms for Determination of Sample Size

The procedures to be used for determining the sample size vary according to whether one wishes to find the best sampling design (minimum δ) for a fixed cost or the least expensive sampling design that can detect a specified δ. For each problem the formulas differ slightly and hence are both considered here.

In all cases one must consider transformations, removal of outliers, robust estimation methods, etc., so that the means are more or less normally distributed. One then needs to obtain good estimates of the variance due to measurement error, σ_t^2, and the variance among the individual sampling units σ^2. This latter quantity may be difficult to obtain if the reason for using composite samples is that analytical errors are very high when measuring a single individual sampling unit. If the data have been transformed then these estimates must also be obtained for the transformed variables. The cost, C_t, of each measurement on a composite sample and the cost, C_k, of each individual sampling unit are also needed. There may be other costs associated with collecting data. What is important in this context is the cost of adding another individual sampling unit to a composite or to make another replicate measurement. The "total cost" being modeled is the variable part of the experiment after any initial setup costs. One also needs to make a decision about what levels of type I and type II errors one can tolerate and the smallest differences that are important to detect.

9.2.4.1 Finding the Best Sampling Plan for a Fixed Cost, *C*

1. Compute the ratio $r = \sqrt{\frac{S_t^2 c_k}{S^2 c_t}}$.
2. For integer values of n and $k = \frac{n}{r}$ ($k \geq 1$), let $s = \frac{C}{kc_k + nc_t}$.
   The number of samples should be truncated to an integer and constrained to be equal to or greater than 2 so that an estimate of the variance can be obtained in the planned design.
3. For each feasible combination of n, k, and s estimate δ, the smallest difference that one can expect to detect in the planned experiment:

$$\delta^2 = \frac{2}{b}\left(\frac{S_t^2}{n} + \frac{S_k^2}{k}\right)\left(t_{\alpha[v]} + t_{2(a-k)[v]}\right)^2.$$

   The combination that yields the smallest δ is the optimal design.

9.2.4.2 Finding the Least Expensive Sampling Plan That Can Detect a Specified Difference

1. Compute the ratio $r = \sqrt{\frac{S_t^2 C_k}{S^2 c_t}}$. For integer values of n and $k = \frac{n}{r}$ (k rounded to an integer and constrained to be ≥ 1), compute

$$\hat{m} = \frac{2}{\delta^2}\left(\frac{S_t^2}{n} + \frac{S^2}{k}\right)\left(t_{\alpha[v]} + t_{2(1-k)[v]}\right)^2.$$

   The degrees of freedom, $v = m - 1$, are a function of m; so this equation must be solved iteratively. The result should be truncated to an integer and constrained to be greater than or equal to 2 so that an estimate of the variance can be obtained in the planned design.
2. For each feasible combination of n, k, and m compute the cost of each group in the planned experiment:

$$C = m\left(kc_k + nc_t\right).$$

Information about the underlying distribution of the individual sampling units is needed in order to meet assumptions of a particular statistical method. According to the central limit theorem, minor deviations from normality are not important, especially if sample sizes are not small.

Chapter 10
Spatial Structures of Site Characteristics and Composite Sampling

10.1 Introduction

Environmental samples are most often collected at sites and therefore cannot be considered stochastically independent of one another. There is a common underlying contamination diffusion process that affects all samples, possibly in varying degrees. As a consequence, the samples collected at a particular site can be viewed as a realization of the corresponding spatial point process. It is then obvious that a statistical analysis of such data involves not only the overall population mean and variance but also parameters of the spatial process such as components of the variability of the process, spatial autocorrelation among sampling locations. In particular, the interest is in the trend, which corresponds to the expectation of the process, and spatial autocorrelation, which is usually characterized by the variogram, semivariogram, or covariogram. There is also an interest in identifying the components of variability, especially the scale of variability in comparison with the scale of sampling, which is measured in terms of the distance between successive sampling locations.

Spatial issues that are important in environmental sampling are (1) what scale of sampling is necessary to adequately sample a particular site and (2) how to design compositing of individual samples in order to limit the effect of micro-scale variability on the spatial patterns at a larger scale. We also model the spatial variability while considering the above two issues. Though there is no unique way to model spatial variability of a spatial process, we propose an interpretation that appears reasonable to us. In this chapter, we discuss the statistical model of sampling and considerations of designing the composite sampling plan.

10.2 Models for Spatial Processes

A spatial process is a stochastic (i.e., random) process indexed by the location of the sample point. These points may or may not form a lattice set. The data may be continuous or discrete. Let $\boldsymbol{s} \in \mathcal{R}^d$ be a generic sampling point in d-dimensional Euclidean space and suppose the potential value $X(\boldsymbol{s})$ at spatial location $\boldsymbol{s}$ is a

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_10, © Springer Science+Business Media, LLC 2011

random variable. Now let $\boldsymbol{s}$ vary over the index set $D \subset \mathcal{R}^d$, which will represent the site being sampled so as to generate the random process

$$\{X(\boldsymbol{s}), \boldsymbol{s} \in D\}. \tag{10.1}$$

Here, $X(\boldsymbol{s})$ is the value of the variable of interest at the spatial location $\boldsymbol{s}$. To analyze spatial data, i.e., data on the spatial process $\left\{X(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathcal{R}^d\right\}$, we need to model $X(\boldsymbol{s})$. To begin with, we assume that $E[X(\boldsymbol{s})]$ and $\mathrm{Var}[X(\boldsymbol{s})]$ exist for every $\boldsymbol{s} \in D$. Let us write

$$E[X(\boldsymbol{s})] = \mu(\boldsymbol{s}),$$
$$\mathrm{Var}[X(\boldsymbol{s})] = C(\boldsymbol{s}, \boldsymbol{s}),$$

and

$$\mathrm{cov}[X(\boldsymbol{s}_1), X(\boldsymbol{s}_2)] = C(\boldsymbol{s}_1, \boldsymbol{s}_2).$$

The mean function $\mu(\cdot)$ is called the trend of the $X(\boldsymbol{s})$ and $C(\cdot, \cdot)$ is called its covariance function. Note that the trend and the covariance function of a spatial process define its first two moments. One standard way of modeling $X(\boldsymbol{s})$ is through its first two moments, especially in some form of stationarity.

The process $\{X(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is said to be strictly stationary if for any positive integer n, any n locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, any n (Borel) sets $B_1, \ldots, B_n \subset D$, and any vector $\boldsymbol{h} \in \mathcal{R}^d$, we have

$$P[X(\boldsymbol{s}_1) \in B_1, \ldots, \boldsymbol{s}_n \in B_n] = P[X(\boldsymbol{s}_1 + \boldsymbol{h}) \in B_1, \ldots, X(\boldsymbol{s}_1 + \boldsymbol{h}) \in B_n]. \tag{10.2}$$

Here, the process is stationary in the sense that every finite dimensional distribution of the process is not changed if all the points are shifted in the same way, that is, in the same direction and by the same distance. Note that, for a strictly stationary process, we have $E[X(\boldsymbol{s})] = E[X(\boldsymbol{s} + \boldsymbol{h})]$ for any $\boldsymbol{h} \in \mathcal{R}^d$. That is, $E[X(\boldsymbol{s})]$ is a constant, and we can write

$$E[X(\boldsymbol{s})] = \mu. \tag{10.3}$$

For any two locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ and any vector $\boldsymbol{h}$, we obtain

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\boldsymbol{s}_2 + \boldsymbol{h}, \boldsymbol{s}_2 + \boldsymbol{h}).$$

In particular, taking $\boldsymbol{h} = -\boldsymbol{s}_2$, we have

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\boldsymbol{s}_1 - \boldsymbol{s}_2, \mathbf{o}),$$

so that the covariance function can be considered a function of the difference $\boldsymbol{s}_1 - \boldsymbol{s}_2$ alone. That is, writing $\boldsymbol{h} = \boldsymbol{s}_1 - \boldsymbol{s}_2$, we write

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\boldsymbol{s}_1 - \boldsymbol{s}_2) = C(\boldsymbol{h}). \tag{10.4}$$

Note that, for a strictly stationary process, $C(-\boldsymbol{h}) = C(\boldsymbol{h})$, $\boldsymbol{h} \in \mathcal{R}^d$.

A process is defined to be increment stationary if it satisfies (10.3) and for any positive integer n, any n locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$, any $n-1$ (Borel) sets $B_1, \ldots, B_{n-1} \subset D$, and any vector $\boldsymbol{h} \in \mathcal{R}^d$, we have

$$\begin{aligned} &P[X(\boldsymbol{s}_2) - X(\boldsymbol{s}_1) \in B_1, \ldots, X(\boldsymbol{s}_n) - X(\boldsymbol{s}_{n-1}) \in B_{n-1}] \\ &= P[X(\boldsymbol{s}_2 + \boldsymbol{h}) - X(\boldsymbol{s}_1 + \boldsymbol{h}) \in B_1, \ldots, X(\boldsymbol{s}_n + \boldsymbol{h}) - X(\boldsymbol{s}_{n-1} + \boldsymbol{h}) \in B_{n-1}]. \end{aligned} \tag{10.5}$$

It is obvious that a strictly stationary process is increment stationary, but the converse need not be true.

It is often more convenient to use the variogram or semivariogram than the covariance function of a spatial process. For a process satisfying (10.3), the semivariogram is defined as

$$\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) = \frac{1}{2} E[X(\boldsymbol{s}_1) - X(\boldsymbol{s}_2)]^2 = \frac{1}{2} \mathrm{Var}[X(\boldsymbol{s}_1) - X(\boldsymbol{s}_2)].$$

The variogram is $2\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2)$. Note that, by definition, $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$ and $\gamma(\mathbf{0}) = 0$, but it is possible that $\lim_{\boldsymbol{h} \to \mathbf{0}} \gamma(\boldsymbol{h}) \neq 0$. If $\gamma(\boldsymbol{h}) \to c_0 > 0$, as $\boldsymbol{h} \to \mathbf{0}$, then c_0 is called the nugget effect. This is because it is believed that micro-scale variation is causing a discontinuity at the origin.

For an increment stationary process, $\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) = \gamma(\boldsymbol{s}_1 + \boldsymbol{h}, \boldsymbol{s}_2 + \boldsymbol{h})$ for any $\boldsymbol{h} \in \mathcal{R}^d$. Letting $\boldsymbol{h} = -\boldsymbol{s}_2$, we have

$$\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) = \gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2, \mathbf{0})$$

and we write

$$\gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) = \gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2). \tag{10.6}$$

A process is said to be intrinsically stationary if it satisfies both (10.3) and (10.6). Every increment stationary process is intrinsic stationary, but the converse need not be true. Second-order stationary processes are intrinsically stationary. For a second-order stationary process, we have

$$\begin{aligned} \gamma(\boldsymbol{s}_1, \boldsymbol{s}_2) &= \frac{1}{2} \mathrm{Var}[X(\boldsymbol{s}_1) - X(\boldsymbol{s}_2)] \\ &= \frac{1}{2} \{\mathrm{Var}[X(\boldsymbol{s}_1)] + \mathrm{Var}[X(\boldsymbol{s}_2)] - 2\mathrm{cov}[X(\boldsymbol{s}_1), X(\boldsymbol{s}_2)]\} \\ &= \frac{1}{2} \{C(\mathbf{0}) + C(\mathbf{0}) - 2C(\boldsymbol{s}_1 - \boldsymbol{s}_2)\} \\ &= C(\mathbf{0}) - C(\boldsymbol{s}_1 - \boldsymbol{s}_2), \end{aligned}$$

which is a function of $\boldsymbol{s}_1 - \boldsymbol{s}_2$. Hence the process is also intrinsically stationary. In particular, if $\boldsymbol{h} = \boldsymbol{s}_1 - \boldsymbol{s}_2$, then

$$\gamma(\boldsymbol{h}) = C\mathbf{o} - C(\boldsymbol{h}). \tag{10.7}$$

This simple relationship between the semivariogram and the covariance function makes the covariogram more convenient than the variogram, though both contain equivalent information.

Recall that $C(\cdot)$, the covariogram, also called the spatial autocovariance function, is defined as the covariance between the X-values at two locations. If $C(\mathbf{0}) > 0$, then we define

$$\rho(\boldsymbol{h}) = \frac{C(\boldsymbol{h})}{C(\mathbf{0})}$$

and call it the correlogram, also known as the spatial autocorrelation function.

Often the second-order properties of a process are assumed to depend only on the distance between two locations and not on the direction of the vector joining them. A second-order stationary process is said to be isotropic if

$$C(\boldsymbol{s}_1 - \boldsymbol{s}_2) = C(||\boldsymbol{s}_1 - \boldsymbol{s}_2||).$$

An intrinsically stationary process is isotropic if

$$\gamma(\boldsymbol{s}_1 - \boldsymbol{s}_2) = \gamma(||\boldsymbol{s}_1 - \boldsymbol{s}_2||).$$

In general, given the semivariogram it is not possible to reproduce the covariance function of a spatial process. A special case in which the semivariogram and the covariance function are equivalent is that of a second-order stationary process with

$$\lim_{||\boldsymbol{h}|| \to \infty} C(\boldsymbol{h}) = 0.$$

In this case, by (10.7),

$$\gamma(\boldsymbol{h}) = C(\mathbf{0}) - C(\boldsymbol{h}),$$

so

$$\lim_{||\boldsymbol{h}|| \to \infty} \gamma(\boldsymbol{h}) = C(\mathbf{0})$$

and

$$C(\boldsymbol{h}) = \lim_{||\boldsymbol{u}|| \to \infty} \gamma(\boldsymbol{u}) - \gamma(\boldsymbol{h}).$$

The quantity $C(\mathbf{0})$ is called the sill of the semivariogram. The smallest value of $||\boldsymbol{r}||$ for which $2\gamma(\boldsymbol{r}(1+\epsilon)) = 2C(\mathbf{0})$, for any $\epsilon > 0$, is called the range of the variogram in the direction $\boldsymbol{r}/||\boldsymbol{r}||$.

In practice, neither the semivariogram nor the covariance function is known. They have to be estimated from the data.

10.2.1 Composite Sampling

Composite sampling becomes a useful technique for reducing the response surface prediction variance when there is considerable variability on a micro-scale (see Starks, 1986). Composite sampling in the response surface prediction context is usually performed by taking individual samples from within a small area, mixing them thoroughly, and making a single measurement upon the composite sample. Compositing is usually used for decreasing the response surface prediction variance. We also consider how compositing affects the covariogram, for it is this effect that will determine the effect of compositing upon the prediction variance (Starks, 1986).

When the individual samples are collected from a site using a regular grid, and when data are collected on composite samples, then it is obvious that the interest will initially be in the relationships between the spatial processes that represent composite sample data and individual sample values. In this section, we discuss these issues.

Sampling issues that we address in our study are (1) what scale, i.e., distance between sample locations, of sampling is necessary to adequately sample a particular site and (2) how to design compositing of individual soil samples at each sample location in order to limit the effect of variability on a very local scale to viewing the spatial pattern of a larger scale.

Specifically, we consider how the choice of k, the composite sample size, and of the individual sample spatial configuration affects the covariogram of the composite samples. The choice of k will be limited by the amount of material that can be logistically handled and that can be homogeneously mixed. However, within these restrictions, k can be chosen by considering how it affects the composite sample nugget variance. Here, it is simpler to consider a single stochastic process rather than separate identifiable small-scale and micro-scale processes.

The model to describe the spatial process that represents the composite sample values in terms of individual sample values is assumed to be of the following form:

$$Y(\boldsymbol{s}) = \sum_{i=1}^{k} X(\boldsymbol{s}_i) + \epsilon,$$

where Y is the measured value on a subsample of a composite of k individual samples collected at locations $\boldsymbol{s}_i$, $i = 1, \ldots, k$, the true concentration of the ith individual sample is $X(\boldsymbol{s}_i)$, and ϵ denotes the measurement and subsampling error. Here $\boldsymbol{s}$ is usually taken as the mean of the k locations $\boldsymbol{s}_i$, $i = 1, \ldots, k$:

$$\boldsymbol{s} = \sum_{i=1}^{k} \boldsymbol{s}_i .$$

The values of ϵ for different composite samples are treated as independent and identically distributed with mean zero and variance σ_ϵ^2. Subsampling and measurement error include such factors as inaccurate recording of the exact sampling locations or of the exact amounts of soil volumes extracted, inability to thoroughly mix the composite, or imprecision in laboratory procedures.

The individual sample values $X(\boldsymbol{s}_i)$ will be regarded as observations on a spatial process $\{X(\boldsymbol{s}), \boldsymbol{s} \in D\}$ which will be represented as

$$X(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \xi(\boldsymbol{s}) + \delta(\boldsymbol{s}),$$

where $\mu(\boldsymbol{s})$ is a large-scale deterministic trend, $\xi(\boldsymbol{s})$ is a small-scale stochastic process, and $\delta(\boldsymbol{s})$ is a micro-scale stochastic process.

The large-scale trend is regarded as varying over the entire study region. The small-scale process occurs on a spatial scale considerably smaller than the entire study region, but larger than the minimum inter-sample distance (or the minimum distance for which there are sufficient data), while the micro-scale process varies within the minimum inter-sample distance. Thus, within-composite variation is due to δ, while μ and ξ affect the between-composite variability.

The model assumes stationarity in Y, at least within small data neighborhoods. Three levels of stationarity can be distinguished for the composite sample process, too. Strong stationarity requires that the joint distribution of $Y(\boldsymbol{s}_1)$, $Y(\boldsymbol{s}_2), \ldots, Y(\boldsymbol{s}_n)$ should be equivalent to the distribution of $Y(\boldsymbol{s}_1 + \boldsymbol{h})$, $Y(\boldsymbol{s}_2 + \boldsymbol{h}), \ldots, Y(\boldsymbol{s}_n + \boldsymbol{h})$ for each vector $\boldsymbol{h}$.

Second-order stationarity requires that the means, $E[Y(\boldsymbol{s})]$, and variances, $\mathrm{Var}[Y(\boldsymbol{s})]$, exist, are constant, and do not depend upon $\boldsymbol{s}$ and also that the covariance between $Y(\boldsymbol{s}_1)$ and $Y(\boldsymbol{s}_2)$ exists and depends only on the vector $\boldsymbol{h} = \boldsymbol{s}_2 - \boldsymbol{s}_1$ joining $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. Under second-order stationarity, the correlogram,

$$\rho(\boldsymbol{h}) = \mathrm{corr}[Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_1 + \boldsymbol{h})],$$

is related to the semivariogram, $\gamma(\boldsymbol{h})$, by

$$\begin{aligned} 2\gamma(\boldsymbol{h}) &= E[Y(\boldsymbol{s}) - Y(\boldsymbol{s} + \boldsymbol{h})]^2 \\ &= 2\sigma^2[1 - \rho(\boldsymbol{h})]. \end{aligned}$$

The process is (second-order) isotropic when $\rho(\boldsymbol{h})$ is a unique function of the distance $\boldsymbol{h} = |\boldsymbol{s}_2 - \boldsymbol{s}_1|$, in which case we write $\rho(\boldsymbol{h})$ for the correlogram.

A third form of stationarity, weak stationarity, requires that the means exist and do not depend on the location $\boldsymbol{s}$ and that the semivariogram be a unique function of the vector $\boldsymbol{h}$. Strong stationarity, together with the existence of the first two moments, implies second-order stationarity, which in turn implies weak stationarity.

Usually, the correlogram becomes small for large $||\boldsymbol{h}||$ and the semivariogram levels off to the sill. A nugget effect, detected by a positive intercept in the sample semivariogram, reflects the presence of micro-scale variation and/or measurement error (Cressie, 1988).

Decomposing spatial variability into a large-scale trend and a small-scale stochastic process can be problematical. This may be partly due to differences in observer judgment (Cressie, 1988) and partly a matter of indeterminacy in both trend model and the covariance structure of the stochastic process if these are unknown beforehand (Armstrong, 1984). A common approach to approximating the trend is the fitting of polynomials in local data neighborhoods, but the degree to which local polynomials fit trend or small-scale stochastic process depends, in part, upon the data neighborhood size. Our approach was to fit the trend over the entire region in one operation, using the bicubic spline method of Dierckx (1981). This spline function is a smoothing, not an interpolative, function and an appeal of this method is that small-scale variation might not be fit if there is a high degree of smoothing. Subsequent to the data decomposition, a cross-validation study was performed in order to examine the behavior of response surface predictions if the actual sampling had been less intense.

The composite sample nugget variance, $\sigma^2_{N_c}$, is

$$\sigma^2_{N_c} = \sigma^2_{\text{sill}_c} - \sigma^2_{\tau_c},$$

where $\sigma^2_{\text{sill}_c}$ is the sill or limiting value of $\gamma(\boldsymbol{h})$ as $||\boldsymbol{h}|| \to \infty$ for the process of composite sample values and $\sigma^2_{\tau_c}$ is the limiting value of the covariance between two composite samples as $||\boldsymbol{h}|| \to \mathbf{o}$.

The composite sample sill, or the variance of the mean of the individual samples, in the absence of measurement error, is

$$\sigma^2_{\text{sill}_c} = \text{Var}\left(\sum_{i=1}^{k} X_i/k\right) = \frac{1}{k^2}\sum_{i=1}^{k}\text{Var}(X_i) + \frac{2}{k^2}\sum_{\substack{i\,,\,i'=1\\ i<i'}}^{k}\text{cov}(X_i, X_{i'})$$

$$= \frac{\sigma^2_\tau + \sigma^2_N}{k} + \frac{2}{k^2}\sum_{\substack{i\,,\,i'=1\\ i<i'}}^{k}\text{cov}(X_i, X_{i'}).$$

The covariance between two composite samples, $Y = (X_1 + X_2 + \cdots + X_k)/k$ and $Y^* = (X^*_1 + X^*_2 + \cdots + X^*_k)/k$, is

$$\text{cov}(Y, Y^*) = \text{cov}\left(\frac{1}{k}\sum_{i=1}^{k} X_i, \frac{1}{k}\sum_{i'=1}^{k} X^*_{i'}\right)$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\mathrm{cov}(X_i, X_i^*) + \frac{2}{k^2}\sum_{\substack{i\,,\,i'=1\\ i<i'}}^{k}\mathrm{cov}(X_i, X_{i'}^*)\,,$$

which converges to

$$\sigma_{\tau_c}^2 = \frac{\sigma_\tau^2}{k} + \frac{2}{k^2}\sum_{\substack{i\,,\,i'=1\\ i<i'}}^{k}\mathrm{cov}(X_i, X_{i'}^*)$$

as the two compositing configurations become coincident. It is then obvious that

$$\sigma_{N_c}^2 = \sigma_{\mathrm{sill}_c}^2 - \sigma_{\tau_c}^2 = \frac{\sigma_N^2}{k}.$$

Thus, the nugget variance for a composite sample decreases inversely with increasing composite sample size k, but does not depend upon the spatial configuration of the individual samples.

The spatial configuration of the individual samples does influence the composite sill variance. In fact, this variance decreases with increasing distance among the individual samples. As the inter-sample distance approaches zero, $\mathrm{cov}(X_i, X_{i'}^*)$ approaches σ_τ^2, and

$$\sigma_{\mathrm{sill_c}}^2 \simeq \frac{\sigma_\tau^2 + \sigma_N^2}{k} + \frac{k-1}{k}\sigma_\tau^2 = \sigma_\tau^2 + \frac{\sigma_N^2}{k}.$$

On the other hand, as the distance between individual samples approaches the range of spatial autocorrelation (i.e., the distance at which the covariogram effectively vanishes),

$$\sigma_{\mathrm{sill_c}}^2 = \frac{\sigma_\tau^2 + \sigma_N^2}{k} = \frac{\sigma_{\mathrm{sill}}^2}{k}.$$

From these results and from data collected at any particular site, we attempt to infer how the composite sampling plans can be effective in reducing the nugget variance and the sill.

10.3 Application to Two Superfund Sites

10.3.1 The Two Sites

Data from the Dallas Lead Site and the Palmerton Site were chosen for this exploratory analysis because of common features of these sites and prior

investigations: the similarity in the contamination processes, to the sampling schemes, and to the extensiveness of prior statistical analyses. The processes sampled at both sites were of heavy metal accumulation from the fallout of air-borne particles emitted from point sources. The sampling schemes were designed to provide information about the data autocorrelation structure so as to enable response surface prediction via kriging, composite sampling was utilized at both sites, and special samples were taken to estimate variability due to certain sources.

10.3.2 Methods

10.3.2.1 The Data Analysis

The data were first examined for inconsistencies. Then, some features of the data were examined through frequency distributions, contour plots, identification of outliers, plots of variance vs. mean, sample semivariograms, and variance component estimates from individual, duplicate, and split sample data. Decomposition of the data into large-scale trend vs. small-scale stochastic process was explored through fitting trend models and examining residual semivariograms.

The sample semivariogram was calculated by

$$g = \frac{1}{2n(\boldsymbol{h})} \sum_{i=1}^{n(\boldsymbol{h})} [Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i + \boldsymbol{h})]^2,$$

where $Y(\boldsymbol{s}_i)$ is an observation on a composite sample with center location, $\boldsymbol{s}_i$, and $Y(\boldsymbol{s}_i + \boldsymbol{h})$ is an observation on a composite sample with center location $\boldsymbol{h}$ away, and $n(\boldsymbol{h})$ is the number of observations $\boldsymbol{h}$ units apart. Intervals of $\boldsymbol{h}$ were determined so that $n(\boldsymbol{h})$ was at least 30, a minimum suggested by Journel and Huijbregts (1978).

10.3.2.2 A Cross-Validation Study

A cross-validation analysis was performed in order to examine how the predictive power of data changed with different sampling intensities. This was performed by resampling the data at intensities lower than the realized intensity and assessing how well data omitted one at a time could be predicted by the resampled data. Observations were randomly assigned to one of two groups: a subset regarded as resampled points to be used in calculating kriging predictions and a subset omitted in the calculation of kriging predictions. The selection of data was performed by first stratifying the data to insure that there was some degree of systematic coverage of the sampled site and then by permuting the observations within strata and choosing permuted points in order until the desired density was realized. All observations, from both the resampled subset and its complement, were predicted from the resampled subset, however. In each of 100 runs per selected sampling intensity, each datum in the non-resampled subset was predicted from the resampled data and each

resampled datum was predicted from the remaining resampled points. The measure of prediction performance was the mean squared error (MSE) of cross-validation,

$$\text{MSE of cross-validation} = \sum_{i=1}^{n} \frac{(y_i - y_i^*)^2}{n},$$

where y_i is datum $i, i = 1, \ldots, n$, and y_i^* is the predicted value of y_i. This statistic was also calculated for the case of all observations being resampled. A measure of inter-sample distance of the resampled points was calculated by the median minimum distance of points to neighbors.

