RESEARCH

Bettina Hüttenrauch

Targeting Using Augmented Data in Database Marketing

Decision Factors for Evaluating External Sources

Springer Gabler

Targeting Using Augmented Data in Database Marketing

Bettina Hüttenrauch

Targeting Using Augmented Data in Database Marketing

Decision Factors for Evaluating External Sources

Springer Gabler

Bettina Hüttenrauch, geb. Krämer
Frankfurt, Germany

Dissertation Johannes Gutenberg-Universität Mainz, 2015

OnlinePlus material to this book can be available on
http://www.springer-gabler.de/978-3-658-14577-4

ISBN 978-3-658-14576-7 ISBN 978-3-658-14577-4 (eBook)
DOI 10.1007/978-3-658-14577-4

Library of Congress Control Number: 2016942900

Springer Gabler

Printed on acid-free paper

This Springer Gabler imprint is published by Springer Nature
The registered company is Springer Fachmedien Wiesbaden GmbH

Preface

This dissertation is the result of my studies on data augmentation in database marketing. I am particularly interested in data and how it can be used to differentiate people in order to automatically generate individualized communication. I am always surprised on how much data we leave behind as digital footprints and – at the same time – how little we know about our customers as database marketing analysts. It is my utmost concern to reduce this information gap.

I have concentrated on the combination of statistics and communication during my media management degree at the University of Mainz. After an internship at Lufthansa German Airlines, I was granted the possibility to conduct my first data augmentation for the database marketing department at Miles & More, the frequent flyer program of Lufthansa. I was employed as a database marketing analyst in 2010 and have been working in this exciting field since.

There are enterprises specialized in data collecting and usage, e.g. Google, Facebook, and Apple. When using their platforms, we experience targeted communication. However, we seldom really notice it, because relevant information is not something one realizes (unless struck as daunting). Much more often, we perceive advertisements as irrelevant, misplaced, or inappropriate. The vast amount of companies does not have a detailed data basis to segment and select customers for differentiated marketing communication.

At work, we had many discussions about which external sources can be used for database marketing purposes and how. There is a general reluctance regarding for example volunteer surveys and social media sources. In these sources, data has been collected for a non-representative subgroup of customers only. Furthermore, most of the data cannot or must not be matched on an exact basis. Data augmentation projects require a considerable amount of know-how, time, and money. It is not approached, unless the return on marketing investment can be anticipated.

To me, these external sources provide a set of opportunities. Their contained information cannot be obtained otherwise. External sources are beneficial, because they are up-to-date, easy to acquire, and cheap. Service providers offering data fusion services are often overpriced and less experienced in the respective data. Without much effort, internal database marketing analysts can do a better job. By highlighting important facts to regard, I want to facilitate the use of data augmentation in companies.

Also, I would like to encourage the academic discussion regarding data augmentation in database marketing. A wide range of augmentation approaches has evolved, both in direct marketing and online marketing. However, the scientific foundation for these approaches is sparse. I believe that the methods and use cases for data augmentation will advance, if the academic discussion is pushed. I would like to make a contribution to this matter.

I am much obliged to thank all persons that supported me during my dissertation project. First and foremost I thank Heinz-Werner Nienstedt for his supervision, support, discussions, and enthusiasm. I was very happy to be able to continue to study at your chair at the Johannes Gutenberg University of Mainz.

At the same time, I worked at Lufthansa Miles & More as a database marketing analyst. I could not have had a better working environment for completing a dissertation, while working at the same time. I thank my superiors for the positive acceptance of my "hobby", the flexibility, and the support in all matters – including the possibility to use SAS. I also thank my team members for inspiring discussions and challenging questions. It was good to be able to get a direct feedback from the future beneficiaries of my work.

It was a great pleasure to get the chance to discuss my work with two of the most important researchers I cited, Susanne Rässler and Pascal van Hattum. Both of them surprised me by being eager to engage in my ideas and plans. Thank you for inviting me to Bamberg and Amersfoort, respectively, and for taking the time to review my work. I was was able to present my working paper at the 2013 EMAC Doctoral Colloquium in Istanbul to Arnaud de Bruyn, Arvind Rangaswamy, and Ujwal Kayande. Thank you for pushing me and for being so motivating. We had a very productive and encouraging atmosphere in our track. I also thank Heather Skinner and Nick Lee for reviewing my work at the Academy of Marketing 20th Doctoral Colloquium in Cardiff.

Finally, I would like to thank my husband, family and friends for motivating and understanding me in any possible way. Thank you for familiarizing with my complex topic, for proof-reading, and for listening. You have always been so proud and never questioned my success. I could not have asked for any more support.

Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA analysis of variance

B2B business-to-business

B2C business-to-consumer

BDSG Bundesdatenschutzgesetz

C2C consumer-to-consumer

CMH Cochran-Mantel-Haenszel

CPL conversion probability lift

CRM customer relationship management

DWH data warehouse

IP internet protocol

KPIs key performance indicators

MAR missing at random

MCAR missing completely at random

MNAR missing not at random

ROMI return on marketing investment

Sh.-W. Shapiro-Wilk test

SWOT analysis of strength, weaknesses, opportunities, and threads

TMG Telemediengesetz

UWG Gesetz gegen den unlauteren Wettbewerb

Definition of Variables

A areas under and between the conversion probability curves

a number of auxiliary variables

b absolute value

c customer indicator variable

CCR correct classification rate of the target value

CL conversion probability lift

CM conversion probability lift magnitude

D donor unit in the data augmentation model

d number of donors

d' number of representing units

F F statistic as derived from the F distribution

h hit indicator variable

IQV index of qualitative variation

j number of values per link variable

k number of values per target variable

l number of link variables

ML model lift

N number of observations

n number of selected customers for a direct marketing campaign

O overlap between recipient unit and donor unit

o number of overlapping elements

P overall population in the data augmentation model

p overall number of elements

R recipient unit in the data augmentation model

r number of customers or recipients

S set of source data mechanism indicator variables

t number of target variables

$TCCR$ total correct classification rate

v number of link variable classes

W W statistic for Shapiro-Wilk's test for normally distributed residuals

X set of link variables

Y set of target variables

Z set of auxiliary variables

Chapter 1

Introduction to data augmentation in marketing

Although marketing specialists spend a considerable amount of time, money, and know-how on relevant marketing campaigns, everybody is confronted with more less than well personalized advertisements every day. Relevance in this context is defined by attracting the positive attention of the recipient to the content or offer. While it can take weeks, if not months, to create these campaigns, the customers receiving the offer only need seconds to decide whether an offer is relevant or not. Especially in direct marketing, where prospective customers are purposely selected, nothing is less desirable to a marketer than an offer being ignored. The right selection and allocation of marketing communication is an every-day database marketing problem.

The data available on customers is not always sufficient to adequately define target groups and to meet the marketing goals. At the same time, external information is available encompassing many relevant facts. By augmenting this data, both companies and customers would profit from the increased relevance. While the information overload would decrease for the customers, wastage could be avoided from a company's point of view. Which external information sources are suitable for data augmentation and

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_1

how they can be used are the central questions of this study. This question has not yet been regarded in academic research so far.

In this chapter, we lay the basis for our study by explicating the problem of irrelevant communication and unused possibilities and deriving our research question. We explain important concepts and the context of data augmentation in database marketing in order to determine the research field and applicability. Eventually, we describe the research approach with which we answer the research question so as to establish practical guidelines for how to assess external sources upfront regarding their suitability for data augmentation in database marketing.

1.1 Research problem and relevance

Data augmentation can increase marketing efficiency. Database marketing analysts are responsible for finding the right target groups for individualized and personalized marketing communication. But the available information in the customer database is limited, so that augmenting data has become a valuable alternative to directly collecting data from customers. In this chapter, we explicate the problem of irrelevant communication and the unused possibilities from external sources in order to motivate the research question. We describe the academic and practical research context and the current state of research regarding the topic. From the practical need, the chances given by external sources, and the lack of attention in academic discussion our research question and desired contribution are derived.

1.1.1 Irrelevant communication and unused chances

Direct marketing has the goal of maximizing the profit of individual customers by increasing their spending volumes, exploiting their willingness to pay, and reducing their communication costs to a minimum, so as to grow the return on marketing investment (ROMI). Ideally, customers receive only relevant information to increase their interest, satisfaction and eventually their loyalty to the company. The need for efficient marketing is derived from the economic environment as described in chapter 2.2.1. The profit related to a customer centric communication approach (in contrast to a product centric approach) has already been recognized in the 20's century. It has been promoted both in practice and in academics (e.g. by Huldi (2002), Link and Hildebrand (1993), Rossi, McCulloch, and Allenby (1996), and Schweiger and Wilde (1993)). The customer focus as a major marketing goal is described in more detail in chapter 2.1.1. The vision of customer relationship management (CRM) is to convey the right information to the right person at the right location and time.

Data is the basis for all direct marketing activities. In order to best reach the customers, a lot of information is required on their preferences, needs and wants, and state in the customer life cycle. The better the available data, the more precise target group selections can be made. Database marketing structures in companies are available, extending the mere collection of transaction data to more sophisticated data mining methods and models (e.g. Adriaans and Zantinge (1998), Küsters (2001), Ratner (2001b), and Weiss and Indurkhya (1998)). The data is stored in a customer database, usually a customer data warehouse (DWH). These structures are explicated in more detail in chapter 4.1.1.

With the existing structures and processes, relevant communication should be very easy to deliver. But all too often, one is negatively surprised on how little companies know. For example when being female, aged 20-30, and living in a metropolitan area, one most likely receives online ads from dating websites offering handsome bachelors in the respective area. It fits the (few) available data, but might not be relevant. There are several reasons why companies lack relevant information.

Targeting for direct mailings, online marketing, or newer media is done semi-manually or implemented for automatic deployment by database marketing analysts. They use the information available in order to select the right target groups for campaigns and promotions. When information is not ready at hand, database marketing analysts make assumptions, build models, and derive predictions in order to target the right customers. If a retailer wants to promote a luxury product, e.g. an expensive watch, it does not have the variable "affinity for expensive watches" in its database. But it has monthly spending, transaction volumes, and comparable products bought. The retailer would target customers with a suitable transaction history for his campaign. However, the affinity for other high end products and expensive watches do not have to be correlated and may lead to irrelevant communication.

Even if detailed information is available, it is most commonly available only for a small, highly active portion of the customers. The big portion of occasional and inactive customers is not well describable by sufficient criteria. Consequently, well targeted promotions are possible only for a small group of customers, which is not sufficient for sales purposes. All other customers receive standardized offers. The resulting wastage is high.

Other information may be available and useful, but may not be allowed to be used. Online behavioral data such as surfing behavior, mailing awareness, and click frequency could help to identify customers generally affine for ads and commercial information. But unless customers have not been asked for their permission to use this behavioral data, companies must not apply it for targeting on an individual basis in Germany. The legal environment for database marketing is described in more detail in chapter 2.2.3.

Additional knowledge about customers is often available in aggregated form only. Surveys are conducted frequently to gain deeper insight into customer segments. For example, grandmothers may love buying presents for their kindergarten grandchildren, or singles working in the finance sector may be prone to take last minute offers for their vacations. But variables like family information, relationship status, or employment are usually not available from the customer database. Thus, although these segments are well describable and campaign actions are clear, it is not possible to identify these segments in order to treat them individually.

The result of these problems is a disadvantage for both customers and companies. If the data is not the right data, cannot be interpreted, must not be used, or is not sufficient in order to differentiate target groups, relevance degenerates to simple gender-age-region-schemas as in the example. Because of these limitations, ad space (available media in terms of channels and platforms) cannot be used efficiently. The ad burden for customers is high, and so are costs per contact when comparing contacts with sales. This is particularly true for below the line media, thus all channels through which customers can directly and individually be reached and where access is limited, e.g. email, letter, SMS, or promoters at the point of sale.

Data augmentation can be the answer to many of the problems described. Supplemental variables not available for the customers can be matched to individual profiles based on link variables present in both the customer database and the external source. They are derived from external sources; e.g. the company website, a customer survey, market media studies, or social media applications. As a result, definite variables can be used for targeting, rather than demographic target group descriptions, derivations, or common knowledge. Values are augmented for all customers, so that not only very active customers can be differentiated. Because data is augmented by groups rather than on a personal level, also sensitive information, e.g. a personal income level, is addable. Even aggregated data can be used, if an appropriate augmentation set up with suitable link variables, is chosen.

Today, we face a situation in which data is collected at various touch points, but little information is actually used to improve the marketing communication. A list of available external sources and their possibilities is given in chapter 2.3. With the exponential growth of data and information, CRM is experiencing a renaissance in academics and management. While the possibility and necessity of the use of external information for CRM has already been recognized (Arnold, 2011; Breur, 2011), a detailed study describing the process and which sources to use is still missing.

Many external sources have disadvantageous features, like being incomplete, partially overlapping with the customer group, not representative, or generally small. To apply the contained target variables to the customer database can lead to biased results. One reason why external sources are not used for database marketing purposes is the anxiety that available sources are not valuable in terms of data utility, meaning that it cannot be assessed upfront whether data augmentation results will be reliable and effective. With our study, we assess different forms of sources in order to give practical insights on which sources to use and how.

1.1.2 Academic and practical research context

The conditions for data augmentation with external sources in database marketing relevant today have been established in different research fields. In statistics and market research, methods for deriving joint information on people from different sources have been developed. The usage of augmentation techniques in marketing has begun with the introduction of widely-used electronic tracking systems and the change to a customer focus. Along with the emergence of social media marketing, local and mobile marketing, the merging of external data and the usage in direct marketing have come to form a new research field. These branches of science converge at data augmentation with external sources in database marketing.

This work ties in with the research projects of Kamakura and Wedel (1997, 2000), Putten et al. (2002a; 2002b), and Gilula, McCulloch, and Rossi (2006). They used data augmentation, or data fusion, in the context of marketing. Especially, it contributes to Hattum and Hoijtink's (2008b, 2008a) idea of data fusion and extends it to a broader range of possible applications. The two comprehensive works of Rässler (2002) and D'Orazio, Di Zio, and Scanu (2006) build the methodological basis for this extension of the data augmentation idea. Fundamental ideas therein, in the studies named previously, and in this work are attributed to Rubin (Rubin, 1976; Little & Rubin, 2002), who fathomed the conditions for data augmentation in the context of missing data theory. This work enlarges the field of data augmentation and statistical matching with a new and practical focus on using existing sources for data augmentation in database marketing.

The literature on data augmentation has been developed in statistical and marketing journals, before it found its way into specialized journals like *Database Marketing & Customer Strategy Management* and the *Journal of Targeting, Measurement and Analysis for Marketing*, which have been consolidated into the *Journal of Marketing Analytics* as of 2013 (Palgrave Macmillan, 2013). Major studies from a statistical point of view can be found in the *Journal of the American Statistical Association* and *Biometrika.* The literature from a marketing perspective is more disperse, including the *Journal of Marketing Research*, the *International Journal of Market Research*, and the *Journal of Direct, Data and Digital Marketing Practice.*

Data augmentation in database marketing is a practice oriented field of research. Database marketing analysts will directly benefit from the solutions. They work in the marketing or analytics department of companies, or for specialized agencies offering data augmentation services, like tns infratest (2012) in Germany or The SmartAgent Company (2013) in the Netherlands. The professional field is constantly growing, and so is the need for tools and methods. While companies like Google, Facebook, and Apple are expected to know a considerable amount of information on any arbitrary person, companies whose main business is not data collection have difficulties finding manageable ways of handling external data. Little is published on data augmentation for practitioners. There is no professional or academic exchange of ideas and approaches for data augmentation.

Data augmentation for database marketing is promoted at many places and with detailed descriptions of internal and external sources – often without going into detail about the processes and challenges related to it (Breur, 2011; Kuhner, 2013; Schiff, 2010). This study examines the contemporary situation for data augmentation with external sources in database marketing. It theoretically analyzes the special features relevant for data augmentation in this field and takes into consideration the challenges related to the using external sources. It is a starting point for a professional and academic conversation regarding the topic, facilitating a much more standardized and sophisticated development of the research field.

1.1.3 Research question and desired contribution

The question arising from the problems encountered in database marketing practice can be answered by the information available from external sources. But under which circumstances can external sources be used for data augmentation? It is necessary that database marketing analysts gain information from the data augmentation results, but not sufficient. In a marketing context, the information itself is only a mediator for the targeting goal, which is to increase conversion probabilities. Thus, the following research question is derived.

> *Which external data augmentation sources are able to increase conversion probabilities?*

The question inherently suggests that it is possible to increase conversion probabilities by augmenting external data. Previous works by Putten et al. (2002a) and Hattum and Hoijtink (2008b) suggested that data augmentation is able to increase conversion rates, at least under certain circumstances. However, it has never been considered what is special about data augmentation using external sources in database marketing and under which circumstances data augmentation results significantly increase conversion probabilities. In particular, it has never been assessed comprehensively which sources are suitable for data augmentation in database marketing and which source characteristics are essential in this assessment.

Most of the literature on data augmentation explicitly or implicitly refers to representative sources. These are convenient, but their features cannot be generalized for all forms of sources available today. A theoretical contribution of this study is a comprehensive description of sources and their formal characteristics. The quality of data augmentation depends on which link and target variables are used and on the predictive power of the link variables regarding a target variable. The augmentation methods also have an influence on the augmentation results.

The managerial contribution of this study is to establish a guide for augmenting external data in database marketing. It defines which characteristics are relevant for assessing external sources and the variables contained. The guide provides information on how to choose the right augmentation method and on how to manage expectations regarding augmentation results. Database marketing analysts are not familiar with using the new sources for data augmentation. By establishing a practical guide for ex ante evaluating data augmentation sources, the idea of data augmentation becomes more tangible. This study provides a starting point for broader usage. Once a process has been established, the techniques can be further refined by practitioners and applied to many different cases.

1.2 Research concepts and context

The research objective of this study is to examine data augmentation with external sources in database marketing for establishing guidelines regarding their usage in practice. In order to answer the research question and to facilitate the understanding for our approach, important concepts and the context of data augmentation in database marketing are explained in this chapter. The exact terminology and definition of data augmentation is given, as well as illustrative examples for available external sources. The different sources available in today's digital world have disadvantageous characteristics, such as being incomplete, not representative, or generally small. The situation of conditional independence between source and target variables is exemplified, and an introduction to the characteristics of external sources is given. We delineate the field of data augmentation in database marketing to other adjacent fields which are differentiated in order to state the applicability of our findings.

1.2.1 Definition of data augmentation and terminology

Data augmentation in database marketing is the process of taking so-called target variables from an external source and adding them to the customer database, based on link variables. It refers to adding supplemental data to the customer database based on similarities of elements instead of a unique identifier. The target variables are the variables of interest in the external source, which are not available from the customer database. Rässler' wording is applied here, who called the external source donor unit and the customer database recipient unit (2002, p. 3). Both units comprise a number of elements, i.e. rows in a database table. They are used synonymously here with customers, persons, or observations. The recipient unit contains all customers relevant to the augmentation frame. The donor unit contains customers and possibly other persons available from the source. A schematic illustration of data augmentation is given in figure 1.1. It is formally described in chapter 4.2.

Sources are augmentable regardless to their overlap in elements to the customer database, if there is a definable set of link variables appearing in both sources (Adamek, 1994). This is further explained in chapter 4.2.5. If a correlation exists between the link variables and the target variables in the donor source, it is possible to form groups of persons being alike as measured by their link variable values (e.g. D'Orazio et al., 2006, p. 2; Gilula et al., 2006; Rässler, 2002, p. 11; Rodgers, 1984), so-called look-alike profiles (Ratner, 2001b). Based on these, target variables can be predicted for the customers in the recipient unit using appropriate methods. These are listed in chapter 3.2.3. Data augmentation is always performed on an individual level, meaning that every customer receives a distinct target variable value fitting best (Rässler, 2002, p. 17). If this *best* value is augmented to the customers, it is referred to as deterministic data integration (Jiang, Sarka, De, & Dey, 2007). The decision on which target value to augment is made based on rules as stated in chapter 4.2.4.

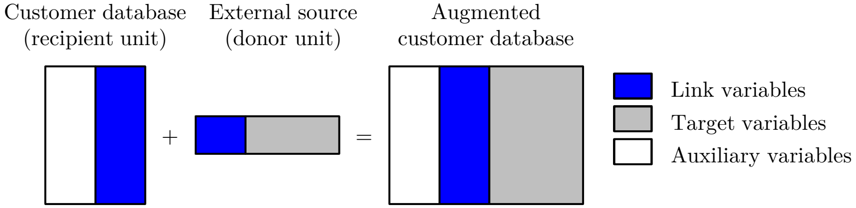


Figure 1.1: Schematic illustration of data augmentation as derived from Putten et al. (2002a, p. 2)

Data augmentation is motivated by the need to analyze data collected in different sources that cannot be observed in a single source for the customers at interest (Rässler, 2002, p. 2). It has the purpose of developing customer profiles with which conversion probabilities can be predicted (Ratner, 2001b). Data augmentation results are not preferable to single source data. Data should be collected directly from the customers whenever possible. But augmentations are a notable alternative whenever single source data is not available or unreasonably difficult to obtain (D'Orazio et al., 2006, p. 1; Kamakura & Wedel, 1997).

The sources are augmented in order to receive a file containing all variables from both sources (Gilula et al., 2006; Rässler, 2002, p. 16f; Rodgers, 1984). The result of data augmentation is a rectangular dataset with information on all variables and elements (Rässler, 2004). The datasets resulting from data augmentation are complete, i.e. every customer is contained in it, concise, i.e. every customer is contained only once, and consistent, i.e. all variables have the same concepts and definitions (Bleiholder & Naumann, 2008). The resulting artificial data can be used like real data. Decisions can be made based on individual values. Depending on the strategic objective, inferences and distributions can be calculated. There is no need to know all the analysis objectives when carrying out the augmentation. However, augmented information should always be labeled and treated as such.

There exists no standard terminology for the act of adding data from external sources to the customer database. The process described in this study most closely resembles the approaches of Putten (Putten et al., 2002a, 2002b) and Hattum and Hoijtink (2008a, 2008b), who used the term data fusion. Data fusion is an umbrella term used in several branches of research, referring to systems that match data in different ways (Arnold, 2011). Additional information is necessary to specify what kind of data fusion it is referred to. Data fusion is oftentimes used in the context of statistical matching, where the focus is on creating one dataset from formerly two, with information from one dataset complementing the other and vice versa. Statistical matching occupies a partially different problem, highlighting the special case in which inference is made on variables never jointly observed. As both do not exactly describe the situation of database marketing, they are misleading terms.

Several other terms are used to describe the event of matching data, among them data enrichment, data integration, file concatenation (D'Orazio et al., 2006, p. 1; Rässler, 2002, p. 2), deterministic data fusion (Breur, 2011), data augmentation, and data enhancement. The meaning and the respective focus of the most relevant terms is given in table 1.1. While data enhancement focuses on the fact that data is advanced or developed, usually in a process evolving over time, data enrichment is mainly based on improving data by refinement. They both have in common that there is not necessarily a relation to existing data in the database. Data imputation relates to unintentionally missing data that needs to be substituted. It is about filling in gaps rather than actually adding new information.

Term	Meaning	Focus
Data augmentation	Adding new information to an existing database	Supplementing information that is already there
Data enhancement	Increasing the quality or value of the data	Advancing, usually in a development process, evolving over time
Data enrichment	Improving the quality, value, or extent of data	Refining, usually in the context of raw data
Data fusion	Matching two databases	Creating one dataset from formerly two, with information from one dataset complementing the other and vice versa
(Missing) data imputation	Completing missing values	Filling missing values with meaningful substitutes
Data integration	Combining sources to form a better overall picture	Merging databases, usually with a unique identifier or with equivalent information
Scoring	Building a model for existing data and applying it to new data	Detecting structures at one point in time and predicting values for cases appearing later

Table 1.1: Data augmentation terminology

Data integration and scoring are terms borrowed from different contexts that could also fit the database marketing situation because of their similar techniques. However, data integration is a more general term, embracing all sorts of data matching situations, including exact matching, record linkage, and improving existing data by comparing equivalent information of different sources (Guigó, 2012). The method of scoring mostly refers to regression techniques, where a model is built for existing data and applied to new data. It is about detecting a structure at one point in time and predicting values for cases appearing later, but belonging to the same customer database.

Data augmentation is the most suitable term for the act of adding data from external sources. The term augmentation file has already been used by early practitioners of data fusion (Radner, 1980). It focuses on supplementing information that is already there, which fits the context of database marketing, where the new information is always regarded in conjunction with the information already available. In fact, the new information is only a clarification of information that has already inherently existed in the customer data. Like in augmented reality, existing information is supplemented and the complete picture is examined (Azuma, 1997).

The term data augmentation is also more narrowly used in another branch of data fusion research. It is based on Bayesian statistics and iterative simulation methods using expectation maximization algorithms (Dyk & Meng, 2001; Little & Rubin, 2002, p. 2; Tanner & Wong, 1997). While likelihood based approaches and modeling techniques are feasible for data augmentation in marketing as well, it is neither limited to the data augmentation idea of the mentioned researchers, nor does it fully comprise it.

1.2.2 Specific characteristics of external information

External information in the context of database marketing comprises every data source that is not directly collected and saved on a personal level with the existing customer data. It is defined by not having unique identifiers for the customers. External information can be created to be augmented to the customer database, like a volunteer survey. It can also be a source publicly available and accessible, such as market media studies or social media data. Because external information sources are manifold, some examples are given.

The desire to use external information arises, when companies advance into business fields deviating from their core business, such as cross-selling or the introduction of new products. It also emerges, when other information than the existing customer transaction data is necessary to delineate a target group, such as preferences, needs, and wants. If a publisher's core business is to sell local newspapers, a new cross-selling idea might be to distribute a special interest magazine. It is possible to approach the target group in a way that demographic criteria are met, but the special interest in the magazine content cannot be derived from the existing customer database. If the publisher further has a small online shop and wants to introduce an entirely new product, such as local concert tickets, the same problem occurs. And if an advertising client of the newspaper wants to issue a special jewelry supplement to couples before Valentine's Day, the publisher has trouble targeting these.

The publisher could engage a specialized market research provider to find answers to the open questions, such as tns infratest (2012) in Germany or The SmartAgent Company (2013) in the Netherlands. They have the means of conducting representative surveys, based on an elaborate methodological set-up, and afterwards performing an augmentation. By conducting a dedicated survey, the goals of the marketing department can explicitly be addressed and the sampling frame can be chosen accordingly. However, there are disadvantages to this approach. The data fusion offers of market research providers are very costly. The expertise in this field is rare and prices are set accordingly. Furthermore, the process is very time consuming. Surveys are one-time activities, with insight not being updated. Accordingly, the data utility is not very high.

Before the publisher decides to engage a specialized market research agency, he could check whether the questions can be answered by other sources available. Special interests and competitive papers in the field of magazines are well reported in nationwide market media studies. In Germany, examples of such surveys are the *Communication Networks*, the *Typologie der Wünsche* (Institut für Medien- und Konsumentenforschung, 2012a, 2012b), and the *VerbraucherAnalyse* (Axel Springer AG, 2012). Market media studies have a focus on media usage and product information, and are representative for the German population.

In order to get insights into the target groups of local concert tickets, the publisher could conduct a volunteer survey online. Short volunteer surveys are common practice nowadays and can grant insights into a self-selected forthcoming subgroup of customers. The information is not predefined by a provider, like in a market media study. Instead, it can be defined by the interviewer. Datasets are available in raw form and can be analyzed in any thinkable way. These surveys are not representative and might not cover the whole customer base.

For the jewelry supplement, the information on the relationship status is of interest. Social media sources contain such information. Social media platforms offer application programming interfaces through which individual information can be exchanged with the networks. The publisher could create a social media application with a registration and permission process. The permissions grant access to personal profile information, liked pages, interests, activities, relationship status, and much more. Information is particularly honest and up-to-date. Even locations can be derived from there. The observed group is usually partially overlapping with the customer group and is not representative.

Depending on where, when, from whom, and for which purpose the data was collected, external data sources differ. They cannot be treated equally and sometimes it is not even advisable to use them at all. Resuming above examples, a market media study, a volunteer survey, and a social media source have different characteristics. A market media study is representative for the overall population, e.g. Germany. Some of the publisher's customers (a representative sample of them) are likely to have taken part in the study, but it also contains other people not relevant to the customer database. A volunteer survey is naturally restricted to online contactable people who visited a website in a certain time period and thus not representative. If it is distributed through the publisher's website, the overlap is high, whereas it is small, if it is placed on another website. Social media sources are known to cover a great portion of people, but definitely omit those not reachable online, those skeptical in terms of data privacy, and those whose circle of friends is not affine for social media. It can be collected for a subgroup of customers or a partially overlapping group. Whenever a source is not originally meant to be used for data augmentation, some effort is necessary to prepare the data.

Whenever information relevant to a direct marketing problem is available, it is desirable to use this external data for data augmentation. But does the overlap of customer database and source matter? Does it disturb the augmentation if people irrelevant to the customer database are included in the source? Answering these questions is a contribution of this study. Once the specific characteristics of such data augmentations have been studied on a theoretical level, a conceptual model is built identifying crucial factors for the quality of the data augmentation results. Exit criteria are defined excluding unsuitable sources from being used for data augmentation.

1.2.3 Independence of source and target variables

Sources like volunteer surveys, social media, or market media studies are easy to access. Elements do not have to be identical, because there is no need to find the one correct match in the source, as long as there are donors being alike. This is attributed to the assumption that both recipients and donors are sampled from the same overall population, so that they exhibit the same relationships and correlation structures (Radner, 1980; Rässler, 2002, p. 3; Rodgers, 1984). A population in this context refers to a unit of elements, i.e. people, conforming to a set of defined criteria. The population is the unit on which statements are made, although only a subgroup or sample of it has been observed (Powell, 1997, p. 66). Here, it is referred to as overall population to contrast it against the source and customer populations. It is equal to the overall market interesting for a company and should not be confused with the more narrow definition of a country population. Most of the external sources are nonprobability samples , i.e. the elements have different probabilities to be included and some do not have the chance to be included at all (Powell, 1997, p. 67). In contrast, a random or probability sample is a form of sampling, where every element has the same known probability to be included in a sample (Powell, 1997, p. 70).

Such sources can cause problems not always directly obvious. For example, a volunteer survey might ask for interest in shoes. A volunteer survey is also referred to as self-selected sample, convenience sample, or accidental sample (Hudson, Seah, Hite, & Haab, 2004; Powell, 1997, p. 68f). It is assumed that this volunteer survey was spread by a renowned online shopping portal with its key selling driver being clothes. The survey group is sampled from the internet population using the online shopping portal. Thus, not every element in the overall population has the chance to be included. It is further assumed that the reached audience has a tendency to be interested in shoes above average. When using the volunteer survey for data augmentation purposes with "interest in shoes" as a target variable, it can be suspected that the number of recipients having been attributed an interest in shoes will be above average. This example is explained in more detail in with according calculations in chapter 4.1.4.

The notion of the source data mechanism refers to the mechanism describing whether a person has been observed in an external source or not, e.g. whether a person has participated in a volunteer survey. Beyond the question *if* somebody has participated, the question *why* somebody participated can be of interest, if this *why* influences the answers of the survey. In the example above, the source data mechanism is not ignorable and the augmentation results can be biased. The central question for all augmentation problems with external sources is whether the source data mechanism can be ignored in a way that it does not bias the augmentation results.

The link variables on which the data is augmented are also of importance. For simplicity reasons, the data in our example is augmented by a single link variable: gender. It might turn out that the high tendency of the survey participants to like shoes is attributed to the fact that many women took part in the survey, who are more interested in shoes than men. This explains the high portion of shoe-lovers, rather than the fact that so many shoe-lovers shop on the online platform. Because the data is augmented based on gender, this effect is corrected for with women receiving only data from women and men respectively. Then, the source data mechanism (participation in the volunteer survey) and the target variable (interest in shoes) are conditionally independent given the link variable (gender). In that case, the source data mechanism can be ignored. One of the theoretical contributions of this study is to assess and analyze the influence of the source data mechanism of external sources on data augmentation results.

1.2.4 Delineation of the research subject

In order to define the focus of the study, it is important to highlight related fields which are not regarded. Data augmentation is a database marketing approach for analyzing "small" data and detecting new information by augmenting external data on a personal, but not exact level. Results are used individually in order to improve the data basis in business-to-consumer (B2C) communication with the goal of strengthening customer loyalty. It supports the objective of database marketing, in close relation to the goals of targeting and direct marketing.

Database marketing

Database marketing is defined by Shaw and Stone (1988, p. 3f) as "an interactive approach to marketing, which uses individually addressable marketing media channels [...] to extend help to a company's target audience, to stimulate their demand, and to stay close to them by recording and keeping an electronic database memory of customer, prospect and all communication and commercial contacts, to help improve all future contacts and to ensure more realistic planning of all marketing". As suggested by the name, the focus of database marketing is on data and on how marketing performance can be improved by using data.

Database marketing, CRM, and direct marketing depict forms of customer oriented marketing. In contrast to database marketing, CRM establishes, optimizes, and retains lasting and profitable customer relationships (Hippner, Leber, & Wilde, 2002), whereas direct marketing comprises all marketing activities and communication channels that target customers individually (Dallmer, 2002). While every area interacts with the others, the perspectives are slightly different. Database marketing focuses on data analysis and data usage, whereas CRM focuses on the customer relationship as a whole and direct marketing focuses on the implementation of targeted campaigns (Blattberg, Kim, & Neslin, 2008, p. 6).

Link and Hildebrand (1993, p. 30) use the RADAR model to describe the database marketing process. Database marketing starts with research (R), where the current situation is analyzed and the goals and methods of a database marketing project are specified. The second step in the database marketing process is the analysis (A) of data collected according to the objectives. The results of the analysis enable the database marketing analyst to detect (D) potential chances and risks, and to propose specific actions (A). From the reactions (R) of the customers, new insights can be attained in order to start another research. The aim of this study is to propose a new method with which the analysis of data can be improved and from which differentiated segments for campaign actions can be detected (AD).

Targeting

Targeting comprises all activities differentiating customers and providing them different suitable offers. Marketing communication is distributed to all customers during their customer life cycle in order to trigger desired customer actions. The focus of this study is on loyalty and retention, rather than acquisition or reactivation. Oftentimes, it is much more efficient to retain existing customers than to acquire new ones (Höhl, 1999; Woo & Fock, 2004) or reactivate inactive customers (Heun, 2002). The purpose of retention is to boost the purchase frequency, to minimize the perceived purchase risk, and thus to decrease the price elasticity (Meffert & Bruhn, 2009, p. 458). The customer value increases with the duration of the customer relationship (Reichheld & Schefter, 2001). It does not make sense to invest in all customers. Rather, every customer relationship's costs and benefits are analyzed in order to identify profitable customers in the long run. Although the concepts depicted here can be transferred to other forms of marketing, including business-to-business (B2B) marketing, this study focuses on the B2C marketing market. This is mainly attributed to the fact that information on end customers is much easier accessible and from different sources than information on business customers.

In order to improve a company's conversions, several parameters in the marketing strategy have to be regarded. Targeting finds answers not only on *whom* to contact (data basis), but also with *what* kind of product or offer, via *which* medium or channel, and at what *point in time*. The form of *how* an offer might look is of importance as well. These parameters interact and cannot be regarded separately, although it is possible to identify factors that are more important than others (Bult & Wansbeek, 1995). Only if all parameters are attended and improved, the overall marketing strategy is prone to succeed. The here proposed improvement of the data basis is only a starting point in improving the targeting strategy.

Targeting can be performed on personal level, internet protocol (IP) based level, media level, or location level. Targeting on personal level requires identifiable and describable customers. In online targeting, the targeting is mainly based on IP addresses and surfing behavior and requires an automated algorithm able to compute and deliver information accordingly. Predictive behavioral targeting combines online surfing behavior and results from online surveys in order to predict preferences for online advertisements (Noller, 2009). Targeting on media level is performed by media planners and results in the right choice of relevant media and slots. Geo-targeting is becoming increasingly popular, as customers are knowingly reachable at different locations via mobile devices (Dialog Marketing Monitor, 2012). From micro targeting, it has long since been known that customers living in different areas differ in terms of their personal incomes and other hard facts (Putten, 2010, p. 84). Sometimes, targeting is defined more narrowly in a sense that different customers or prospective customers are offered different prices for the same product (Feinberg, Krishna, & Zhang, 2002). In our sense, this is one aspect of targeting among others.

Data augmentation

Data augmentation or data fusion in a broader sense comprises all activities which add new information to an existing database. Depending on where this information comes from, how it is added to the database, and how it is analyzed later, different forms of data augmentation can be differentiated. Data augmentation can serve several purposes. It is possible to combine sources containing similar information in order to construct more detailed values of variables, to subjoin elements that were addressed in different sources, and to add new variables to existing sources (Bakker, 2012; Guigó, 2012). In our study, we focus on the latter.

The task of database marketing is to identify data that is relevant for customer analysis and to use multivariate analysis methods and data mining tools to identify and differentiate segments (Schmidberger & Babiuch-Schulze, 2009). Although both data augmentation and data mining are tools for database marketing, they differ. The goal of data mining is to develop models that independently identify meaningful patterns in big sources of data (Hagedorn, Bissantz, & Mertens, 1997). It aims at discovering new insights within one single database. Contrarily, the aim of data augmentation is to combine two databases in a way that a new database is generated with additional information. It can be a preliminary stage to data mining, supplying more data to mine in (Putten et al., 2002b, p. 1).

The term exact matching, or associative data fusion (Breur, 2011), relates to information that is matched on the basis of a unique identifier present in both databases (Rodgers, 1984). The term data augmentation relates to information that is matched based on link variables and has a predictive nature. In the context of data management, exact matching is also called merging or joining. Unique identifiers could be the social security number or name and address (Rässler, 2004). Exact matching identifies individuals, whereas data augmentation identifies similar customers. In a way, exact matching is an idealistic form of data augmentation, because it matches explicit information with 100% certainty. The goal of data augmentation is always to receive a match as close as possible to these attributes.