10.3.2.3 The Behavior of the Composite Sample Variogram

The behavior of the composite sample nugget and the composite sample sill was graphically examined for different choices of composite sample design factors and underlying covariance structure of the data. The behavior of the composite sample nugget was examined for different composite sample sizes, while the behavior of the composite sample sill was examined for different choices of inter-sample distance and rectangular configurations of individual samples.

10.3.3 Results

10.3.3.1 Decomposition of Data Variability

Cubic spline models were fitted to the Dallas Lead Site and the Palmerton Site data in order to decompose the spatial variability into regional trend and a stochastic process. Sills were considerably lowered by spline models with relatively few knots as compared to semivariogram sills when no trend was removed. Omnidirectional semivariograms were calculated for the Dallas Lead Site data as there had been no indication of anisotropy in these data. The Palmerton Site semivariogram was collapsed over sampling phase, since the difference in the nugget from increasing the composite sample size from 4 to 9 appeared to be small relative to noise in the semivariogram. This judgment was made from viewing the semivariograms with no trend removed and from the pooled duplicate sample variances.

The proportions of variability in the log(ppm) data attributable to different sources were calculated from the spline model results and the pooled sample variances. The proportions of variability due to combined subsubsampling and measurement error, to micro-scale variation, and to the combined effect of large-scale trend and local discontinuities associated with outliers were calculated as

$$\frac{\hat{\sigma}_a^2}{\hat{\sigma}_z^2}, \quad \frac{\hat{\sigma}_n^2 - \hat{\sigma}_a^2}{\hat{\sigma}_z^2}, \quad \text{and} \quad \frac{\hat{\sigma}_z^2 - \hat{\sigma}_n^2}{\hat{\sigma}_z^2},$$

respectively, where $\hat{\sigma}_a^2$, $\hat{\sigma}_n^2$, and $\hat{\sigma}_z^2$ are the estimated measurement error, nugget, and sample variances of the log(ppm) data, respectively. The estimated proportion of variability due to subsampling and measurement error was small, being less than 1%. The proportion of variability due to micro-scale variation was larger, in the range of 15–25%, while the proportion of variability due to both the large-scale trend and outliers was the largest, in the range of 75–84% (see Table 10.1).

Table 10.1 The estimated components and proportions of variability attributable to different sources in the Dallas Lead Site and the Palmerton Site data

			Variance component estimate	
Site	Metal	Total[a]	Nugget[b]	Subsampling and measurement error
Dallas Lead – DMC area	Pb	1.315	0.313	0.00528
Dallas Lead – RSR area	Pb	1.277	0.314	0.00528
Palmerton	Cd	1.238	0.199	0.00275
Palmerton	Pb	0.803	0.201	0.00453
Palmerton	Zn	1.266	0.206	0.00380
			Percent of variability estimate	
Site	Metal	Trend[c]	Micro-scale	Subsampling and measurement error
Dallas Lead – DMC area	Pb	76.2	23.4	0.4
Dallas Lead – RSR area	Pb	75.4	24.2	0.4
Palmerton	Cd	83.9	15.9	0.2
Palmerton	Pb	75.0	24.5	0.6
Palmerton	Zn	83.7	16.0	0.3

[a] Sample variance of log(ppm)
[b] Estimated by semivariogram of spline model residuals collapsed over distance
[c] Includes local discontinuities of outliers

10.3.3.2 Cross-Validation Study Results

The purpose of this study was to assess the spatial scale necessary to capture the large-scale trend. Therefore, the identified outliers were omitted from the cross-validation analysis, since they probably could not be predicted well from nearby data.

Omnidirectional spherical semivariograms were fit to the Dallas Lead Site data and to the Palmerton Site data and were used in calculating the kriging predictions. A linear anisotropic semivariogram provided a good fit to the Palmerton Site semivariogram data, but when the data were thinned below approximately half of the full data set, the kriging predictions with this model were unstable. When more than half of the data were resampled, the predictions based upon the spherical semivariogram and upon the linear anisotropic semivariogram were very similar. Neighborhoods

of points with highest correlation were used in calculating kriging predictions, with neighborhood sizes chosen to be 20 for the Dallas Lead Site data and 15 for the Palmerton Site data. A Lagrange multiplier for a constant mean was used in the calculation of kriging coefficients as it provided a lower MSE of cross-validation than kriging predictions calculated without Lagrange multipliers.

The MSE of cross-validation decreased in a nonlinear pattern with increasing percentages of data in the resampled subset. There was relatively little loss in prediction accuracy with up to about 60% omitted from the data subset. For example, when the resampled subset constituted 42–45% of the non-outlier data set, the increase in the MSE of cross-validation was just 10–17% over the MSE of cross-validation when all non-outliers were resampled. The median minimum inter-sample distance among resampled points exhibited a similar pattern with increasing resampling percentages (Fig. 10.1). Standard errors of the mean for the 100 runs were 0.0005–0.007 and 0.0002–0.006 for the MSE of cross-validation and for the median minimum inter-sample distance, respectively.

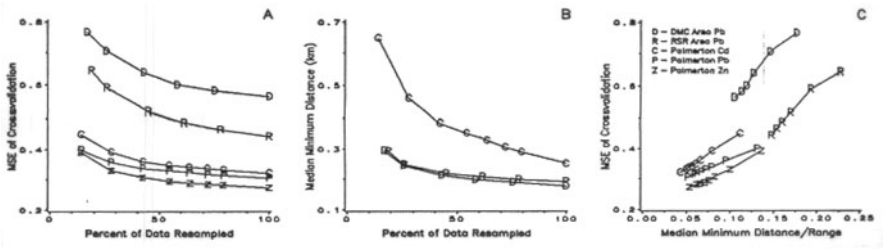

Fig. 10.1 The cross-validation results: (**a**) MSE vs. the percent of data resampled, (**b**) median minimum inter-sample distance vs. the percent of data resampled (Palmerton Site Cd, Pb, Zn data were nearly identical so only the Cd data are presented), and (**c**) MSE vs. the median minimum inter-sample distance scaled by the range of autocorrelation estimated by omnidirectional spherical semivariogram models. Values are the means of 100 runs except for the case of 100% of the data being resampled

Standardizing inter-sample distances by estimated ranges of large-scale autocorrelation presented a different view of the scale of sampling from that anticipated in the sampling design. If the range of autocorrelation was estimated from omnidirectional spherical semivariogram models fitted to the data with only outliers omitted, then the median minimum inter-sample distances were 11–15 and 4–5% of the estimated ranges of autocorrelation, respectively (Fig. 10.1). Calculating the inter-sample distance on the scale of the large-scale process thus gave a sampling scale much less than the 1/3–2/3 of the range of autocorrelation thought to have been required for sampling a small-scale stochastic process.

The capability of estimating semivariograms was not lost when much of the data was omitted, though there may have been too few points to reliably estimate a semivariogram when only 17–19% of the Dallas Lead Site non-outlier data were resampled. The sample semivariogram patterns when data were omitted in

resampling were similar to the semivariogram patterns of all non-outlier data (Bolgiano et al., 1990).

10.3.3.3 The Effect of Compositing upon the Sill

The composite sample sill decreases with increasing distance, h, between individual sample locations and approaches $\frac{1}{k}$th of the individual sample sill as h approaches the range of autocorrelation. Therefore, the decrease in the sill with increasing h becomes steeper as the composite sample size increases. For spherical and Gaussian covariance models (Journel and Huijbregts, 1978), the effect upon the sill of h depended upon the relative size of the nugget variance and the composite sample size. The choice of h has a larger effect upon the sill when the nugget variance is relatively small as compared to when it was relatively large (see Fig. 10.2).

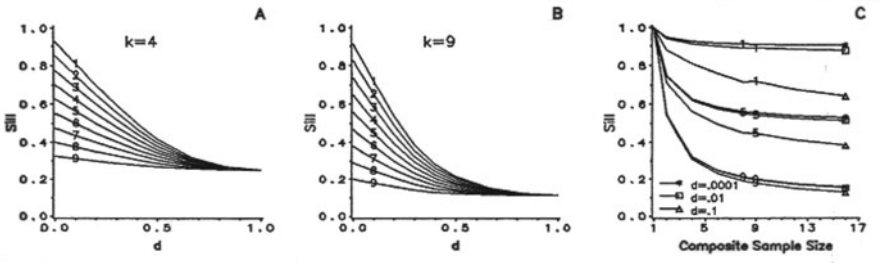

Fig. 10.2 The nugget variance vs. h, the inter-sample distance for $k = 4$ (**a**) and for $k = 9$ (**b**) and vs. the composite sample size (k) for rectangular configurations of individual samples. A spherical covariogram model has been assumed, with the nugget comprising $\frac{a}{10}$th of a standardized sill of 1.0, where a is the integer plotted on curves; h has been standardized relative to a range parameter of 1.0 (i.e., the covariance between the values at two points that are $h = 1$ units apart is 0). Results for a Gaussian covariance model are also similar

10.3.4 Discussion

This retrospective analysis of data collected at the Dallas Lead and the Palmerton Superfund Sites examined the nature of variability in the data and sampling considerations for sites possessing the variability scales hypothesized for these data. The size of variability sources and the spatial scale on which they occur are important factors in designing an efficient sampling scheme and accurately predicting a contaminant concentration response surface. Being able to allocate resources in light of anticipated sizes of variability sources can contribute to cost-effective sample design (Provost, 1984).

Traditionally, the sampling approach to geostatistics has implicitly assumed that there exists a stochastic process nested within the regional trend that can be captured by a systematic design (Yfantis et al., 1987; Flatman et al., 1988). However, the

Dallas Lead Site data and the Palmerton Site data did not appear to exhibit small-scale variability. Instead, the variability appeared to consist of a large-scale trend with discontinuities caused by either local contamination processes or local soil disturbance, by variability occurring on a micro-scale, and by a very small measurement error. The cross-validation results implied that sampling of these large-scale processes might have been suitably achieved with a larger scale of sampling than the scale that was utilized.

One purpose of sampling a small-scale stochastic process is to achieve stationarity in the data or at least to be able to assume approximate stationarity in local data neighborhoods. The importance of the stationarity assumption may depend upon the use for which the response surface prediction is intended and the importance of an accurate variance estimate. The kriging predictions may be less sensitive to stationarity violations than are kriging variance estimates, as the kriging weights are likely to be similar using different, but reasonable, semivariogram models, while the kriging variance estimate depends upon the assumed semivariogram. However, for these data, it appears that calculation of kriging coefficients by other than a pure nugget correlation model requires that the data correlation structure be modeled from the large-scale trend. Cressie (1989) also appears to have utilized the large-scale autocorrelation structure to perform kriging.

If the data correlation structure is to be estimated from data sampled on the large scale, then the guidelines for sampling on the small scale may not apply. The grid locations at the Dallas Lead Site and at the Palmerton Site were arranged to be 2/3 and 1/3, respectively, of the anticipated range of autocorrelation, which are scales typically suggested for geostatistical studies (Flatman et al., 1988). However, if a systematic sampling scheme is designed to capture the regional trend, then the distance between sample locations might be much smaller than 1/3–2/3 of the large-scale autocorrelation range.

If local contamination processes are important components of the entire contamination process, as was evident at the Dallas Lead Site, then sampling can be designed to detect hotspots (Gilbert, 1987), as well as to capture the large-scale trend. At the Dallas Lead Site, the local hotspots were largely associated with industrial sites, and sampling of such locations might be planned rather than being randomly encountered. Prediction of the extent of hotspot contamination around the local industries would likely be overestimated if the autocorrelation structure of the large-scale process were employed for interpolation (Gensheimer et al., 1986). Perhaps the best solution to predicting the extent of local hotspot contamination is realized by further sampling near those locations.

Spatial variability has traditionally been modeled in geostatistics as a stochastic process that is not location dependent. However, as a process is studied in greater detail, underlying causal mechanisms may become apparent, and so the designation of deterministic or stochastic is often applied relative to our scale of reference (Wilding and Drees, 1983). Much of the pattern of heavy metal contamination at the examined sites seems to be related to spatial features, such as the alignment of the Palmerton Site contamination contours with the ridge and valley topography and the lead level decrease in the Trinity River flood plain at the Dallas Lead Site

DMC area. Further, at both sites, questions have been raised about the heavy metal contamination being attributable to both smelters and motor vehicles (Carra, 1984; Starks et al., 1987). The empirical approach of geostatistics ignores information such as topography, airflow, erosion, and transportation networks that might be used to understand causal mechanisms. As the demand for a higher degree of model explanation vs. model empiricism (Lehmann, 1990) is required for characterization of hazardous waste sites, spatial data analysis and response surface prediction might shift focus to the modeling of causal factors.

The Palmerton Site data provided an opportunity to examine the efficacy of composite sampling upon reducing the variogram parameters of the nugget variance and the sill. Variability in the data semivariograms and the lack of sufficient data at small distances obscured any possible differences in nugget variances between the two sampling phases. The nugget variance of the Palmerton Site composite sample data was not insignificant, estimated at about 16–25% of the total variability in the logged data. Individual and duplicate sample data indicated that compositing four individual samples may have decreased the nugget variance relative to not compositing, while compositing nine individual samples did not lower duplicate sample variances appreciably. Conclusions based upon these data, however, are inclusive, since there appeared to be considerable data variability in small samples and there may have been a possible confounding effect of compositing sample size with inter-sample distance. Compositing of four and nine individual samples was achieved by configuring the individual sample position in square grids with inter-sample distance, d, of 4.24 and 2.125 m, respectively. However, these observations are consistent with theoretical results. The inverse relationship between the nugget variance and the composite sample size indicates that compositing has the greatest effect in reducing the nugget variance when the individual sample nugget variance is large, and there are decreasing returns to lowering the nugget variance as the composite sample size increases. Increasing the composite sample size from 1 to 4 decreases the nugget variance by 400%, but increasing the composite sample size from 4 to 9 only decreases the nugget variance by 14%.

The difference in sills between phase 1 and phase 2 Palmerton Site data reflected, in part, a difference in the spatial location of samples. The first phase samples tended to be collected from locations with high contamination, while second phase samples tended to be collected from locations with lower contamination. Since the effect of compositing upon the sill cannot be discerned from the data, the numerical results might serve to guide judgment upon the effectiveness of compositing in reducing the sill. The effectiveness of compositing in reducing the sill depends upon the configuration and nugget variance of individual samples. If the nugget variance comprises 50% of total stochastic variability, d is very small, then for a spherical variogram, increasing the composite sample size from 1 to 4 would decrease the sill 37.5%, while increasing the composite sample size from 4 to 9 would decrease the sill 11%. This decrease would be lower or higher if the nugget comprised less or more of the total stochastic variability, respectively. As with the effect of compositing upon the nugget variance, there is a decreasing return upon lowering the sill by increasing the composite sample size. Increasing d might be a means to decreasing the sill unless

the nugget is large, though there is an upper limit on d for the composite sample to be considered representative of a point in space.

Additional considerations may affect the selection of a composite sample design. There is no statistical cost to increasing the composite sample size when d is very small relative to the distance between composite samples, unless mixability becomes more difficult. Therefore, the benefit of increasing precision might be weighed against the cost of increasing the composite sample size. Starks (1986) also minimized the mean squared error in predicting the spatial average of the sample support by the composite sample value in selecting a composite sample design. However, we found that this criterion did not appear to be useful for this purpose unless there was an identifiable stochastic process on a very small scale, because the value of this criterion was nearly constant, assuming a Gaussian covariance model, except for d near the range of autocorrelation.

Cost-effective sampling is likely to be achieved when the sampling design reflects knowledge about the sizes of variability components in the process of interest and the spatial scales on which they occur. If the regional trend and variation among nearby individual samples contribute significantly to spatial variability, then the choices of a sampling scale in measuring the important features of the large-scale trend and compositing of individual samples might be important to achieving cost-effectiveness in hazardous waste site characterization.

10.4 Compositing by Spatial Contiguity

10.4.1 Introduction

Sampling to determine the extent of pollution traditionally involves taking measurements at every sampling location, often on a grid. Composite sampling is an alternative approach that forms composite samples from a number of individual samples, tests the composite sample, and retests aliquots of the individual samples when the test on the composite sample indicates that one or more of its constituent samples may be polluted. Used in this way, composite sampling is most efficient when the overall contaminant levels are relatively low or when the contamination is spatially clumped, for otherwise excessive retesting of the constituent samples will be necessary.

The Center for Statistical Ecology and Environmental Statistics carried out a simulation study (i) to compare the cost (number of measurements) of various retesting schemes and (ii) to study how spatial pattern in the data affects the overall performance of composite sampling in the hotspot identification (action level) case.

In the presence/absence case, measurements indicate the presence or absence of contamination, but the contaminant levels are not available or are not important. The method is to form and test composite samples and then to retest aliquots of the individual samples comprising any composite that tests positive for pollution. A variety of strategies have been proposed for carrying out the retesting. These

include the classical Dorfman (1943) scheme, the Sterrett (1957) scheme, the Gill and Gottlieb (1974) scheme, and a scheme based on entropy (Hwang, 1984).

A hotspot can be defined to consist of contiguous locations at which the contaminant concentration exceeds a certain level c. This value may be an action level that would require some remedial action. Instead of analyzing every sample at every location, composite sampling combined with a suitable retesting strategy can be used to determine the particular locations where pollution exceeds the level c. The method is to form and analyze composite samples and then to reanalyze aliquots of the individual samples of any composite that returns a value greater than c/k, where k is the number of samples in the composite. If a composite sample returns a measurement smaller than c/k then (barring measurement error) one is assured that every constituent individual sample is below the action level and no retesting is needed. In the contrary case, some retesting is required because one or more of the component samples might, though not necessarily, exceed the action level. When required, the retesting can be carried out according to various strategies as described below.

It is not possible to form composite samples using existing data. Instead, a conceptual composite sample can be formed and a value calculated by averaging. Thus simulations can be carried out on realistic data. The algorithms necessary to carry out these simulations are of two types, composite sample forming and retesting. The computer programs to implement these algorithms are given in Appendix A of Bolgiano et al. (1989).

10.4.2 Retesting Strategies

The four retesting strategies mentioned above are designed for use in the presence/absence case. Briefly, these strategies are as follows:

1. Exhaustive retesting: This procedure exhaustively tests every individual sample from composites that test positive for pollution.
2. Sequential retesting: This procedure sequentially tests individual samples from a positive testing composite. This stage is continued until a sample tests positive for pollution, when a composite sample is formed from the remaining individual samples. The process is repeated as often as necessary.
3. Binary split retesting: This procedure divides the individual samples from a positive testing composite sample into two groups, as nearly equal in size as possible. A composite sample is formed from each group and is tested for pollution. The process is repeated as often as necessary.
4. Entropy-based retesting: This procedure starts with a pool of unclassified samples from which composite samples are sequentially formed and tested. When one of these "original" composite samples tests positive, then a "secondary" composite sample is formed using one half (or as nearly as possible) of the individual samples from the "original" sample. If the "secondary" composite sample tests positive, then the remaining individual samples from the "original"

composite sample are returned to the pool of unclassified samples. On the other hand, if the "secondary" composite sample tests negative, then a composite sample formed from the remaining individual samples would have to test positive. These individual samples are treated as belonging to an original (but smaller) composite sample that tested positive, and the process continues.

The above procedures are modified for use in case of an action level. The modification of the exhaustive retesting procedure is straightforward. Samples are tested sequentially until the sum exceeds c (or, equivalently, until the average exceeds c/k), then the remaining individual samples are composited.

The sequential retesting procedure was similarly modified so that the individual samples are tested until the sum of the concentrations exceeds c. The remaining samples are then composited. This scheme has been further modified. The total amount of pollution in the un-retested individual samples can be calculated from the retested samples, and the unretested samples need not be composited. The modified procedure tests individual samples sequentially until the remaining pollution is less than c. This modified procedure can also be considered a modification of the exhaustive retesting procedure.

The modification of the binary split retesting procedure for the action level case is straightforward. A composite sample that exceeds the action level c/k for the composite is split as nearly as possible into two composite samples which are tested, and the procedure continues.

The entropy-based retesting procedure proceeds as described in strategy 4 except that the composite sample must exceed c/k in order for them to be considered testing positive for pollution.

10.4.3 Composite Sample-Forming Schemes

In order to assess the effect of spatial patterns on the relative costs of the retesting strategies, various methods of forming the composite samples have been examined by Bolgiano et al. (1989). These are chosen to determine if information on the spatial structure can be used to reduce the amount of retesting required in a composite sampling program.

1. Random order: Selecting and compositing observations at random ignores all spatial structures. Different runs of this algorithm give different composite samples and different results.
2. Natural order: The data were perhaps collected or numbered in some systematic manner that reflect its spatial structure. Composite samples are formed using the data in the order it is received.
3. Circular sectors I: Composite samples (CSs) are formed based on the distance from the center of the data. First, a value k is fixed for the number of individual samples in each composite. If k does not evenly divide the total number of available samples, then the remainder ("left-over") samples are grouped into a

single circular CS located at the center of the data. Next, a circular ring of CSs is formed using just enough CSs so that the remaining number of CSs can be divided by 4. Finally, circular rings are formed with four CSs each.

4. Circular sectors II: CSs are formed in sectors as above, except that the "left-over" samples are along the boundary of the region. First, a single circular CS is formed. Then, rings of four CSs are formed. The "left-over" CSs are in the next ring. Finally, the "left-over" samples are in the last ring.
5. Circular sectors III: CSs are formed in sectors as above. The "left-over" samples are in the central circle. One CS is in the next ring. Then rings of four composite samples are formed and finally the "left-over" CSs are in the last ring.
6. Vertical strips: A square grid is superimposed upon the study region so that on average 10 cells would be necessary to form a single CS. This partitions the space into vertical strips, and CSs are formed by proceeding up the first strip, down the second strip, and so forth until all CSs are formed. The "left-over" samples, if any, comprise the last CS.
7. Horizontal strips: This is the same as Algorithm 6, but with the roles of horizontal and vertical interchanged.

Each of the above composite sampling schemes has been combined with the retesting procedures and the routines are run on four data sets using several composite sample sizes. The results also depend on the action level. Tables of output are given in Appendix B of Bolgiano et al. (1989). After inspecting the output, the graphs given in Section 6 of Bolgiano et al. (1989) were produced, using a simulation based on the Dallas Lead study (see Isaaks, 1984; Flatman, 1984). Different action levels can be thought of as different pollution levels. Low action levels correspond to high pollution levels and high action levels to low pollution. That is, if a data set had the same spatial pattern with twice the pollution level, the result would be the same as using the current data with the action level divided by 2. (See Bolgiano et al. (1989) for more details.) These algorithms are not thought to be optimal in any sense, but they show that spatial patterns in the data can have a significant effect on the results. The most important observation made in this study is that it is more efficient to form composite samples using individual samples collected from spatially contiguous locations. In general, it is more efficient to form composite samples from relatively homogeneous individual samples. If spatial patterns cannot provide sufficient information to identify similar individual samples, other methods have to be devised. Use of ranked set sampling for this purpose is an available choice and is described in the following section.

10.5 Compositing of Ranked Set Samples

10.5.1 Ranked Set Sampling

Ranked set sampling (RSS) involves the drawing of m random samples with m units in each sample from a population. Then, the m units of each sample are ranked by a

visual inspection or any other rough and inexpensive method not requiring the exact measurements of the variable of interest. The unit with the smallest rank is quantified from the first sample, the unit having the second smallest rank is quantified from the second sample, and this process of quantification continues until the unit with the largest rank is quantified from the mth sample. This procedure involves the quantification of m units and, as such, it yields m measurements, one from each set of the ordered sample. The whole procedure is repeated r times. It means, in other words, that in each of the r cycles, m^2 units are randomly selected from a population and r measurements are obtained corresponding to each rank. This method of selection, thus, provides mr quantified values in total, though m^2r units are randomly selected from the population. These mr quantified values constitute a ranked set sample. Takahasi and Wakimoto (1968) and Dell (1969) provide a mathematical formulation for this sampling method introduced earlier by McIntyre (1952).

Let $X_{11}, \ldots, X_{1m};\ X_{21}, \ldots, X_{2m};\ \ldots;\ X_{m1}, \ldots, X_{mm}$ be independent random variables all having the same cumulative distribution function (cdf) $F(x)$. Further, let $X_{i(1)}, \ldots, X_{i(m)}$ denote the order statistics of $X_{i1}, \ldots, X_{im}$ $(i = 1, \ldots, m)$, respectively. Let $X_{1(1)}, \ldots, X_{i(i)}, \ldots, X_{m(m)}$ denote the ranked set sample, where $X_{i(i)}$ denotes the ith order statistic (as no error in ranking is supposed here, there is no difference between the judgment ordered and the actual ordered sample) in the ith sample. The randomly drawn samples are shown below:

$$\begin{array}{cccccc}
\text{Set} & & & & \\
1 & X_{11} & X_{12} & \cdots & X_{1m} \\
2 & X_{21} & X_{22} & \cdots & X_{2m} \\
\vdots & & & \vdots & \\
m & X_{m1} & X_{m2} & \cdots & X_{mm}
\end{array}$$

After the units in each row are ranked, they appear in the following arrangement:

$$\begin{array}{cccccc}
\text{Set} & & & & \\
1 & X_{1(1)} & X_{1(2)} & \cdots & X_{1(m)} \\
2 & X_{2(1)} & X_{2(2)} & \cdots & X_{2(m)} \\
\vdots & & & \vdots & \\
m & X_{m(1)} & X_{m(2)} & \cdots & X_{m(m)}
\end{array}$$

Since the ith ranked unit in row i will be quantified, the quantified elements lie along the diagonal as shown below:

$$\begin{array}{cccccc}
\text{Set} & & & & \\
1 & X_{1(1)} & * & \cdots & * \\
2 & * & X_{2(2)} & \cdots & * \\
\vdots & & & \vdots & \\
m & * & * & \cdots & X_{m(m)}
\end{array}$$

The mean of the ranked set sample (considering one cycle only) is denoted by $\bar{X}_{(m)}$ where

$$\bar{X}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{i(i)}.$$

For convenience, we represent $X_{i(i)}$ by $X_{(i:m)}$. Then

$$\bar{X}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)}$$

and

$$E[\bar{X}_{(m)}] = \mu.$$

This shows that $\bar{X}_{(m)}$ is an unbiased estimator of the population mean (μ). If the whole procedure of drawing random samples is repeated r times, then the ith order statistic from the ith sample in the jth cycle is denoted by $X_{(i:m)j}$. The unbiased estimator of the population mean (μ) is given by

$$\bar{X}_{(m)r} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{r} X_{(i:m)j}}{mr}$$

and $$E(\bar{X}_{(m)r}) = \mu,$$

since $$\sum_{i=1}^{m} \mu_{(i:m)} = m\mu,$$

where $\mu_{(i:m)}$ represents the expected value of the ith order statistic.