Record linkage and statistical matching are differentiated here due to their similar, but yet different focuses. According to Cibella, Guigó, Scanu, and Tuoto (2012), record linkage is the act of fusing two sources with identical elements. In contrast to exact matching, a unique identifier is missing or the unique identifiers are deficient. The task of record linkage is to find the best, if possible true, match. By contrast, statistical matching is the act of fusing two sources without any overlap or with the overlap being negligible (D'Orazio et al., 2006, p. 2; Rässler, 2002, p. 3). It is usually assumed that the two sources were independently and randomly sampled from an overall population (D'Orazio et al., 2006, p. 3; Rässler, 2002, p. 20), so that the overlap does not matter and does not influence the data fusion results. There has been some literal confusion on the definitions of record linkage and statistical matching in the data augmentation literature. D'Orazio et al. (2006, p. 2) refer to record linkage, when the elements are at least partially overlapping and only those of the overlap are sought to be found. On the other hand, they also refer to statistical matching, when the observations are identical and a unique identifier is missing. Rässler requires the overlap between elements for statistical matching to be "at least small, if not zero" (2002, p. 6). None of the research conducted so far explicitly examined the case in which the data augmentation source is neither identical (as in record linkage) nor independently and randomly sampled from an overall population (as in statistical matching).

Data augmentation can have a micro or macro objective. The micro approach has the objective of generating a new value for every observation in the dataset. The result is a synthetic dataset that can be used like a primary source, containing information on all observations and all variables. The macro approach has the objective of obtaining the joint distribution among variables that have not been jointly observed. A macro view on the data can usually only be calculated from micro data, which is why the two approaches are not necessarily applied separately (D'Orazio et al., 2006, p. 2f). In database marketing, the micro approach is of primary interest. Moreover, the preservation of distributions is not always desirable, as the main focus is on generating the best possible value for each customer. This is not necessarily associated with the idea of generating values in a way that the marginal and overall distributions are preserved (chapter 4.1.3).

Big data context and differences

In order to contrast our data augmentation approach against the term big data, a circumscribable definition of big data is needed. We refer to one of the most recent and considered books describing big data by Simon (2013). The philosophy of big data is to use all available data, especially from new technologies at a scale exceeding all up-to-date volumes, and differentiating between relevant and irrelevant information (Simon, 2013, p. 4ff). The essence of big data is the capability of handling unstructured data (Simon, 2013, p. 35). Unstructured data comprises all forms of poly-structured data (e.g. non-relational data or text), semi-structured data (e.g. XML), and meta data not representable in traditional relational databases (Simon, 2013, p. 32ff). Unstructured data is mainly created outside of a company, while internal data still adhere to traditional relational structures (Simon, 2013, p. 39). However, it does not replace, but only complement traditional data management (Simon, 2013, p. 55f).

More precisely, big and small data differ in terms of condition, location, and population. If data is clean and ready to process, it is considered well-conditioned. Location describes the residence of data tables in relational databases, which can be a single rectangular dataset or many different tables respectively. Population refers to individuals in the database and their characteristics. They can be either a random sample or non-random samples, be primary (collected for the marketing goal) or secondary (not collected for the intended purpose), and be stable or unstable. Big data is characterized by being ill conditioned, located in many different tables, and having secondary and unstable features (Ratner, 2003, p. 8f).

While big data necessarily requires database structures and tools able to process tremendous amounts of data, the term big data does not refer to the volumes alone. So-called small data, as present in relational customer databases, can still be big in terms of data volumes (Simon, 2013, p. 55). This understanding has changed to earlier definitions of big data. Ratner (2003, p. 8) defined big data by the number of observations being analyzed, where datasets with 50,000 individuals is considered big, and a dataset of up to 200 individuals is considered small.

From the definition, it becomes clear that data augmentation in our sense is not a big data issue. We solely rely on relational databases. All sources available must meet such a structure or must be pre-processed and aggregated in order to fit this frame. Nevertheless, some intentions of big data and data augmentation usage are similar. Data augmentation copes with data from different tables and the sources can have ill conditioned, secondary, and unstable features. Augmentation techniques can also be used in a big data context (in fact, they should). However, data types, data structures, methods, and tools used in our study are those handling small data in the sense of Simon.

1.3 Research approach

The goal of our study is to understand the characteristics of external sources and to explore the influence of these characteristics on the quality of data augmentation results. A case study approach with simulated missing target variables has been chosen in order to answer the research question. Thereby, it is possible to give answers to questions not answerable in practice. We shortly summarize the definition of a case study and explain the data origin and the characteristics of the data basis. An evaluation of the suitability of the case study approach is given. It comprises general requirements for answering the research question so as to overcome certain limitations of the case study approach, a comparison of three alternative approaches that could have been chosen, as well as a motivation why the case study approach is the best research environment possible for our intention.

1.3.1 Case study with simulated sampled sources

In a case study, data augmentation is explored in a particular context in order to "retain the holistic and meaningful characteristics" (Yin, 2009, p. 4) of the data augmentation setting. Case studies in general are suitable for research problems that have to be regarded in or cannot be separated from their context (Perry, 2000). In the case of data augmentation, the customer database with its link variables is the context in which the data augmentation is carried out. Although a case study is a "detailed examination of a single example" (Abercrombie, Hill, & Turner, 1984, p. 34, as cited in Flyvbjerg, 2006), the data is rich enough to understand multiple aspects of the subject (Baxter & Jack, 2008). The unit of analysis, or case, is the unit that is measured and analyzed (Yin, 2014, p. 29). Here, the unit of analysis is the result of one data augmentation. The cases enable analysis *within* cases, *between* cases, and *across* cases. That way, a holistic picture can be obtained, while associations between variables, as well as influences of certain parameters can be regarded (Baxter & Jack, 2008).

The modification of a case study is easier and cheaper than that of real world systems. Various approaches can be adapted and compared. Because of its reduction and simplicity, the implications of modifications are easily analyzable and interpretable (Dekker, 1993). It can even be carried out for situations that have not yet been established in the real world (Albright, Winston, & Zappe, 2011, p. 919).

The data augmentation situation with missing target values is simulated so that the results can be compared to the true values and derivations can be made for practical applications. In real world applications, the true values are not known. If the situation of missing target variables is simulated, a hit rate can be calculated and it can be compared to the values that would have been obtained when augmenting data by chance, given the existing target variable distributions. That way, the results of a case study can be evaluated internally. Internal evaluation refers to validating the augmentation results and derived KPIs in comparison to the true values. External evaluation refers to assessing the utility of the results in terms of return on marketing investment. Different options and characteristics of sources can be compared and overall tendencies can be observed.

The cases are chosen with a theoretical sampling (Eisenhardt, 1989) or *information-oriented* sampling (Flyvbjerg, 2006) approach. Sometimes, the term sampling is defined more narrowly as random sampling. In our study, the term sampling refers to the fact that various sources are chosen based on feasible and available combinations of source characteristics. The data augmentation sources to be tested have different characteristics resembling real world cases, such as an online source, a social media source, a representative customer survey, or a market media study. We perform data augmentation with these sources for various target variables and differing methods. From these multiple variations of a data augmentation situation, valuable insights are derived. When conducting a case study, it is assumed that the selection of particular cases offer more interesting and illuminating insights than if choosing cases randomly (Flyvbjerg, 2006).

The data for the case study is a real-world sample from the customer database of a renowned German company. The name of the company is omitted due to data protection and confidentiality reasons. The real-world origin guarantees realistic distributions of variables and correlations among link and target variables and varying source data mechanisms. Observations are anonymized and variables are pseudonymized. No personal data such as name or address information is used. The data basis is of sufficient size, so that several samples can be drawn and the augmentation results still offer adequate measures from which conclusions can be drawn. This is described in more detail in chapter 5.2. In order to receive meaningful results for the different data augmentation versions, a fully rectangular dataset is used for the population, without missing values.

Link and target variables are defined from expert knowledge and depending on the context of the available data. They are shown in chapter 5.2.

The link variables comprise socio-demographic information like age, gender, and residential region, as well as seven behavioral and preferential variables. The target variables comprise socio-demographics variables, like income or general propensity to buy, as well as behavioral and preferential variables from three different branches and nine different products. The information present is reduced to categorical variables, as if the information came from external sources, for example market research. For source protection reasons the real variable names are changed to generic titles. A comprehensive number of variables and observations is needed in order to perform different augmentations. Thus, the number of variables and extent of information is bigger than it would be in real world applications. The case study is carried out using SAS 9.2 business analytics and business intelligence software.

1.3.2 Suitability of the case study method

We have chosen a case study design with sampled sources in order to answer the research question. There a certain limitations related to a case study design, as well as to sampling semi-artificial sources, i.e. sources that would be possible, but are not actually derived from a different dataset as in practical data augmentation applications. In the following, we argue how the case study approach is equipped with enough detail and diversity to give general and transferable insights into the research question. We show how it outperforms other possible research approaches.

General requirements

There are general requirements regarding the study design for answering the research question and the hypotheses to be tested. Data augmentation is afflicted with the fact that it can never be known whether the augmented values are true, unless the customers are directly asked. This is usually not feasible. Only if marketing campaigns using data augmentation results perform better than before data augmentation, it can be assumed that the augmented values are at least partially correct. Thus, in order to give indications on the quality of the augmentation and to answer the research question, a situation must be created in which the true values are known.

The study design must comprise several sources and target variables in order to enable between and across case analysis. Each source must be of sufficient size, so that many link variable classes are available with a significant number of donors representing each class. These sources need to differ regarding their characteristics. In order to deduct general statements, they should vary on the whole range of possible values. Different target variables need to be augmented from each source, so that the effects of different target variables can be examined.

A correlation test of source and target variables must be possible in order to assess the source data mechanism type as described in chapter 4.2.5. In a practical application, this could be achieved by an auxiliary source, i.e. a representative survey asking for all variables relevant to the data augmentation situation (link variables, target variables, customership, and source usage). In the study design, this problem can be overcome by creating the sources instead of using existing ones. Then, the sampling mechanism must be carefully chosen in order to resemble real-world sources as much as possible.

Finally, the results of the augmentations with several sources must be consistent and comparable, as well as generalizable. Consistency refers to the reliability of the study set-up. Comparability is achieved best, if all variables not in the focus of the study are kept equal. Analytical generalizability refers to the ability to apply the results to any other data augmentation use case (Yin, 2014, p. 38). However, a trade-off exists between the restriction of the study to certain settings in order to achieve comparability and giving insight into the general applicability of the results.

Alternative study approaches

There are different possible ways to approach the research question: a series of real-world data augmentations with appropriate documentation, a case study with sampled sources, or a full simulation of the model frame. The three alternative approaches differ in various ways. Their properties are illustrated in table 1.2 and afterwards described in more detail.

Property	Documentation of real-world augm.	Case study with sampled sources	Full simulation
Data basis	several data sources	one data source	no existing data source, but simulation of it
Data quality and richness	poor	rich	to be designed
Variability of recipient unit	possible	not possible	possible
Link and target variables	all different	same for all	to be designed
True values for target variables are known	no, only determinable through customer survey	yes	yes
Distributions	realistic	quasi-realistic	artificial
Sampling of sources	taken from real world	information-oriented	random
Diversity of sources	low	high	very high
Calculability of source data mechanism	only possible from auxiliary data	possible (ex post)	possible (ex ante)
Controllability of source characteristics	no control	partial control	full control
Number of observations for comparison	limited to less than hundred	thousands	quasi unlimited
Comparability	low	high	high
Consistency	low	high	high
Transferability to practice	possible	possible with limitations	difficult
Costs	high	low	medium

Table 1.2: Alternative study designs for answering the research question

Our research question would not be solvable by other research methods, such as a survey, an experiment, or an analysis of historical data. A survey, as well as analysis of historical data, would only be feasible, if data augmentation with external sources was already carried out. It is, however, not commonly used in business yet. An experimental setting would be too artificial. It would not be easy to imitate the link and target variable types available in practice and their relationships. Previous studies conducted for data augmentation in database marketing were primarily carried out on a case study basis (Hattum & Hoijtink, 2008a, 2008b; Krämer, 2010; Putten et al., 2002a, 2002b), although they did not use sampled sources. Rather, they only referred to one specific example.

Documentation of data augmentation series from real-world A good way to answer the research question would be to conduct a very high number of data augmentations, document parameters and results of each, and make an aggregated statement afterwards in order to answer the research question. Marginal and joint distributions, as well as source data mechanisms, would be realistic and to analyze a wide range of real world data augmentations would do justice to the superior goal of generalizability and transferability. However, the range of possible data augmentations is endless and to define a study as big as to cover it is virtually not possible. There are also a number of conceptual and practical reasons why conducting hundreds of data augmentations is not feasible.

First and foremost, the true target values are not known. A strategy to overcome this preclusion could be to conduct a customer survey after every augmentation in order to receive the true values for comparison purposes. Without even regarding the methodological obstacles related to this approach, it would probably not be possible to receive answers of *all* customers. The same applies to the calculation of the source data mechanism, which would only be possible from an auxiliary source. It would hardly be possible to receive sources with all types of mechanisms. Because the data augmentations would be performed at different points in time, the consistency would be low. Although it would theoretically be possible to regard different recipient units, this would even further stretch the model frame and necessary examinations.

There is also a simple practical problem: to conduct such an extensive series would not be feasible from a cost and time perspective, because it would take years to conduct it, with inestimable costs related to it. The number of augmentations would probably be limited to less than a hundred, involving several data sources with the related data preparation and harmonization effort. Consequently, although theoretically the best way to gain insights into realistic augmentation problems, the number of augmentations would be too low to make any sort of substantiated general statement, while the consistency would be low.

Case study with sampled sources A case study with sampled sources overcomes the problem of unmanageable ranges of applications by restricting the study frame to one population and one recipient unit, while at the same time regarding several forms of donor units. In order to establish comparability, the basic setting, like link and target variables, is controlled. The study frame has to be chosen in a way that it is meaningful and largely transferable to other, at least similar, situations. Flyvbjerg argued that detailed, context-dependent knowledge of the researcher can be even more valuable at times than "predictive theories and universals" (2006, p. 224). While the effects of the variables altered in the case study can be generalized, the results cannot be transferred to other contexts with different overall options. However, if the case is carefully chosen and has an extreme or critical character, it increases its generalizability (Flyvbjerg, 2006).

Because the data used for case study purposes is a real world dataset, the marginal and joint distributions are quasi realistic. The quasi restriction is made, because the donor units are not taken from real world, but sampled from the overall population based on the requirements of the study. In a case study, variables are not randomly manipulated like in a stochastic simulation (Yin, 2009, p. 11f). The information-oriented sampling approach is an advantage, because the range of possible sources can easier be covered when choosing sources based on desired categories. Additionally, an unlimited number of sources can be sampled from the overall population, leading to a wide range of results, which enable a good basis for generalization.

Due to the case study set up, the true target values are known and the source data mechanisms are estimable ex post, i.e. by comparing the resulting source to the target variables. This is the main reason to choose an artificial study set up over a documentation of realistic augmentations. Because all augmentations are derived from and applied to the same database, comparability and consistency are high. Finally, the costs to conduct the study are reasonable.

Full simulation A full simulation of the data augmentation situation provides more flexibility than the previous alternatives. From table 1.2, it can be seen that recipient unit, donor units, link and target variables would be formable tailored to the need of the research question. Existing known marginal and joint distributions could help to form close to realistic datasets. The source data mechanisms would be created based on prior information. A calculation would thus not be necessary anymore, because it is ex ante defined which source data mechanism is modelled. With multiple imputations, the whole range of mechanisms and target variable would be realizable. Just like the case study approach, the simulation approach is comparable, consistent, and economical.

However, the flexibility of the full simulation approach can also be a drawback. As data augmentation in database marketing deals with real people, the marginal and joint distributions are very complex, difficult to predict, and full of deviations and unexplained errors. The link and target variables only capture a small portion of all relationships. To artificially simulate all these relationships from scratch is difficult, if not impossible. Although millions of augmentations would be possible, the gist of the results would only be correct, if distributions and correlations were chosen in a way resembling the reality extraordinarily well. The data could easily be too clean and thus the findings might overestimate the possibilities in practice.

If this fit to reality is not mastered, the results cannot be transferred to practical applications and would not have any value.

Creating the best research environment possible

From the previous evaluation of alternatives, it can be seen that both the documentation of data augmentation series from real-world and the full simulation have drawbacks prohibiting a meaningful use. In the former approach, two central calculations are not possible or only with a disproportionate effort: the calculation of the hit rate and the assessment of the conditional association between source data mechanism and target variables, given the link variables. In the latter approach, all variables and samples would have to be simulated, leading to an unmanageable amount of possibilities in terms of distributions and relationships, while not knowing how to simulate human characteristics and behavior best. In the case study approach with simulated sources, both problems can be overcome by using a real world dataset with realistic distributions and correlations. The missing data situation is simulated by taking away the target variables from the recipient unit, only to augment them thereafter and compare them to the true values.

Case studies cover the research questions for the given context in a comprehensive way in order to generalize it to the whole unit of analysis (Yin, 2014, p. 25). The generalization is more self-evident, if the case study is sufficiently broad, so that many other contexts at least resemble the case study context. For example, if only data from one branch is regarded, the generalization to other branches is questionable. Therefore, we use real purchasing data from four different branches and various consumption categories in our case study. Another drawback would be, if the population and customer structure of the case study is very different to other applications. To that end, our population is stratified to represent the German population in terms of gender and age, as it is described in chapter 5.2.1. That way, all demographic strata are examined.

Our data is especially rich in quantity, so that many different sources can be sampled and many different augmentations can be regarded. We use sampled donor units representing all kinds of sources, but also identical, partially overlapping, and distinct sources. For every source data mechanism, we examine nine different data augmentations sources, in order to foreclose the risk of one deficient or peculiar source influencing the overall statements. We also sample sources for which the suitability for data augmentation has already been proven, in order to contrast the results to the sources that have not been examined in detail yet.

Nevertheless, by conducting a case study, we accept certain limitations regarding the analytical generalizability. The results of a case study always need to be regarded in the context of the study. The more a potential context differs from the chosen case study context, the lower the certainty that a data augmentation in this context exhibits the same features. If different parameters are applicable, the results cannot be transferred (Robinson, 2004, p. 11). We are aware of the fact that all decisions made in building the case study influence the augmentation results. However, our goal is to compare different characteristics of sources, which is a relative objective. In a real world application, these relative tendencies are of interest in the *respective* context, while absolute values can differ.

We strongly believe the case study approach can give valuable insights into the problems related to data augmentation in practice. Data augmentation is very costly and time-consuming. It requires advanced database marketing skills and is a decisive investment, which is only approved by the management, if a monetary success can be anticipated. At the same time, the variety of sources and their individual properties need to be handled with suitable methods. Every data augmentation approach is unknown territory and uncertainty of effectiveness and efficiency is high. These obstacles are common reasons for abandoning the data augmentation idea at early project stages. With this study, we shed light on different data augmentation sources and on how their characteristics influence the data augmentation results. The guidelines to be developed enable data augmentation decisions to be made faster and at lower costs. To provide a starting point for this examination, the case study approach with sampled sources as proposed here is the best and most feasible research method. It can overcome obstacles like the unknown true target values or the inestimable source data mechanism. Furthermore, it can give a comprehensive insight into the general influence of source characteristics on the augmentation results, which cannot be derived from practical sources that are not diverse and numerous enough to allow for such general insights.

1.4 Structure of the paper

Our study is divided into seven parts. The first two chapters describe the foundation and relevance for our study. In chapter 2, the strategic motivation for data augmentation in marketing is derived from an analysis regarding the strength and weaknesses within the company and the opportunities and threads arising from the external environment. Herefrom, the managerial necessity for data augmentation in database marketing is derived, which poses the starting point for our research. In academia, the problem of data augmentation in database marketing has only been regarded sporadically or on a universal level. In chapter 3, a literature review on data augmentation is given, retracing the evolution of data augmentation studies in different fields and demonstrating the research work already done regarding the process of data augmentation in marketing. Chapter 4 contains the theoretical contribution of our study. We explain the specifics of data augmentation in marketing and describe the data augmentation model mathematically. From the established theory, a conceptual model and test design for the case study approach is derived in chapter 5. With this set-up, a series of augmentations is carried out and the results are documented for an overall examination. The analysis of results is divided into an analysis of the general data augmentation key performance indicators (KPIs) derived from the case study (chapter 6) and an overall analysis of results and hypothesis tests (chapter 7). We apply existing and develop new KPIs for the assessment of data augmentation results in a simulated setting, where the true values are known. These measures are the methodological contribution of our study. The overall test results and findings form the managerial contribution of our study, which are summarized in a practical guide on how to upfront assess possible data augmentation sources regarding their aptitude. We conclude our study with appropriate limitations to our study and data augmentation in general (chapter 8) and a summary of the study findings (chapter 9). In this chapter, we substantiate our approach and structure.

Our study begins with an analysis of the context for data augmentation in database marketing. We inspect internal conditions within the business organization, including marketing goals, targeting in marketing practice, and conversion as the crucial marketing measure (chapter 2.1). From these conditions, certain needs arise that are not always fulfillable with traditional marketing tools. At the same time, the company environment poses chances and challenges (chapter 2.2). The economic framework of data augmentation expediates the usage in database marketing. The technological framework is an enabler, while certain constraints are derived from the legal framework when it comes to using personal data. The sociological and psychological framework regards the perception of data augmentation from a customers' perspective, which in turn has implications for data augmentation. Opportunities arise also from the sources available inside and outside an organization (chapter 2.3). Implications for data augmentation in database marketing are derived by bringing together marketing needs and available sources in an analysis of strength, weaknesses, opportunities, and threads (SWOT). From the SWOT (chapter 2.4), data augmentation is derived as a relevant direct marketing strategy for companies.

The context for data augmentation is followed by a literature review describing approaches, theories, and methods concerning data augmentation – not only in marketing. The evolution of data augmentation studies is retraced from the beginning to recent studies (chapter 3.1). Many augmentation methods are derived from traditional missing data problems and from statistical matching theory. While much literature is available on data augmentation in well-conditioned environments, none of the researchers have regarded the unfavorable conditions of the prevalent external sources available to database marketing analysts today. The literature review includes a description of the data augmentation process, which has been established by previous researchers and is recaptured and adapted for the special case of database marketing (chapter 3.2). It comprises data screening and data preparation steps, the choice of the best data augmentation method, execution, and internal and external evaluation of augmentation results.

As data augmentation in database marketing has special features and conditions, the methodological framework is devised next. It consists of a description of data augmentation specifics in marketing and the data augmentation model. The specifics include the customer database as a recipient unit, possible donor units and their characteristics, as well as available variables (chapter 4.1). Special attention is paid to the conditional independence of source and target variables and the micro validity in terms of target variable values. We develop a data augmentation model, in which the data augmentation process is described from a theoretical point of view (chapter 4.2). Populations and samples, as well as variables, are differentiated. A univariate pattern approach is suggested and resulting target values and the uncertainty inherent in data augmentation are formalized. Furthermore, the ignorability of the source data mechanism is mathematically described and a restricted class of acceptable source data mechanisms is developed.

After having laid out the theoretical basis for data augmentation in marketing, a test design is established for evaluating different source characteristics and for answering the research question. A conceptual model comprises all relevant relationships, from which hypotheses are derived (chapter 5.1). Model lift effects and how they can eventually influence conversion probability lifts are described. We answer the research question by performing a case study with simulated sources. The data basis for the experiment is further described and used methods are explained (chapter 5.2).

The test design is followed by the data analysis. For every augmentation in the case study, a set of descriptive data and measures in preserved (chapter 6.1). The pre-screening phase is described, including a quality check and derived managerial implications (chapter 6.2). The accuracy and precision of the data augmentation results is evaluated by existing classification measures (chapter 6.3). A so-called model lift describes by how much the data augmentation results increase the knowledge on the customers as compared to not having that information (chapter 6.4). The final KPI of interest is the conversion probability lift (CPL), which describes by how much the conversion probability of a selected target group is increased when using data augmentation results (chapter 6.5).

In the second part of the data analysis, the source data mechanism antecedents from the conceptual model are validated and evaluated (chapter 7.1). Different tests are used to perform this validation. The final part of the data analysis comprises the analysis of influencing factors in data augmentation (chapter 7.2), the tests of the hypotheses, and an examination of which data augmentation method is used best in which data augmentation context (chapter 7.3). The chapter is finished with a practical guide for ex ante evaluating data augmentation sources (chapter 7.4). After having derived insights from the case study, we give advice on how to find relevant information, how to check the suitability of potential data augmentation sources, and how to choose a good data augmentation method.

There are certain limitations related to data augmentation in marketing, which are stated in chapter 8.1. As our study only verifies the hypotheses stated for a defined use case, further research opportunities arise from our work (chapter 8.2). They mainly concern the ignorability of the source data mechanism, a deeper exploration of steps in the proposed data augmentation process, other data augmentation opportunities, and uplift models.

Chapter 2

Strategic motivation for data augmentation

When visiting a marketing conference, it is very likely that at some point the allegory of the "Tante-Emma-Laden"[1] is mentioned. People in the middle of the 20th century shopped in such a small general store. "Tante Emma" stands for customer focused individual information. Through her personal contact to all customers, she exactly knew her customers' needs, wants, and preferences. However, "Tante Emma's" capabilities were not primarily due to her outstanding targeting abilities. Rather, mobility, variety, and competition were so limited that people did not have a choice.

Today, the competitive environment is much tougher. People are more mobile and better informed, so that loyalty has become a valuable asset. Personalized, individualized, and relevant communication has become the superior marketing goal, because marketers and managers have recognized the that a customer directed communication increases loyalty and boosts sales. Such communication is only possible, if the companies have knowledge on their customers, much like "Tante Emma". But today, the face-to-face

[1] In English, "Tante-Emma-Laden" is translated as "mom-and-pop store" or corner shop. The literal translation of "aunt Emma" personifies the seller as a friendly and knowledgeable counterpart.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_2

information acquisition process is largely replaced by digital means. Every transaction process is recorded, every scan of a loyalty card produces data, and every click on a website can be tracked. One of the main challenges today is the collection and integration of data in order to form a holistic picture and to supply customers with as relevant information as possible.

In this chapter, we analyze the strengths and weaknesses of companies regarding individual communication. We name the opportunities and threats arising from the economic, technological, legal, sociological and psychological environment. Data augmentation can be a solution to the information lacks of companies. With a SWOT analysis, we eventually motivate data augmentation as a relevant strategy for marketing and management today.

2.1 Internal conditions within the company

The internal conditions of data augmentation in database marketing are divided into the three sub parts marketing goal, practice, and measures. In the following, we describe the trend of customers having become the focus of marketing, because treating customers individually leads to a maximum profit for the company. Accordingly, knowledge about the customers needs to be acquired. This knowledge is used in marketing practice for targeting customer segments and addressing them personally. The success of marketing campaigns is measured by conversions, which quantify the success of the marketing efforts. The internal conditions are an important factor in the SWOT analysis that we conduct in chapter 2.4. From them, strengths and weaknesses of the company are derived, facilitating the benefit of data augmentation for the marketing strategy.

2.1.1 Customer focus as a major marketing goal

Only relevant marketing communication attracts the attention of customers. A customer is a person who has already had a contact with a company, e.g.

bought a product or service, has taken interest in doing so, or subscribed for promotional communication. Relevance is achieved by attractive offers for products customers are interested in – presented to the customers at a suitable point in time, when they are receptive to these offers. However, with the low costs of advertising and direct marketing today, especially for online communication, customers are often contacted with all kinds of irrelevant promotions. Customers ignore these advertisements (Cuthbertson & Messenger, 2013). If the information conveyed by companies is not relevant, the overall attention to a company's marketing activities decreases. Relevance is not only important for the company itself, but also in its competitive environment. The so-called consumer addressability, the ability to customize communication based on individual customer information, has been shown to be a competitive advantage (Chen & Iyer, 2002).

Relevance can only be achieved, if customers are treated individually tailored to their interests and needs. The perspective of marketing and communication has changed from product to customer centric (Kelly, 2007; Garg, Rahman, & Kumar, 2010). Customers are regarded from a 360° perspective: buying behavior, demographic and socioeconomic profiles, lifestyles, attitudes, and media exposure (Kamakura & Wedel, 1997). With regards to the customers, markets and target groups can be defined and identified, offers can be created and communicated, and buys and follow-up buys can be initiated (Meffert, 2000, p. 12).

The customer focus aims at fostering each individual customer relationship in a way that it is maximally profitable for the company. Loyalty is one of the key aspects, so that customer value and equity are benchmarks for the operative and strategic directions of marketing (Helm & Günter, 2006; Huldi, 2002). An important marketing goal is to retain and to develop customers. Whereas the focus of retaining customers is on preventing them from defecting to competing market participants, the focus of developing customers is to increase their value for the company. Both implicate a constant effort, especially in so-called ”non-contractual” relationships, where every new sale has to be earned (McCrary, 2009). An increase in revenue is for example achieved by exploiting an individual's willingness to pay. Exemplary actions for these strategies are cross-selling, up-selling, and increasing the efficiency of customer contacts (Ang & Buttle, 2009).

These goals can only be achieved, if companies know as much as possible on their customers (Behme & Mucksch, 2001). Data is the basis for treating customers individually in order to increase their performance and value. It is gathered, analyzed, interpreted, and used for customer differentiation by a database marketing team. An extensive database is necessary, providing all information needed to offer the best product at the best time to the best price. The optimal allocation of resources is one of the main goals of database marketing. The demand for more information, more individualization, and more personalization is in the interest of both the customers and the companies. Customers expect offers tailored to their expectance. The companies expect efficient media usage and high conversion rates.

Data augmentation is one database marketing tool to acquire the data necessary for customer differentiation. It can provide information not otherwise available regarding product preferences, general purchasing power, preferred communication channels, life-cycle related information indicating the right time to offer a product, and much more. It enables a better description of the customers, complementing the 360° view necessary to provide relevant marketing communication.

2.1.2 Targeting in marketing practice

In order to implement individual customer strategies, a direct marketing approach is needed. Scovotti and Spiller (2006, p. 199) define direct marketing as follows: "Direct marketing is a data driven interactive process of directly communicating with targeted customers or prospects using any medium to obtain a measurable response or transaction via one or multiple channels". In direct marketing, every communication event is dedicated to a target group, a selected group of prospective customers for which the cost-benefit ratio of marketing is highest. Targeted advertising is the essence of below the line marketing. In contrast to above the line media, wastage can be avoided (Bruhn, 2009, p. 191; Greve, Hopf, & Bauer, 2011).

When customers are selected for a marketing campaign, it is referred to as targeting. Targeting is the ability to differentiate customers based on data in order to provide them individual and personalized offers. The word is derived from the verb "to target", which means to reach exactly the group at which an offer is aimed. It has been a research topic since the late 1990's. Dong, Manchanda, and Chintagunta (2009) give a good overview on the first studies concerning targeting. Targeted marketing reduces the cost of production, distribution, and promotion (Bull & Passewitz, 1994). Hopf (2011) and Kelly (2008) state that targeting is valuable to consumers, because it gives them the sentiment that products are immediately available, findable, accessible, understandably prepared, and personally selected. They gain the impression that products and benefits are real and original, that someone cares for them, and that their data is safe at all times.

The vision of targeting is to develop it to a one-to-one marketing process, in which every customer receives relevant products, services, and information at the optimal point in time (Link & Hildebrand, 1993, p. 29). However, such a perfect targeting is possible with a disproportional effort (Freter, 1997, p. 46). The next best solution is to find meaningful customer segments, which can be reached through consequent data usage (Liehr, 2001). In order to profitably treat these segments, they have to be identifiable, quantifiable, addressable, and of adequate size (Rapp, 2002a, p. 67). They should be stable, responsive in a way that response is similar among segment participants, and actionable in a way that the marketing strategy is consistent with the company objectives (Hattum & Hoijtink, 2010).

Individualization is needed on every touch point. While (e-)mail is the most important channel for direct marketing (Bult & Wansbeek, 1995), many of the marketing tools today have the possibility to personalize and target (Iyer, Soberman, & Villas-Boas, 2005); e.g. newsletter, websites, and mobile or social platforms. As channels become more fragmented, with the new ones not necessarily replacing, but adding to the old ones, an overall customer view is needed (Rhee, 2010). Much of the communication to the customers is done electronically and is controlled through so-called contact optimization technologies. These are based on certain business rules and contact policies, and are mainly fed by customer data (J. Berry, 2009).

Targeting has the goal to select or segment customers. *Selecting customers* refers to choosing the best target group for a given offer or communication, often given certain constraints like target group size, budget, or required conversion rate. *Segmenting customers* refers to allocating customers to several target segments in order to distinguish them, usually in terms of offers, prices, or creative appeal. Both aim at reaching a maximal conversion rate. The conversion rate is calculated by the number of conversions, divided by the number of recipients of a marketing campaign. In the first case, customers are selected according to *one specific* value of each selection variable to reduce the target group size. In the latter case, the optimal distribution of customers *among all* variable values is of interest.

Data is the basis for targeting, thus all targeting related areas of research focus on getting better knowledge from data; e.g. customer segmentation, data mining, and data fusion. The customer base is a crucial core asset, because it has a central function in the success of the company (Wirtz, 2009, p. 54). Core assets in general are distinguished by having an intrinsic value, being rare, not imitable, and not substitutable (Barney, 1991). The efficient use of customer data is a core competence of a company. The combination of the core asset customer database and the core competence database marketing can lead to a significant competitive advantage.

To purposefully segment customers is not easy. These segmentations are only possible if sufficient data is available (Behme & Mucksch, 2001). The information is sufficient, if it is relevant to the purchase behavior, if it has informative value in terms of media and channels to be used and accessibility, and if it supports the identification and measurement of customers. It is profitable and not volatile (Freter, 1997, p. 90ff). Marketing data often faces the problem of missing data in one way or another (Kamakura & Wedel, 2000). The classical market segmentation criteria used for segmentations, such as demographics and other identification data, are usually not directly relevant to the purchase behavior (Brogini, 1998, p. 113ff). A brand and its product are not preferred over competitive products because of demographic criteria like age and gender (Petras, 2007), but because they have a certain function and benefit for the customer. Descriptive data collected with the customers is often incomplete, inconsistent, and mostly aged (Kelly, 2007). Consequently, intelligent typologies aligned with the customer purchase behavior are coming into focus (Homburg & Sieben, 2005; Leitzmann, 2002). Such information is available from transaction data (Kelly, 2007).

But transaction data also ignores important information. So-called soft facts are worth knowing in order to optimally reach customers with campaigns. Customer databases have limited information on the following data categories (Breur, 2011; Dialog Marketing Monitor, 2012; Liehr, 2001).

- *General characteristics and preferences of the customers:* needs and motives, characteristics of the customers, information on media usage, attitudes towards daily routine, work life, leisure time, and family
- *Product and purchase related preferences:* product purchase motivation in the competition environment, attitudinal and evaluative data, such as quality perception and brand advocacy
- *Post-purchase behavior and opinions:* satisfaction with products and services bought, likelihood to recommend

In order to predict future customer preferences, today's data needs to be analyzed (Putten, 2010, p. 16) and enriched. While some targeting goals can already be achieved by mining the existing data, some of the necessary information is only available elsewhere. DWHs usually contain hard facts only. With these, it is difficult to conform to the requirements of successful customer segmentation. However, it is possible to acquire them externally and augment them to the customer database (Hippner, Rentzmann, & Wilde, 2002). Missing information can be retrieved from various sources, e.g. market research or other internal and external sources.

2.1.3 The marketing measure: conversion

The main KPI for targeted campaigns is the conversion. A conversion is the reaction to a marketing communication, e.g. a sale, a response, the participation in a raffle, or any other specified desired customer activity. The desired action is specified in advance and has to be measurable. Conversions can only be calculated for direct marketing campaigns with a clearly defined and identifiable target group (Rossi et al., 1996). Reactions need to be able to be traced back to the customer.

Conversions are the ultimate goal of marketing. They pay off in terms of sales, qualifying sales leads, and building customer relationships (Roberts & Berger, 1999, p. 9f) and are monitored closely. The increasing challenges in the economic framework, as described in chapter 2.2.1, require advertising efficiency (Laase, 2011). One goal of marketing is to improve targeting performance, or targetability (Chen, Narasimhan, & Zhang, 2001). The ability to target customers can have a higher and more durable impact on marketing performance than other marketing activities (Chen et al., 2001).

For planning purposes, marketers try to predict the conversion probability for individuals, for specific segments, or for the campaign as a whole. Those customers with the highest conversion probability are selected for direct marketing campaigns, taking into consideration the costs. Ideally, a return on investment can be calculated from a model for individual customers (Ratner, 2001a). In every target group selection, there are more and less prospective customers. The overall conversion probability is a mixture of the relative concentration of target customers (Smith, Boyle, & Cannon,

2010) in a selection, depending on the individual conversion probability of the prospective customers versus the individual conversion probability of the ones mistakenly selected. The challenge of predicting conversions is to identify the most profitable customers, rather than those being most likely to respond to promotional offers (McCrary, 2009).

The conversion probability is calculated taking into account as many factors as possible. It can be divided into a baseline probability of purchasing and time, contact, and purchase history related probability factors (Moe & Fader, 2004). The baseline probability of purchasing is relatively stable and can be attributed to the characteristics of customers. The others factors are volatile and change depending on the context. The marketing instruments used to stimulate it are different from those stimulating the "ad hoc" conversion probability. Data augmentation results are generally able to improve the knowledge on the baseline probability of purchasing, rather than on purchase situation related factors.