The expression for the variance of $\bar{X}_{(m)r}$ is given by

$$\mathrm{Var}(\bar{X}_{(m)r}) = \frac{1}{m^2 r}\sum_{i=1}^{m} \sigma^2_{(i:m)},$$

where $\sigma^2_{(i:m)}$ denotes the variance of the ith order statistic. Also,

$$\mathrm{Var}(\bar{X}_{(m)r}) = \frac{1}{mr}\left\{\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2\right\},$$

where σ^2 denotes the population variance.

The RSS protocol requires the independent ordering of the randomly selected units of each sample (set) separately with respect to the magnitude of the

characteristic of interest without knowing their exact measurements. This requirement causes a problem in conducting a real trial in a field because the randomly selected units may be so widely spaced that the judgmental ordering of them becomes very difficult. With these facts in view, each set has to be formed on the basis of the sampling units selected randomly from a smaller area instead of the whole area under consideration. In order to implement it some locations are first randomly selected on a target area and then a "square cross" or a "circular frame" is placed at each such location. Quadrats of fixed size are either put at the marked positions on each square cross or placed randomly within a circle. (See Gore et al. (1992) for a detailed discussion on the issue.)

Sometimes even the judgmental ordering of the randomly selected units becomes difficult to be performed on the basis of the variable of interest (X). In order to overcome this problem, the ordering is carried out on the basis of some other easily available variable (Y) which is known as a concomitant variable. It helps in accomplishing the judgmental ordering conveniently because it is supposed to be correlated with the main variable of interest. To carry out the ordering m bivariate samples of size m are drawn randomly first. The X associated with the smallest ordered Y is quantified from the first sample, the X corresponding to the second smallest Y is quantified from the second sample, and so on. Lastly, the X associated with the largest Y is selected from the mth sample for the quantification. The whole cycle is repeated r times to obtain mr quantified units of the variable X.

10.5.2 Relative Precision of the RSS Estimator of a Population Mean Relative to Its SRS Estimator

We compare the variance of the mean of a ranked set sample with that of a mean based on a random sample of mr observations and not with a random sample based on m^2r observations. It is so because only the cost of quantification is considered. A random sample of size mr is obtained by randomly selecting one unit from each sample of size m in each cycle and then the unit is quantified. The mean based on the sample obtained by simple random sampling (SRS) is denoted by $\bar{X}$ where

$$\bar{X} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{r} X_{ij}}{mr}.$$

$\bar{X}$ is also an unbiased estimator of the population mean (μ) with the variance

$$\mathrm{Var}(\bar{X}) = \sigma^2/mr.$$

The expression for computing the relative precision (RP) is given by

$$
\begin{aligned}
rm\text{RP} &= \frac{\text{Variance of mean with random sampling}}{\text{Variance of mean with ranked set sampling}} \\
&= \frac{\text{Var}(\bar{X})}{\text{Var}(\bar{X}_{(m)r})} \\
&= \frac{\sigma^2/mr}{\frac{1}{mr}\left\{\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2\right\}} \\
&= \frac{1}{1 - \frac{1}{m}\sum_{i=1}^{m}(\tau_{(i)}/\sigma)^2} \quad \text{where} \quad \tau_{(i)} = \mu_{(i:m)} - \mu.
\end{aligned}
$$

The relative cost (RC) and the relative savings (RS) are computed as shown below:

$$
\text{RC} = \frac{1}{\text{RP}}, \quad \text{RS} = 1 - \text{RC}, \quad \text{or} \quad \text{RS} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\tau_{(i)}}{\sigma}\right)^2.
$$

Here the sample size of each rank is constant and the expression for RP does not appear to depend on the number of cycles considered. But the estimate of the relative precision would depend on m and r. McIntyre (1952) had conjectured that for "typical unimodal distributions" RP would not be much less than $\frac{m+1}{2}$ under the assumption of perfect ranking. However, Takahasi and Wakimoto (1968) have shown that the RP is bounded below by 1 and above by $(m + 1)/2$ for all continuous distributions with finite variances and the upper limit is realized only in case of a rectangular distribution. Further, Dell and Clutter (1972) have shown that the method of sampling provides an unbiased estimator of the population mean and the variance of the estimator is smaller than or equal to that of the corresponding SRS estimator even in the presence of error in ranking. But the magnitude of RP gets diminished by the imperfect ranking. There is no gain due to RSS if a judgmental ranking is the same as random ordering. Stokes and Sager (1988) have mentioned that the amount of the improvement of RSS over SRS in the case of perfect ranking is due to the fact that a ranked set sample consists of order statistics which are independent whereas in case of a simple random sample, these are always positively correlated. Patil et al. (1992) have surveyed the literature on the method of sampling and outlined its new applications in environmental investigations.

10.5.3 Unequal Allocation of Sample Sizes

The magnitude of the relative precision of the RSS estimator of the population mean relative to its SRS estimator also depends on the characteristics of the population under consideration. As such, it could be increased by resorting to unequal allocation keeping in view the nature of the population under consideration. To deal with asymmetric distributions, McIntyre (1952) suggested to allocate sample

sizes for various subpopulations proportional to their standard deviations. Halls and Dell (1966) applied the method and found considerable reduction in the magnitude of variance of the sample mean. But contrary to this finding, Martin et al. (1980) obtained either no gain or little gain by this unequal allocation of sample sizes in their investigation.

In order to describe the method, let us suppose that $r_1, r_2, \ldots, r_m$ denote the number of times units having the rank $1, 2, \ldots, m$ are quantified consecutively. Here, $r_1 + r_2 + \ldots + r_m = n$ (total sample size), $r_i \geq 1$ for all i. The value of r_i is determined in proportion to the standard deviation of the ith group. If T_i denotes the sum of the measurements for the units having the ith rank, then the RSS estimator $(\bar{X}_{(m)u})$ of the population mean is given by

$$\bar{X}_{(m)u} = \frac{1}{m} \sum_{i=1}^{m} \frac{T_i}{r_i}, \qquad E(\bar{X}_{(m)u}) = \mu,$$

and

$$\mathrm{Var}(\bar{X}_{(m)u}) = \frac{1}{m^2} \sum_{i=1}^{m} \frac{\sigma^2_{(i:m)}}{r_i},$$

where $\sigma^2_{(i:m)}$ denotes the variance of the ith order statistic. Note that $\mathrm{Var}(\bar{X}_{(m)u})$ is estimated, provided $r_i \geq 2$. If $r_1 = r_2 = \cdots = r_m = r$, the RSS design is said to be balanced; otherwise, it is unbalanced. In the present situation, $0 \leq \mathrm{RP} \leq m$; see Takahasi and Wakimoto (1968).

10.5.4 Formation of Homogeneous Composite Samples

RSS may also be utilized advantageously for forming more homogeneous composite samples compared to those based on random groupings. With m samples of size m, we form m composite samples by physically mixing the units of the same rank. Likewise, we get mr composite samples on the basis of m^2r units. These samples, in turn, provide mr measurements. The standard deviation of these measurements is expected to be smaller than that of the same number of measurements obtained after quantifying the composite samples consisting of physically mixed units selected randomly, m at a time, out of m^2r available units in most of the cases. For example, in case of 68 total units, 2 samples each of size 2 are randomly selected 17 times (i.e., $m = 2$ and $r = 17$). Then the two units of each sample are ordered as the small and the large. This results in 34 units for each of the 2 groups. Out of these units, two units are mixed physically at a time in each group. This gives, in fact, 17 composite samples in each group which need to be quantified. Thus, one has to make 34 measurements altogether. Contrary to this, for the usual composite sampling protocol based on random groupings, 2 units are randomly selected at a time out of 68 units and physically mixed before resorting to quantification. This,

Table 10.2 Sample size, mean, and standard deviation (SD) for individual samples, composite samples, and composites of ranked samples for grid A

Set size	Item	Sample size	Mean	SD
	Individual samples	184	200.72	902.9
2	Composite samples (random compositing)	92	200.72	627.9
2	Composites of ranked samples	92	200.72	618.4
	Individual samples	180	183.8	870.7
3	Composite samples (random compositing)	60	183.8	490.6
3	Composites of ranked samples	60	183.8	470.4
	Individual samples	176	187.8	880.2
4	Composite samples (random compositing)	44	187.8	509.8
4	Composites of ranked samples	44	187.8	321.5

also, yields 34 measurements. On comparing the standard deviations of these two sets of composite samples it is expected that the standard deviation of the measurements of the composite samples based on the ranked units should have smaller value than that of those based on the random groupings in most of the cases. In other words, the composite samples formed utilizing the RSS protocol are expected to be more homogeneous than those based on random groupings. The results are summarized for grids A and C in Table 10.2.

Chapter 11
Composite Sampling of Soils and Sediments

11.1 Detection of Contamination

11.1.1 Detecting PCB Spills

The US Environmental Protection Agency (EPA) has set reporting requirements for polychlorinated biphenyl (PCB) spills and views PCB spills as improper disposal of PCBs. The EPA has determined that PCB spills must be controlled and cleaned up. Components of the cleanup process may include, among other things, sampling and analysis to determine the materials spilled. The level of action required is dependent on the amount of spilled liquid, PCB concentration, spill area and dispersion potential, and potential human exposure. A sampling design is proposed by Boomer et al. (1985) for use by EPA enforcement staff. The proposed design involves sampling on a hexagonal grid which is centered on the spill site and extends just beyond its boundaries. Compositing strategies, in which several samples are pooled and analyzed together, are recommended.

In practice, the contaminated area from a spill will be irregular in shape. In order to protect against underestimation of the spill area, sampling within a circular area surrounding the contaminated area is recommended. The detection problem can be modeled as follows: try to detect a circular area of uniform contamination whose center is randomly placed within the sampling circle. Figure 11.1 illustrates this model. Two general types of design are possible for this detection problem: grid designs and random designs. Random designs have two disadvantages compared to grid designs for this application. First, random designs are more difficult to implement in the field, since the resulting pattern is irregular. Second, grid designs are more efficient for this type of problem than random designs. A grid design is certain to detect a sufficiently large contaminated area which some random designs may not. For example, the suggested design with a sample size of 19 has a 100% chance of detecting a contaminated area of radius 2.8 ft within a sampling circle of radius 10 ft. By contrast, a design based on a simple random sample of 19 points has only a 79% chance of detecting such an area.

Therefore, a grid design is recommended. A hexagonal grid based on equilateral triangles has two advantages for this problem. First, such a grid minimizes the

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_11, © Springer Science+Business Media, LLC 2011

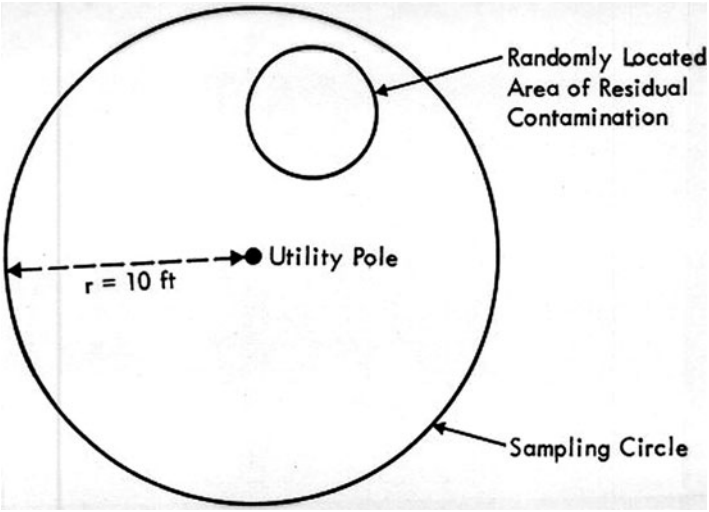

Fig. 11.1 Randomly located area of residual contamination with the sampling circle (*source:* Boomer et al. 1985)

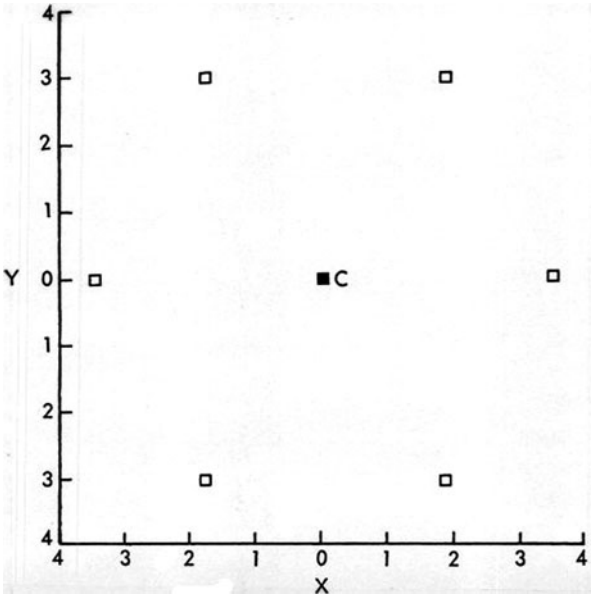

Fig. 11.2 Location of sampling points in a 7-point grid (*source:* Boomer et al. 1985)

circular area certain to be detected. Second, some previous experience (Mason, 1982; Matern, 1960) suggests that the hexagonal grid performs well for certain soil sampling problems. The smallest hexagonal grid has 7 points, next 19 points, the third 37 points, as shown in Figs. 11.2, 11.3, and 11.4, respectively.

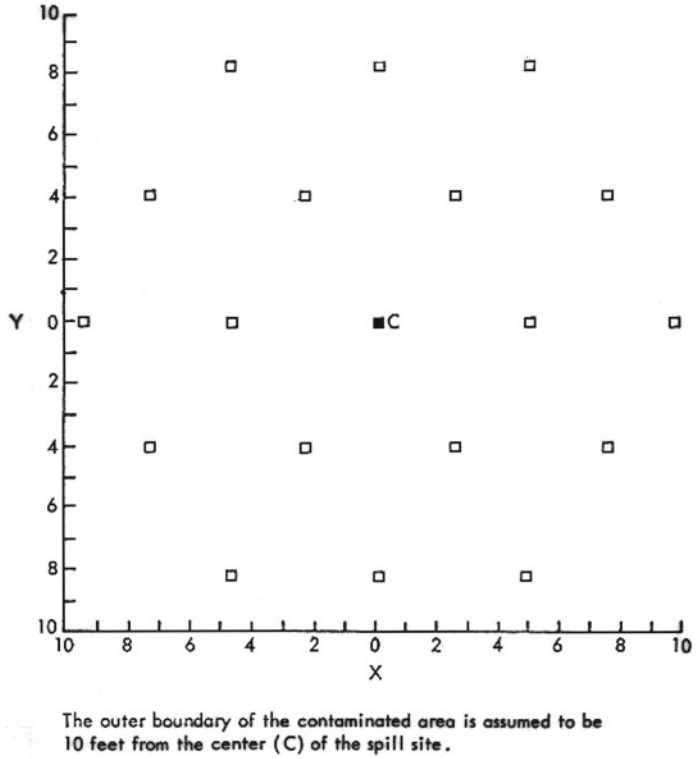

Fig. 11.3 Location of sampling points in a 19-point grid. (*source:* Boomer et al. 1985)

In general, the grid has $3n^2 + 3n + 1$ points. To completely specify a hexagonal grid, the distance s between adjacent points must be determined. The distance s is chosen to minimize the size of the contaminated circle which is certain to be sampled. Values of s so chosen, together with the number of sampling points and the radius of the smallest circle certain to be sampled, are shown in Table 11.1.

11.1.2 Compositing Strategy for Analysis of Samples

Once samples are collected at a site, the goal of the analysis may be to determine whether at least one sample has a PCB concentration above the allowable limit. Thus, it is not important to determine precisely which samples are contaminated or even exactly how many samples are contaminated. The cost of analysis can therefore be substantially reduced by employing compositing strategies, in which groups of samples are thoroughly mixed and evaluated in a single analysis. If the PCB level in a composite is sufficiently high, it can be concluded that a contaminated sample is present; if the level is sufficiently low, then all individual samples can be declared

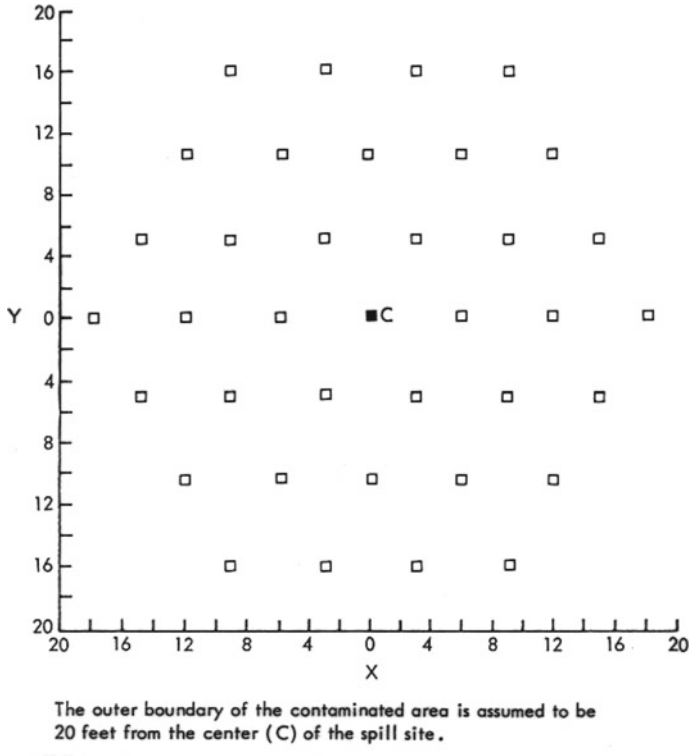

Fig. 11.4 Location of sampling points in a 37-point grid. (*source:* Boomer et al. 1985)

Table 11.1 Parameters of hexagonal sampling designs for a sampling circle of radius r feet

Number of points	Distance between adjacent points, s(ft)	Radius of smallest circle certain to be sampled
7	0.87r	0.5r
19	0.48r	0.28r
37	0.3r	0.19r

Source: Boomer et al. (1985)

clean. For intermediate levels, the constituent samples must be analyzed individually to make a determination.

The applicability of compositing is potentially limited by the size of the individual specimens and by the sensitivity of the analytical method at low PCB levels. First, the individual samples should be large enough so that composites can be formed while leaving enough material for individual analyses if needed. The second limiting factor is the detection limit of the analytical method. If the detection limit is 1 part per million (ppm), and the assumed permissible level is 10 ppm, no more than 10 specimens should be composited at a time.

In compositing specimens, the location of the sampling points to be grouped should be taken into account. Boswell and Patil investigated this problem with a simulation study and have found that contiguous specimens should be composited, if feasible, in order to maximize the potential reduction in the number of analyses produced by the compositing strategy. Some possible compositing strategies are indicated graphically in Figs. 11.5, 11.6, 11.7, and 11.8.

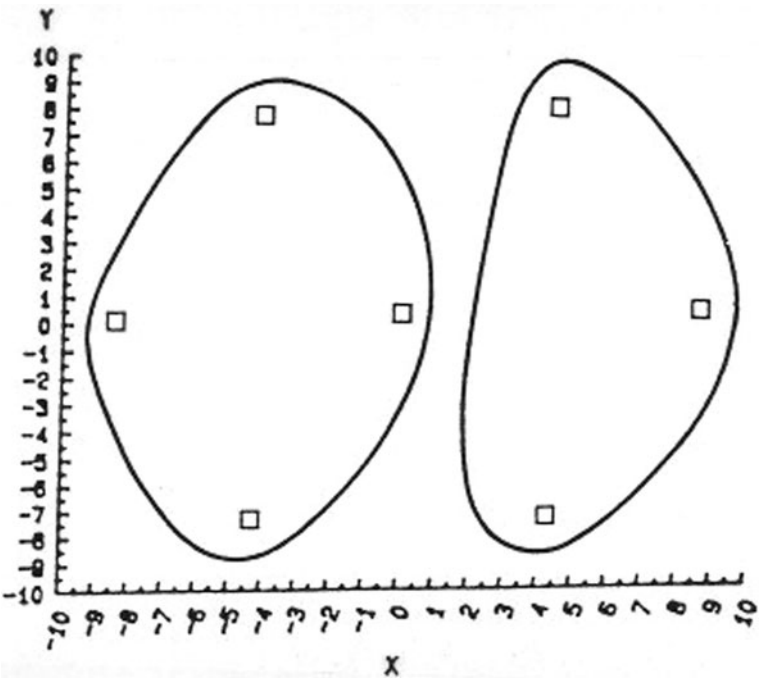

Fig. 11.5 Location of sample points in a 7 sample point plan with detail of a two-group compositing design (*source:* Boomer et al. 1985)

11.2 Estimation of the Average Level of Contamination

11.2.1 Estimation of the Average PCB Concentration on the Spill Area

In addition to detecting the presence of PCBs on the spill site, it may also be important to estimate the average PCB concentration. Composite sampling has an advantage over individual sample measurements in this problem, too. While maintaining the same precision as that of the mean of all individual sample values (had they been measured), the mean of the composite sample values estimates the average PCB concentration at a substantially reduced cost of analysis.

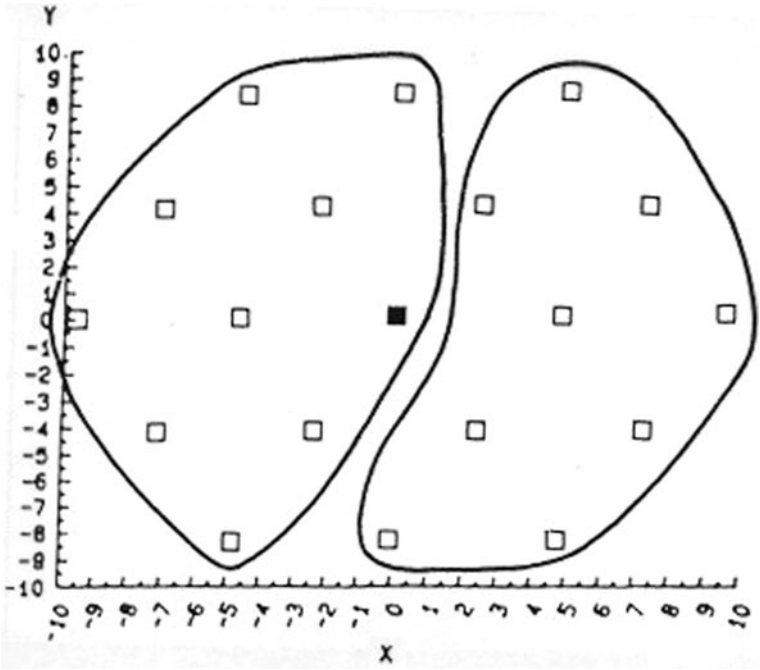

Fig. 11.6 Location of sample points in a 19 sample point plan with detail of a two-group compositing design (*source:* Boomer et al. 1985)

For n composites of size k, the average of the composite sample measurements is an unbiased estimator of the population mean, with a standard error of $\sigma/\sqrt{(nk)}$, where σ is the population standard deviation, n is the number of composite samples, and k is the composite sample size (Rohde, 1976; Elder et al., 1980). Note that the standard error of the composite sample estimator is the same as that of the individual sample estimator computed from nk individual samples, but the number of analytical measurements is n in the case of composite samples as opposed to nk in the case of individual samples.

11.2.2 Onsite Surface Soil Sampling for PCB at the Armagh Site

A preliminary study was carried out (Gore et al., 1992) to evaluate the performance of composite sample techniques for characterizing PCB concentration. Since sampling and chemical analyses had already been carried out, the study reported here is a retrospective one, in which compositing is "simulated" by averaging the recorded measurements for individual samples. In the absence of measurement error, these averages exactly reproduce the measurements on composite samples that would have been obtained by physical compositing. However, measurement error may

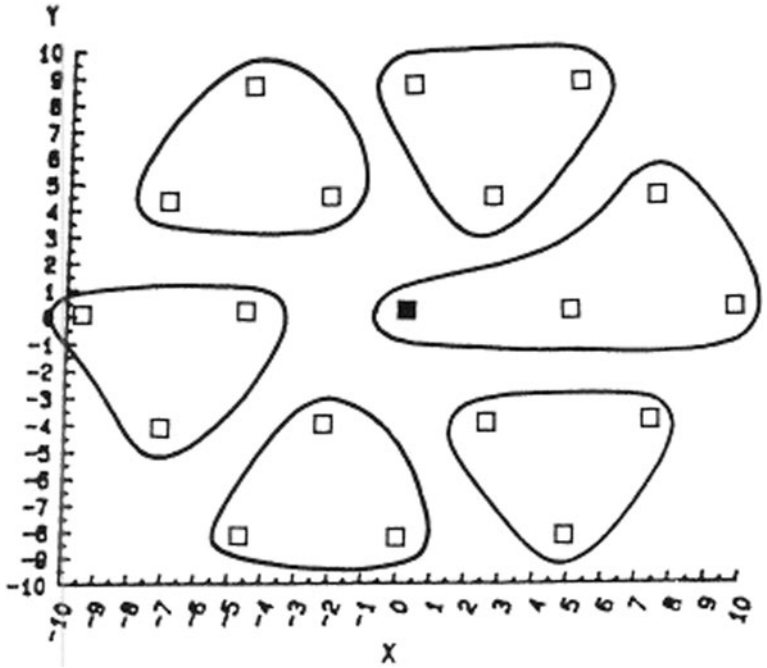

Fig. 11.7 Location of sample points in a 19-sample point plan, with detail of a two-group compositing design (*source:* Boomer et al. 1985)

not be negligible for PCB concentrations, and further study should be necessary to assess the impact of such error on the cost-efficiency of composite sample techniques.

11.2.3 The Armagh Site

11.2.3.1 Location and Features

The Armagh compressor station is located in West Wheatfield Township, Indiana County, about 1.25 miles south of US Route 22. The map in Fig. 11.9 shows the location of the Armagh site in the State of Pennsylvania. The site includes one compressor building along with several other buildings on 79 acres. There are two known liquid pits. The surrounding area contains 64 residences within 1 mile of the station. Some of these houses have private wells; however, a public water supply line was recently installed. None of the private wells has been found to be contaminated as a result of the Texas Eastern activities. There is one wetland situated within one-half mile of the site. Richard Run, which flows to the south of the site, is classified as a cold water fishery. There are no public recreational facilities near the station.

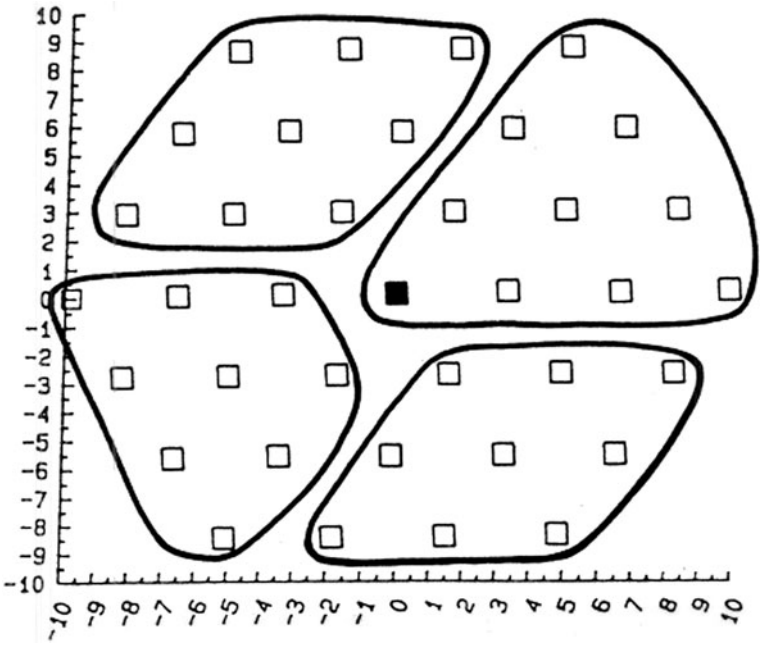

Fig. 11.8 Location of sample points in a 37-sample point plan, with detail of a four-group compositing design (*source:* Boomer et al. 1985)

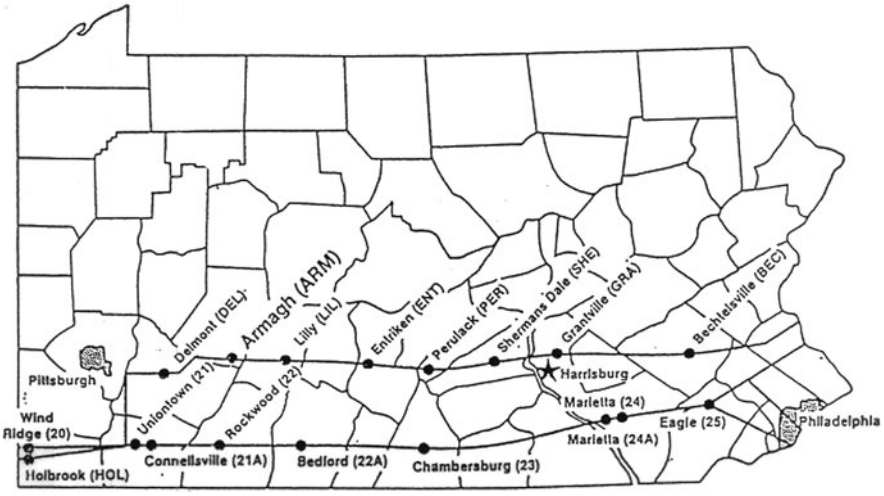

Fig. 11.9 Locations of the Pennsylvania sites

11.2.3.2 Onsite Soils

Onsite soils are defined as being within the confines of the station site fencing and are therefore accessible only by Texas Station personnel and authorized site visitors. The areal extent of excavation is expected to be determined by the 10 parts per million (ppm) PCB contour lines which are generated based upon the onsite soil characterization data. The cleanup criterion for onsite soils is specified by an average overall PCB concentration of 5 ppm. The objective of the onsite surface soil sampling was to characterize the presence of PCBs at the Armagh site. Sampling locations around potential sources of contamination were selected for sampling in phase I. As part of phase II sampling, samples were collected at points around each phase I sampling location having a total PCB concentration greater than 10 ppm for onsite surface soils.

11.2.3.3 Onsite Surface Soil Sampling

Potential sources of PCB had been identified, and a rectangular grid was laid out about each such source. Four different onsite grids were identified by the alphabetic codes "*A*" through "*D*." Grid points were identified by a two-digit row number and an alphabetic column code. Sampling of the surface soil was done at selected grid points in two distinct phases. The second phase was undertaken to fill in locations not covered during phase I. Grid "*D*" was not sampled during phase I, but only during phase II. Phase II locations were generally farther away from the potential PCB source, and the measured PCB concentrations tended to be lower during this phase.

A total of 130 onsite surface soil samples were collected during phase I and 228 during phase II as follows.

	Phase I	Phase II
Grid A	78 samples	106 samples
Grid B	16 samples	16 samples
Grid C	36 samples	32 samples
Grid D		74 samples

The map in Fig. 11.10 shows the grids and sampling locations on the Armagh site. The distance between consecutive rows as well as between consecutive columns was 25 ft. For the purpose of computerization and to facilitate analysis using statistical computer packages, the alphabetic column codes were converted into numeric codes with *A* into 1, *B* into 2, and so on. Row and column codes for grids *B*, *C*, and *D* were shifted to synchronize them with the codes of the grid A. This synchronization enables plotting of all the sampled grid points on the same graph. Schematic plots, which are not to the scale, showing grid points that were sampled in either phase are given in Fig. 11.11a–g.

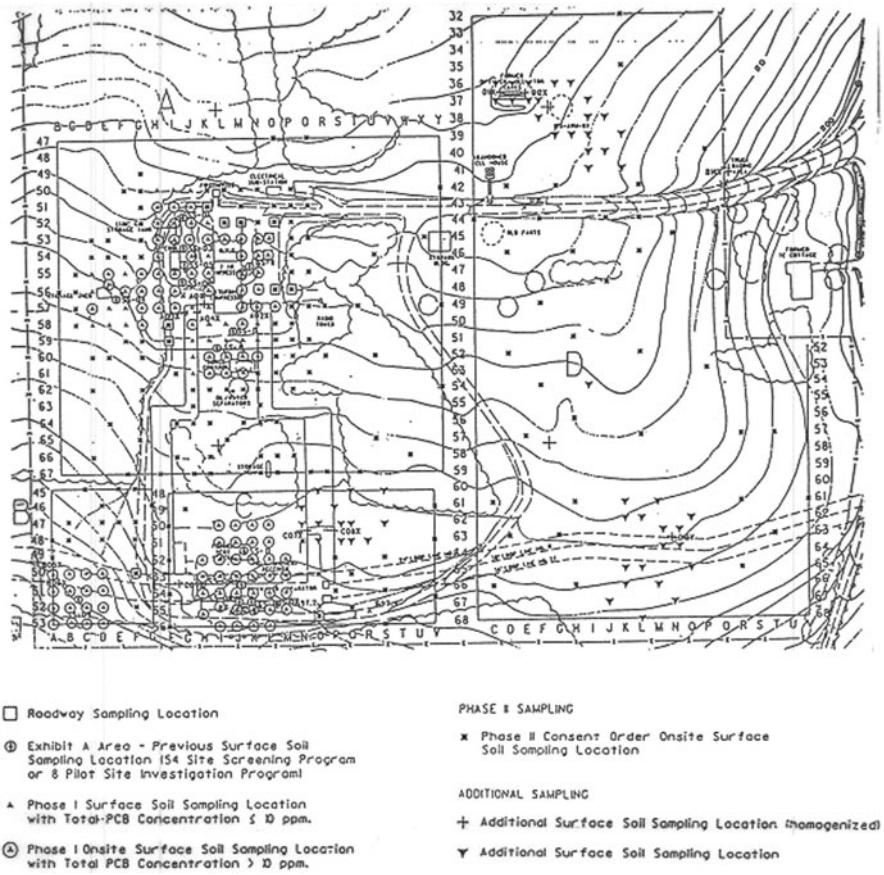

Fig. 11.10 Grid layout and sampling locations for onsite surface soil sampling at the armagh site

Soil samples were taken from a 0 to 6 in. depth. After removing vegetation, rocks, and other debris, the sample at each grid point was thoroughly mixed to obtain a homogeneous sample for analysis and quantification. Duplicate and triplicate samples were taken at some locations, but these have not been included in the analysis. The discrepancy between the measurements on primary, duplicate, and triplicate samples can be useful in studying measurement errors and will be investigated separately in a subsequent report.

11.2.4 Simulating Composite Samples

11.2.4.1 Choice of the Composite Sample Size

Boswell and Patil (1990) have investigated strategies for composite sample formation when samples are spatially correlated. After comparing four different

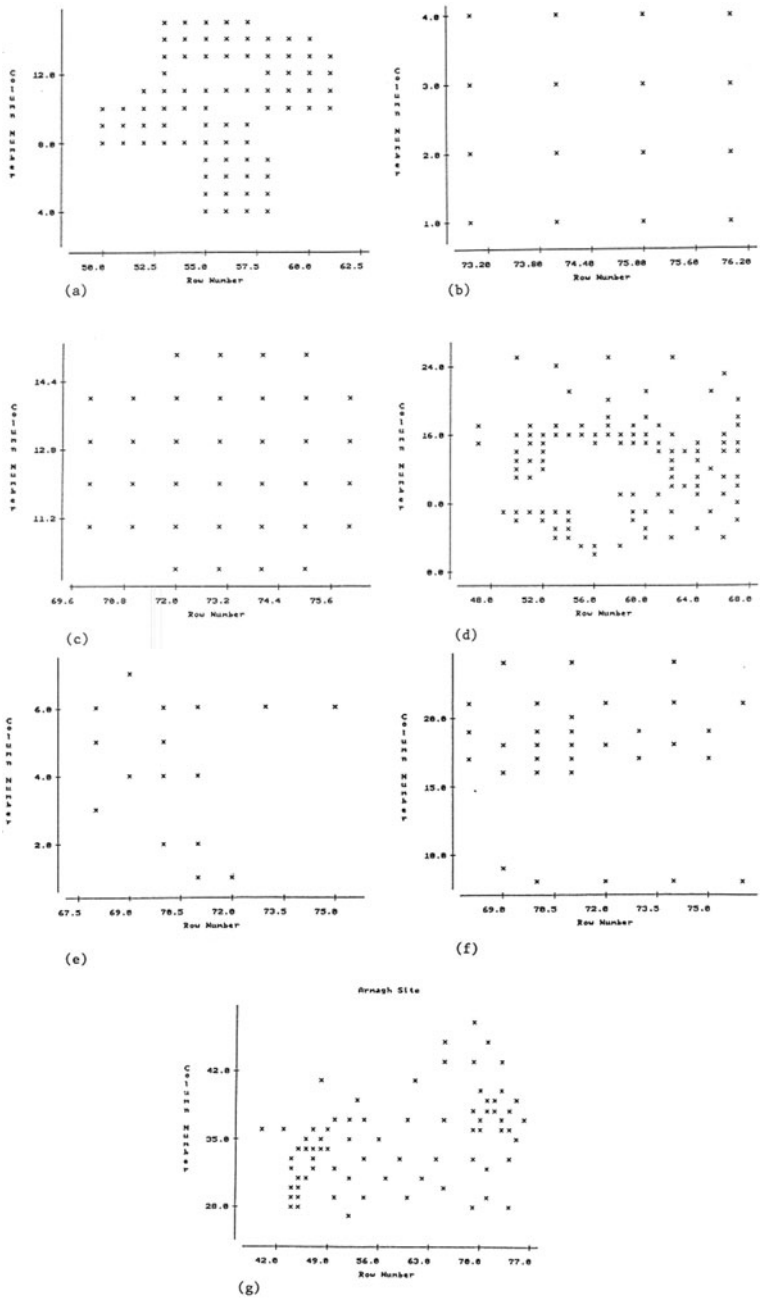

Fig. 11.11 (**a**) Sampling locations on grid A, phase I; (**b**) sampling locations on grid B, phase I; (**c**) sampling locations on grid C, phase I; (**d**) sampling locations on grid A, phase II; (**e**) sampling locations on grid B, phase II; (**f**) sampling locations on grid C, phase II; and (**g**) sampling locations on grid D, phase II

compositing strategies for classification of individual samples, Boswell and Patil arrived at the conclusion that, when there is spatial dependence among the sampling locations, compositing samples from neighboring points, as nearly as possible in a square region, increase the cost-efficiency of composite sampling. Due to the spatial dependence, these samples are likely to exhibit greater homogeneity than randomly selected samples.

In order to maximize within-composite homogeneity, it was decided that all composites would be formed within a sampling phase and also within a grid. This may also be desirable from the management and operational point of view. These considerations led to the decision to composite individual samples taken from contiguous locations belonging to the same grid and sampled in the same sampling phase.

11.2.4.2 Composite Sample Formation

After the composite sample size was determined through considerations described above, it was necessary to identify the sampling locations to be composited. There is considerable subjectivity involved since not all the grid points were included in the sampling plan and the sampled grid is not exactly rectangular. However, precaution was taken to avoid selection bias in the composite sample formation. First, even though the PCB concentrations at the sampled locations were known, the formation of composite samples was based only on the geographical positions of the sampling locations. Second, a few alternative composite sample formation protocols were implemented for comparison with that used for the analysis reported here. Since the estimate of the mean PCB concentration does not depend on the compositing protocol, the estimate of the variance was used as the criterion for this comparison. Unbiased estimates of the population mean μ and the population variance σ^2 were calculated. Relevant tabulation is given in Table 11.2.

We observe that composite sample estimates of the population mean are identical to the corresponding individual sample estimates. Since individual samples from

Table 11.2 Unbiased estimates of μ and σ

	Measurements of Individual samples			Composite samples		
	N	$\bar{X}$	S_x	n	μ	σ
Phase I						
Grid A	78	363.32	1355.6	20	363.32	1793.72
Grid B	16	64.56	40.0	4	64.56	36.12
Grid C	36	1075.14	2076.5	9	1075.14	2974.74
Phase II						
Grid A	106	81.46	198.0	26	81.46	208.94
Grid B	16	26.01	23.3	4	26.01	36.46
Grid C	32	70.2	79.3	8	70.2	84.1
Grid D	74	36.59	91.1	19	36.59	97.28

contiguous sample locations are composited, they are expected to be somewhat homogeneous. As a consequence, the variation between composites is expected to be larger than the variation within composites. The composite sample estimate of the variance is accordingly larger than the individual sample estimate in most of the cases. However, since the composite sample estimates involve 25% measurement cost as compared to the individual sample estimate, composite sampling is preferred to exhaustive testing. On the other hand, confidence intervals for population means will be wider if computed from composite sample measurements rather than from individual sample measurements. The conclusions drawn from composite sample data will then be more conservative than those drawn from individual sample data. In either case, composite sampling schemes perform better than exhaustive testing.

11.2.5 Locating Individual Samples with High PCB Concentrations

To illustrate the method described in Section 3.2, in case of the Armagh site, we note that the highest PCB concentration in a composite sample was 4897.5 ppm. Since the size of this composite was 4, the highest PCB concentration in an individual sample cannot exceed 19,590 ppm. Exhaustive testing of the constituent samples resulted in the highest PCB concentration in an individual sample, namely, 10,000. Since there was a composite sample with PCB concentration of 3999.5 ppm, it could contain an individual sample with PCB concentration exceeding 10,000 ppm. Upon measuring every individual sample in this composite, it was indeed found to be the case, as there was an individual sample with PCB concentration of 10,700 ppm. This implies that no composite sample with a PCB concentration of 2675 ppm or less can contain an individual sample with PCB concentration exceeding 10,700 ppm. Since there was no composite sample with a measurement exceeding 2675 ppm, the sampling location with the largest PCB concentration was identified. Note that it required only 8 measurements in addition to the 90 composite sample measurements.

Figure 11.12a–d shows a scatterplot of individual sample measurements plotted against the simulated composite sample measurements. The two rays from the origin indicate the upper and the lower bounds on the largest individual sample measurement for every composite sample measurement. Thus, corresponding to the composite sample with a measurement of 4897.5 ppm, the upper bound for an individual sample measurement in this composite is 19,590 ppm, while the lower bound for the same is 4897.5 ppm, which is the same as the composite sample measurement. Since 4897.5 ppm was the largest composite sample measurement, individual samples in this composite were measured separately, and an individual sample with a PCB concentration of 10,000 ppm was identified. A horizontal line through the point identifying this individual sample indicates that there is only one composite which can possibly contain an individual sample with more than 10,000 ppm of PCB concentration. Exhaustive testing in this composite located an individual sample with a PCB concentration of 10,700 ppm. There is no other composite

that can contain an individual sample with a PCB concentration exceeding 10,700 ppm, as is evident from Fig. 11.12b. The exhaustive testing of two composites have thus identified the individual sample with the largest PCB concentration. This search can easily be extended to identify more individual samples with high PCB concentration. Figure 11.12b–d shows how exhaustive testing of only three composites identify the individual samples with the four largest PCB concentrations. In other words, with only 12 measurements in addition to the 90 measurements on the simulated composite samples, we were able to identify the 4 individual samples with the highest PCB concentrations.

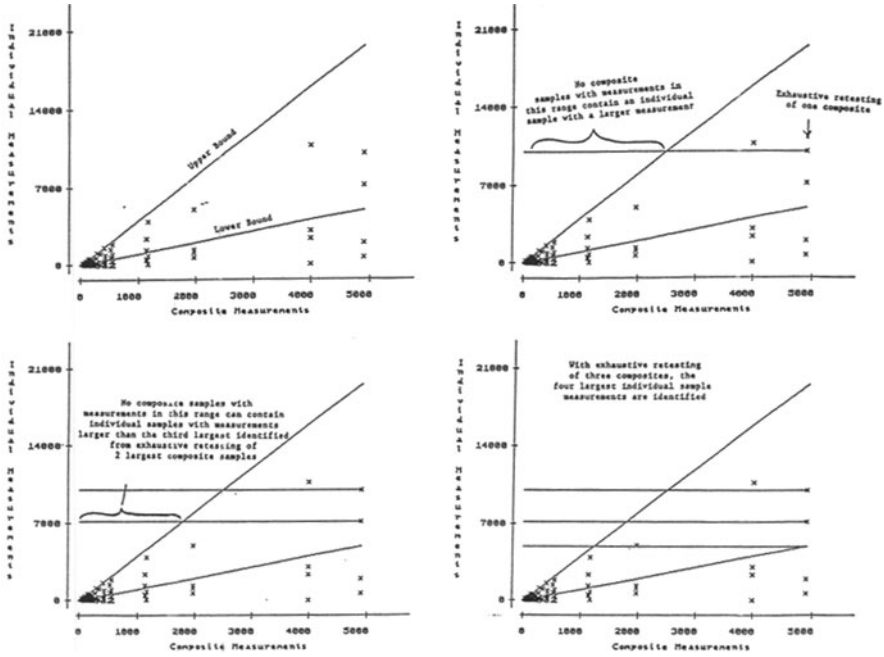

Fig. 11.12 The sweep-out method for identifying the individual samples with extremely large values

11.3 Estimation of Trace Metal Storage in Lake St. Clair Post-settlement Sediments Using Composite Samples

During 1985, Canadian and the US agencies and institutions undertook a cooperative study of Lake St. Clair sediments (Rossman, 1988). The objectives of the sediment sampling program were to describe the organic and metal contaminants stored in the sediments and to provide information on the permanence of storage. Sediment cores were collected from 36 locations within the lake (see Fig. 11.13). Core samples were collected by inserting core liners as deeply as possible into the

sediments. Penetration into the sands was 3–9 cm and into the silts and clays was 6–36 cm. Cores were extruded at 1 cm intervals to a depth of 10 cm and at 1–2 cm intervals for sediment depths greater than 10 cm. After noting the sediment texture, the intervals were stored frozen in polyethylene bottles. In the laboratory, frozen

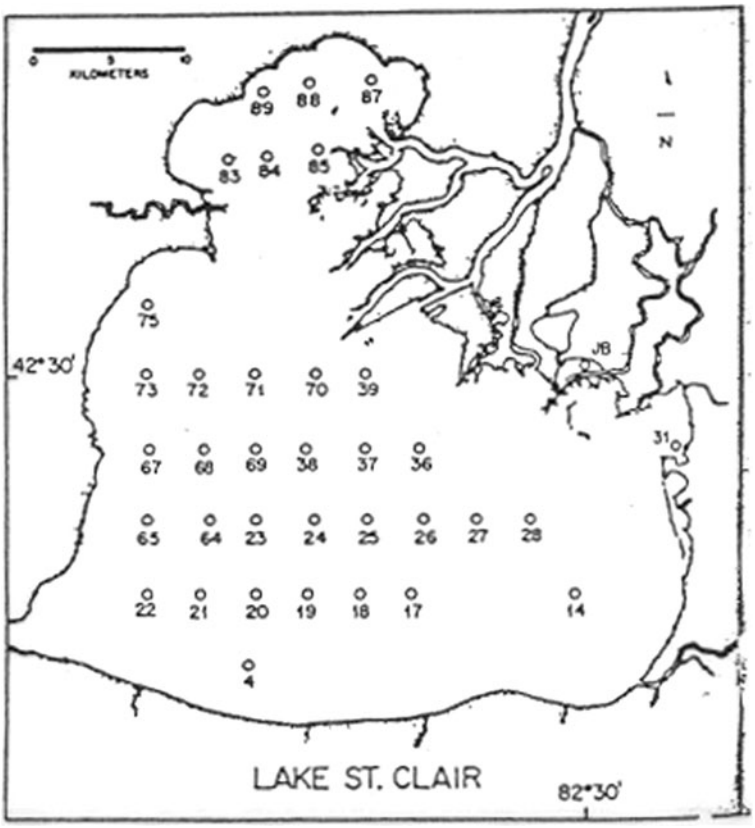

Fig. 11.13 Stations at which cores were recovered during sampling of Lake St. Clair in 1985 *source:* Rossman 1988

Table 11.3 Mean coefficients of variation (mg/kg) for analysis of Lake St. Clair sediments

Element	Coefficient of variation	Limit of detection
Bi	8.3	0.054
Cd	18	1.8
Cr	4.5	4.8
Cu	3.4	2.4
Ni	4.3	5.4
Pb	8.0	7.4
Sb	21	0.047
Zn	5.0	4.8

Source: Rossmann, 1988)

Table 11.4 Results of analysis of municipal-digested sludge compared to previous analyses of the sludge and the USEPA true average and range (mg/kg)

	Lake St. Clair study			Previous studies		USEPA reported concentration	
Metal	*N*	Mean	Standard deviation	Mean	Standard deviation	Mean	Range
Bi	10	23.6	2.26	24.2	2.43	–	–
Cd	10	18.7	1.23	18.2	0.351	20.772	2.49–39.1
Cr	10	198	13.0	195	0.577	204.46	115–294
				217	4.04		
Cu	18	1040	26.0	1040	0.707	1095.3	831–1360
				1020	0.0		
Ni	10	181	5.03	182	1.44	198.31	164–233
				188	0.342		
Pb	10	532	17.6	517	0.0	518.76	305–733
	10	546	10.9	517	12.0		
Sb	9	5.13	0.970	6.99	0.0707	–	–
				5.94	0.156		
Zn	9	1230	35.6	1220	15.4	1320	1190–1450

Source: Rossman (1988)

samples were weighed and freeze-dried without external heat. After weighing the dried samples to determine water content, the samples were stored in polyethylene bottles. These samples were then gently ground with a mortar and pestle. Composite samples were formed with subsamples of sediment from each section of a core proportional to its contribution to the total mass of sediment in the core. Composite samples were thoroughly mixed and stored in polyethylene bottles.

A 29 subsample of each composite was weighed into a 250 ml flask, spiked with polonium-209, and extracted into 100 ml of hot (80°C) 10 (v/v) hydrogen peroxide. Each time the sample volume was reduced to 5–10 ml. When the reaction with the hydrogen peroxide subsided, more hydrochloric acid was added to bring the volume to 50 ml. This process was repeated two more times during the 40-h extraction period. After 40 h, the solution was allowed to evaporate to a volume of 5–10 ml. This extraction technique dissolved all components of the sediment except silicate minerals. The extract was then separated from the insoluble residue, and the filtered extract was transferred into a 100-ml volumetric flask and brought to volume.

Cadmium, chromium, copper, nickel, lead, and zinc were analyzed by standard lame techniques using an atomic absorption spectrophotometer. Quantification was with standard curves. Bismuth and antimony were analyzed by flameless atomic absorption. Except for cadmium and antimony, the coefficient of variation for each metal was below 10% (Table 11.3). Detection limits are those obtained for the ranges of concentration found in the samples. For the eight metals considered in the study, a total of 288 analyses were done. Of these, only three results for composite samples were below the detection limit. Thus the composite results were within the certainty of the analysis. Analyses of the USEPA municipal-digested sludge and a previously analyzed Lake Michigan sediment sample were used to monitor the quality of the analyses. All results were within the given USEPA range of acceptance or were reasonably close to the previous results (Tables 11.4 and 11.5).

Table 11.5 Results of analysis of a standard lake mud during the Lake St. Clair study compared with previous analysis of the standard lake mud (mg/kg)

	Lake St. Clair study			Previous studies	
Metal	*N*	Mean	Standard deviation	Mean	Standard deviation
Bi	3	0.372	0.234	0.296	0.0240
Cd	3	5.74	0.0781	–	–
Cr	3	54.7	0.131	56.2	0.17
Cu	3	40.1	0.0961	39.8	1.25
Ni	3	30.1	0.0709	36.3	1.55
Pb	3	68.4	2.33	79.0	1.80
	3	77.7	0.687		
Sb	3	0.451	0.0301	0.550	0.0328
				0.520	0.0329
Zn	3	146	2.33	168	14

source: Rossmann (1988)

Chapter 12
Composite Sampling of Liquids and Fluids

12.1 Comparison of Random and Composite Sampling Methods for the Estimation of Fat Content of Bulk Milk Supplies

The fat content of milk is determined on a composite sample which is formed from samples using all deliveries during a specified period of time. Milk samples taken at the time of collection are transported to the processing plant and assembled into composites. Since it is a known fact that composite samples provide an unbiased estimate of the population mean, the interest is in comparing the precision of composite sample estimator with that of the individual sample estimator. An important consequence of compositing is the loss of information on individual sample measurements, and hence the loss of information on sample-to-sample variability within a composite. While comparing between individual samples to locate any differences, compositing may add a variance component to the analysis. Williams and Peterson (1978) developed a method to provide a framework for assessing the precision of sampling schemes through estimation of components of variation associated with the sampling process. They identified four components: variance due to real difference between collections from a supplier within a compositing period (biological variance σ_d^2), variance among samples taken from the same collection (sample variance, σ_s^2), variance among measurements on the same sample (testing variance, σ_t^2), and the variance associated with the formation of a composite sample (compositing variance, σ_c^2).

12.1.1 Experiment

Sixty-one herd milk supplies in three different creamery locations were sampled during the trials. Samples were taken by the regular tanker drivers. The milk in the bulk tanks was agitated for 1 min prior to pumping off and two samples were taken (first after starting and second toward the end of pump off) from each delivery. On arrival at the creamery samples were analyzed in duplicate for fat, and in addition, each one was used in the formation of a 2-week composite. After each fortnightly

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_12, © Springer Science+Business Media, LLC 2011

interval, the two composite samples were analyzed in duplicates for fat. Fat content was determined by the Milko Tester Automatic.