2.2 External conditions around the company

In the following, the external conditions for data augmentation in marketing are explicated. The economic environment poses several challenges that companies need to adjust to. The technological development offers opportunities that companies can turn to their account, with risks arising from missing out the trends. The legal environment provides the framework for all company activities. Respecting the borders is mandatory. The sociological and psychological environment comprises expectations, needs, apprehensions, and anxieties of the consumers. We have outlined the external conditions already in our first data augmentation study (Krämer, 2010). It is retraced and adjusted here where applicable. The external conditions are relevant to the SWOT analysis that we conduct in chapter 2.4. Therefrom, chances and risks are derived, leading to explicit potentials and limitations for the application of data augmentation in marketing.

2.2.1 Economic environment

The economic environment poses many challenges that companies need to cope with. Companies of all branches are faced with increasing complexity and dynamics (Huldi, 2002) and the number and efficiency of competitors aggravate the competition environment (Link, 2000). Frequently mentioned challenges are cost pressure, diminishing marginal utility, increasing speed of innovation, shortened product life cycles, and raising product homogeneity (Tiedtke, 2000). These developments are not new, but they keep governing many management decisions (Hippner, Leber, & Wilde, 2002). Due to the increasing opening and liberalization of markets and the resulting international expansion strategies, more market segments exist. Through globalization, new key markets are made accessible, and successful business models are transferred abroad (Meffert & Bruhn, 2009, p. 457). Only if realistic potentials for the own company are recognized and effectively implemented, is it possible to successfully operate and increase the company value on the long run (Huldi, 2002).

As a result, all company divisions are striving for success and have to deliver measurable results (Blattberg et al., 2008, p. 458f). Especially in marketing, where the central tasks are mainly determined by other divisions, the transparency on all steps and expenditures is important (marketinghub, 2009). For cost cutting reasons, advertising budgets are examined carefully. Online media, which feature both low costs and high transparency in terms of advertising impact, are chosen increasingly often. The percentage of advertising expenses dedicated to online advertisements is raising constantly (Dialog Marketing Monitor, 2009). The newer trends of mobile marketing and social media marketing are equally expedient and are increasingly used for marketing purposes (Dialog Marketing Monitor, 2012).

Several strategies have been developed to encounter the economic challenges. The success probability of individual activities and their measurement is raised by the coordinated and targeted steering of database marketing (Schweiger & Wilde, 1993), so that the overall allocation of resources is improved. With the markets being fragmented, the product homogeneity and the willingness to switch between sellers, companies must differentiate their products and personal communication (Schweiger & Wilde, 1993; Blattberg et al., 2008, p. 7). The increasing need for individualization in conjunction with price competition and complexity, but also an unmanageable number of customers and the consequent inability of companies to address all these customers in person, is encountered with the standardization and automation of products, services, and communication actions. The so-called mass customization concepts comprise modulated features and communication packages, which can be individually designed, but are then provided automatically (Piller, 2006). Mass customization is supposed to live up to the needs and wants of the customers, while at the same time being cost efficient (Meffert & Bruhn, 2009, p. 459ff).

2.2.2 Technological environment

Through the technological development in the past years, data collection and usage is enabled and impelled. New information and communication technologies have replaced the personal communication between companies and customers (Meffert & Bruhn, 2009, p. 460). Most of today's transactions, call center inquiries, and other contacts at various touch points are recorded electronically (Breur, 2011; Kelly, 2007). These electronic footprints allow for detailed customer analysis and a holistic customer picture. The wealth of customer data has been recognized and referred to as the gold of the 21st century (Hebestreit, 2009; Singh, 2013). Many of these developments pose the possibility of individual content and addressability (Bensberg, 2002, p. 164f). Based on data, behavior can be analyzed, learned from, and triggered. Nowadays, most companies have integrated CRM systems and a high-capacity DWH (Hippner, Rentzmann, & Wilde, 2002).

Because of the increasing capacities and the progressive digitalization, the data volume stored at companies is constantly growing (Baker, Harris, & O'Brien, 1989; Behme & Mucksch, 2001, p. 9). New data sources in terms of new channels, new technologies, and new customer touch points are continuously emerging (Breur, 2011; Hipperson, 2010). At the same time, real-time CRM is possible today enabling greater intelligence through better performance of analytics, growing data volumes, and higher speed of deployment (Acker, Gröne, Blockus, & Bange, 2011). Storage costs are rapidly decreasing (Breur, 2011; Dull, Stephens, & Wolfe, 2001; Kelly, 2007). Network connections for technically linking sources are inexpensive (Bleiholder & Naumann, 2008). Likewise, the cost of data collection is decreasing.

The usage of online media has rapidly increased in the last 20 years. As reported by the *ARD/ZDF Online Study*, 76% of the German population used online media in 2012 (Eimeren & Frees, 2012), as opposed to 44% in 2002 (Eimeren, Gerhard, & Frees, 2002) and 7% in 1997 (Eimeren, Oehmichen, & Schröter, 1997). As a saturation of consumption is almost reached, the growth started to level in 2010 (Eimeren & Frees, 2012) and has been below one percentage point from 2011 to 2012 (tns Infratest, 2012). Additionally, mobile devices have taken the form of little computers, containing just as much information as personal computers with its internet access and many more useful applications. Social media is regarded as another media channel with its own rules and conditions. All of these media experience a rapid dispersion and are quickly adapted especially by young people (Dialog Marketing Monitor, 2012). The establishment of new devices leads to a multiplication of usage situations and new market models.

2.2.3 Legal environment

There are three major legal questions related to data augmentation in database marketing: Is data augmentation legally sound? Which data categories are permitted to be augmented? Who may be contacted with a targeted campaign after having augmented the data? The following explanation is based on German law and would have to be reviewed for applications in other countries.

Privacy is very important for all data augmentation activities (Breur, 2011). The Bundesdatenschutzgesetz (BDSG), the German federal data protection act, governs all questions related to personal data, i.e. data related to natural persons (BDSG, § 3) directly or indirectly enabling the recognition of individuals (Arndt & Koch, 2002). It guards every individual from being affected in his or her personal rights (BDSG, § 1). The *preventive ban subject to permit* applies, meaning that collecting, processing, and using personal data is acceptable only, if a law explicitly permits or mandates it, or if the concerned person has agreed (Arndt & Koch, 2002). The BDSG was adapted by the federal government and is effective since September 2009 (Schaar, 2011). The amendments comprise, amongst others, rules and regulations regarding address trading, market research and opinion polls, and third party address processing (Eickmeier & Hansmersmann, 2011). Companies have to respect the BDSG as soon as they collect, process, or use data with data processing equipment (BDSG, § 1).

Two data categories do not adhere to the BDSG. Anonymous data enables the recognition of individuals only with a disproportionate high effort in terms of time, cost, and manpower (BDSG, § 3). Anonymous data can be used for developing the optimal marketing mix without concerns regarding data protection (Freter, 1997, p. 446). Likewise, aggregated data is not considered personal data, because no conclusion can be drawn from them on individuals. However, if an individual person is collated to a particular group on which certain details are known, this is considered a personal reference (Arndt & Koch, 2002).

It is explicitly appreciated by the second amendment to the BDSG to fuse data with the objective of avoiding unnecessary advertising in targeted marketing (BDSG, § 28). Companies are allowed to store a reasonable amount of additional data on their customers in order to use it for the selection of targeted campaigns (Plath & Frey, 2009). The notion of *store* relates to modifying existing data in a way that rightfully collected data is added (Däubler, Klebe, Wedde, & Weichert, 2010, p. 461). The notion of *rightfully collected* means that the data was collected with the person concerned (BDSG, § 4), at a place where the data is publicly available, or where the responsible authority would be allowed to publish them (BDSG, § 28). The notion of *a reasonable amount* has to be regarded for each case individually. The collection of data directly from a person requires a permission (BDSG, § 4), unless a law explicitly permits or mandates it. There are constraints to the collection of additional data, which require a considerable care of so-called sensitive data. Sensitive data comprises race and ethnical family background, political opinions, religious beliefs, union memberships, health, and sexuality (BDSG, § 3).

The legitimacy of data collection is not equal to the legitimacy of personal contact by advertising. To this effect, the BDSG was even tightened by the new amendments. While postal mailings are still admissible without a permission of the recipient (Lambertz, 2009), commercial emails or SMS require an active consent, referred to as opt-in (Pauli, 2009). It means that a tick box on a website, for example, cannot be prefilled, but has to be actively ticked by the customer. This consent may not be interlinked with the general signing of a contract with the company (BDSG, § 28).

The basis for all questions related to competition law is the Gesetz gegen den unlauteren Wettbewerb (UWG), the German act against unfair practices. It protects the competitors, consumers, and other market participants. It prohibits, amongst other things, unfair and misleading business activities, as well as unacceptable disturbance (UWG, § 1/3/5/7). While the BDSG and the UWG pursue different purposes, they come to an agreement in terms of advertisement (Plath & Frey, 2009). Exceptions to the BDSG are stated in the UWG, such as the personal contact by email in connection with the purchase of a product or service, the advertisement of similar goods and services, or if the customer is fully aware of how to object to the personal contact (UWG, § 7). For these exceptions, no permission is necessary before contacting a customer personally by email. Above all, the UWG prohibits marketing actions suitable to exploit the inexperience of children and teenagers in terms of business (UWG, § 4).

Another category of data, online usage data, is regulated in the Telemediengesetz (TMG), the German telemedia act. Online websites are permitted to save online usage data for advertising and market research purposes. These profiles, however, can only be saved in pseudonomized form (TMG, § 15). Once a unique identifier is missing, the data does not fall under the regulations of the BDSG.

2.2.4 Sociological and psychological environment

Since the 1980's, an increasing information overload has been observed, meaning that only about two percent of information a customer is confronted with is perceived (Kroeber-Riel, 1988). Especially advertisement is ignored: It is only perceived by customers, if it is relevant and if there is a benefit related to it (Mahrdt, 2009, p. 12f; McKay, 2009). The perception probability rises with increasing accordance between the communicated benefit and the personal interests of the customers (Schweiger & Wilde, 1993). Direct marketing is supposed to separate spam from real information (Roberts & Berger, 1999, p. 15). Spam is not directed to any particular target group. Rather, it aims at sending out as many emails as possible via spamming botnets (Kanich et al., 2008).

Before deciding for a product, customers are by all means interested in information. When deciding for a product, they expect an offer to be superior to others in terms of characteristics, quality, and price (Link, 2000). But customers do not only rely on information provided by companies and media. They select which information they want to receive through which channel (Leitzmann, 2002). The internet makes products and prices transparent (Feinberg et al., 2002) and is becoming increasingly popular as an information source, thereby reducing information asymmetries (Bensberg, 2002). A self-reliant information search is possible. Recommendations and reviews from other customers, so-called consumer-to-consumer (C2C) communication, are becoming stronger influences for buying decisions. Nurturing consumer communities, facilitating conversations, and listening to what people say will have a much stronger focus in the future (Hipperson, 2010).

As especially online information is getting more individualized, and personalized, data security is becoming a more delicate topic. Customers want to know which personal information is saved and used. Data security topics are regularly discussed with great passion in public. A German survey in 2009 showed that most people do neither trust the government nor private companies when it comes to data security (Insitut für Demoskopie Allensbach, 2009). There is a limit to how much data can be collected from a single person (Baker et al., 1989), as customers would like a company to hold as little data on them as possible (Ozimek, 2010).

Nevertheless, to publish personal information on the web is both contemporary and common: it is perceived as welcome, if not desirable, to present oneself on the web (Wagner, 2010). It is noticeable that customers are motivated to articulate their preferences, opinions, and interests, if they are directly related to individual products or personal benefits (Hippner, Leber, & Wilde, 2002; Tiedtke, 2000). In future, advertisers expect an increasing digital media competence, which comprises the self-selection of advertising contents and the involvement of customers in the product development process (Dialog Marketing Monitor, 2009).

2.3 Sources for data augmentation

There are various established and potential data sources available for data augmentation in database marketing. The availability of the sources is the most important chance in our SWOT framework in chapter 2.4. Data is collected at various touch points and through various channels. Sources for data augmentation can be data publically available, data available within the company, or data dedicatedly accumulated for data augmentation purposes. The notion of external sources does not refer to *where* data was collected. External sources might well be available in-house, e.g. from a website, from an in-house survey, or from a social media application. However, they are considered external, because the data cannot directly be merged to customer profiles. An algorithm or pattern is necessary in order to augment the customer database with the external data.

While some of the sources are established data augmentation sources, others are not commonly used yet, e.g. those not being identical to the customer database or a representative sample thereof. Uncertainty arises as to whether these potential sources can be used for data augmentation. The focus of this paper is to analyze the different characteristics of data augmentation sources and their implications on the validity and significance of the results. That way, also potential sources not commonly used today can be evaluated regarding their suitability for data augmentation.

The most important traditional and newer data sources are depicted in the following, highlighting their specific potential for data augmentation. Features and downsides of the source types are explained. We compare all sources regarding access and availability, usefulness of available variables, data preparation effort necessary to augment the data, costs, and timeliness. To know the characteristic features helps in recognizing them in the methodological framework described in chapter 4.

2.3.1 Public and official sources

Public and official sources are defined by fully covering or representing a national population. These sources fully include or represent the customers (of that country), but also more people from a bigger population. They are easily accessible and methodologically sound, i.e. the quality of the data is high and data collection has taken place in a statistically proper way. It is usually carried out by the federal statistical office or be renown market research institutes. Examples of these providers in Germany are *forsa*, *Institut für Demoskopie Allensbach*, or *TNS*. Publically available sources not fulfilling the criteria of fully including or representing the customers are not considered here.

National census data Official statistics or national census data are at hand for every registered person in a country. Common information categories interesting from such register data are income, education, housing, employment, consume propensity, and living condition information. Census data is only collected once in a few years, which makes it an impractical sources for data augmentation. Additionally, census data is often not available in enough detail regarding link variables to be used for data augmentation. A census source is usually impractical and unnecessary as a sample (Powell, 1997, p. 67). However, it is a thinkable source and is mentioned here in order to contrast it to other sources.

Market media studies It is possible to obtain market research data from market research institutes or publishing companies, such as *Axel Springer AG*, *Bauer Media Group*, and *FOCUS Magazin Verlag GmbH*. These market media studies comprise a broad range of information categories, and thus pose many information opportunities. The comprised information can be divided into socio-demographic, psychographic, consumer preferential, interest, and behavioral information. Variables are predetermined, but big market research studies include so many variables that they usually satisfy the information needs. External market research has been used and described for marketing purposes. For example, Putten et al. (2002a, p. 3) used a survey on a whole branch and Krämer (2010) used a German population representative market media study.

As public sources are not designed to fit the individual purposes of data augmentation in marketing, they need to be critically examined in terms of their suitability regarding concepts and definitions of the link variables. There is data harmonization effort related to using public sources for data augmentation purposes. Furthermore, some comprehensive so-called single source studies often use specific data fusion methods in order to reduce response burden (Gilula et al., 2006; Kamakura & Wedel, 1997). To use these fused sources for another data augmentation requires critical reflection.

Public market research information is generally easy to obtain, well-conditioned, and the market research data market is very transparent. The quality is generally high, but nevertheless should be assessed depending on the issuing institute, because market research vendors are not neutral sources, but profit-oriented companies. The qualitative judgment is dependent on the individual information goals of the companies (Hippner & Wilde, 2001). General qualitative assessment criteria are utility (e.g. as a basis for decision), completeness (in terms of information sought), timeliness, and verity. Additionally, accuracy and reliability are relevant if the data source is a sample (Berekoven, Eckert, & Ellenrieder, 2009, p. 24ff).

2.3.2 Company-owned sources

Company-owned sources are only interesting for data augmentation, if they are not already incorporated in the infrastructure of the customer DWH and equipped with a unique identifier. There are two reasons why information can be unavailable on customers. Either the information is not available for some customers, but for others. Then it is technically possible to obtain the information, but the customers without data have not had any transactions from which that data can be derived. Or there are data sources that are not connected with the customer DWH, so that the information is not available for any of the customers.

Existing customer DWH In general, the existing customer DWH is the recipient unit. However, data augmentation can be a tool, if information is available for a certain subgroup of customers only. This kind of data augmentation, also referred to as scoring, is used in order to get information on all customers, although only a fraction of the customers has a value for a specific variable. The results of these augmentations are used in order to acquire new customers for specific products or for cross-selling purposes. This data is not only available in real-time and free of cost, but also derived from the analytical systems optimized for database marketing purposes, so that data preparation efforts are minimal. The usefulness of variables augmented by scoring is limited, because variables are not entirely new to the customer DWH, but only new for a subgroup of the customers.

Operational data Some of the operational data is not (yet) available to analytical systems in a way that they can be used in database marketing. In many companies, data is collected at various touch points. The operational systems of companies were formerly not laid out to serve the database marketing purpose. There might also be reasons for not making data available on a personal level, for example confidentiality. Common examples range from call center reportings and in-store information to web analytics data. Whenever these sources have sufficient link variables, they can be augmented to the customer database instead. The data can be made available in real-time, thus preventing problems of timeliness.

Online tracking data A special kind of operational data are online sources. The company website can be tracked and much information can be derived from surfing behavior. Tracking customer transactions was formerly only possible if they had a loyalty card or if the kind of transaction required the transaction to be saved. Today, all transactions in the internet or even unfinished transactions and nonbinding product searches are traceable. This data can be reused to offer products and target customers.

Furthermore, many variables not directly linked to purchase behavior are easily and accurately available online (Moe & Fader, 2004). Even if it is possible to directly link the surfing behavior to customer profiles, for example if customers are logged in to a website, this can be not allowed due to legal constraints. More details regarding legal requirements can be found in chapter 2.2.3. Consequently, surfing data is collected on an anonymous basis and can be used by data augmentation. The data preparation effort can be high for tracking data, because information needs to be converted from an unstructured form to relational data structures, in order to be used for the data augmentation methods presented here.

Company-owned sources have a high overlap with the customer database, but might not capture all customers. To that effect, the donor unit is a subset of the recipient unit. Depending on whether the source is the existing customer DWH or an operational or online tracking source, the availability of suitable link variables differs. Whether the available information is useful needs to be decided from case to case. If no confidentiality constraints exist, data augmentation is only one, sometimes short-term, strategy to making this data available. Another more durable strategy would be to link the data directly to the customer database with a unique identifier.

2.3.3 Accumulated sources

All sources specifically created and designed for data augmentation purposes are referred to as accumulated sources. Because accumulated sources are designed by a department looking for specific insights, basically any information can be comprised. They are a way of obtaining information from a customer sample not otherwise collectable in a regular business relationship, e.g. information on education, occupation, and households (Hattum & Hoijtink, 2008a). Other interesting variables are referred to as marketing mix related reaction parameters, because they are able to segment customers with similar reactions to marketing mix instruments (Freter, 1997,

p. 92). Accumulated sources can include descriptive variables like attitudes, perceived image of the company, or customer satisfaction (Liehr, 2001).