12.1.2 Estimation Methods

Testing variance σ_t^2 could be estimated from the data for each collection and also from the composite data. Since the composite data were more complete than the individual collection data, a hierarchical orthogonal ANOVA was possible with the data, from which testing variance was estimated. Sampling variance was estimated for each herd collection. This involved the subtraction of the estimate of testing variance from an estimate of testing plus sampling variance. Compositing variance was estimated as follows: in the hierarchical ANOVA of composite data, the expectation of the mean square for between composites within herd-periods is given by

$$\sigma_t^2 + 2(\sigma_c^2 + \sigma_s^2/\bar{N}),$$

where $\bar{N}$ is the harmonic mean number of collections per herd-period. Compositing variance can be estimated from this formula by inserting estimates of sampling and testing variances. The estimate of testing variance used was the one from the composite data. Estimates of the biological variability are based on the direct calculation of day-to-day variation in fat percentage for each herd-period. The expectation of this variance is

$$\sigma_d^2 + \sigma_s^2/2 + \sigma_t^2/4$$

from which an estimate of biological variance (σ_d^2) can be obtained by subtraction of multiples of estimates of σ_s^2 and σ_t^2.

12.1.3 Results

Estimates of testing, sampling, compositing, and biological variances for three locations are shown in Table 12.1. Two estimates of testing variance are presented for each location, based on the variation between duplicate determinations of fat percentage for individual collections and composite samples, respectively. These values are small relative to the sampling and compositing components. The sampling variances are somewhat smaller than the compositing variances, while both differ between locations. The biological components of variability were about 10 times as large as sampling or compositing components. Random and composite sampling schemes are compared in Table 12.2.

Table 12.1 Estimates of testing, sampling, compositing, and biological variances for three locations

	Testing variance $\times 10^{-4}$				Sampling variance $\times 10^{-4}$			
Location	Individual collections	df	Composite data	df	No. of observations	Estimate	Compositing variance $\times 10^{-4}$	Biological[a] variance $\times 10^{-4}$
A	1.24	1272	0.99	220	635	11.8	18.3	180
B	0.50	4342	1.60	468	2171	4.3	12.8	224
C	0.21	2532	0.74	360	1257	28.8	34.0	187

[a] Period 11 was excluded from this analysis for locations A and B as it gave large values that differed very much from other periods

Source: Connolly and O'Connor (1981)

Table 12.2 Comparison of accuracy of random and composite sampling schemes for determination of fat percentage under a seven-collection per composite system

Variance for location under random sampling ($\times 10^{-4}$)			
Days sampled at random	A	B	C
1	167	198	190
2	71	83	82
3	39	45	46
4	23	26	28
5	13	14	17
6	7	6	9
Composite method variance ($\times 10^{-4}$)	21	15	39
Days sampled to give equivalent precision	4–5	4–5	3–4

Source: Connolly and O'Connor (1981)

12.1.4 Composite Compared with Yield-Weighted Estimate of Fat Percentage

In addition to satisfying variability criteria, sampling methods should also be unbiased. Fat percentage for a period can be estimated by the composite sampling method, or if information is available in every collection, by a weighted mean of the fat percentage for each collection, weighted by the milk yield in the collection. In the current data, the difference and the ratio of the weighted to the composite estimate were both analyzed in ANOVA which examined the effects of herds and periods (Table 12.3). There was no consistency over locations. In addition to these average effects, the difference between the methods varied significantly over herds and periods for two of the locations.

12.2 Composite Sampling of Highway Runoff

Storm water runoff from highways is monitored by manual grab sampling or automatic water quality samplers in conjunction with flow measuring instruments. Discrete runoff samples can be used to characterize the changes in concentration of

various pollutants through a storm, but are usually mixed in some way to form a composite sample so that average concentrations can be used to calculate total mass loadings of pollutants. Because runoff characteristics are continuously changing, sampling at discrete points is limited in accuracy. Small storms may pass unsampled, peaks in concentration may occur between samples, or large storms may exceed the container capacity of the sampler. For these reasons, it is desirable to continuously accumulate a composite runoff sample for determination of total pollutant loadings. Wullshleger et al. (1976) suggest four methods of combining discrete samples to obtain a composite according to the time they were taken and the flow rate or the volume they represent. Another method is to use a device that continuously removes a fixed fraction of the storm water runoff proportional to the flow rate and automatically accumulates it in a composite sample. Clark et al. developed such a device and took samples from Interstate-5, I-5, in Seattle between February and September 1979. A summary of the analysis of their data is given in Table 12.3.

Table 12.3 Highway runoff water quality comparisons

	1–5 sampling site in Seattle			
	Composite concentration			
(1)	Average (2)	Range (3)	Range of discrete sample concentrations (4)	Average of national composites (2) (5)
pH	6.1	5.1–6.9	4.5–7.1	–
Conductivity	87.0 μmho/cm	30.0–146.0	31.0–409.0	–
COD	137.0	75.0–211.0	8.0–914.0	147.0
TSS	145.0	43.0–320.0	30.0–1120.0	261.0
VSS	38.0	12.0–100.0	2.0–696.0	77.0
TOC	27.0	4.0–47.0	BDL–83.0	41.0
Pb	0.8	0.2–1.5	0.1–5.5	0.96
Zn	0.40	0.2–1.0	0.03–1.9	0.41
Cu	0.03	BDL–0.07	BDL–0.15	0.10
TKN	1.11	0.64–1.96	0.18–3.96	2.99
NO_3-NO_2-N	0.82	0.52–1.65	0.05–2.20	1.14
Total P	0.34	0.20–0.55	0.12–1.08	0.79

Note: BDL = below detectable limit. All concentrations in mg/l unless stated otherwise
Source: Clark et al. (1981)

A fully automated discrete sampling system was established with a mechanical sampler. A composite sampler was also developed with the following considerations:

1. The composite sampler must produce a representative sample, with the average characteristics of the runoff from an entire storm
2. The resulting sample was to be used in calculating the entire storm amount, and no other flow measuring device was being used

3. The sample volume had to be sufficiently small to store
4. The sampler should sample solids in the storm water and must not be incapacitated by litter and debris
5. The sampler should need minimal maintenance and should not require electrical power
6. The cost of the composite sampler should be significantly lower than the conventional discrete sampler

After it was developed, the composite sampler was tested and compared with the conventional discrete sampler, and the following conclusions were drawn. The composite sampler was capable of accurately removing a fixed amount of fraction of the total flow in the channel proportional to the flow rate. Operation of the composite sampling system was simple and required a minimal amount of maintenance. One unit of the composite sampler costs $900 as opposed to a cost of $6440 for the conventional discrete sampler. Figures 12.1, 12.2, and 12.3 show how composite samples collected with the composite sampler performed in comparison with the discrete sampler.

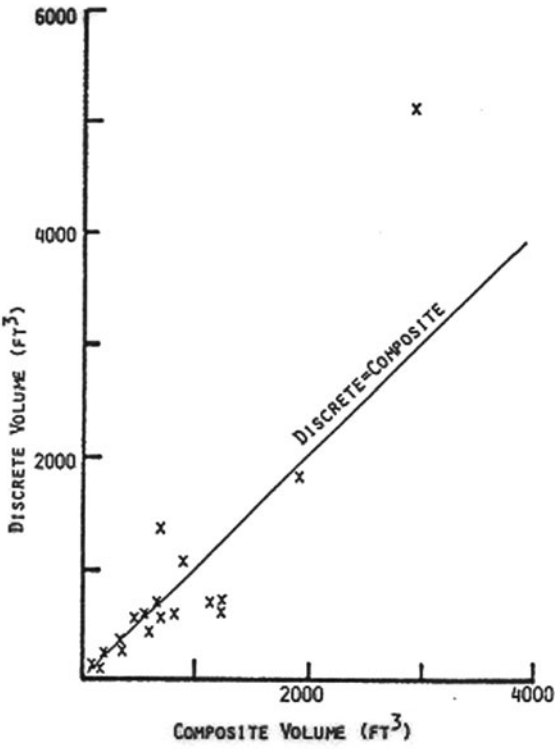

Fig. 12.1 Discrete vs. composite runoff volume (1 ft^3 = 0.028 m^3) (*Source*: Clark et al., 1981)

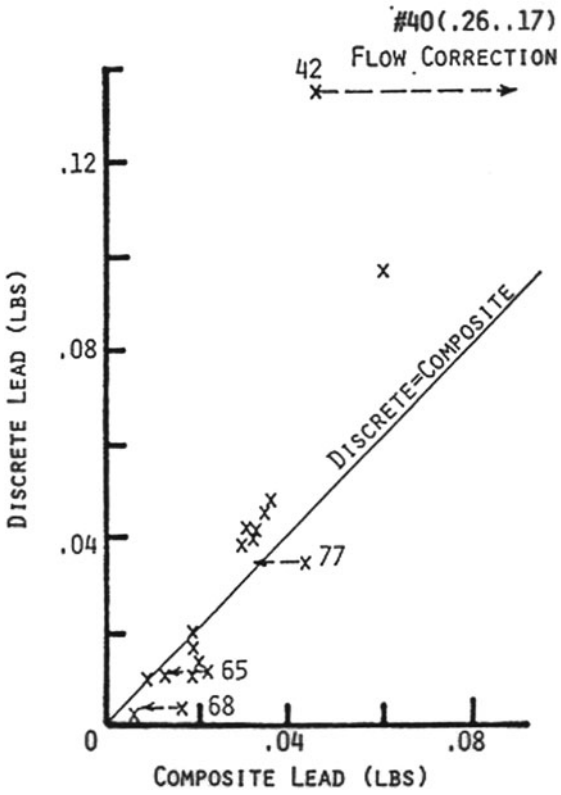

Fig. 12.2 Discrete vs. composite results for COD (1 lb = 0.453 kg) (*Source*: Clark et al., 1981)

12.3 Composite Samples Overestimate Waste Loads

Schaeffer et al. (1983) have reported two case studies that attempt to make a comparison between grab and composite samples while evaluating wastewater treatment plants. Wastewater treatment plant performance is monitored by the collection and analysis of samples from the process stream for physical, chemical, and microbiological constituents. Samples may be broadly classified as "grab" or "composite." A grab sample represents the composition of the flow at a given instant in time, irrespective of the flow volume. A composite sample represents an average composition of the flow over time and may or may not be proportional to the flow. Flow proportional (FP) sampling can be accomplished in one of the following ways: fixed time with sample volume proportional to flow (VP) or fixed volume with time proportional to flow (TP). Non-flow proportional (NFP) composites are usually taken as a fixed volume at fixed times.

Schaeffer et al. have reported the results of the analysis of effluent samples at St. Charles and Freeport. At St. Charles, all effluent samples were taken at the outlet of the final chlorination process for clarified water. Grab samples were taken every

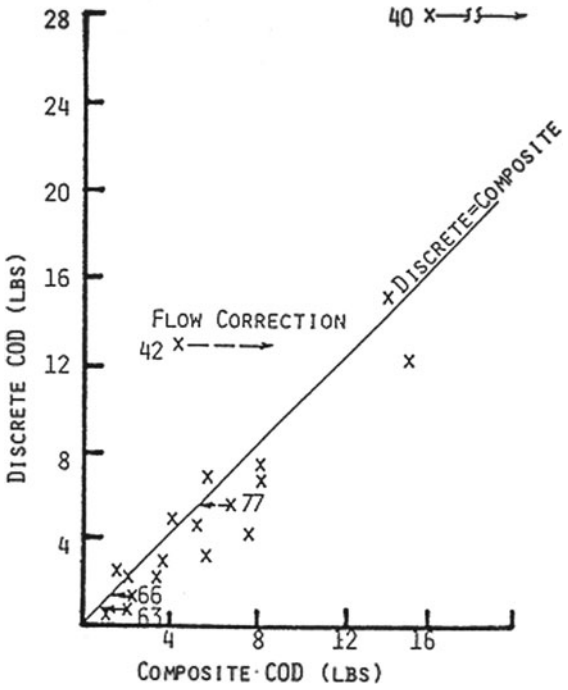

Fig. 12.3 Discrete vs. composite lead (l lb = 0.453 kg) (*Source*: Clark et al., 1981)

Table 12.4 Freeport effluent concentrations and loads (the standard deviations (SD) are computed directly from sample data as s_c; variance correction for compositing, S_w^2, and autocorrelation, τ, are not included)

	Concentrations (ppm)			Loads (ppm $\times m^3/s$)		
Parameter	Maximum	Mean	Skew	Maximum	Mean	Skew
N	Minimum	SD	Kurtosis	Minimum	SD	Kurtosis
Hourly grabs						
NH_3	18.0	12.8	0.3	4.3	2.5	0.0
167	9.4	1.8	3.1	1.2	0.8	2.0
TSS	1480.0	971.2	0.3	390.0	190.0	0.4
167	704.0	125.3	4.9	86.0	64.0	3.1
Daily time-proportioned composites						
NH_3	24.3	11.4	0.2	4.9	2.3	0.2
100	2.8	3.5	4.6	0.4	0.7	5.2
TSS	1158.0	862.3	0.1	316.5	172.1	0.2
99	658.0	102.9	2.8	81.5	40.2	3.9
Daily volume-proportioned composites						
NH_3	22.5	11.6	0.0	4.6	2.3	0.1
100	0.6	3.4	4.7	0.1	0.7	4.5
TSS	1222.0	884.8	0.0	314.6	176.6	0.1
99	674.0	116.2	2.9	81.2	41.6	3.5

Source: Schaeffer et al. (1983)

Table 12.5 St. Charles effluent concentrations and loads (standard deviations (SD) are computed directly from sample data as s_c; variance corrections for compositing, s_w^2, and autocorrelation, τ, are not included)

	Concentrations (ppm)			Loads (ppm $\times$m^3/s)		
Parameter	Maximum	Mean	Skew	Maximum	Mean	Skew
N	Minimum	SD	Kurtosis	Minimum	SD	Kurtosis
Hourly grabs						
NH_3	15.5	7.3	0.1	2.9	1.3	0.1
168	2.0	4.0	2.1	0.3	0.7	1.8
TSS-105	122.0	28.0	1.2	23.0	4.9	1.9
191	0.0	24.6	3.8	0.0	4.7	4.4
TSS-180	32.0	8.0	0.9	6.3	1.4	2.0
163	0.0	7.1	3.3	0.0	1.3	5.0
Daily time-proportioned composites						
NH_3	17.5	9.6	0.1	3.3	2.2	0.1
50	5.0	2.8	2.8	1.1	0.5	2.8
TSS-105	13.0	39.0	2.2	30.1	8.7	1.9
65						
Daily volume-proportioned composites						
NH_3	17.5	12.6	0.0	4.3	2.8	0.8
63	6.0	3.0	2.3	1.7	0.5	4.3
TSS-105	99.0	42.0	0.6	24.1	9.5	0.4
66	7.0	22.5	3.2	1.1	5.1	3.3

Source: Schaeffer et al. (1983)

hour as well as every 24 h and composite samples were taken for 24 h per day. Time proportional samples were taken with an ISCO sampler from the bottom of the flume every 15 min, and flow proportional samples from the center of the flow with a Lakeside Trembler sampler. At Freeport, samples were taken at the same stream locations and in the same temporal pattern. The flow proportional composite samples were collected using a BIF Sanitrol sampler.

Samples from both treatment plants were analyzed for total suspended solids (TSS) and ammonia (NH_3). Flows were monitored continuously at both facilities. Table 12.4 summarizes the data for Freeport, and Table 12.5 the data for St. Charles. The tables give the number of observations, the mean, standard deviation, skewness, kurtosis, minimum, and maximum. Table 12.6 summarizes certain statistical information developed from the analysis of the data.

Table 12.6 Composite pair differences – volume minus time-proportioned concentrations and loads

	Concentrations (ppm)		Loads (ppm $\times$m^3/s)	
Parameter (*N*)	Mean	*t*	Mean	*t*
Freeport NH_3 (100)	0.24	1.97	0.05	2.00
Freeport TSS (99)	22.50	3.70	4.47	3.83
St. Charles NH_3 (49)	2.99	9.42	0.65	9.00
St. Charles TSS (65)	2.54	0.90	0.71	1.08

Source: Schaeffer et al. (1983)

Chapter 13
Composite Sampling and Indoor Air Pollution

Asthma is one of the most common respiratory diseases. The occurrence of asthma is associated with atopy. Many studies have reported an association between sensitization to dust mite allergens and asthma. Dermatophagoides mites and cats produce a variety of allergens. Measurement of a specific allergen can be used to assess allergen exposure. It has been proposed that greater than 10 μg of total allergen per gram of dust should be regarded as high and represent a risk for acute attacks of asthma in a majority of mite-allergic individuals, concentrations as low as 2 μ/g should be regarded as moderate and represent a risk of sensitization, and less than 2 μg/g presents little risk for a majority of atopic individuals.

Quantification of specific allergens in dust from human dwellings provides important information for determining allergen exposure. The fact that indoor allergens are not equally distributed in the dust of human dwellings makes it difficult to estimate allergen exposure with a high degree of certainty. A composite sample may provide a more reliable estimate of indoor allergen exposure and minimize error associated with unequal distribution of allergens on discrete objects. In many applications, the use of composite samples has effectively reduced the number of measurements necessary to provide reliable estimates of contamination. Likewise, composite samples of household dust may provide useful information while minimizing the sample collection effort and analytical test costs.

13.1 Household Dust Samples

Dust samples from three specific objects and composite samples from the same three objects were collected from the living rooms and bedrooms of 15 homes by a single technician using a special filter sampling device (ALK Laboratories, Inc., Milford, CT) connected to a *Red Devil*™ vacuum (Model 503; Royal Appliances Mfg. Co., Cleveland, OH). Discrete and composite samples were collected from floor, furniture (upholstery/bed), and window coverings in both the living room and a bedroom of each home. Discrete samples were collected by vacuuming the specific objects for 10 min. Composite samples were collected in a defined sequence by vacuuming the three objects for 5 min each (see Table 13.1). In this way, the composites were

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_13, © Springer Science+Business Media, LLC 2011

Table 13.1 Sampling order used to form the composite samples

1	Living room	U, W, F
2	Living room	W, U, F
3	Living room	F, U, W
4	Bedroom	B, W, F
5	Bedroom	W, B, F
6	Bedroom	B, F, W

U=upholstery, W=window, F=floor, B=bed
Source: Lintner et al. (1992)

formed at the time of sample collection by allowing the vacuum cleaner to do the physical mixing of the dust from several objects. An alternative to this would be to collect separate dust samples from each discrete object and manually mix the dust from each of the samples in the laboratory to form a composite. The reason for forming composite samples is to estimate a patient's overall exposure to allergen in a specific indoor environment.

The results of this study seem to indicate that the actual measurement of a composite sample will be approximately the average of the values that would be obtained from separate measurements on discrete samples (see Tables 13.2 and 13.3). However, if an object has a significantly higher allergen content than other objects, the composite sample measurement tends to be higher than the average of the discrete sample measurements. Also, in order to effectively use composite sampling, only items which are likely sources of allergen should be used to form a composite sample.

Table 13.2 Comparison of mite allergen measurements from composite and discrete dust samples

Allergen	Comparison	*P* value
DER I	Bedroom vs. living room	0.12
	Living room, composite vs. discrete[a]	0.07
	Bedroom, composite vs. discrete[b]	0.55
Der p I	Bedroom vs. living room	0.01
	Living room, composite vs. discrete[a]	0.23
	Bedroom, composite vs. discrete[b]	0.53
Der f I	Bedroom vs. living room	0.57
	Living room, composite vs. discrete[a]	0.09
	Bedroom, composite vs. discrete[b]	0.33

[a] Contrast for the living room composite:
composite = floor/4 + upholstery/2 + window/4
[b] Contrast for the bedroom composite:
composite = floor/3 + bed/3 + window/3
Source: Lintner et al. (1992)

Table 13.3 Comparison of cat allergen measurements from composite and discrete dust samples (signs test)

Allergen	Comparison	*P* value
Fel d I	Bedroom, composite vs. discrete Composite = floor/3 + bed/3 + window/3	0.75
Fel d I	Living room, composite vs. discrete Composite = floor/3 + bed/3 + window/3	0.55

Source: Lintner et al. (1992)

Chapter 14
Composite Sampling and Bioaccumulation

The human body, or any living organism for that matter, when exposed to a polluted environment, accumulates contaminants in its tissue. It is, therefore, very useful to sample tissue from a sample of such organisms under investigation, in order to evaluate the amount of accumulation, called bioaccumulation, since biological processes cause the accumulation of a particular contaminant in the organism. It is a common observation that the tissue from a single member is not sufficient for making measurement. It is therefore necessary for technological reasons to composite the tissue samples extracted from several organisms so that a measurement is possible.

Compositing tissue samples extracted from several selected organisms represents an attempt to estimate the average concentration. If

$$X_1, X_2, \ldots, X_k$$

represent the contaminant concentration of k tissue samples from k individual organisms, then these samples can be pooled to obtain a single composite measurement:

$$Y = \sum_{i=1}^{k} w_i X_i,$$

where, for $i = 1, \ldots, k$, wq_i is the proportion of the contribution from the ith individual to the composite. Rohde (1976) showed that the expected value and the variance of Y are given by

$$E(Y) = \mu, \quad \text{Var}(Y) = \sigma^2/k + k\sigma_w^2\sigma^2,$$

where μ is population mean; σ^2 is population variance; σ_w^2 is variance of the compositing proportions; and k is the number of individual samples in each composite.

If the w_i's are all equal, $w_i \equiv 1/k$, then the numerical value of Y is equal to the average of the k sample values, that is, $Y = \bar{X}$. In this case, by analyzing only one composite sample, an estimate of the mean of k individual samples is obtained. However, due to compositing, the information on the individual sample variability

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4_14, © Springer Science+Business Media, LLC 2011

is lost. This is true for a single composite sample. Replicate composite samples can be used in bioaccumulation monitoring programs to obtain a more accurate estimate of the population mean and to increase the precision of this estimate.

The comparison between a single composite and replicate individual samples can be extended to replicate composite samples (see Rohde, 1976, 1979). The mean of n composite sample values $Y_1, Y_2, \ldots, Y_n$ is given by

$$\bar{Y} = \sum_{j=1}^{n} Y_j/n.$$

The expected value and the variance of $\bar{Y}$ are given by

$$E(\bar{Y}) = \mu, \ \mathrm{Var}(\bar{Y}) = \sigma^2/nk + k\sigma_w^2\sigma^2.$$

In particular, if the composite samples comprise samples of equal mass so that $w_i \equiv 1/k$ and hence $\sigma_w^2 = 0$, then

$$\mathrm{Var}(\bar{X}) = \sigma^2/k, \mathrm{Var}(\bar{Y}) = \sigma^2/nk,$$

where n is the number of replicate samples (individual or composite) used in the estimate of the population variance (σ^2) and k is the number of individual samples constituting each composite sample. In this case, it is easy to verify that

$$\frac{\mathrm{Var}(\bar{X})}{\mathrm{Var}(\bar{Y})} = n.$$

Thus, it can be seen that a collection of replicate composite tissue samples will result in a more efficient estimate of the mean. It should also be noted that for unequal proportions of composite samples, the variance of the composite sample mean increases with σ_w^2 and in extreme cases may even exceed the variance of the individual sample mean. A table of values for σ_w^2 that lead to such an increase is given by Schaeffer and Janardan (1978). Using the Dirichlet model for compositing probabilities, Rohde (1979) has shown that

$$\frac{\mathrm{Var}(\bar{X})}{\mathrm{Var}(\bar{Y})} = \frac{n+1}{2}$$

as the increase in the precision that can be achieved at an additional cost of compositing.

14.1 Example: National Human Adipose Tissue Survey

The National Human Adipose Tissue Survey (NHATS) is an annual survey to collect and analyze a sample of adipose tissue specimens from autopsied cadavers and surgical patients. The primary objectives of NHATS include the following:

- To identify chemicals that are present in the adipose tissue of individuals in the US population
- To estimate the average concentration levels of selected chemicals in adipose tissue of individuals in the U.S. population and in various demographic subpopulations
- To determine if any of the four factors (namely, geographic region, age, race, and sex) affect the average concentration levels of selected chemicals detected in the US population

Every year approximately 800–1200 adipose tissue specimens are collected using a multistage sampling plan. First, the 48 contiguous states are stratified into four geographic areas, which form four strata. Next, a sample of metropolitan statistical areas (MSAs) is selected from every stratum with probabilities proportional to MSA populations. Finally, several cooperators (hospital pathologists or medical examiners) are chosen from every selected MSA and asked to supply a specified quota of tissue specimens. The quota specifies the number of specimens needed in each of the categories defined by the donor's age, race, and sex. The categories are:

- Age groups: 0–14 years, 15–44 years, and 45+ years;
- Race: Caucasian and non-Caucasian; and
- Sex: male and female.

The sampling plans were designed to give unbiased and efficient estimates of the average concentration levels of selected chemicals in the entire population and in various subpopulations defined by the demographic variables described above. Levels are characterized by the average or median chemical concentrations; prevalence is the proportion of individuals with chemical concentrations exceeding specified criterion levels.

14.2 Results from the Analysis of 1987 NHATS Data

The analysis was performed on data obtained from 48 composite samples formed from 865 adipose tissue specimens from sampled cadavers and surgical patients. Thus, each composite contained an average of 18 specimens. Not all of the chemicals provided sufficient data to perform a meaningful analysis. Two criteria were used to determine which chemicals should be analyzed. First, a chemical must be detected in at least 50% of the composites. Second, a minimum of 30 measurements were considered necessary to achieve sufficient precision of the estimates. Thus, of the 16 chemicals, there were 9 that met both criteria for performing the analyses.

For each of the nine chemicals analyzed, Table 14.1 lists the estimated average concentration in the entire population and in the three age groups.

Table 14.1 Estimated average concentrations (pg/g) with relative standard errors (%) for selected dioxins and furans from FY87 NHATS composite samples

Compound	Entire nation	Age group 0–14	15–44	years 45+
Population percentages	100	23	46	31
Dioxins				
2,3,7,8-TCDD	5.38	1.98	4.37	9.40
	(6)	(41)	(12)	(4)
1,2,3,7,8-PECDD	10.7	3.30	9.33	18.2
	(4)	(22)	(7)	(4)
1,2,3,4,7,8/	75.1	23.4	70.9	120
1,2,3,6,7,8-HXCDD	(4)	(23)	(6)	(3)
1,2,3,7,8,9-HXCDD	11.7	6.13	10.8	17.1
	(4)	(18)	(7)	(4)
1,2,3,4,6,7,8-HPCDD	110	45.7	99.8	174
	(3)	(11)	(5)	(3)
1,2,3,4,6,7,8,9-OCDD	724	215	692	1150
	(4)	(17)	(7)	(5)
Furans				
2,3,7,8-TCDF	1.88	1.97	1.45	2.45
	(7)	(11)	(15)	(7)
2,3,4,6,7,8-PECDF	9.70	1.87	8.00	18.0
	(8)	(100)	(15)	(8)
1,2,3,6,7,8-HXCDF	5.78	1.80	4.59	10.5
	(13)	(83)	(26)	(13)

Source: Orban et al. (1990)

Glossary and Terminology

Bayesian approach: Usually an optimal statistical procedure depends on the population parameters, which are often unknown. Assuming a prior distribution of the unknown parameters, it is sometimes possible to predict a value of the parameter and hence employ the near-optimal procedure. Prior experience, local knowledge, and expert opinion usually lead to a prior probability distribution of the parameter(s) of the random variable under observation. The method which uses the prior distribution of the parameter(s) to optimize the statistical decision is known as a Bayesian approach to the concerned problem.