Representative customer surveys A survey is usually carried out by a third party service provider and consists of a representative subset of the customers that has to answer a questionnaire asking for the target variables. Because of data privacy regulations, the answers cannot be matched on an exact basis with the customer profiles. The only way to receive target variables for individual customers is to augment it. Market research information has the advantage of anonymity. Surveys reduce desirability bias, compared to information stated in front of a company, so that the collected information is possibly more valuable and more truthful.

Volunteer surveys An inexpensive alternative of conducting a representative survey is to conduct a volunteer survey. Customers self-select who wants to be interviewed. Volunteer surveys have a cost advantage, because no complicated arrangements have to be made in order to achieve representation. However, Pineau and Slotwiner (2003) showed that results from volunteer online surveys cannot easily be used to draw inferences on the overall population. They used different typical marketing categories to indicate differences between the internet community recruited by volunteer surveys and the overall population, thus implying that the conditional independence assumption is not valid for these sources in general. The error in terms of representation in the context of volunteer survey is referred to as self-selection bias (Hudson et al., 2004).

Social media data Social media brings new opportunities for data categories, e.g. all kinds of personal and sometimes sensitive information are collectable. Especially sentiments about products and services or purchase intentions are present in social media. Extracting and interpreting this information, e.g. by text mining, can have tremendous effects on sales

(Breur, 2011). Also, the degree of social connectedness has an impact on purchase probabilities (Naseri & Elliott, 2011). Casteleyn, Mottart, and Rutten (2009, p. 439) stated that the ”heartbeat of today's society” becomes obvious in networks like *Facebook*. Often, social networks have more accurate and much more detailed data on the (social media) population than other empiric research institutes. The information, however, is only accessible to selected researchers (Heinrich, 2011). Companies need to cope with other solutions, like data available from public social media profiles or social media applications, with which the users are asked for their permission to access distinct data categories. That way, transparency about used information is given and companies are legally enabled to collect private social media data. Groups like these are not of representative nature (Casteleyn et al., 2009).

Surveys and social media data differ in terms of their characteristics. For surveys, variables can be freely defined, so that information potentials are unlimited. One of the advantages of surveys is that formats can be chosen in accordance with the customer database. The survey usually includes a subset of the customer database. Costs can be high, if the survey is supposed to be representative. A drawback of market research data is that it is conducted at a single point in time. It should be processed and augmented right after being analyzed, because most data augmentation models do not account for time differences and problems related to this.

To use social media data can lead to high data preparation efforts. The information available needs to be interpreted. The liking of a brand page, for example, can mean a lot of things, only one of them being the fact that someone intends to purchase this brand. Because there is no single best way of obtaining social media data, access can be summarized as being limited or at least complicated. The cost of social media data varies accordingly. Timeliness is not a problem, as data extraction can take place just in time for data augmentation.

A number of American and German studies have been conducted to show correlations between social media activities and various behavioral characteristics. If target variables are correlated with the fact that someone uses social media channels, these sources can lead to biased augmentation results. Kutter (2013) showed that this risk is low. However, it is worth examining the conditional dependencies between sources and target variables, as described in chapter 4.1.4.

People update their social profiles in order to maintain a good image of them online – much more than in a customer database. It should always be kept in mind that social media profiles are designs of what people want to express about themselves. They do not necessarily reflect what people are like, but more what they want to be like (Casteleyn et al., 2009). Joining groups or liking pages are acts of self-portrayal. It can generally be assumed that the stated information is accurate (Abel, 2011). Restrictions apply in terms of verity and social desirability. We do not go into detail on which restrictions apply to the usage of social media data. The topic is discussed in the respective media psychological literature.

2.3.4 Comparison of sources

Whether a source is generally suitable for data augmentation depends on several factors. It is a precondition that a source needs to be accessible and available, needs to include useful target variables and link variables able to predict these target variables. This is explained in more detail in chapter 4.2. However, sources can differ in how easily they are accessible, in how many target variables are included, and in how much data preparation effort is necessary in order to harmonize the link variables of source and recipient unit. Sources are the more suitable, the less they cost and the more recent data is. The more obstacles there are, the higher the resistance to attempt a complicated data augmentation project. The influence of the overlap between the source and the customer database on the data augmentation results is subject of this paper. It is explained in more detail in chapter 4 and examined in chapter 6 and chapter 7. The sources listed are compared in table 2.1 regarding these factors.

	Public sources		Company-owned sources			Accumulated sources		
Criterion	Census data	Market media	Existing DWH	Operat. data	Tracking data	Repr. survey	Volunt. survey	Social media
Access and availability	+/−	+	+	+/−	+/−	+	+	+/−
Usefulness target variables	−	+	+/−	+/−	+/−	++	++	+/−
Readiness link variables	−	+/−	++	−	−	++	++	+/−
Costs*	+	+/−	+	+/−	−	−	+/−	+
Availability data and updates	−	−	+/−	+	+	−	−	+
Overlap with customer database	full	full (repr.)	none	partial	partial	full (repr.)	sub-group	partial

* Low costs result in a high rating

++	+	+/−	−
perfect	good	specific	limited

Table 2.1: Rating of available data augmentation sources

Access and availability can be limited or change over time. Survey data, e.g. from market media studies or accumulated surveys, as well as the existing customer DWH are more easily accessible than national census data or those sources that need to be technically connected with the customer DWH, such as operational, tracking, or social media data. Official statistics, for example, would be a great source of very precise and substantial information, but are not available to private enterprises. Sources less available involve other parties or systems. External market research studies, especially the German market media studies, are useful, but are subject to changes that cannot be influenced. Additionally, legal constraints apply to the use of many sources. Social media data are easily accessible, but need to be handled with care when it comes to data privacy issues. However, none of the sources really has restricted access or unavailability – otherwise they would not be potential data augmentation sources.

The usefulness of available target variables is case-specific. In general, the more target variables there are in a source, the more profitable the augmentation results. Accumulated surveys, both representative and volunteer surveys, are perfect in terms of useful target variables, because the target variables of the survey can be defined by the company. Market media studies are not self-designed, but contain so many variables that they are considered good in terms of the usefulness. All other sources only contain specific variables, while national census data is rather limited in terms of target variable variety.

Definitions and formats of the link variables can be designed for accumulated surveys, as well as for existing customer DWH data used for scoring purposes. The preparation effort can be a precluding criterion. While official statistics and survey data are pre-screened and structured, web or social media data need to be reduced and transformed before augmenting them. National census, operational, and tracking data needs to be critically examined in terms of availability and readiness of link variables. Only if suitable link variables are given, at least after data preparation, a source can be used for data augmentation.

Costs for the acquisition of data augmentation sources can vary, depending on the expenses related to collecting or buying data, connecting and preparing it. Company-owned sources are generally less costly than public sources in terms of acquisition, but more costly in terms of connection and preparation. This can be a make-or-buy decision, because the costs need to be compared to the cost of collecting data in-house, if possible. Unfavorable formats like in tracking data can lead to additional expenditures. Costs are also related to the process of data augmentation and needed technology. Accumulated data is more costly, if a survey is conducted, than if data only has to be ”tapped”, like social media data.

In theory, sources can only be augmented without error, if the data of the source and the customer database are observed at the same point in time (D'Orazio et al., 2006, p. 4). Customers' ideas, needs, and behaviors change continuously and augmented data ages (Ozimek, 2010). Thus, the length of time between data collection date and data augmentation influences the quality of the augmentation results. The utility of augmented data decreases over time (Even, Shankaranarayanan, & Berger, 2010; Ozimek, 2010). The error related to the usage of outdated information is referred to as timeliness error. Data permanently available and updated, like operational from company-owned sources or user generated content from social media sources, is generally more beneficial for data augmentation purposes than survey data, which is only collected once or at most yearly. For returning marketing problems and tasks, it is more valuable to acquire sources that can be augmented at any point in time with recent information. As the information in the customer database changes just as quickly when variable values are added or adapted during business processes, a so-called on-demand computation can be established, for which data sources are maintained and updated separately and only brought together at the time when augmented information is needed (Jiang et al., 2007).

2.4 Implications for data augmentation

From the external and internal conditions and the available information sources, a strategy is derived for data augmentation in database marketing. A SWOT is conducted in order to illustrate which strengths of the company can be used in order to benefit from opportunities and encounter risks, as well as which weaknesses can be turned into strengths by exploiting external chances. It shows which risks should be avoided. Data augmentation is the logical deduction from these considerations and can be implemented as solution for many problems. An overview of the factors relevant in the SWOT is given in table 2.2.

SWOT	Helpful	Harmful
Internal origin	*Strengths* Customer DWH for collecting all relevant information for customer communication. Knowledgeable database marketing team Customers can be targeted directly and individually	*Weaknesses* Customers cannot be differentiated on a one-to-one basis Not enough information available to purposefully segment customers Conversion probability is not known before implementing a marketing campaign
External origin	*Opportunities* Several source types available including information worth knowing Technological progress enables affordable data storage and efficient data usage Customers expect benefits and tailor-made offers, are interested in information facilitating their purchase decision	*Threads* Economic situation is afflicted with dynamics, complexity, competition intensity, and market fragmentation Personal data is particular protected by law and must be handled with care Customers are price-sensitive, tired of receiving irrelevant communication, and anxious about what happens to their data

Table 2.2: SWOT analysis implying strategic suggestions for data augmentation

The chances offered by additional sources can help in overcoming the lack of segmentable customer information and turning it into strength. Customers expect differentiated marketing tailored to their needs. Only if communication is targeted to highly selected groups, it reaches a maximum of acceptance (Hattum & Hoijtink, 2008a; Laase, 2011). However, there exists a discrepancy between desired and received information (Liehr, 2001). In order to offer a surplus to self-generated information and in consideration of the information overload already in place, advertising contents have to be relevant. With the increasing need to know more about the customers in order to address them individually, the necessity of combining different sources is raising (Kamakura & Wedel, 1997). This is made possible by data augmentation. It is anticipated that the successful outcome of a data augmentation leads to a better overall customer experience (Breur, 2011).

Relevance is not only of interest to the customers, for which personal benefits, time and cost savings, and quality are important. It is also in a company's' interest. Every customer contact has an economic potential in terms of return on investment (e.g. revenue increase, customer profitability, competitive advantage). The customer specific treatment and communication is thus directly creating value for the company (Rapp, 2002b).

By using the technological advancements in database marketing, companies can increase their position in the competition environment. Thanks to storage capacities and advanced analytical systems, data can be processed in shorter time. Those companies able to react to the customer needs better and faster than the competition can gain a significant competitive advantage (Fogarty, 2008). Today, technology and know-how enabling data augmentation are available. The database marketing know-how itself becomes a core asset, because it builds on a conglomerate of interdigitating activities not easy to imitate (Schweiger & Wilde, 1993).

The trends in media usage and technological advancements can be used for incorporating augmentation results in the regular communication process. The increase of online communication channels supports the cost saving and efficiency goals. The internet has low variable costs, while globalization and technical progress drive the demand of online applications. The more intensive customers use the internet, the bigger the amount of detailed information and opinions usable for improving communication and services. With the help of database marketing, modulated tailored communication instruments can be developed, automatically assigning suitable offers to customers on the basis of their information and contact history, distributed independent from channels. Data is the basis for individualization, automation, and mass customization.

Data augmentation is especially useful, because it respects the legal requirements. Data used for data augmentation purposes can be anonymous data, so that no direct inference to individuals is possible. Certain data sources containing unique identifiers must not be used, even if it was possible. On the other hand, to use data in order to deliver relevant advertising is explicitly appreciated by law, appraising the general approach of data augmentation. When moving within the legal borders, companies are able to create a lasting competitive advantage.

The ability to target customers can overcome the thread of customers being anxious about their data. Although all activities of companies, including data augmentation, are legally sound, there is a general anxiety of customers regarding data protection. Companies should therefore take these concerns seriously, address them, and establish trust, whenever marketing activities are planned. Trust is a combination of controllability, transparency, and security of action, which are facilitated by quality and stability (Winand & Pohl, 2000). Database marketing helps to create trust by implementing a clean permission handling, which is updated in real time, and by providing information for individual customer contacts. Once the customers recognize their preferences in the offers, they are apt to articulate further interests. To realize the one-to-one vision by addressing customers with perfectly individualized and personalized offers could disconcert customers, rather than satisfying them. Chen and Iyer (2002) showed that it might not be desirable for every company in a competitive environment to perfect individualization. To aim at micro segments that are augmented with external data therefore seems to be a reasonable goal.

All in all, data augmentation is a smart strategy in the current marketing environment. It adds information to the customer database not otherwise available, while respecting existing law and customer concerns. Technological advancements and internal developments like powerful CRM infrastructures and knowledgeable database marketing teams favor this approach. The additional data can be used to reach the marketing goals of individualization and relevance, leading to a competitive advantage.

Chapter 3

Literature review on data augmentation

The need to combine data from different sources is almost as old as the ability to store data itself. Some interesting information is collected at one place by one instance, other interesting information in another place by some other instance. If there is an identifier to serve as a key, joining tables is a matter of technology. But if no such identifier is present, the statistical fusion of such tables can be complicated. Nevertheless, it can answer interesting questions. Researchers have found solutions for combining such sources, like income figures and tax returns in official statistics (Okner, 1972), media usage and product usage from surveys (Baker et al., 1989; Rässler, 2002; Wendt, 1977), or customer segments and mailing recipients in marketing (Hattum & Hoijtink, 2008b).

Data augmentation with external sources in database marketing is based on concepts of missing data theory, statistical matching, and other predecessors in marketing. Our approach to data augmentation ties in with the methods developed for these purposes. The different approaches have different focuses and features. They can be adapted to serve our purpose as well. Previous researchers thought about where the data came from and

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_3

how it was sampled, which variables were best suited for augmentation purposes, which methods to use best for the augmentation process, and how to evaluate the results achieved. They faced data preparation problems and those regarding the uncertainty involved in data augmentation.

In this chapter, we recapture the different data augmentation strategies already described and point out similarities and differences to our approach. We list concepts, definitions, and methods relevant to our approach. The existing literature lays the basis for the specific theory developed for data augmentation in database marketing in chapter 4 and for answering our research question. We summarize the process steps for data augmentation described by previous researchers tailored to our augmentation needs.

3.1 Evolution of data augmentation studies

Since the 1980's, a versatile area of data augmentation research has emerged. Its techniques have been applied in different disciplines and for various purposes. The first approaches were undertaken by Okner (1972) and Wendt (1977). Okner matched federal statistics sources to achieve better insight into household incomes. Wendt matched print media information, television viewing behavior, and consumer information for media planning purposes. The basis for data augmentation was laid in the field of official statistics, survey and register statistics. Official statistics has a macro perspective, because it does not matter whether values for individuals are reproduced correctly, as long as the distributions of the overall population are reproduced well and inferences can be made on variables never jointly observed (Okner, 1972; Radner, 1980; Rodgers, 1984). Among the methods majored in the area of official statistics are those based on distance functions, using cells, constraints or weights, but also regression techniques and iterative expectation maximization algorithms (Radner, 1980; Rodgers, 1984).

Three areas of data augmentation are important for our data augmentation approach: missing data theory, statistical matching, and data augmentation approaches in marketing. Many techniques are derived from missing data problems (Rässler, 2002, p. 6f). In the following, we clarify fundamental thoughts of this subject in order to show how data augmentation problems are a special form of missing data problems. In the field of statistical matching, important concepts and methods for augmenting data have been developed. Rässler (2000, 2002, 2004) and Kamakura and Wedel (1997) conducted extensive research in this field. The field of data augmentation in marketing is depicted in order to show where our data augmentation approach fits in this research area and how it is able to develop and enhance existing studies, especially by Ratner (2001a, 2001b, 2003), Gilula et al. (2006), and Hattum and Hoijtink (2008b, 2008a).

3.1.1 Missing data problems and imputation

Missing data theory was developed in order to handle observations with missing values, such as refused answers in a survey or industrial series with mechanical breakdowns (Little & Rubin, 2002, p. 3). Instead of excluding the observations with missing data from the analysis, missing data techniques intend to find a suitable substitute for missing values, so that all observations can be used. The so-called complete case analysis, which reduces the datasets to observations without missing values, leads to bias when making inferences on the entire population (Little & Rubin, 2002, p. 3f). The main assumption is that the missing data hides true values meaningful for analysis (Little & Rubin, 2002, p. 8). A good overview on handling missing data problems is given by Madow, Olkin, and Rubin (1983).

In general, missing data imputation techniques can be distinguished into deductive, deterministic, and stochastic imputations (Nordholt, 1998). Deductive imputations are used, if there is information available from which the correct missing value can be deducted; e.g. the current age group from a given birth date. Deterministic imputations predict a meaningful value for the missing value. Stochastic imputations are extending deterministic imputations by adding a random component to the predicted values, so that existing variance is reflected. Our approach resembles the deterministic imputation technique most.

Missing data theory distinguishes between missing data patterns and missing data mechanisms. A missing data pattern describes which values are missing. The mechanisms specifies the relationship between the missingness and the values of missing and non-missing variables (Little & Rubin, 2002, p. 4). While data augmentation in marketing refers to a specific missing data pattern, the whole range of missing data mechanisms is feasible.

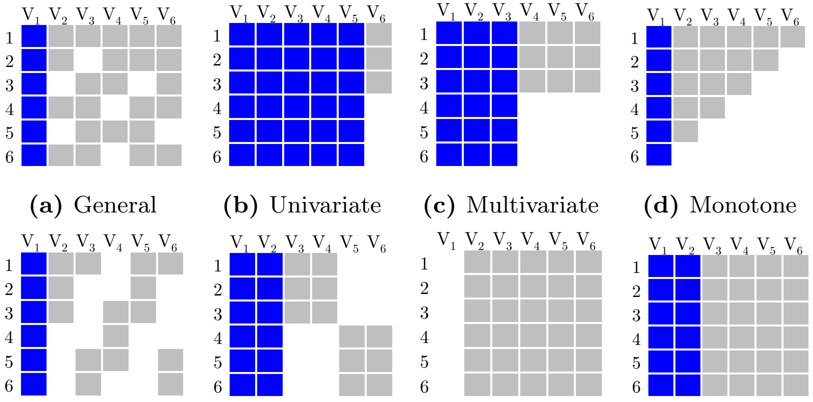
Missing data patterns

Common missing data patterns are shown in figure 3.1. Variables available for all observations are shaded blue. The values of other variables are either gray, if they have been observed, or white, if they are missing. In the general situation of missing values (3.1a), values are missing unintentionally and have a haphazard pattern. In such a pattern, data is randomly missing for observations or variables, so that no rule can be established as to which data is missing. Data can be missing due to unit or item nonresponse, or undercoverage (Dempster & Rubin, 1983). The result of missing data imputation is a fully rectangular dataset, from which further insight can be derived (3.1h).

If the data has an even missing data pattern, observations and variables can be grouped or sorted in a way that a rule can be established describing the pattern. Identification problems and certain difficulties associated with data augmentation can be avoided by creating special patterns of missingness (Rässler, 2004).

(b) If values are missing for one variable only, the pattern is referred to as *univariate* missing data pattern (3.1b).

(c) If they are uniformly missing for several variables, the dataset has a *multivariate* missing data pattern (3.1c). It is referred to as sub-



(a) General **(b)** Univariate **(c)** Multivariate **(d)** Monotone

(e) Split **(f)** Matching **(g)** Factor anal. **(h)** Result

Figure 3.1: Examples of missing data patterns as derived from Little and Rubin (2002, p. 5) and Rässler (2004)

sampling (Kamakura & Wedel, 2000), if an extensive survey is conducted among a subgroup of interviewees (Little & Rubin, 2002, p. 4f).

(d) *Monotone* patterns (3.1d) can result from attrition in longitudinal studies, if respondents drop out of panels (Little & Rubin, 2002, p. 5f).

(e) For *split* questionnaire survey design (3.1e), a set of core questions is answered by all interviewees and serves as link variables. All other questions are answered by subgroups and are later fused to form a complete dataset (Baker et al., 1989; Raghunathan & Grizzle, 1995). The same technique is applied for panel studies and is referred to as time sampling. Interviewees are asked questions at different points in time in a rotating fashion in order to save costs and reduce response burden (Kamakura & Wedel, 2000).

(f) File *matching* (3.1f) or statistical matching attempts to analyze the relationship between variables that have never been jointly observed

(Rässler, 2004), i.e. between the two gray blocks in the figure. Statistical matching is explained in more detail in chapter 3.1.2.

(g) *Factor analysis* (3.1g) can be approached as a missing data problem, where essentially the latent variable defining the classes is not known for any of the observations (Little & Rubin, 2002, p. 8) and instead, all target variables are available. In this case, the link variable is missing and needs to be estimated.

Data augmentation in our context has a univariate or multivariate missing data pattern. It is argued in chapter 4.2.3 that multivariate data patterns can be divided into several univariate missing data problems in order to gain more accurate insights into the individual variables with missing data. Data augmentation in database marketing is not confined to these patterns. However, the pattern of univariate missing data is relevant to our data augmentation problem.

The intention of augmenting data is not to perform an analysis on the overall population including observations with missing data, like in missing data problems. Rather, the data augmentation approach is to analyze only the part of the population that has the missing data, not the one the data is augmented from. While traditional missing data problems often have a macro perspective, data augmentation in database marketing has a micro perspective, as described in chapter 4.1.3.

Missing data mechanisms

For the process that causes missing data (Rubin, 1976), i.e. the missing data mechanism (Little & Rubin, 2002, p. 4ff), it is of interest which data is missing from a dataset. These mechanisms describe whether their missingness is related to any of the observed or unobserved data. The three categories of missing data have first been formalized by Rubin (1976) and are depicted in more detail by Little and Rubin (2002). They refer to the randomness of missing data and its influences on the augmentation process.

MCAR If the data is missing completely at random (MCAR), the missing values are a random sample of the dataset (Rässler, 2000). The missingness does not depend on the observed or unobserved data (Little & Rubin, 2002, p. 12). This does not mean that the pattern of the missing values itself is random, but rather that their missingness is not related to any other variables. If a representative group is interviewed for a survey on an overall population, the missing data mechanism is MCAR. MCAR sources are easiest to handle in terms of data imputation.

MAR If the data is missing at random (MAR), the missingness depends only on the observed part of the data. The values are not missing completely at random, but are conditioned on observed variables in the dataset (Rässler, 2000). If younger people are more likely to refuse to take part in a survey, but the age is observed or known, the missing data mechanism is MAR. As stated by Cochran (1983), there is virtually always a reason why data is missing for one person and not for the others. It is a valid assumption that data is usually not missing *completely* at random.

MNAR If the distribution of missing values depends on the missing values, they are missing not at random (MNAR) (Little & Rubin, 2002, p. 12). Then, the missingness is dependent on the values that would have been observed (Rässler, 2000). If people with a high income are more likely to have missing values on the income variable than people with a low income, there is an association between the missing data mechanism and the values that would have been observed. Using data augmentation results from such a source would lead to underestimation of the overall income. If data is not missing at random, the missing data mechanism must be incorporated in a data augmentation model. However, various models are possible and uncertainty arises from no knowing the correct one. The augmentation results highly depend on the choice of the best model (Allison, 2013).

In the context of data augmentation in database marketing, data is not initially missing. The situation of missingness is designed, because the desired target variables are not available in the customer database. When adding the external source to the customer database, a univariate or multivariate pattern evolves. The target values are missing for all customers, while they are available for all donors. In this context, the relationship between the availability of data, i.e. whether a person has been observed in a source or not, and the values of the target variables is of interest. The missing data pattern can be MCAR, MAR, or MNAR. This is described in more detail in chapter 4.2.5.

3.1.2 Statistical matching

The literature on statistical matching comprises many techniques relevant to data augmentation. It poses the first link between data augmentation techniques and use cases in marketing, particularly in media planning. Statistical matching concentrates on matching two data sources and making inference on the variables never jointly observed. Rubin (1986) was one of the first to address this problem theoretically, after Rodgers (1984) had already given a good evaluation of the problems encountered in statistical matching. Statistical matching relies on two independent and representatively sampled sources, with the complexity arising from the fact that conditional independence needs to be assumed between the target variables of one source and the other, because it is not possible to fully estimate the association between those variables from the available data (Little & Rubin, 2002, p. 7).

Kamakura and Wedel (1997) went into detail about fusing two independent samples in order to gain insight on the inferences between product and media usage, a problem commonly addressed in media research. Among the antecessors in this field were Baker et al. (1989) and Adamek (1994), mainly using hot deck imputation techniques in order to solve the file concatenation problem. Kamakura and Wedel compared different imputation techniques and extended the ideas of Rubin (1976, 1986) from continuous to discrete variables, thereby addressing a major difference between marketing applications and other approaches. They also dealt with the idea of one dataset having been collected by the initiator of the data imputation and the other one gathered externally.

Rässler (2000, 2002, 2004) addressed both problems from official statistics and media research. Her most important contribution was a compendium on statistical matching (Rässler, 2002). It outlines the basic idea of statistical matching and defines approaches to different forms of statistical matching, depending on the variable scales and the knowledge intention. Besides specifying different imputation techniques, she developed four levels of validity of matching results. These levels are depicted in more detail in chapter 3.2.4. By doing so, she formulated a major difference between statistical matching problems and data augmentation in database marketing. The goal of statistical matching is always to make statements on a population as a whole. It is not of interest whether individual values are correctly matched. In contrast, data augmentation in database marketing aims at finding accurate values for individuals or the smallest groups possible.

Another important compendium was written by D'Orazio et al. (2006), whose main focus is on applications in official statistics. It explicitly addresses the concepts of auxiliary information and finite populations, both relevant to our study. Like any problem concerned with human populations, our study has a finite population scope.

3.1.3 Data augmentation in marketing

Data augmentation with external sources in marketing started with the augmentation of geographical information or so-called micro household data (Putten, 2010, p. 84). Micro household data can be bought from specialized address data brokers. These fusions based on a single variable (the geo-location), however, do not offer much insight into individual customer behaviors and preferences. They offer a moderately accurate segmentation based on neighborhoods. In our sense, data augmentation is a much more complex and diverse approach to gaining new information on existing customers, based on various combinations of link variables.

The idea of data augmentation in marketing has become more popular in the last years. Ratner (2001a, 2001b, 2003) proposed valuable ideas relevant to data augmentation problems, e.g. how to find good predictive variables and how to use "look-alike profiles" (Ratner, 2001b, p. 66). His work focused on finding information within the customer database and applying it to other customers. Putten et al. (2002a) introduced a generalized model for data fusion in marketing. Gilula et al. (2006) expanded the idea of Kamakura and Wedel (1997) to the categorical variable case that is common in marketing. They used prior information in order to test the conditional independence assumptions and were able to show that even a little inexpensive dataset including all variables could substantially improve data augmentation results.

The use of external information in direct marketing has been applied by Hattum and Hoijtink (2008b), who conducted data augmentation in order to be able to segment customers into five groups, explicitly addressing the problem of micro validity. Furthermore, they introduced a concept for internally and externally evaluating data augmentation results, thereby building a foundation for comparing effectiveness and (cost) efficiency of data augmentation in marketing.

As data augmentation in marketing comes more into focus, detailed questions arise regarding the subject. Hattum and Hoijtink were able to prove that their data augmentation results increased response rates of a questionnaire and afterwards elevated average sales leads. However, the significance of the results could not be calculated (neither for individual values, nor for the overall results), because outcomes where not tested against a comparison group, but against historical measures. Furthermore, the study did not reflect on which sources can be used for data augmentation. Their source was a representative survey among a customer group, a MCAR source, which is an easy and special case when it comes to external information. So far, it is not known how much this restraint can be relaxed to other external sources. MAR sources are more often available for data augmentation. To understand whether and how these can be used is a relevant question for database marketing analysts.

3.2 Process of data augmentation

The process of data augmentation in database marketing is dividable into data screening and preparation, data augmentation, and evaluation. During the data screening step, the data augmentation frame is defined. It describes all possible sets of elements, link variables, and target variables relevant to the augmentation process. All samples are drawn therefrom. The purpose of data preparation is to receive two rectangular datasets for further processing. After having prepared the data, the best method is chosen. Once the method has been identified, the data augmentation is carried out, and results are filed and documented for further usage. The results are usually finished to serve as selection criteria in target group definitions. It is important to evaluate the augmentation results, both internally and externally, especially regarding effectiveness and efficiency of the new data.

3.2.1 Data screening

There are two possible starting points for a data augmentation project: the need for additional information or the availability of a source. If the need for additional information is the cause, it makes sense to screen available public and company owned sources for target variables containing that information, before potentially deciding to accumulate, i.e. create, a source not yet at hand. In this case, the first data screening step consists of the comparison and choice of possible sources. If the availability of a source is the cause, as many target variables as possible are extracted from the source in order to enrich the data basis for future targeting tasks.

Either way, elements and variables of the source are catalogued, describing relevant characteristics like formats, scales, and domains. During the data screening phase, all elements and variables relevant to data augmentation are selected and are further treated in the data preparation step. The selection of relevant variables reduces the columns of recipient and donor unit to a minimum number of characteristics. The main goal of variable screening is to remove autocorrelations, bad predictors, and variables that are not relevant to the business objectives (Putten et al., 2002b). The data screening and preparation steps can be skipped, if the source is an accumulated source. In that case, link and target variables, as well as elements can be chosen to best suit the data augmentation design.

Link variable selection

Link variables are common to both donor and recipient unit. Whereas the term link variable is also applied by Liehr (1999), other terms used are matching variables (Kamakura & Wedel, 1997) or common variables (Gilula et al., 2006). When augmenting data to an existing customer database, it is always assumed that the link variables in the donor unit and the recipient unit have similar concepts and definitions, so that they are fusible (D'Orazio et al., 2006, p. 163). If the link variables in the recipient and donor unit represented different information, it would be precarious to use them for data augmentation. Initially, it is only important that they have the same meaning. They do not have to be formally identical, as they can be homogenized during the data harmonization step (Rässler, 2002, p. 17).

Database marketing analysts face the problem that the perfect model for data augmentation is not known. Thus, it is only possible to try and find the best subset of possible link variables (Ratner, 2003, p. 56). The selected link variables should be largely independent of each other (Sharot, 2007). In general, the more link variables are available, the better the target values can be augmented to the recipients. However, it also depends on the explanatory power of the variables. Baker et al. (1989) emphasize that a good link variable selection is crucial to the success of data augmentation. They showed that link variables with a high correlation to the target variables produce good augmentation results. Those link variables with higher variances are generally more suitable for data augmentation (Kim, Baek, & Cho, 2004), because target variable differences can easier be detected.

If there are link variables with little extra explanatory power, it is possible that they add more variance to the results than they add information to the augmentation. It is not the individual link variable's correlation with the target variable that is important to the data augmentation model, but the link variable's relative contribution to the model, taking into consideration all link variables. Good models are characterized by the fact that the total effect of the link variables in the model is even greater than the sum of their individual effects (Ratner, 2003, p. 56f). Thus, it is best to find a number of necessary variables by omitting all variables that are either irrelevant, i.e. not correlated to the target variable, or redundant, i.e. not adding any extra information to the target variable (Ratner, 2001a).

From a computational point of view, the fewer the link variables, the easier the augmentation procedure. A complex multivariate framework can make it hard to make inferences or interpret augmentation results (D'Orazio et al., 2006, p. 167). This applies especially for categorical variables. The more link variables and the more values contained in the domain of the target variable, the harder it is to find accurate augmentation procedures. The calculation of conditional independence – if at all possible from the given data – is almost impractical with a high number of link variables. This topic is discussed in more detail in chapter 7.1. Different combinations of link variables should be evaluated with a test dataset. Among the elimination techniques used most often are forward elimination, backward elimination, and stepwise elimination (Ratner, 2001a). The more of these variations are performed, the better, because subsets are always judicially chosen by the database marketing analyst (Ratner, 2003, p. 57f).

The link variables should be a mixture of demographic variables and other variables with predictive power regarding the target variables (Baker et al., 1989). It has been shown in previous data augmentation projects that link variable sets including product usage and transaction history data are much more meaningful than those sets only comprising socio-demographic variables in terms of their ability to discriminate between target variable values (Liehr, 2001).

In certain situations, it makes sense to introduce so-called critical variables. Critical variables must be matched exactly. They function as cell dividers, rather than as link variables; e.g. data from men is always only augmented to male recipients. In that case, a separate augmentation is performed within each cell (Sharot, 2007; Rässler, 2002, p. 19). Before declaring link variables as critical variables, it should be checked that these variables discriminate extraordinarily well between target variable values, or that the division makes sense from a conceptual point of view. If the cells are not chosen well, correlations among variables are weakened in the augmented data (Baker et al., 1989). In an ideal augmentation, the differentiation between critical variables and matching variables becomes obsolete, because all link variables perfectly predict the target variables (Sharot, 2007).

During the data screening step, both the recipient unit and the donor unit are reduced to the relevant link variables, so that these are similar to both units. It might be necessary in the following data preparation step to deconstruct or combine variables in order to harmonize the concepts or formats. If the logical choice of link variables yields a number of link variables too big for further processing, it is possible to aggregate these variables or to reduce their dimensions.

Target variable selection

The target variables contain the new information relevant to the marketing problem. Whereas the term target variable is also applied by Gilula et al. (2006), other terms used are specific variables (Rässler, 2002, p. 16) or critical variables (Kamakura & Wedel, 1997). Target variables differ greatly among data augmentation applications, because they highly depend on the branch, the company, and marketing goals. The need for these variables can be derived from every day work or might be the result of expert interviews in departments for which the marketing department is working.

As stated before, the link variables should be able to discriminate between the target variable values. This is a limiting factor in choosing target variables. If none of the link variables is able to discriminate between the target variable values, the respective values cannot be augmented. The predictive power can be estimated from a test dataset. An example of how this can be done is given in chapter 5.2.2. When aiming at aggregated results, it might be reasonable to augment variables that do not have a strong relationship (Baker et al., 1989). However, if link and target variables have only a weak predictive relationship, target values will be close to being randomly distributed among recipients (Adamek, 1994). The error related to weak predictive relationships is referred to as prediction error.

Element selection

The selection of relevant elements reduces the rows of the customer database to a relevant number of customers as defined by the data augmentation frame. Only those customers with meaningful values for all link variables are relevant. Often, inactive customers in the customer database do not have sufficient values for all relevant link variables; e.g. regarding the transaction history. If, for example, an external market research source is used as a donor unit, it covers the German, and German speaking, population (or that from any other country) in private households aged 14 to 69 (Hofsäss

& Engel, 2006, p. 75). The customer database has to be adjusted to fit this sampling frame, because it cannot be assumed that customers with different nationalities or from other age groups can be regarded the same way as those belonging to the sampling frame. This does not mean that the source has to be identical to the recipients in the customer database. However, they have to be sampled from the same overall population.

The unit on which data is measured needs to be similar. It is straight forward that the units in customer databases are customers. The units in surveys are usually persons as well, but can also be households. Online and click stream data is often based on IP addresses, whereas customer profiles from online shopping portals are not seldom used by multiple persons. The different persons using IP addresses and online shopping profiles have different needs, wants, and interests. Furthermore, it does not only matter that the units in a source are persons, but also that information is stored on a personal basis. Social media data is often stored based on relationships or activities, so that interesting information on individuals might have to be translated to the person unit first in order to have an integral donor unit.

3.2.2 Data preparation

The objective of the data preparation step is to harmonize the data in the recipient and donor unit in a way that they agree in terms of concepts and definitions. Issues like missing or deficient values, types and frequency of data errors, and aggregation levels of the data are regarded when preparing the data (Hippner & Wilde, 2001).

Data harmonization

If data augmentation sources are not designed to be added to the customer database, several preparation steps have to be performed in order to achieve the required conditions. The process of transforming the link variables observed in the donor and recipient unit in a way that they are in accordance with the same concepts and definitions is referred to as harmonization or homogenization (D'Orazio et al., 2006, p. 163f). Variables might differ in terms of definitions, formats, scales, date, and mode of data collection. If variables cannot be harmonized, they should be omitted from the augmentation (D'Orazio et al., 2006, p. 167). Errors related to the harmonization of variables are referred to as harmonization errors.

Definitions Definitions cannot be adjusted. This problem is referred to as semantic heterogeneity. It means that either two different things are called by the same name, or the same things are called by different names in different sources (Dey, Sarkar, & De, 1998). Link variables defined differently in the donor and the recipient unit can only be used if the analyst decides that they refer to the same idea, nevertheless. This is a responsible task, as sources and variables can almost always be interpreted in different ways (Ozimek, 2010) and it is usually not possible to ask the data collectors which definition they had in mind.

Formats and scales The term recoding or recategorization (D'Orazio et al., 2006, p. 167) comprises transformation, grouping, and scale conversion methods. Recoding is necessary in order to align the formatting of variables in donor and recipient unit, to improve the information content, or to adjust the data for the requirements of specific data augmentation methods (Küsters, 2001). Sometimes, existing variables need to be combined in order to represent the same information as variables in the other source.

- During a *schema transformation*, variables are formally transformed into a structurally or semantically similar schema (Leser & Naumann, 2007, p. 116). It can comprise changes of labels, notations, or similar.

- If a mathematical formula is applied to a variable (e.g. sum or exponential function), it is referred to as *transformation*. In order to be mathematically transformed, variables need to be metric. Knowledge on the variables and its relationships is necessary in order to perform a mathematical transformation (Hippner & Wilde, 2001).

- It is possible to group categorical variables in order to increase their information value. With the help of *conditional mapping*, each value of the original variable (if) is assigned a new value (then).

- If a variable is transformed to another scale level, it is referred to as *scale conversion*. For example, the age is usually coded as age groups in surveys. Scale conversion to a lower scale level involves a loss of information (Backhaus, Erichson, Plinke, & Weiber, 2008, p. 10), but nevertheless might increase the information value of variables (Küsters, 2001). Sometimes, the loss of information is introduced on purpose in order to better generalize augmentation results or absorb outliers or other unexplainable noise (Adriaans & Zantinge, 1998, p. 44; Liehr, 1999; Weiss & Indurkhya, 1998, p. 59).