Binary factor: If a factor that is likely to affect the sample values has two possible levels, then it is called a binary factor. That is, a binary factor affects the sample values through its presence or absence in the samples.

Binary split retesting: After a composite sample tests positive, indicating that at least one of the constituent individual sampling units possesses the trait, the composite sample is split into two composite subsamples, as equal in size as possible, and each composite subsample is subjected to measurement. Each composite subsample that indicates the presence of the trait is similarly subjected to binary split and subsequent measurement. The procedure continues until every individual sampling unit that formed the original composite sample is classified.

Classification error: The error of misclassifying an individual sample. That is, either classifying a sampling unit possessing the trait as not possessing it or classifying a sampling unit not possessing the trait as possessing it.

Classification problem: The problem of classifying every (individual) sampling unit into one of the two possible categories, usually identified by the "presence" and "absence" of the trait, even if the measurement is not necessarily of the presence/absence type.

Cleanup evaluation: A statistical investigation to determine if a cleanup activity has been effective in that whether or not a previously hazardous site is not hazardous any more, after the cleanup activity was undertaken.

G.P. Patil et al., *Composite Sampling*, Environmental and Ecological Statistics 4,
DOI 10.1007/978-1-4419-7628-4, © Springer Science+Business Media, LLC 2011

Composite sample measurement: The measurement on the variable of interest obtained from a composite sample. Note that, as in the case with individual sampling units, a composite sample measurement need not be the same as the corresponding composite sample value unless the measurement is made without error. Also, if the composite sample is made homogeneous by mixing it thoroughly, then the composite sample measurement is expected to be a simple or weighted average of the constituent individual sample values, provided the measurement is made without error.

Composite sample size: The number of individual sampling units that are used to form a single composite sample.

Composite sample value: The value of the variable of interest for a composite sample. If the composite sample is thoroughly mixed, then the composite sample value is expected to be a simple or weighted average of the constituent individual sample values.

Composite sampling: A sampling procedure where several individual sampling units are selected and procured, but are not immediately subjected to measurement. Composite samples are formed by pooling and physically mixing a predetermined number of sampling units or subunits for making measurement.

Composite subsample: A composite sample formed from a subset of individual samples that constituted a composite sample.

Conjugate prior distribution: If the posterior distribution of parameters belongs to the same family of distributions as their prior distribution, then the prior distribution is called a conjugate prior distribution.

Continuous measurement: A measurement that gives the numerical value of the variable of interest is called a continuous measurement.

Covariogram: Consider the spatial process $\{Z(s), s \in D\}$, where $D \subset R^d$. Suppose

$$\operatorname{cov}(Z(s_1), (s_2)) = C(s_1 - s_2)$$

depends only on the difference $s_1 - s_2$ for all $s_1, s_2 \in D$. The function $C(\cdot)$ is called a covariogram or a stationary covariance function.

Curtailed retesting: When a composite sample of size k has indicated the presence of the trait, some form of retesting is employed. If the first $k-1$ of the constituent individual sampling units indicate the absence of the trait, then the kth individual sampling units is classified as possessing the trait without actually making a measurement on this sampling unit.

Data quality objectives (DQO) Process: A statistical procedure to ensure that the data collection will be most effective in the sense of collecting maximal information at a minimal cost.

Entropy-based retesting: The procedure assumes a large collection of unclassified individual sampling units. The procedure begins by forming a composite sample of a predetermined size k. If the composite sample tests negative, then all its constituent individual samples are classified as not possessing the trait. However, if this composite sample tests positive, indicating that at least one constituent individual sample possesses the trait, then the composite sample is split into two composite subsamples as equal in size as possible, and subjected to measurement. If the first of the two composite subsamples indicates the absence of the trait, then the second is assumed to posses the trait and is, therefore, subjected to further binary split. On the other hand, if the first composite subsample indicates the presence of the trait, then the individual sampling units that form the other composite subsample are not classified and are returned to the pool of unclassified individual sampling units. Continuing in this way, each of the k individual sampling units used to form the composite sample is either classified as not possessing the trait or is returned to the pool of unclassified individual sampling units, except for exactly one individual sampling unit that is classified as possessing the trait. At this stage, another composite sample of size k is formed from the pool of unclassified individual sampling units. The procedure continues until all the individual sampling units are classified. Although this classification procedure is not hierarchical like the other classification procedures, it is optimal in that it maximizes the entropy.

Equal and unequal allocations in ranked set sampling: In the ranked set sampling protocol, the total sample size can be allocated to different ranks in several ways. If all the ranks are selected with equal frequency, we call it an equal allocation; otherwise, there is an unequal allocation.

Exhaustive retesting: This procedure begins by forming a composite sample of a predetermined size k. If the composite sample tests negative, then all the k constituent individual sampling units are classified as not possessing the trait. On the other hand, if the composite sample tests positive, indicating that at least one of the constituent individual sampling units possesses the trait, then every individual sampling units is separately subjected to measurement and is classified accordingly.

Identification of sample maximum: The procedure that identifies the individual sampling unit having the largest measurement. This procedure identifies the sample maximum with certainty, but the total number of measurements required to do so is not fixed.

Individual sample measurement: The measurement on the variable of interest obtained from an individual sampling unit. Note that an individual sample measurement is different from the corresponding individual sample value unless the measurement is made without error. Also note that, while every individual sample has a value for the variable of interest, every individual sample may not provide a measurement, since some individual sampling units are not necessarily subjected to measurement.

Individual sample value: The value of the variable of interest for an individual sampling unit. Note that every individual sampling unit has a fixed value for the variable of interest, even though only a few selected individual sampling units are subjected to measurement.

Kriging: Kriging is a minimum-mean-squared-error method of spatial prediction that (usually) depends on the second-order properties of the spatial process under study. Matheron (1963) named this method after D. G. Krige, a South African mining engineer who developed empirical methods for determining true ore-grade distributions from distributions based on sampled ore grades (Krige, 1951).

Linear model for composited data: A linear model is used to express the relationship between individual sample values and composite sample values.

Nugget effect: Suppose $\gamma(h)$ is the semivariogram for a spatial process $\{Z(s), s \in D\}$. That is, $\gamma(h) = E[Z(s+h) - Z(s)]^2$. It is then easy to note that $\gamma(0) = 0$. If $\gamma(h) \to c_0 \neq 0$ as $h \to 0$, then c_0 is called the nugget effect.

Optimal composite design: A design for forming composite samples from individual sampling units in order to maximize the efficiency of the inference drawn from the composite sample data.

Posterior distribution: The conditional probability distribution of population parameters, given the observed value(s) of the random variable(s), is called the posterior distribution of population parameters. The posterior distribution is derived from the prior distribution of the parameters and the observed value(s) of the random variable(s).

Prediction of sample maximum: The procedure that predicts the largest individual sample value. This procedure has a fixed number of measurements, but may fail to identify the largest individual sample value with a positive probability. That is, there is a positive, though usually small, probability that the predicted sample maximum is not the actual sample maximum in that there is some individual sampling unit having a measurement larger than the predicted sample maximum.

Presence/absence measurement: A measurement that indicates the presence or absence of the trait under study.

Prevalence: The proportion of (individual) sampling units that possess the trait under study. Note that the prevalence is the true proportion in the population and will differ from the observed proportion in any particular case.

Prior distribution: The belief, usually based on some prior information, local knowledge, and expert opinion, about the possible variation in the values of population parameters is sometimes expressed in terms of a probability distribution of these parameters. Such a postulated probability distribution of population parameters is called their prior distribution.

Random weights: The composite sample value is a simple or weighted average of the constituent individual sample values. If the proportions of individual

sampling units or subunits that are used to form a composite sample are not fixed, then these proportions are treated as random variables. In this case, the composite sample value is weighted average of the constituent individual sample values with random weights.

Ranked set sampling: A method of sampling, where large samples are initially selected for judgmentally ranking their members without involving costly laboratory procedures and are followed by subsequent quantification of a few individual sampling units with selected ranks.

Retesting: After obtaining measurement on a composite sample, some or all of the constituent individual sampling units may be subjected to measurement, either individually or in the form of composite subsamples. This stage of measuring some or all of the individual sampling units that have already been subjected to measurement as part of a composite sample is called retesting.

Semivariogram: See **Variogram**.

Sequential retesting: This procedure begins by forming a composite sample of a predetermined size k. If the composite sample tests negative, then every constituent individual sampling unit is classified as not possessing the trait. On the other hand, if the composite sample tests positive, indicating that at least one of the constituent individual sampling units possesses the trait, then the individual sampling units are sequentially subjected to measurement until an individual sampling unit tests positive. At this stage, all the unclassified individual sampling units are pooled into a single composite subsample, which is then measured for the trait. If the trait is present in the composite subsample, then the same procedure is repeated, until all the individual sampling units that formed the original composite sample are classified.

Sill: Let $C(\cdot)$ be the covariogram of the spatial process $\{Z(s), s \in D\}$ (see **Covariogram** and **Variogram**). It is easy to establish that the semivariogram function satisfies $\gamma(h) = C(0) - C(h)$. If $C(h) \to 0$ as $\| h \| \to \infty$, then $\gamma(h) \to C(0)$. The limit $C(0)$ is then called the sill of the semivariogram.

Site characterization: Characterization of a (waste) site as hazardous or not hazardous.

Spatial autocorrelation: Any dependence, as measured by a correlation coefficient, among sampling units that form a sequence of points on the sampling site in a specific direction is called the spatial autocorrelation for the corresponding spatial process. Note that the spatial autocorrelation usually depends on both the direction and the spatial lag.

Spatial contiguity: When the locations of certain sampling units form a contiguous set on the sampling site, we call these sampling units spatially contiguous. Spatial contiguity usually ensures that the values of the concerned sampling units are close to each other.

Spatial structures: Any structure in the values of the variable of interest on the sampling site as depending on the locations of sampling units.

Subsampling a composite sample: It is sometimes desired to investigate the homogeneity of a composite sample. In such a case, a subsample is extracted from the composite sample for making measurement. This procedure is called subsampling of the composite sample.

Sweep-out method: A method used to identify the individual sampling unit having the largest measurement. In this procedure, any individual sampling unit that is not likely to have the largest measurement is eliminated from the potential search so as to avoid unnecessary measurements.

Variogram: Consider the spatial process $\{Z(s), s \in D\}$, where $D \subset R^d$. Suppose

$$\text{Var}(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2)$$

depends only on the difference $s_1 - s_2$ for all $s_1, s_2 \in D$. The function $2\gamma(\cdot)$, which is a function only of the difference $s_1 - s_2$, is called a variogram and $\gamma(\cdot)$ is called a semivariogram of the spatial process $\{Z(s), s \in D\}$.

Bibliography

Ahmed A. S., Webster L., Pollard, Pat., Davies, I. M., and Moffat, C. F. (2006). Description & evaluation of a sampling system for monitoring Hydrocarbons in sediments Fisheries Research Services Internal report No. 09/06, Fisheries Research Services Marine Laboratory, Aberdeen.

Ahn, C. Y., Joung S. H., Park, C. S., Kim, H. S., Yoon, B. D., and Oh, H. M. (2008). Comparison of sampling and analytical methods for monitoring of cyanobacteria-dominated surface waters. Hydrobiologia 596:413–421.

Aigner, M. and Schughart, M. (1985). Determining defectives in a linear order. J. Stat. Plan. Infer. 12:359–368.

Alam, M. J., Renter, D., Taylor, E.,Mina, D., Moxley, R., and Smith, D. (2009). Antimicrobial susceptibility profiles of salmonella enterica serotypes recovered from pens of commercial feedlot cattle using different types of composite samples. Curr. Microbiol. 58:354–359.

Anderson T. H. and Taylor, G. (2001). Nutrient pulses, plankton blooms, and seasonal hypoxia in Western long island sound. Estuaries 24(2):228–243.

Anscombe, F. J. (1956). On estimating binomial response relations. Biometrika 43:461–464.

Armstrong, M. (1984). Problems with universal kriging. Math. Geol. 16:101–108.

Arnold, S. F. (1977). Generalized group testing procedures. Ann. Stat. 5:1170–1182.

Back, P. E. (2007). A model for estimating the value of sampling programs and the optimal number of samples for contaminated soil. Environ. Geol. 52:573–585.

Baker, A. S., Kuo, S., and Chae, Y. M. (1981). Comparisons of arithmetic average soil pH values with the pH values of composite samples. Soil Sci. Soc. Am. J. 45:828–830.

Baldock, F. C., Lyndal-Murphy, M., and Pearse, B. (1990). An assessment of a composite sampling method for counting strongyle eggs in sheep feces. Aust. Vet. J. 67:165–167.

Barry D. P. K., Dillon, P. P. P., and Chassagne A. (2008). Preliminary risk assessment for the Salisbury storm water ASTR project. CSIRO: water for a healthy country National Research Flagship.

Bartlett, M. S. (1935). Mathematical appendix to a paper, by G. E. Blackman, entitled "A study by statistical methods of the distribution of species in grassland associations." Ann. Bot. 49:749–777.

Beaufort County Public Works (2009).Year 2008–2009 report, Beaufort County storm water monitoring, Beaufort County, South Carolina.

Becker, H. B. (1977). Composite sampling in aquatic environments. Plankton Committee, CM 1977/L:5, International Council for the Exploration of the Sea.

Bell, R. M. and Ellickson, P. L. (1989). Does pooling saliva for cotinine testing save money without losing information? J. Behav. Med. 12:503–507.

Bennett, R. A. (1971). A comparison of preservatives for composite samples of milk for protein testing. Technical note. Aust. J. Dairy Technol. 26:37.

Berger, T. B., Mehravari, N., Towsley, D., and Wolf, J. (1984). Random multiple-access communication and group testing. IEEE Commun. 32:769–779.

Berner, T. O. (1987). Sample design for the fiscal year 1988. Draft Final Report, Contract 68-02-4243, Office of Toxic Substances.

Beyene, A., Legesse, W., Triest, L., and Kloos, H. (2009). Urban impact on ecological integrity of nearby rivers in developing countries: the Borkena River in highland Ethiopia. Environ. Monit. Assess. 153:461–476.

Bhattacharyya, G. K., Karandinos, M. G., and DeFoliart, G. R. (1979). Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies. Am. J. Epidemiol. 109:124–131.

Bicking, C. A. (1967). The sampling of bulk materials. Mater. Res. Stand. 7:95–116.

Birch, G. F., Fazeli, M. S., and Matthai, C. (2005). Efficiency of an infiltration basin in removing contaminants from urban stormwater. Environ. Monit. Assess. 101:23–38.

Blackman, G. E. (1935). A study by statistical methods of the distribution of species in grassland associations. Ann. Bot. 49:749–777.

Blomqvist, P. (2001). A proposed standard method for composite sampling of water chemistry and plankton in small lakes. Environ. Ecol. Stat. 8:121–134.

Bohrnstedt, G. W. and Goldberger, A. S. (1969). On the exact covariance of products of random variables. J. Am. Stat. Assoc. 64:1439–1443.

Bolgiano, N. C., Boswell, M. T., Patil, G. P., and Taillie, C. (1989). Evaluation of the kriging model for abundance estimation of marine organisms. Technical Report Number 89–0601. Center fpr Statistical Ecology and Environmental Statistics, Pennsylvania State University, University Park, PA, 16802.

Bolgiano, N. C., Boswell, M. T., Patil, G. P., and Taillie, C. (1989). Task 4: Report on evaluation of selected statistical methods with potential for addressing superfund site characterization problems. Final Interim Report to SRA Technologies, SRA Technologies/EPA Prime Subcontract to Penn State, Subcontract No. 40400–5–01, October 31, 1989.

Bolgiano, N. C., Patil, G. P., and Taillie, C. (1990). Spatial statistics, composite sampling, and related issues in site characterization with two examples. In *Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs*, H. Lacayo, R. J. Nadeau, G. P. Patil, and L. Zaragoza, eds. USEPA Statistical Policy Branch, Washington DC. pp. 79–114.

Bolla, M. (1991). Relations between spectral and classification properties of multigraphs. DIMACS Technical Report 91-27, DIMACS Center, Rutgers University, Piscataway, NJ.

Boomer, B. A., Erickson, M. D., Swanson, S. E., Kelso, G. L., Cox, D. C., and Schultz, B. D. (1985). Verification of PCB spill cleanup by sampling and analysis. EPA Contract No. 68-02-6938, TC 8501-A37.

Boswell, M. T. (1978). Composite sampling and its application to the estimation of plankton density—A basic review of concepts and methods. In Studies of Distribution Functions of Fish and Shellfish in the Marine Environment. Final Report: Contract NMFS 03-7-043-35116.65–75.

Boswell, M. T. and Patil, G. P. (1987). A perspective of composite sampling. Commun. Stat. Theor. Methods 16:3069–3093.

Boswell, M. T. and Patil, G.P. (1990). Composite sampling using spatial autocorrelation for Palmerton hazardous waste site: A preliminary report. In *Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs*, H. Lacayo, R. J. Nadeau, G. P. Patil, and L. Zaragoza, eds. USEPA Statistical Policy Branch, Washington DC. pp. 20–42.

Boswell, M. T., Burnham, K. P., and Patil, G. P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies. In *Handbook of Statistics*, Vol. 6, P. R. Krishnaiah and C. R. Rao, eds. Elsevier, New York, NY. pp. 469–488.

Brown, G. H. and Fisher, N. I. (1972). Subsampling a mixture of sampled material. Technometrics 14:663–668.

Brown, G. H. and Robson, D. S. (1975). The estimation of mixing proportions by double sample using bulk measurement. Technometrics 17:119–126.

Brown, K. W., Flatman, G. T., Englund, E. J., Shoener, E., Starks, T. H., Rohde, S. C., Schnell, M. H., Fisher, N. J., Sparks, A. R., and Gruber, D. K. (1989). Documentation of EMSL-LV contribution to Palmerton, PA, Superfund remedial investigation U.S. EPA, Washington, DC.

Brown, K. W., Mullins, J. W., Richitt, E. P., Flatman, G. T., Black, S. C., and Simon, S. J. (1985). Assessing soil lead contamination in Dallas, Texas. Environ. Monit. Assess. 5:137–154.

Bruchet, A., Cognet, L., and Mallevialle, J. (1984). Continuous composite sampling and analysis of pesticides in water. Water Res. 18:1401–1409.

Brumelle, S., Nemetz, P., and Casey, D. (1984). Estimating the means and variances: Comparative efficiencies of composite and grab samples. Environ. Monit. Assess. 4:81–84.

Burrows, P. M. (1987). Improved estimation of pathogen transmission rates by group testing. Phytopathology 77:363–365.

Bush, K. A., Federer, W. T., Pesotan, H., and Raghavarao, D. (1984). New combinatorial designs and their applications to group testing. J. Stat. Plan. Infer 10:335–343.

Callaghan, M. O., Gerard, E. M., Waipara, N. W., Young, S. D., Glare, T. R., Barrell, P. J., and Conner, A. J. (2004). Microbial communities of solanum tuberosum and magainin-producing transgenic lines. Plant Soil 266:47–56.

Cameron, D. R., Nyborg, M., Toogood, J. A., and Laverty, D. H. (1971). Accuracy of field sampling for soil tests. Can. J. Soil Sci. 51:165–175.

Cameron, J. M. (1951). The use of components of variance in preparing schedules for sampling of baled wool. Biometrics 7:83–96.

Carra, J. S. (1984). Lead levels in blood of children around smelter sites in Dallas. In *Environmental Sampling for Hazardous Wastes*, G. E. Schweitzer and J. A. Santolucito, eds. American Chemical Society, Washington, DC. pp. 53–66.

Carey, J. M. and Keough, M. J. (2002). Compositing and subsampling to reduce costs and improve power in benthic infaunal monitoring programs. Estuaries 25(5):1053–1061.

Carson, J. Jr. (2001). Analysis of composite sampling data using the principle of maximum entropy. Environ. Ecol. Stat. 8:201–211.

Carter, R. E. and Lowe, L. E. (1986). Lateral variability of forest floor properties under second-growth Douglas-fir stands and the usefulness of composite sampling techniques. Can. J. For. Res. 16:1128–1132.

Casey, D. B. (1982). Measuring Sample Maximums: An Application to Water Quality Monitoring. M.S. Thesis, Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, Canada.

Casey, D. B., Nemetz, P. N., and Uyeno, D. (1985). Efficient search procedures for extreme pollutant values. Environ. Monit. Assess. 5:165–176.

Cassie, R. M. (1963). Microdistribution of plankton. Oceanogr. Mar. Biol. Ann. Rev. 1:223–252.

Cassie, R. M. (1971). Sampling and statistics. In *A Manual on Methods for the Assessment of Secondary Productivity in Fresh Waters*, W. T. Edmondson and G. G. Winberg, eds. Blackwell, Oxford.

Chang, G. J., Hwang, F. K., and Lin, S. (1982). Group testing with two defectives. Disc. Appl. Math. 4:97–102.

Chen, C. C. and Hwang, F. K. (1989). Detecting and locating electrical shorts using group testing. IEEE Trans. Circuits Syst. 36:1113–1116.

Chiang, C. L. and Reeves, W. C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. Am. J. Hygiene 75:377–391.

Chung, C. F. and Garrett, R. G. (1984). Note on optimal composite sample size selection. In Current Research, Part B, Geological Survey of Canada, Paper 84-1B, 351–354.

Clark, D. L., Asplund, R., Ferguson, J., and Mar, B. W. (1981). Composite sampling of highway runoff. J. Environ. Eng. 107:1067–1081.

Cline, S. M. and Severin, B. F. (1989). Volatile organic losses from a composite water sampler. Water Res. 23:407–412.

Coffey, R. D., Parker, G. R., Laurent, K. M., and Overhults, D. G. (2000). Sampling animal manure. Co – operative extension service, college of Agriculture, University of Kentucky.

Connolly, J. and O'Connor, F. (1981). Comparison of random and composite sampling methods for the estimation of fat content of bulk milk supplies. Irish J. Agric. Res. 20:35–51.

Connolly, J. and O'Connor, F. (1982). Examination of some factors affecting the accuracy of bulk-milk chemical composition using composite and random sampling methods. Irish J. Agric. Res. 21:19–26.

Correll, R. L. (2001). The use of composite sampling in contaminated sites – a case study. Environ. Ecol. Stat. 8:185–200.

Courtin, P., Feller, M. C., and Klinka, K. (1983). Lateral variability in some properties of disturbed forest soils in southwestern British Columbia. Can. J. Soil Sci. 63:529–539.

Cressie, N. (1988). Spatial prediction and ordinary kriging. Math. Geol. 20:405–421.

Cressie, N. (1989). Geostatistics. American Statistician 43:197–202.

Cressie, N. (1991). *Statistics for spatial data*. Wiley, New York, NY.

Curiale, M. S. (2000). Validation of the use of composite sampling for Listeria monocytogenes in ready– to-eat meat & Poultry products. American means Institute Foundation.

Daniel, T. C., Wendt, R. C., McGuire, P. E., and Stoffel, D. (1979). A comparison of composite sampling techniques for monitoring runoff. Trans. ASAE 22:1310–1312.

Davis, C. E., Grizzle, J. E., and Bryan, J. A. (1973). Estimation of the probability of post transfusion hepatitis in hemophilia treatment. Biometrics 29:386–392.

Degasperi, C. (2004). Major lakes phytoplankton study: comparison of composite sampling techniques. Department of Natural Resources & Parks, Water & Land Resources Division, King County.

Dell, T. R. (1969). The theory and some applications of ranked set sampling. Ph.D. Thesis, Department of Statistics, University of Georgia, Athens, GA.

Dell, T. R. and Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. Biometrics 28:545–553.

Dierckx, P. (1981). An algorithm for surface-fitting with spline functions. IMA J. Numer. Anal. 1:267–283.

Dores, E. F. G. C., Spadotto, C. A., Weber, O. L. S., Carbo, L., Vecchiato, A. B., and Pinto, A. A. (2009). Environmental behaviour of metolachlor and diuron in a tropical soil in the central region of Brazil. Water Air Soil Pollut. 197:175–183.

Dorfman, R. (1943). The detection of defective members of large populations. Ann. Math. Stat. 14:436–440.

Doug, C., Fowler, B., Wright, R. H., Mao, L., and Fields, J. (2008). Composite sampling & Analysis of low concentration organic compounds & Trace Elements in surface waters at the Tracy fish collection facility, Tracy California. Vol. 41. US Department of interior, Bureau of reclamation, mid-pacific region, technical service center.

Douglas, R. W., Menary, W., and Jordan, P. (2007). Phosphorus and sediment transfers in a grassland river catchment. Nutr. Cycl. Agroecosyst. 77:199–212.

Dow, R. P., Coleman, P. H., Meadows, K. E., and Work, T. H. (1964). Isolation of St. Louis encephalitis viruses from mosquitoes in the Tampa Bay area of Florida during the epidemic of 1962. Am. J. Trop. Med. Hyg. 13:462–468.

Drechsler, H. D. and Nemetz, P. N. (1978). The impact of composite sampling and other data aggregation procedures on pollution detection in the pulp and paper industry. Can. J. For. Res. 8:328–340.

Du, D. Z. and Ko, K. I. (1987). Some completeness results on decision trees and group testing. SIAM J. Alg. Disc. Meth. 7:159–166.

Duncan, A. J. (1962). Bulk sampling: Problems and lines of attack. Technometrics 4:319–344.

Edland, S. D. and van Belle, G. (1994). Decreased sampling costs and improved accuracy with composite sampling In *Environmental Statistics, Assessment, and Forecasting*, C. R. Cothern and N. P. Ross, eds. Lewis, Boca Raton, FL.

Einax, J. W. and Kraft, J. (2002). Small_scale Variability of metals in soil and composite sampling. ESPR-Environ. Sci. Pollut. Res. 9(4):257–261.

El-Baz, A. and Nayak, T. (2004). Efficiency of composite sampling for estimating a lognormal distribution. Environ. Ecol. Stat. 11:283–294.

Elder, R. S. (1977). Properties of composite sampling procedures. Ph.D. Dissertation. Virginia Polytechnic Institute and State University, Blacksburg, VA.

Elder, R. S., Thompson, W. O., and Myers, R. H. (1980). Properties of composite sampling procedures. Technometrics 22:179–186.

Englund, E. J. and Flatman, G. T. (1986). Annual report of research in environmental statistics, geostatistics, and chemometrics. MS.

Environment Agency (1999). Monitoring of Radio active releases to water from Nuclear facilities: Technical guidance note Miz. National compliance assessment service, Lancaster.

EPA QAMS (1991). Data quality objective process. In *Orientation to Quality Assurance Management*, U.S. Environmental Protection Agency, Quality Assurance Management Staff (RD-680), 401 M Street, SW, Washington, DC 20460.

Ersoy, A., Yunsel, T. Y. and Cetin, M. (2004). Characterization of land contaminated by past heavy metal mining using geostatistical methods. Arch. Environ. Contam. Toxicol. 46:162–175.

Fabius, J. (1973). Two characterizations of the dirichlet distribution. Ann. Stat. 1:583–587.

Federer, W. T. (1984). Cutting edges in biometry. Biometrics 40:827–839.

Federer, W. T. (1987). On screening samples in the laboratory and factors in factorial investigations. Commun. Stat. Part A Theor. Methods 16:3033–3049.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications*. 2nd Edition. Wiley, New York, NY.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. 3rd Edition. Wiley, New York, NY.

Fine, P. E. M. (1975). Quantitative studies on the transmission of Parahistomonas wenrichi by ova of Heterakis gallinarum. Parasitology 70:407–415.

Fine, P. E. M. and Sylvester, E. S. (1978). Calculation of vertical transmission rates of infection, illustrated with data on an aphid-borne virus. Am. Nat. 112:781–786.

Finucan, A. M. (1964). The blood testing problem. Appl. Stat. 13:43–50.

Fish, D. J. (2006). Practical consideration of gas sampling & gas sampling systems & standards Canadian gas association.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. Phil. Trans. (A) 222:309–368.

Fisher, R. A. and Yates, F. (1963). *Statistical Tables for Biological, Agricultural and Medical Research*. Hafner, New York, NY.

Flatman, G. T. (1984). Using geostatistics in assessing lead contamination near smelters. In *Environmental Sampling for Hazardous Wastes*, G. E. Schweitzer and J. A. Santolucito, eds. American Chemical Society, Washington, DC.

Flatman, G. T., Englund, E. J., and Yfantis, A. A. (1988). Geostatistical approaches to the design of sampling regimes. In *Principles of Environmental Sampling*, L. H. Keith, ed. American Chemical Society. Washington, DC. pp. 73–83.

Fonseca, J. P. C. Da (1991). Ecological diversity and ecological systems complexity: local or global approach? Rev. Ecol. Biol. Sol. 28:51–66.

Garcia, R. (1965). Collection of Dermacentor Andersoni (stiles) with carbon dioxide and its application in studies of Colorado tick fever virus. Am. J. Trop. Med. Hyg. 14:1090–1093.

Garey, M. R. and Hwang, F. K. (1974). Isolating a single defective using group testing. J. Am. Stat. Assoc. 69:151–153.

Garg, N. K. and Mohan, S. (1987). Group testing protocol with capture for random access communication. IEEE Commun. 35:849–854.

Garner, F. C., Stapanian, M. A., and Williams, L. R. (1988). Composite sampling for environmental monitoring. In *Principles of Environmental Sampling*, L. H. Keith, ed. American Chemical Society, Washington, DC. pp. 363–374.

Garner, F. C., Stapanian, M. A., Yfantis, E. A., and Williams, L. R. (1990). Probability estimation with sample compositing techniques. MS.

Garrett, R. G. and Sinding-Larsen, R. (1984). Optimal composite sample size selection, applications in geochemistry and remote sensing. J. Geochem. Explor. 21:421–435.

Garry Struthers Associates Inc. (1997). Sampling & Analysis Plan: Field sampling plan-part B. Wenatchee Tree Fruit Research Center (TFREC) test plot remediation, Wenatchee, Washington, DC.

Gastwirth, J. L. and Hammick P. A. (1989). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: Application to estimating the prevalence of AIDS antibodies in blood donors. J. Stat. Plan Infer. 22:15–27.

Gensheimer, G. J., Tucker, W. A., and Denahan, S. A. (1986). Cost-effective soil sampling strategies to determine amount of soils requiring remediation. In *Proceedings of the National Conference on Hazardous Wastes and Hazardous Materials*, March 4–6, 1986, Atlanta, GA. pp. 76–79.

Gibbs, A. J. and Gower, J. C. (1960). The use of a multiple-transfer method in plant virus transmission studies—some statistical points arising in the analysis of results. Ann. Appl. Biol. 48:75–83.

Gilbert, R. O. (1987). *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York, NY.

Gill, A. and Gottlieb, D. (1974). The identification of a set by successive intersections. Info. Control 24:20–35.

Gilstein, C. Z. (1985). Optimal partitions of finite populations for Dorfman-type testing. J. Stat. Plan. Infer. 12:385–394.

Göransson, E., Johnson, R. K., and Wilander, A. (2004). Representativity of a mid-lake surface water chemistry sample. Environ. Monit. Assess. 95:221–238.

Gore, S. D. and Patil, G. P. (1993). Identifying extremely large values using composite sample data. J. Environ. Stat. (to appear). Environ. Ecol. Stat. 1, 227–245

Gore, S. D., Patil, G. P., and Taillie, C. (1992). Studies on the applications of composite sample techniques in hazardous waste site characterization and evaluation: II. Onsite surface soil sampling for PCB at the Armagh Site. 92–0305.

Gore, S. D., Patil, G. P., and Taillie, C. (2001), Identifying the largest individual sample value from a two-way composite sampling design. Environ. Ecol. Stat. 8:151–162.

Gore, S. D., Patil, G. P., Sinha, A. K., and Taillie, C. (1993). *Certain Multivariate Considerations in Ranked Set Sampling and Composite Sampling Designs. In Multivariate Environmental Statistics*, G. P. Patil and C. R. Rao, eds. Elsevier, New York, NY. pp. 121–148.

Graff, L. E. and Roeloffe, R. (1972). Group testing in the presence of test error: An extension of the Dorfman procedure. Technometrics 14:113–122.

Graff, L. E. and Roeloffe, R. (1974). A group-testing procedure in the presence of test error. J. Am. Stat. Assoc. 69:159–163.

Graham, A. (1981). *Kronecker Products and Matrix Calculus: with Applications*. Ellis Horwood, Chichester.

Griffiths, D. A. (1972). A further note on the probability of disease transmission. Biometrics 28:1133–1139.

Hach Company (2004). Sigma 900 standard portable sampler: instrument manual. Catalog number 8991.

Hagstrom, L. and Stapleton, M. (2005). Pitfalls of composite sampling for contaminated land investigations. Annual conference of waste MINZ, New Zealand.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. Ann. Hum. Genet. 20:309–311.

Halls, L. K. and Dell, T. R. (1966). Trial of ranked set sampling for forage yields. J. For. Sci. 12:22–26.

Hannon, P. and Roy, W. D. (2003). Resource report Minetech International Limited.

Hao, F. H. (1988). The optimal procedures for quantitative group testing. Discrete Appl. Math. 26:79–86.

Hardwick, J., Page, C., and Stout, Q. F. (1998). Sequentially deciding between two experiments for estimating a common success probability. J. Am. Stat. Assoc. 93:1502–1511.

Harmel, R. D., King, K. W., and Slade, R. M. (2003). Automated storm water sampling on small watersheds. Am. Soc. Agric. Eng. 19(6):667–674.

Harris, D. J. and Keffer, W. J. (1974). Waste water sampling methodologies and flow measurement techniques. Technical Report 907/9-74-005, U.S. Environmental Protection Agency.

Hathaway II, J. E. (2005). Determining the optimum number of increments in composite sampling. A project submitted to the faculty of Brigham Young University in Partial Fulfillment of the Requirements for the Degree of Master of Science.

Hathaway, J. E., Schaalje, G. B., Gilbert, R. O., Pulsipher, B. A., and Matzke, B. D. (2008). Determining the optimum number of increments in composite sampling. Environ. Ecol. Stat. 15:313–327.

Hayes, J. H. (1978). An adaptive technique for local distribution. IEEE Trans. Commun. 26:1178–1186.

Heart, R. G. Horner, R., Jones, J., Josselyn, M., Pitt, R., and Stenstrom, M. K. (2008). Sample collection methods for runoff characterization at Santa Susana Field Laboratory. SSFL CDO expert panel report.

Heisig-Mitchell, J. Barker, D. L., and Le Blauc, N. E. (2004). Application of Automated systems for clean composite sampling, 4th National Monitoring conference, National water quality monitoring council, Chattanooga, TN, May 17–20, 2004.

Heyman, U., Ekbohm, G., Blomqvist, P., and Grundstrom, R. (1982). The precision of abundance estimates of plankton from composite samples. Water Res. 16:1367–1370.

Hoenig, J. M. (1984). On group testing for diseases. Can. J. Fish. Aquat. Sci. 41:1269.

Hoenig, J. M., Heisey, D. M., Lawing, W. D., and Schupp, D. H. (1987). An indirect rapid methods approach to assessment. Can. J. Fish. Aquat. Sci. 44:324–338.

Holloway, T. T., Lawrence, K. D., Joslin, J. E., and Crisp, N. H., (1991). Looking in the dark for black cats that aren't there, with a light bright enough to find them if they were. EPA Stat. 9:1–3.

Houlahan, J. E. and Findlay, C. S. (2004). Estimating the 'critical' distance at which adjacent land-use degrades wetland water and sediment quality. Landsc. Ecol. 19:677–690.

Hrbacek, J. (1971). Special sampling systems. In *A Manual on Methods for the Assessment of Secondary Productivity in Fresh Waters*. W. T. Edmondson, ed. Blackwell, Oxford. pp. 15–17.

Hu, M. C., Hwang, F. K., and Wang, J. K. (1981). A boundary problem for group testing. SIAM J. Alg. Dis. Meth. 2:81–87.

Hueck, H. J. (1976). Active surveillance and use of bioindicators. In *Principles and Methods for Determining Ecological Criteria on Hydrobiocenoses*. Pergamon Press, New York, NY. pp. 275–286.

Huffman, D. A. (1952). A method for the construction of minimum redundancy codes. Proc. Inst. Radio Eng. 40:1098–1101.

Hui, C. A., Takekawa, J. Y., and Warnock, S. E. (2001). Contaminant profiles of two species of shorebirds foraging together at two neighboring sites in South San Francisco bay, California. Environ. Monit. Assess. 71:107–121.

Huibregtse, K. R. and Moser, J. H. (1976). Handbook for sampling and sample preservation of water and wastewater. Technical Report Number 600/4-76-049, U.S. Environmental Protection Agency.

Hull, M. S., Cherry, D. S., and Neves. R. J. (2006). Use of bivalve metrics to quantify influences of coal-related activities in the Clinch River watershed, Virginia. Hydrobiologia 556:341–355.

Hwang, F. K. (1972). A method for detecting all defective members in a population by group testing. J. Am. Stat. Assoc. 67:605–608.

Hwang, F. K. (1975). A generalized binomial group testing problem. J. Am. Stat. Assoc. 70:923–926.

Hwang, F. K. (1976a). Group testing with a dilution effect. Biometrika 63:671–673.
Hwang, F. K. (1976b). An optimum nested procedure in binomial group testing. Biometrics 32:939–943.
Hwang, F. K. (1978). A note on hypergeometric group testing procedures. SIAM J. Appl. Math. 34:371–375.
Hwang, F. K. (1980). An explicit expression for the cost of a class of Huffman trees. Disc. Math. 32:163–165.
Hwang, F. K. (1984). Robust group testing. J. Qual. Tech. 16:189–195.
Hwang, F. K. and Xu, Y. H. (1987). Group testing to identify one defective and one mediocre item. J. Stat. Plan. Infer. 17:367–373.
Hwang, F. K., Lin, S., and Mallows, C. L. (1979). Some realizability theorems in group testing. SIAM J. Appl. Math. 37:396–400.
Hwang, F. K., Pfeifer, C. G., and Enis, P. (1981). An optimal hierarchical procedure for a modified binomial group-testing problem. J. Am. Stat. Assoc. 76:947–949.
Hwang, F. K., Song, T. T., and Du, D. Z. (1981). Hypergeometric and generalized hypergeometric group testing procedures. SIAM J. Alg. Disc. Meth. 2:426–428.
Isaaks, E. H. (1984). Risk Qualified Mappings for Hazardous Waste Sites: A Case Study in Distribution Free Geostatistics. M.S. Thesis, Department of Applied Earth Sciences, Stanford University.
Isaaks, E. H. and Srivastava, R. M. (1988). Spatial continuity measures for probabilistic and deterministic geostatistics. Math. Geol. 20:313–341.
Izenman, A.-J. (2001). Statistical & legal aspects of the forensic study of Illicit drugs. Stat. Sci. 16(1):35–57.
Izenman, A. J. (2003). Sentencing Illicit drug traffickers: how do the courts handle random sampling issues ? fifth International conference on Forensic Statistics, Venice International University, Venice, Italy, August 30th to September 2, 2002.
Jackson, T. J., Wade, T. L., Sericano, J. L., Brooks, J. M., Wong, J. M., Garcia-Romero, B., and McDonald, T. J. (1998). Galveston Bay: Temporal changes in the concentrations of trace organic contaminants in national status and trends oysters (1986–1994). Estuaries 21(48):718–730.
Janardan, K. G. and Schaeffer, D. J. (1979). Sampling frequency and comparison of grab and composite sampling programs for effluents. Draft MS.
Janardan, K. G., Schaeffer, D. J., and Kerster, H. W. (1980). Modeling composite sampling with application to trace organics monitoring. Model. Simul. 11:985–989.
Japan International Cooperation Agency & forest Research & Development Agency (2005). Manual on soil sampling & analysis for AIR CDM projects.
Jenkins, T. F., Grant, C. L., Brar, G. S., Thorne, P. G., Ranney, T. A., and Schumacher, P. W. (1996). Assessment of sampling error associated with collection & analysis of soil samples at explosives contaminated soils special report 96-15. cold regions research & Engineering Laboratory, US Army Corps. of Engineers.
Jenkins, T. F., Grant, C. L., Brar, G. S., Thorne, P. G., Schumacher, P. W., and Ranney, T. A. (1997). Sampling error associated with collection & analysis of soil samples at TNT-contaminated sites. Field Anal. Chem. Technol. 1(3):151–163.
Johnson, G. D. and Patil, G. P. (2001). Cost analysis of composite sampling for classification. Environ. Ecol. Stat. 8:91–107.
Johnson, N. L. and Kotz, S. (1987). Effects of errors in inspection on a binary method for isolating nonconforming items. Working Paper Series MS/S 87-005, Management Science and Statistics, University of Maryland, College Park, MD.
Johnson, N. L. and Kotz, S. (1988). Estimation from binomial data with classifiers of known and unknown imperfections. Naval Res. Log. 35:147–156.

Johnson, N. L. and Kotz, S. (1990). Randomly weighted averages: Some aspects and extensions. Am. Stat. 44:245–249.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1985). Statistical effects of imperfect inspection sampling. I. Some basic distributions. J. Qual. Technol. 17:1–31.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1986). Statistical effects of imperfect inspection sampling. II. Double sampling and link sampling. J. Qual. Technol. 18:116–138.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1987). Statistical effects of imperfect inspection sampling. III. Screening (group testing). Working Paper Series MS/S 87-025, Management Science and Statistics, University of Maryland, College Park, MD.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1988). Statistical effects of imperfect inspection sampling. III. Screening (group testing). J. Qual. Technol. 20:98–124.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1989a). Dorfman-Sterrett screening (group testing) schemes and the effects of faulty inspection. Commun. Stat. Part A Theor. Methods 18:1469–1484.
Johnson, N. L., Kotz, S., and Rodriguez, R. N. (1989b). Statistical effects of imperfect inspection sampling. IV. Modified Dorfman screening procedures. Working Paper Series MS/S 87-009, Management Science and Statistics, University of Maryland, College Park, MD.
Johnson, N. L., Kotz, S., and Wang, Q. (1989). Randomized-sequential group testing procedures. Working Paper Series MS/S 89-025, Management Science and Statistics, University of Maryland, College Park, MD.
Johnson, N. L., Kotz, S., and Wu, X. (1991). *Inspection Errors for Attributes in Quality Control*. Chapman and Hall, New York, NY.
Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic, New York, NY.
Kerr, J. D. (1971). The probability of disease transmission. Biometrics 27:219–222.
King, K. W. and Harmel, R. D. (2004). Comparison of time based sampling strategies to determine Nitrogen loading in plot –scale run off. Trans. Am. Soc. Agric. Eng. 47(5):1457–1463.
King, K. W., Harmel, R. D., and Fausey, N. R. (2005). Development & sensitivity of a method to select time-& flow paced storm event sampling intervals for headwater streams. J. Soil Water Conserv. 60(6):323–331.
Klamath Blue Green Algae working group (2009). Standard operating procedures: Environmental sampling of cyanobacteria for cell enumeration, identification & Toxin analysis developed for the 2009 AIP Interim measure 12, water quality monitoring activities, Klamath River.
Kleijnen, J. P. C. (1987). Review of random and group screening designs. Commun. Stat. Part A Theor. Methods 16:2855–2900.
Knicker, H., (2007). How does fire affect the nature and stability of soil organic nitrogen and carbon? A review. Biogeochemistry 85:91–118.
Kotz, S. and Johnson, N. L. (1982). Errors in inspection and grading: Distributional aspects of screening and hierarchical screening. Commun. Stat. Part A Theor. Methods 11:1997–2016.
Kotz, S. and Johnson, N. L. (1988). Effects of inspections errors on Dorfman screening procedures with random group size: a note on Kemp and Kemp's Simple Inspection Scheme for Two Types of Defect. Working Paper Series MS/S 88-004, Management Science and Statistics, University of Maryland, College Park, MD.
Kotz, S., Shisong, M. and Johnson, N. L. (1986). Effects of inspection errors on curtailed Dorfman-type procedures. Commun. Stat. Part A Theor. Methods 15:831–838.
Kratochvil, B. and Taylor, J. K. (1981). Sampling for chemical analysis. Anal. Chem. 53:924–938.
Kulkarni, V. M. (1999). A novel method to evaluate efficacy of mill sanitation biocides: Reducing sugar and titrable acidity analysis of final molasses. Sugar Tech 1(3):54–62.
Kumar, S. (1965). A group testing problem. Ann. Math. Stat. 36:727–728.
Kumar, S. (1970). Multinomial group-testing. SIAM J. Appl. Math. 19:340–350.
Kumar, S. (1971). Trinomial in the three categories. Ann. Inst. Stat. Math. 24:171–181.
Kumar, S. and Sobel, M. (1971). Finding a single defective in binomial group testing. J. Am. Stat. Assoc. 66:824–828.

Kurfürst, U., Desaules, A., Rehnert, A., Muntau, H. (2004). Estimation of measurement uncertainty by the budget approach for heavy metal content in soils under different land use. Accred Qual Assur 9:64–75.

Kurtz, D. and Sidi, M. (1988). Multiple-access algorithms via group testing for heterogeneous population of users. IEEE Trans. Commun. 36:1316–1323.

Kussmaul, K. and Anderson, R. L. (1967). Estimation of variance components in two-stage nested designs with composite samples. Technometrics 9:373–389.

Lacayo, H., Nadeau, R. J., Patil, G. P., and Zaragoza, L., eds. (1990). *Proceedings of the Workshop on Superfund Hazardous Waste: Statistical Issues in Characterizing a Site: Protocols, Tools, and Research Needs*, February 21–22, Arlington, VA.

Lane, C. R. (2007). Assessment of isolated wetland condition in Florida using epiphytic diatoms at genus, species, and subspecies taxonomic resolution. EcoHealth 4:219–230.

Lane, C. R. and Brown, M. T. (2006). Energy-Based land use predictors of proximal factors and benthic diatom composition in Florida freshwater marshes. Environ. Monit. Assess. 117:433–450.

Lateef, A., Oloke, J. K., and Gueguimkana, B. (2005). The prevalence of bacterial resistance in clinical, food, water and some environmental samples in Southwest Nigeria. Environ. Monit. Assess. 100:59–69.

Lau, T. (1991). On dependent repeated screening tests. Biometrics 47:77–86.

Lawing, W. D. and Hoeing, J. M. (1984). On group testing for diseases. Can. J. Fish. Aquat. Sci. 41:1269–1270.

Le, C. T. (1981). A new estimator for infection rates using pools of variable size. Am. J. Epidemiol. 114:132–135.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. Ecol. Monogr. 62:67–118.

Leckie, S. E., Prescott, C. E., Grays ton S. J., Neufeld, J. D., and Mohn W. W. (2004). Characterization of humus Microbial communities in adjacent forest types that differ in Nitrogen availability. Microb. Ecol. 48:29–40.

Leczynski, B. A., Mack, G. A., and Berner, T. O. (1987). Population estimates from fiscal year 1982 specimens. National human adipose tissue survey broad scan analysis. Final Report, NHATS-SS-09, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Leczynski, B. A., Mack, G. A., Berner, T. O., Hersey, J. C., and Unger, A. (1988). Statistical analysis of the FY82 NHATS broad scan analysis data. Draft Final Report, NHATS-SS-04, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Lee, J. K. and Sobel, M. (1972). Dorfman and R1-type procedures for a generalized group-testing problem. Math. Biosci. 15:317–340.

Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. Stat. Sci. 5:160–168.

Leitão, S., Pinto, P., Pereira, T., and Brito, M. F. (2007). Spatial and temporal variability of macroinvertebrate communities in two farmed Mediterranean rice fields. Aquat. Ecol. 41: 373–386.

Lejon, D. P. H., Chaussod, R., Ranger, J., and Ranjard, L. (2005). Microbial community structure and density under different tree species in an acid forest soil (Morvan, France). Microb. Ecol. 50:614–625.

Lele, S. and Boswell, M. T. (1986). A general technique for generating pseudo multivariate random variables: A Markov process approach. TR 86–1203, Center for Statistical Ecology and Environmental Statistics, Penn State University.

Li, C. H. (1962). A sequential method for screening experimental variables. J. Am. Stat. Assoc. 57:455–477.

Life Systems, Inc. (1988). Composite sampling for acid rain and human and environmental health measurement—expert panel meeting. Report to Office of Acid Deposition, U.S. EPA, Washington, DC (TR–898–57A).

Liggett, W. S., Inn, K. G. W., and Hutchinson, J. M. R. (1984). Statistical assessment of subsampling procedures. Environ. Int. 10:143–151.

Lin, F. O. (1974). Application of group-testing methods to medically oriented problems: optimal properties of a recursive group-testing procedure for the problem of at most D defective. Ph.D. Thesis, University of Minnesota.

Lindström, A. (2008). Distribution of Paenibacillus larvae spores among adult honey bees (Apis mellifera) and the relationship with clinical symptoms of American foulbrood. Microb. Ecol. 56:253–259.

Lintner, T. J., Maki, C. L., Brame, K. A., and Boswell, M. T. (1992). Sampling dust from human dwellings to estimate the prevalence of indoor allergens. 92–0805.

Lisitza, M. A., DeFoliart, G. R., Yuill, T. M., and Karandinos, M. G. (1977). Prevalence rates of Lacrosse virus (California encephalitis group) in larvae from overwintered eggs of Aedes triseriatus. Mosquito News 37:745–750.

Lock, W. H. (1998). Composite sampling national environmental health forum monographs, soil series No.-3.

Løes, A. K. and øgaard, A. F. (2001). Long-term changes in extractable soil phosphorus (P) in organic dairy farming systems. Plant Soil 237:321–332.

Loyer, M. W. (1983). Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. 37:57–59.

Mack, G. A. and Panebianco, D. L. (1986). Statistical analysis of the FY82 NHATS broad scan analysis data. Draft Final Report, NHATS-SS-04, Office of Toxic Substances, U.S. EPA, Washington, DC.

Mack, G. A. and Robinson, P. E. (1985). Use of composited samples to increase the precision and probability of detection of toxic chemicals. In *Environmental Applications of Chemometrics*, J. J. Breen, and P. E. Robinson, eds. American Chemical Society, Washington, DC. pp. 174–183.

Mack, G. A. and Stanley, J. (1984). Program strategy for the national human adipose tissue survey. Final Report, NHATS-ST-01, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Mack, G. A., Leczynski, B., Chu, A., and Mohadjer, L. (1984). Survey design for the national human adipose tissue survey. Draft Final Report, NHATS-SD-01, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Mahoney, D. F. and Mirre, G. B. (1971). Bovine babesiasis: estimation of infection rates in the tick vector Boophilus microplus (Canestrini). Ann. Trop. Med. Parasitol. 65:309–317.

Marsalek, J. (1975). Sampling techniques in urban runoff quality studies, in water quality parameters. In STP 573, American Society for Testing and Materials, Philadelphia, PA. pp. 526–542.

Mashal, K., Al-Qinna Yahya, and Ali, M. (2009). Spatial distribution & environmental implications of Lead & Zinc in urban soils & street dusts samples in Al-Hashimeyeh Municipality (2009). Jordan J. Mech. Ind. Eng. 3(2):141–150

Mason, B. J. (1982). Preparation of soil sampling protocol: techniques and strategies. ETHURA, McLean, VA, under subcontract to Environmental Research Center, University of Nevada, for U.S. Environmental Protection Agency, Las Vegas, NV.

Massachusetts Water Resources Authority (2007). Sampling procedures/protocols. Industrial pie treatment program.

Matern, B. (1960). Spatial variation. Medd for Statens Skogsforskings Institut 49:1–144.

Matheron, G. (1963). Principles of geostatistics. Econ. Geol. 58:1246–1266.

McGann, T. C. A. (1973). Analysis of aged composite milk samples for payment purposes. Dairy Industries, November, 507–511.

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. Aust. J. Agric. Res. 3:385–390.

Mclaine, P., Shields, W., Farfel, M. Jr., J. J. C., and Dixon S. (2006). A coordinated relocation strategy for enhancing case management of lead poisoned children: outcomes and costs. J. Urban Health Bull. NY Acad. Med. 83(1):111–128.

Meals, D. W. (2004). Water quantity response to riparian restoration in two Vermont agricultural watersheds. Southeastern Regional conference on stream Restoration.

Mehravari, N. (1986). Generalized binary binomial group testing. SIAM J. Alg. Disc. Meth. 7:159–166.

Mekonnen, G., Wilcox, C. J., Bachman, K. C., Thatcher, W. W., and Martin, F. G. (1976). Genetic parameters for a.m.-p.m. composite monthly milk samples. J. Dairy Sci. 59:15–16.

Melo, A. S. and Costa, S. S. (2008). Beta diversity in stream macroinvertebrate assemblages: among-site and among-microhabitat components. Hydrobiologia 598:131–138

Messner, M. J., Clayton, C. A., Michael, D. I., Neptune, M. D., and Brantley, E. P. (1990). Retrospective design solutions for a remedial investigation. (Supplement to: Quantitative decision making in superfund: A data quality objectives case study. Hazard. Mater. Control 3(3).

Mitchell, T. J. and Scott, D. S. (1987). A computer program for the design of group testing experiments. Commun. Stat. Part A Theor. Methods 16:2943–2955.

Monestiez, P., Audergon, J., and Habib, R. (1990). Spatial dependencies and sampling in a fruit tree: A geostatistical approach. Technical Report No. 163 Department of Statistics and Department of Applied Earth Sciences, Stanford University, Stanford, CA.

Monestiez, P. and Switzer P. (1991). Semiparametric estimation of nonstationary spatial covariance models by metric multidimensional scaling. Technical Report No. 165 Department of Statistics and Department of Applied Earth Sciences, Stanford University, Stanford, CA.

Montgomery, H. A. C. and Hart, I. C. (1974). The design of sampling programmes for rivers and effluents. Water Pollut. Contron (London) 73:77–101.

Moon, C. S., Chang, Y. S., Kim, B. H., Shin, D., and Ikeda, M. (2005). Evaluation of serum dioxin congeners among residents near continuously burning municipal solid waste incinerators in Korea. Int. Arch. Occup. Environ. Health 78:205–210.

Moon, J. W. and Sobel, M. (1977). Enumerating a class of nested group testing procedures. J. Comb. Theor Ser. B 23:184–188.

Mundel, A. B. (1984). Group testing. J. Qual. Tech. 16:181–188.

Myrick, T. E., Berven, B. A., and Haywood, F. F. (1981). Determination of concentrations of selected radionuclides in surface soil in the U.S. Health Phys. 45:631–642.

National Oceanic and Atmospheric Administration (NOAA) (1989). A summary of data on chemical contaminants in sediments from the National Status and Trends Program. Progress Report, National Status and Trends Program for Marine Environmental Quality. NOAA Technical Memorandum NOS OMA 49, National Ocean Service, Rockville, MD.

National Oceanic and Atmospheric Administration (NOAA) (1991). A summary of data on tissue contamination from the first three years (1986–1988) of the mussel watch project. Progress Report, National Status and Trends Program for Marine Environmental Quality. NOAA Technical Memorandum NOS OMA 59, National Ocean Service, Rockville, MD.

National Oceanic and Atmospheric Administration (NOAA) (1991). National status and trends program Mollusk chemistry data. 2 Diskette's 3.5″ , 6 Tab-delimited ASCII Text files: Mussel Watch 86–88+ 86–89, 2 NS and T Site Names-Loc., Benthic Surv. 84–86, BS Surv. Sta. Locations.

Nebenzahl, E. (1975). Binomial group testing with two different success parameters. Stud. Sci. Math. Hung. 10:61–72.

Nebenzahl, E. and Sobel, M. (1973). Finite and infinite models for generalized group-testing with unequal probabilities of success for each item. In *Discriminant Analysis and Applications*, T. Cacoullos, ed. Academic, New York, NY. pp. 239–289.

Nelson, M. A., Cash, L. W., Trost, K., and Purtle, J. (2005). Water sampling, analysis & annual load determinations for TSS, Nitrogen & Phosphorus at the L'Angnille river near Palestine. Publication No. MSC – 327, Arkansas water resources center.

Neptune, D. (1990). Practical statistical applications to planning superfund site remediation. EPA Stat. 6:1–3.

Neptune, D., Brantly, E. P., Messner, M. J., and Michael, D. I. (1990). Quantitative decision making in Superfund: A data quality objectives case study. Hazard. Mater. Control 3:19–27.

Nisselson, H. (1987). Evaluation of the compositing scheme used in EED's NHMP. Draft Final Report, NHATS-SS-08, Office of Toxic Substances, U.S. EPA, Washington, DC.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. J. Official Stat. 5:223–239.

Olivares-Rieumont, S., Lima, L., De la Rosa, D., Graham, D. W., Columbie, I., Santana, J. L., and Sánchez, M. J. (2007). Water hyacinths (Eichhornia crassipes) as indicators of heavy metal impact of a large landfill on Almendares River near Havana, Cuba. Bull. Environ. Contam. Toxicol. 79:583–587.

Olson, K., Blair, C., Padmanabhan, R., and Beatty, B. (1988). Detection of dengue virus type II in Aedes albopictus by nucleic acid hybridization with strand-specific RNA probes. J. Clin. Microbiol. 26:579–581.

Onate, B. T. (1953). Some statistical aspects of the use of composites in soil sampling. Philippine Agric. 43:241–257.

Orban, J. E. and Lordo, B. (1989). Statistical methods for analyzing NHATS composite sample data: Evaluation of multiplicative and additive methodologies. Draft Final Report, NHATS-SS-11, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Orban, J., Leczynski, B. A., and Lordo, R. (1987). Estimation of prevalence using composited samples. Draft Final Report, NHATS-SS-10, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Orban, J. E., Lordo, R., and Schwemberger, J. (1990). Statistical methods for analyzing composite sample data applied to EPA's human monitoring program. MS.

Paasivirta, J. and Paukku, R. (1989a). Herrings from Vironlahti, Gulf of Finland 1987. MS (in Russian).

Paasivirta, J. and Paukku, R. (1989b). Use of composited samples to optimize the monitoring of environmental toxins. Chemosphere 19:1551–1562.

Panebianco, D. L. (1986a). A review of hospital nonresponse and its effect on standard errors of sample estimates in NHATS. Draft Final Report, NHATS-SS-05, Office of Toxic Substances, U.S. EPA, Washington, DC.

Panebianco, D. L. (1986b). Sample design for the fiscal year 1987 NHATS. Draft Final Report, NHATS-SD-02, Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

Pasternack, B. S., Sobel, M., and Thomas, J. (1987). Group-testing, halving procedures, and binary search. Commun. Stat. Part A Theor. Methods 16:2851–2871.

Patel, M. S. (Ed.) (1987). Experiments in Factor Screening. Dekker, New York

Patel, M. S. and Ottieno, J. A. M. (1987). Optimum two stage group-screening with unequal group sizes and with errors in decisions. Commun. Stat. Part A Theor. Methods 16:2799–2820.

Patil, G. P. (2000). Editorial: Marching together in the new millennium. Environ. Ecol. Stat. 7: 5–19.

Patil, G. P., Boswell, M. T., Bolgiano, N. C., Taillie, C., Gore, S. D., and Lovison, G. (1990). Tutorial: Theory and application of composite sampling in environmental work. Notes for a short course presented at ASA Winter Conference on Statistics and the Environment, New Orleans, LA.

Patil, G. P., Sinha, A. K., and Taillie, C. (1999). Ranked set sampling: a bibliography. Environ. Ecol. Stat. 6:91–98.

Patil, G. P., Sinha, A. K., and Taillie, C. (1992). Ranked set sampling from a finite population in the presence of a trend on a site. 91–1204 To appear in Journal of Applied Statistical Science.

Peech, M. (1965). Hydrogen-ion activity. In *Methods of Soil Analysis*, C. A. Black, D. D. Evans, L. E. Ensminger, J. L. White, F. E. Clark, and R. C. Dinauer, eds. American Society of Agronomy, Madison, Wisconsin. Agronomy 9:914–926.

Pennington, J. C. et al. (2005). Distribution & Fate of Energetics on DOD Test & training ranges: Interim report 5. Cold Regions Research & Engineering Laboratory, US Army Corps. of Engineers. ERDC/TR- 05–2.

Petersen, R. G. and Calvin, L. D. (1965). Sampling. Agronomy 9:54–72.

Peterson, M. J., Smith, J. G., Southworth, G. R., Ryon, M. G., and Eddlemon, G. K. (2002). Trace element contamination in benthic macroinvertebrates from a small stream near a uranium mill tailings site. Environ. Monit. Assess. 74:193–208.

Petisco, C., García-Criado, B., Vázquez de Aldana, B. R., Zabalgogeazcoa, I. Mediavilla, S., and García-Ciudad, A. (2005). Use of near-infrared reflectance spectroscopy in predicting nitrogen, phosphorus and calcium contents in heterogeneous woody plant species. Anal. Bioanal. Chem. 382:458–465.

Pfeifer, C. G. and Enis, P. (1978). Dorfman-type group testing for a modified binomial model. J. Am. Stat. Assoc. 73:588–592.

Piegorsch, W. W. and Edwards, D. (2002). What shall we teach in environmental statistics? Environ. Ecol. Stat. 9, 125–150.

Pipes, W. O. and Minnigh, H. A. (1990). Detection of microorganisms at very low densities in drinking water by composite sampling. Paper presented at the 2nd International Conference on Statistical Methods for the Environmental Sciences, Como, Italy.

Poussart, J. N., Ardö, J. and Olsson, L. (2004). Verification of soil carbon sequestration: sample requirements. Environ. Manage. 33(Suppl. 1):S416–S425.

Preciado, H. F. and Li, L. Y. (2006). Evaluation of metal loading and bioavailability in air, water and soil along two highways of British Columbia, Canada. Water Air Pollut. 172:81–108.

Proctor, C. H. (1990a). Sampling terms of reference. In *Statistical Sampling: Past, Present and Future—Theoretical and Practical*, M. J. Kowalewski and J. B. Tye, eds. STP 1097, American Society for Testing and Materials, Philadelphia, PA. pp. 55–60.

Proctor, C. H. (1990b). Statistical considerations in bulk sampling. MS.

Provost, L. P. (1984). Statistical methods in environmental sampling. In *Environmental Sampling for Hazardous Wastes*, G. E. Schweitzer and J. A. Santolucito, eds. American Chemical Society, Washington, DC. pp. 79–96.

Rabosky, J. G. and Koraido, D. L. (1973). Gaging and sampling industrial waste waters. Chem. Eng. 80:111–120.

Rajagopal, R. and Williams, L. R. (1989). Economics of sample compositing as a screening tool in ground water quality monitoring. Ground Water Monit. Rev. 9:186–192.

Ramsey, J. M., Bown, D. N., Aron, J. L., Beaudoin, R. L., and Mendez, J. F. (1986). Field trial in Chiapas, Mexico, of a rapid detection method for malaria in anopheline vectors with low infection rates. Am. J. Trop. Med. Hyg. 35:234–238.

Reed, J. F. and Rigney, J. A. (1947). Soil sampling from fields of uniform and nonuniform appearance and soil types. J. Am. Soc. Agron. 39:26–40.

Reiner, E. J., Clement, R. E., Okey, A. B., and Marvin, C. H. (2006). Advances in analytical techniques for polychlorinated dibenzo-p-dioxins, polychlorinated dibenzofurans and dioxin-like PCBs. Anal. Bioanal. Chem. 386; 791–806.

Rodil, R., Martinez, E., Carro, A. M., Lorenzo, R. A., and Cela, R. (2004). Applying supersaturated Experimental designs to the study of composite sampling for monitoring pesticide residues in water. LCGC North Am. 22(3):272–286.

Rohde, C. A. (1976). Composite sampling. Biometrics 32:273–282.

Rohde, C. A. (1979). Batch, bulk and composite sampling. In *Sampling Biological Populations*, R. M. Cormack, G. P. Patil, and D. S. Robson, eds. International Co-operative Publishing House, Fairland, MD. pp. 365–377.

Rohlf, F. J., Akcakaya, H. R., and Ferraro S. P. (1991). Optimizing composite sampling protocols. Technical Report prepared for USEPA, Applied Biomathematics, 100 North Country Road, Setauket, NY 11733.

Roskopf, R. F. (1968). A composite-grab of water pollution control sampling. J. Water Pollut. Control Fed. 40:492–498.

Ross, N. P. and Stokes, L. (1999). EDITORIAL: Special issue on statistical design and analysis with ranked set samples. Environ. Ecol. Stat. 6:5–9.

Rossmann, R. (1988). Estimation of trace metal storage in Lake St. Clair post-settlement sediments using composite samples. J. Great Lakes Res. 14:66–75.

Ruark, G. A., Mader, D. L., and Tattar, T. A. (1982). A composite sampling technique to assess urban soils under roadside trees. J. Arboricult. 8:96–99.
Ryan, J. J., Pilon, J., and Leduc, R. (1982). Composite sampling in the determination of pyrethrins in fruit samples. J. Assoc. Off. Anal. Chem. 65:904–908.
Samuels, S. M. (1978). The exact solution to the two-stage group-testing problem. Technometrics 20:497–500.
Sander, P. and Öberg, T. (2006). Comparing deterministic and probabilistic risk assessments. JSS-J Soils Sediments 6(1):55–61.
Schaeffer, D. J. and Janardan, K. G. (1978). Theoretical comparisons of grab and composite sampling programs. Biometrics 20:215–227.
Schaeffer, D. J., Kerster, H. W., and Janardan, K. G. (1980). Grab versus composite sampling: A primer for the manager and engineer. Environ. Mgmt. 4:157–163.
Schaeffer, D. J., Kerster, H. W., and Janardan, K. G. (1982). Monitoring toxics by group testing. Environ. Mgmt. 6:467–469.
Schaeffer, D. J., Kerster, H. W., Bauer, D. R., Rees, K., and McCormick, S. (1983). Composite samples overestimate waste loads. J. Water Pollut. Control Fed. 55:1387–1392.
Schaeffer, D. J., Park, J. B., Kerster, H. W., and Janardan, K. G. (1980). Sampling and the regulatory maze in the United States. Environ. Mgmt. 4:469–481.
Seshadri, N. and Srikantakumar, P. R. (1985). On group testing for binary-feedback multi access channel. IEEE Commun. 33:574–577.
Sharma, A. K., Rodriguez, L. A., Mekonnen, G., Wilcox, C. J., Bachman, K. C., and Collier, R. (1983). Climatological and genetic effects on milk composition and yield. Dairy Sci. 66:119–126.
Sheehan, P., Dewhurst, R. E., James, S., Callaghan, A., Connon, R., and Crane, M. (2003). Is there a relationship between soil and groundwater toxicity? Environ. Geochem. Health 25: 9–16.
Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley, New York, NY.
Skublov, G. T., Marin, Y. B., Semikolennykh, V. M., Skublov, S. G., and Tarasenko, Y. N. (2007). Volkhovite: A new type of tektite-like glass. Geol. Ore Deposits 49(8): 681–696.
Smith, M. D., Kahn, H. D., and Cameron, K. (1991). Statistical analysis of dioxin and furan measurements in environmental samples from the pulp and paper industry. Report presented at the Environmental Protection Agency Statistics Conference, March 11–14.
Snowdon, P. and Waring, H. D. (1984). Composite samples for foliar analysis. Aust. For. Res. 14:235–242.
Snyder, D. C. and Larson, K. E. (1969). Tables for group screening. J. Qual. Technol. 1:10–16.
Sobel, M. (1960). Group testing to classify efficiently all units in a binomial sample. In *Information and Decision Processes*, R. E. Machol, ed. McGraw Hill, New York, NY. pp. 127–161.
Sobel, M. (1966). Binomial and hypergeometric group testing (abstract). Ann. Math. Stat. 37:1865.
Sobel, M. (1968a). Binomial and hypergeometric group-testing. Stud. Sci. Math. Hung. 3: 19–42.
Sobel, M. (1968b). Optimal group testing. In *Proceedings of the Colloquium on Information Theory, Vol. II*, A. Rényi, ed. Janos Bolyani Mathematical Society, Budapest. pp. 411–488.
Sobel, M. and Elashoff, R. M. (1975). Group testing with a new goal, estimation. Biometrika 62:181–193.
Sobel, M. and Groll, P. A. (1959). Group testing to eliminate efficiently all defectives in a binomial sample. Bell Syst. Tech. J. 38:1178–1252.
Sobel, M. and Groll, P. A. (1966). Binomial group-testing with an unknown proportion of defectives. Technometrics 8:631–656.
Sobel, M. and Tong, Y. L. (1976). Estimation of a normal percentile by grouping. J. Am. Stat. Assoc. 71:189–192.

Sokal, R. R. and Rohlf, F. J. (1981). *Biometry, the Principles and Practice of Statistics in Biological Research*, 2nd Edition. W. H. Freeman, San Francisco, CA.

Spijker, J. (2005). Geochemical patterns in the soils of Zeeland: natural variability versus Anthropogenic impact. Ph.D. thesis submitted to Utrecht University, The Netherlands.

Splitstone, D. E. (2001). Sample support and related scale issues in composite sampling. Environ. Ecol. Stat. 8:137–149.

SR Technics Ireland Limited (2008). Annual environment report for year 2008. Dublin Airport Co. Dublin.

Starks, T. H. (1986). Determination of support in soil sampling. Math. Geol. 18:529–537.

Starks, T. H., Sparks, A. R., and Brown, K. W. (1987). Geostatistical analysis of Palmerton soil survey data. Environ. Monit. Assess. 9:239–261.

Sterrett, A. (1957). On the detection of defective members of large populations. Ann. Math. Stat. 28:1033–1036.

Stevens, L. J. and Combs, C. A. (1980). Group testing to eliminate defectives when prior probabilities for the proportion defective are known. J. Indust. Math. Soc. 30:95–102.

Stokes, S. L. and Sager, T. (1988). Characterization of a ranked set sample with application to estimating distribution functions. J. Am. Stat. Assoc. 83:374–381.

Storey, R. G., Williams, D. D., and Fulthorpe, R. R. (2004). Nitrogen processing in the hyporheic zone of a pastoral stream. Biogeochemistry 69:285–313.

Strawn, D. G., Hickey, P., Knudsen, A., and Baker, L. (2007). Geochemistry of lead contaminated wetland soils amended with phosphorus. Environ. Geol. 52:109–122.

Sudia, W. D., Newhouse, V. F., Beadle, L. D., Miller, D. L., Johnston, J. G., Young, R., Calisher, C. H., and Maness, K. (1975). Epidemic Venezuelan equine encephalitis in N. America in 1971: Vector studies. Am. J. Epidemiol. 101:17–35.

Sudia, W. D., Newhouse, V. F., Calisher, C. H., and Chamberlain, R. W. (1971). California group arboviruses: Isolations from mosquitoes in North America. Mosquito News 31:576–600.

Swallow, W. H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75:882–889.

Swallow, W. H. (1987). Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission. Phytopathology 77:1376–1381.

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann. Inst. Stat. Math. 20:1–31.

Telliard, W. A. (1998). Evaluating field techniques for collecting effluent samples for trace metals analysis. USEPA office of water, engineering & analysis division, Washington, DC.

Tetra Tech, Inc. (1987). Bioaccumulation monitoring guidance: strategies for sample replication and compositing. Final Report, EPA Contract No. 68-01-6938. (TC3953-03).

Texas Eastern Gas Pipeline Company (1989a). Results of the Phase II surface soil and sediment sampling activities at the Armagh site, Pennsylvania, Vol. I. Roy F. Weston, West Chester, PA 19380.

Texas Eastern Gas Pipeline Company (1989b). *Results of the Phase II surface soil and sediment, sampling activities at the Armagh site, Pennsylvania, Vol. II: Appendices*. Roy F. Weston, West Chester, PA 19380.

Thomas, J., Pasternack, B. S., Vacirca, S. J., and Thompson, D. L. (1973). Application of group testing procedures in radiological health. Health Phys. 25:259–266.

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. Biometrics 18:568–578.

Tiemeyer, B., Kahle, P., and Mousa, B. L. R. (2008). Measurements & modeling of water & solute fluxes in a artificially drained lowland catchments. 10th International Drainage workshop of 1C 1D working group on Drainage, Helsinki, Finland/Tallinn, Estonia, July 6th–11th 2008.

Tosic, R. (1982). An optimal group-testing procedure. Stud. Sci. Math. Hung. 17:319–323.

Towatana, P., Voradej, C., and Leeraphante, N. (2003). Reclamation of abandoned shrimp pond soils in Southern Thailand for cultivation of Mauritius grass. Environ. Geochem. Health 25:365–386.

Townsend, S. A. (2006). Hydraulic phases, persistent stratification, and phytoplankton in a tropical floodplain lake (Mary river, northern Australia). Hydrobiologia 556:163–179.

Ungar, P. (1960). The cutoff point for group testing. Commun. Pure Appl. Math. 13:49–54.

US Army, Alaska, & US Army Engineer district, Alaska (2001). Remediating & monitoring white phosphorus contamination at Eagle River Flats (Operable Unit C) Fort Richardson, Alaska. FY 00 report.

USEPA (1983). Quality assurance program plan for the Office of Toxic Substances. Office of Pesticides and Toxic Substances, U.S. EPA, Washington, DC.

USEPA (1985). Verification of PCB spill cleanup by sampling and analysis. Technical Report EPA-560/5-026, Office of Toxic Substances, U.S. EPA, Washington, DC.

USEPA (1987). Bioaccumulation Monitoring Guidance: Strategies for Sample Replication and Compositing, Volume 5. Technical Report EPA 430/09-87-003, Office of Marine and Estuarine Protection, U.S. EPA, Washington, DC.

USEPA (1989). Methods for Evaluating the Attainment of Cleanup Standards: Volume 1: Soils and Solid Media.

USEPA (1990). The 104 mill study. TECH REPORT-USEPA/Paper Industry Cooperative Dioxin Study, Statistical Findings and Analyses, U.S. EPA, Washington, DC.

USEPA (2002). Guidance on choosing a sampling design for Environmental data collection for use in developing a quality assurance project plan. EPA QA/G-5S.

Van Belle, G., Griffith, W. C. and Edland, S. D. (2001). Contributions to composite sampling. Environ. Ecol. Stat. 8:171–180.

VanAssche, W. (1987). A random variable uniformly distributed between two random variables. Sankhya Ser. A 49:207–211.

Walker, W. W. (2002). Long–term watershed monitoring: statistical models & examples. Watershed monitoring & reporting session, Gulf of Mexico Hypoxia workshop, St. Louis, October 16–18, 2002.

Wallin, T. R. and Schaeffer, D. J. (1979). Illinois redesigns its ambient water quality monitoring network. Environ. Mgmt. 3:313–319.

Walsh, M. E., Collins, C. M., Bailey, R. N., and Grant, C. L. (1997). Composite sampling of sediments contaminated with white phosphorus. Special Report 97-30. Cold Region Research & Engineering Laboratory, US Army Corps. of Engineers.

Walsh, M. R. (2003). Eagle River Flats Remediation Project Comprehensive Bibliography 1998 to 2003. Cold Region Research & Engineering Laboratory, US Army Corps. of Engineers. ERDC/CRREL TR-03-15.

Walsh, M. R., Walsh, M. E., and Collins, C. M. (1999).Enhanced natural remediation of white-phosphorus- contaminated wetlands through controlled pond draining. Cold regions research & engineering laboratory, U S Army Corps. of Engineers.

Walter, S. D., Hildreth, S. W., and Beaty, B. J. (1980). Estimation of infection rates in populations of organisms using pools of variable size. Am. J. Epidemiol. 112:124–128.

Watson, G. S. (1961). A study of the group screening method. Technometrics 3:371–388.

Watson, G. S. (1987). Reflections on group screening. Commun. Stat. Part A Theor Methods 16:2795–2798.

Watson, M. A. (1936). Factors affecting the amount of infection obtained by aphis transmission of the virus Hy. III. Philos. Trans. Roy. Soc. London, Ser. B. 226:457–489.

Wegman, E. J. and DePriest, D. J., eds. (1986). Application of the Gibbs distribution to image segmentation. In *Statistical Image Processing and Graphics*. Marcel Dekker, New York, NY. pp. 3–24.

Weideman, C. A. and Raghavarao, D. (1987). Non-adaptive hypergeometric group testing designs for identifying at most two defectives. Commun. Stat. Part A Theor Methods 16:2991–3006.

West, M. (1986). Bayesian model monitoring. J. R. Stat. Soc. Ser. B 48:70–78.

Westrick, J. J. (1979). Collection of automatic composite samples without atmospheric exposure. J. Water Pollut. Control Fed. 51:2945–2951.

Wilding, L. P. and Drees, L. R. (1983). Spatial variability and pedology. In *Pedogenesis and Soil Taxonomy, I: Concepts and Interactions*, L. P. Wilding, N. E. Smeck, and G. F. Hall, eds. Elsevier, New York, NY. pp. 83–116.

Williams, C. J. and Peterson, R. G. (1978). Variation in estimates of milk fat, protein and lactose content associated with various bulk milk sampling programs. J. Dairy Sci. 61:1093–1102.

Williams, L. R. (1990). Cost-effective sampling and analytical strategies for Superfund—How do we implement them? MS.

Williams, L. R., Leggett, R. W., Espegren, M. L., and Little, C. A. (1989). Optimization of sampling for the determination of mean radium-226 concentration in surface soil. Environ. Monit. Assess. 12:83–96.

Wolf, J. K. (1985). Born again group testing: multi access communications. IEEE Trans. Info. 31:185–191.

Worlund, D. D. and Taylor, G. (1983). Estimation of disease incidence in fish populations. Can. J. Fish. Aquat. Sci. 40:2194–2197.

Wullshleger, R. E., Zanoi, A. E., and Hanson, C. A. (1976). Methodology for the Study of Urban Storm Generated Pollution and Control. Technical report prepared by Envirex, Inc. for the Environmental Protection Agency, Contract 68-03-0335.

Xiao, H. (1991). Nonparametric procedures for conditional CDF approximation in spatial fields. Technical Report No. 164, Dept. of Statistics and Dept. of Applied Earth Sciences, Stanford University, Stanford, CA.

Yao, Y. C. (1988). An improvement over the information lower bound in binomial group testing. Prob. Eng. Inform. Sci. 2:313–320.

Yao, Y. C. (1989). A simple characterization of non-randomized admissible procedures in group testing. Stat. Prob. Lett. 8:119–122.

Yao, Y. C. and Hwang, F. K. (1988). Individual testing of independent items in optimal group testing. Prob. Engineer. Inform. Sci. 2:23–29.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. Math. Geol. 19:183–205.

Zamyadi, A., Gallic Hand, J., and Duchemin, M. (2007). Comparison of methods for estimating sediment & Nitrogen loads from a small agricultural watershed. Can. Biosyst. Eng. 49: 1.27–1.36.

Zbawicka, M., Wenne, R., and Skibinski, D. O. F. (2003). Mitochondrial DNA variation in populations of the mussel Mytilus trossulus from the Southern Baltic. Hydrobiologia 499:1–12.

Zimmerman, G. M. (1985). Use of an improved statistical method for group comparisons to study effects of prairie fire. Ecology 66:606–611.

Index

T