Date It is generally not possible to correct the data for not having been generated at the same point in time. Only certain variables can be corrected for time differences, e.g. the age variable. But as data augmentation does not require the units to be identical, it might be more meaningful to retain the correlation between time related information and other information – instead of adjusting only some variables for time differences. For example, the age of a 19 year old student should remain correlated with the interest in college entrance information, instead of harmonizing him to the 21 year old he might be today (the student is probably not interested in college entrance information anymore). Furthermore, whenever there are interrelations between those variables and other variables, these interrelations will be distorted by only changing some of the variables.

Mode If the same information is collected through different channels, the modes of data collection differ. The method of data collection, e.g. the method of interviewing in a survey, can influence the answers of interviewees. The error related to differences due to varying modes is referred to as mode effect (Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011). Data cannot be corrected for mode effects and thus knowing about different modes should lead to a cautious use of information.

Treatment of missing and deficient values

Sometimes values are missing for a number of elements and variables. When referring to missing data in the data preparation context, we are not concerned with the missing data problems described in chapter 3.1.1, lending concepts and methods to data augmentation. Rather, it is referred to customers with incomplete link variable values. A recipient cannot receive information based on link variables, if not all link variables are observed for this person. If values are missing here, the respective elements or variables can be eliminated from further application (Bankhofer & Praxmarer, 1998). Alternatively, missing values can be imputed, if it is possible to find a reasonable value for the gap. Missing values can be replaced by imputation with location parameters, such as an arithmetic mean, median, or mode (Bankhofer & Praxmarer, 1998; Hippner & Wilde, 2001). Most missing values, however, can only be imputed reasonably by a more complex model (Küsters, 2001). Those models are based on either logical variable combinations or methods verifying multivariate structures (Backhaus et al., 2008, p. 11). Missing values for ordinal or categorical variables can sometimes be coded validly as missing, if the missingness itself has a meaning (M. Berry & Linoff, 2004, p. 74) – analogous to ”not applicable” in a survey. This is a very clean way of treating missing values, if the missing items category leads to meaningful conclusions in the interpretation, e.g. the missingness of a telephone number can be connected to the social status of a customer.

If values are detected to be deficient, i.e. if the values do not fit the domain of definition, if plausible relationships are violated, or if there are correlations between variables, but singular values do not correspond to these correlations (Hippner & Wilde, 2001), they can similarly be corrected or eliminated from the analysis. The treatment of deficient values because of violation of domain of definition or plausible relationships is advisable. The violation of correlation structures should be monitored more closely, as different correlation structures could be indicators for new and interesting discoveries (M. Berry & Linoff, 2004, p. 592ff; Küppers, 1999, p. 27).

We want to make clear at this point that the correction of missing or deficient values should be kept to a minimum in data augmentation applications. Replacing missing or deficient link variable values and further using these link variables for data augmentation can lead to inestimable uncertainty and bias, as data augmentation is an imputation process itself. The satisfactory situation of the resulting complete dataset is intriguing, because it leads to false confidence and can produce false models leading to false conclusions (Dempster & Rubin, 1983). There is always a trade-off between the additional knowledge opportunities and the uncertainty created.

3.2.3 Choice of the best data augmentation method

Several data augmentation methods have been developed in the past and can alternatively be chosen for data augmentation problems. They have differing features and varying degrees of complexity. Without claiming to be exhaustive, some of the most common methods are introduced in the following.

Conditional mode or median imputation

Mode and median imputations are methods of imputing target values by measures of central tendency. They are conditional, because for a given link variable class, the most frequent value is augmented to the recipients for categorical variables. The median can be augmented for ordinal variables, depending on the data augmentation goal. A link variable class is defined by a unique combination of link variable values similar for all elements contained in this class. In this context, the link variable classes are referred to as adjustment cells (Little & Rubin, 2002, p. 62). Mode or median imputations are non-parametric, because they do not require a certain kind of distribution. They are explicit modeling techniques, as the assumptions used for drawing from the predictive distribution of the variables are explicit (Little & Rubin, 2002, p. 59).

More rules can be established in order to improve the augmentation, depending on the upfront knowledge of the augmentation setting. If the overlap is low, a threshold can be introduced in order to account for uncertainty. Because the source is a nonprobability sample, a donor class with few donors might not properly reflect the target value range of that class in the overall population. The database marketing analyst can decide to augment the mode or median only, if at least a minimum number of donors occupy the respective class. An example for conditional mode imputation using thresholds is shown in table 3.1. In this example, recipients in the age group 50-60 are not augmented a target value, because only one donor represents that class. In order to further reduce uncertainty, a minimum acceptable percentage of elements carrying the mode value in a class can be specified, so that only frequent values are augmented. In the example shown in table 3.1, recipients in the age group 40-50 years are not augmented a target value, because both values are equally likely. In contrast, if recipient unit and donor have a 100% overlap, the mode or median is augmented even if only one donor is available. If only one donor with the respective variable combination is available, the classes are narrow enough for the correct match to be found.

If the rules are very strict in order to reduce uncertainty, it is likely that no value can be augmented for infrequent link variable classes. In that case, it is possible to iteratively collapse the link variable classes by omitting one link variable at a time, as shown in the last row of table 3.1. Because no values were augmented for age groups 40-50 and 50-60, the age group is collapsed to a 20-year range. For recipients aged 40-60, value 1

Link variable class	Values observed	#	% 0	% 1	Target value	Reason
Age 20-30	0 0 1 1 1 1	6	33%	67%	1	The probability for $y = 1$ is highest.
Age 30-40	0 0 0 0 1	5	80%	20%	0	The probability for $y = 0$ is highest.
Age 40-50	0 0 1 1	4	50%	50%	none	The probability for each value is equal. (threshold: 60%)
Age 50-60	1	1	0%	100%	none	The number of donors observed is too small. (threshold: 3 donors)
Age 40-60	0 0 1 1 1	5	40%	60%	1	After the age group has been collapsed to 20-year ranges, the probability for $y = 1$ is highest.

Table 3.1: Example for values to be augmented given a certain target value distribution using conditional mode imputation

can then be augmented. That way, classes become broader and matches can be found with the same threshold criteria. However, all of these rules (threshold=60%, threshold=3 donors, collapse of age groups) are arbitrary. They depend on the decision of the database marketing analyst and results depend on these decisions. No advice can be given on how to best make these decisions.

Conditional mode or median imputation is an easy way to find basic tendencies in a sufficiently large population. But there are obvious problems related to imputations based on measures of central tendency. The variance is underestimated for these imputations (Rässler, 2000). All customers in a certain class are assigned the same target value, so that the results are meaningful only if there is a sufficiently large number of classes and if these classes are largely homogeneous. While this is more or less true for any augmentation method, the conditional mode imputation approach has further constraints due to the many arbitrary decisions to be made during the imputation process.

Nearest neighbor hot deck

Nearest neighbor hot deck is one of the traditional approaches to data augmentation. The target value is taken from the donor that is closest to the recipient in terms of a distance measure (Hattum & Hoijtink, 2008b) in a geometric hyperspace (Breur, 2011). Calculating a distance function is more obvious in a continuous variable case. In the discrete case, the categories need to be binarized or transformed in order to make the calculations of distance functions possible. After all link values have been converted to metric variables, the distance between elements is measurable. If no exact neighbor is found, it is possible to collapse the categories to a lower level of detail in order to find exact matches (Kamakura & Wedel, 1997).

Nearest neighbor methods are distinguished depending on whether target variable values are selected with or without replacement. If donor and recipient are identical in terms of elements, values can be allocated without replacement so that the overall distance is minimal. Every donor value is used exactly once and not every recipient might receive the closest values. Because values are drawn *without* replacement, the minimum overall distance is not necessarily equal to the sum of individual minimum distances. Representative sources can be regarded like identical sources, because every donor represents a definable number of observations. Values are quasi allocated without replacement from a source, using every donor x times. If the best value is desired for every individual customer, and if sources are not identical in terms of elements, drawing with replacement is the better choice. When drawing *with* replacement, the minimum overall distance is equal to the sum of individual minimum distances.

Nearest neighbor hot deck by means of a distance function should not be confused with the earlier approaches of hot decking by sorting. When performing hot deck imputation by sorting, all observations are sorted in a defined order by the link variables and missing values for the target variable are taken from the previous or following observation (Baker et al., 1989). The sorting order is crucial for this method and significantly influences the results, even if serpentine sorting is used for ordinal variables in order to minimize distances (Carlson, Cox, & Bandeh, 2012). If many categorical variables are present, it is not recommendable to use hot deck imputation by means of sorting.

Hot deck procedures duplicate an existing value and fuse it to another observation (Ford, 1983; Hattum & Hoijtink, 2008b). In contrast to conditional mode or median imputation, nearest neighbor hot deck augments values from close neighbors, not only from the same link variable class. That way, small differences in the link variables are overcome. If an exact match does not exist, because there is more than one donor with the respective variable combination, and the target variable values of the donors differ, the most frequent value is usually used.

There are restrictions to the nearest neighbor hot deck procedure as well. A distance function is not easily calculated for categorical variables, consigning the problems related to using methods developed for numerical data on categorical data. Nearest neighbor hot deck and conditional mode or median imputation techniques are based on rules stated by the database marketing analyst to the best of his or her knowledge. Because they are based on the database marketing analyst's decisions, the quality of the augmentation is highly dependent on these decisions (Kamakura & Wedel, 1997), e.g. in terms of the choice of the distance measures and sequence and the definition of the levels of the augmentation procedure in terms of categories and variables. These methods are not able to detect new structures in the data.

Multivariate methods

Regression methods and latent class analysis are two forms of multivariate methods. They include several variables in order to predict an outcome variable, based on a statistical model. Multivariate methods are able to detect new structures of underlying distributions of variables in the data, because the influence and interaction of predicting variables can be determined and analyzed. Once these have been verified, the same model parameters can be applied to other subpopulations.

Regression methods have been used for missing data problems as described by Little and Rubin (2002, p. 59) and Schafer (1997, p. 197), before having been introduced to data augmentation. Instead of directly taking values from existing donors, as it is done with conditional mode imputation and nearest neighbor hot deck, regression methods are built on a model describing the relationship between several input variables (the link variables) and an output variable (the target variable). Values are even predictable for link variable classes not present in the donor unit.

In contrast to linear regression, logistic regression is suitable for categorically predicted variables. Good descriptions of logistic regression are given by Agresti (2002) and Backhaus et al. (2008, p. 243ff). The logistic function is a probability function of whether an event will occur or not (Backhaus et al., 2008, p. 249). For the overall approximation of Y given X, the likelihood function has to be maximized, thus calculating the parameters estimated for all x in X. Logistic regression methods can be differentiated into binary logistic regression and multinomial or polychotomous logistic regression (Backhaus et al., 2008, p. 244; Ratner, 2003, p. 169ff). Both alternatives are needed for data augmentation in database marketing, depending on whether target variables are binary or have multiple values. If the predictor variables are categorical, a binarization is needed, because logistic regression requires numerical input variables.

Latent class analysis is based on the assumption that there are not observed, thus latent classes, which are similar in terms of their target variable values. The latent classes are the target values and the explanatory variables are the link variables. Hattum and Hoijtink (2008b) carried out a data fusion based on latent class analysis. Their approach consists of two steps. First, the fusion value specific probabilities are estimated from the donor unit. Secondly, the estimated model parameters are used to fuse data to the recipient unit using the classification rule of latent class analysis. The latent class approach is intuitive, because the target variables can be fully explained by the factors of the latent classes (Kamakura & Wedel, 2000). However, it can only be used if the number of classes, thus the number of differing segments, is small. Hattum and Hoijtink (2008b) showed that latent class analysis results perform similar well as logistic regression results.

Multivariate methods do not depend on the overlap between donors and recipients. Whether target values are predicted correctly depends on the predictive power of the link variables. Variations not explained by the link variables can influence the quality of the results. If predictions worked better for one dataset than for another, a model overfitting problem exists. Model overfitting can be minimized by building a model on a training dataset and applying and validating it on a test dataset for which the true values are known, first.

Multivariate methods are generally able to handle large numbers of variables. If carried out correctly, valuable insights into the data structure are possible (Baker et al., 1989). On the other hand, regression models must be built with respect to all conditions expected for the methods in order to receive meaningful results. This can pose a problem in marketing, if the available data does not satisfy these rules. Possible interactions between independent variables harm the results (Baker et al., 1989). Furthermore, maximum likelihood methods strongly depend on the predictive power of the link variables, which can be a constraining factor.

Model based multiple imputation methods

Multiple imputation is one of the newer methods used for data augmentation. In order to account for the variability in the data, a set of plausible values that indicate the uncertainty about the right value is imputed. Resulting values are combined to form an aggregated figure with according probabilities. Multiple imputation refers primarily to the fact that values are imputed various times. A method still has to be specified for calculating the values. The number of imputations is set to a very high number in order receive as many different results as possible. The more imputations are carried out, the better the values represent the underlying distribution. The result is a vector of possible values with the values reflecting both variation and uncertainty (Allison, 2012; Rässler, 2000, p. 69). Multiple imputation is used in the context of maximum likelihood estimation using Bayes techniques (Little & Rubin, 2002, p. 97ff) including Markov Chain Monte Carlo methods (Rässler, 2000). Details on these methods can be found in Schafer (1997), Little and Rubin (2002), Rässler (2002), D'Orazio et al. (2006), and references therein.

The feature of multiple imputation approaches is that the uncertainty caused by missing data is represented better than in methods where only a single value is imputed (Herzog & Rubin, 1983). Multiple imputation has the advantage that it does not rely on the assumption of conditional independence, as various imputations can be carried out with different parameters of conditional associations (Rubin, 1987, p. 187). The algorithm converges at the best solution possible. A disadvantage of multiple imputation results is that they are not reproducible, unless the statistical program allows for defining a seed variable. Multiple imputation methods are applicable for augmenting data in marketing. However, they are very complex and exceed the scope of this study.

Testing, calibration, and choice

In order to choose the best method, relevant methods are selected depending on the scales of variables and the goal of the augmentation. A test design is applied prior to the actual data augmentation in order to choose the model to be applied to the recipient unit. A cross validation design is proposed by Hattum and Hoijtink (2008a) in which the donor unit is divided into multiple datasets. One dataset is defined as the training dataset on which the model is built. The other datasets are used as test datasets, so that models can be fitted to a training group and then be applied to a test group. Its target variable values are augmented by the model developed for the training dataset and compared to the true values.

3.2.4 Execution and internal evaluation of results

Once the decision towards a data augmentation method has been made, the augmentation can be carried out. The model is applied to the recipient unit and new variables are added to every customer with corresponding probability measures. The details of the augmentation model are described in chapter 4.2. The new values are filed for further application into the customer DWH.

In order to provide insight into the quality of the augmentation, it is evaluated. Hattum and Hoijtink (2008b) differentiate the evaluation of data augmentation results into internal and external evaluation criteria. Internal criteria are useful to judge the reliability and validity of data augmentation. External criteria are used to evaluate the effectiveness and efficiency of the added information. With external evaluation, it can be retroactively decided whether a data augmentation leads to a return on marketing investment. In the following, the internal evaluation criteria for reliability and validity of the augmentation results are described. Some external evaluation criteria are summarized in the successive chapter.

Reliability of data augmentation results

The goal of data augmentation is to create target values that are more accurate than the results of a random allocation of missing values (Baker et al., 1989). This overall reliability goal is directly influenced by each individual customer's augmentation results. The reliability of the results is testable successively during the data augmentation application. It can be summarized later in order to get a clear picture of the augmentation's reliability. The reliability is mainly dependent on the number of link variables, the correlation between link variables and target variables, the characteristics of the source, the overall applicability of definitions, and the accuracy and maintenance of the used data. We describe possible errors in the data augmentation process, e.g. errors in the primary source, errors in the data preparation phase, or even errors in the actual data augmentation process. A good general introduction into error sources during data integration can be obtained from Zhang (2012), from which most error sources are cited.

Reliability of the source and data preparation The reliability of the source comprises element errors and variable errors. Elements are erroneous, if they are not consistent with the population frame. If a source is representative, additional sampling errors can occur. Self-selection bias refers to the fact that volunteer surveys are not representative, because survey participants decide on the participation, rather than the survey provider. Variables are erroneous if they are measured incorrectly, if they have been collected at a previous point in time, or if the mode of data collection influences the variable values. The reliability of the data preparation refers mainly to the correct harmonization of link variables in donor and recipient unit, so that they adhere to the same concepts, definitions, and formats.

Reliability of the augmentation The reliability of the augmentation can be divided into the reliability of the model and the reliability of the method. The reliability of the model compromises coverage errors, identification errors, and correlation errors. Coverage errors refer to the incapability of the model to fit the target population. Identification errors refer to the incapability of the model to identify the right elements. There are several method specific errors thinkable influencing the reliability of the method. Exit criteria for nearest neighbor hot deck, for example, are equal distances between a recipient and several donors.

Reliability of the results The reliability of the augmentation results is influenced by all previous errors. Augmentation results are erroneous, if the assigned target values are wrong. They can be false negatives (type I errors), if an interesting value is not assigned although the customer has that value, or false positives (type II errors), if an interesting value is assigned although the customer does not have that value (Jiang et al., 2007). Prediction errors result from weak predictive relationships between link and target variables, so that the probability of the two error types is high. The lower the degree of overlap between donor unit and recipient unit, the less likely there is a good match for a recipient, given the link variables (Adamek, 1994). The respective error is referred to as matching bias. Matching bias does not necessarily result in wrong values, but the probability for wrong values increases.

Validity of data augmentation results

Rässler (2004) established four levels of validity for evaluating augmentation results.

Preserving individual values With the first level of validity, it is assessed how well original values are reproduced. Whenever the added value is equal to the original value, the match is called a hit. The hit rate is the overall measure for the validity of individual variables. Hits only occur for discontinuous variables, because the probability of a hit for continuous variables equals zero (Rässler, 2004). The hit rate is one of the most important measures for data augmentations in marketing, because it matters whether individuals received a correct value. Hits can only be analyzed in simulated settings. In practice, the original values are not known.

Preserving joint distributions For the second level of validity, the joint distribution of all variables should be preserved in all samples, as well as in the augmented dataset. This is especially important for cross tabulation, because otherwise it cannot be guaranteed that the association between variables never jointly observed is valid (Rässler, 2004). An ideal data augmentation would preserve individual values as well as joint distributions. However, in an imperfect setting, a trade off exists between preserving individual values or joint distributions when choosing an augmentation model.

In order to pursue the goal of targetability improvement for marketing purposes, the decision is made for the preservation of individual values.

Preserving correlation structures The third level of validity refers to preserving correlation structures between variables and higher moments of distributions (variance, skewness, and kurtosis). Just as for the second level, an ideal data augmentation would preserve correlation structures, but an imperfect data augmentation might not (Rässler, 2004). It has been shown that data augmentations are generally not able to reproduce all major correlations in the data (Baker et al., 1989). In our study, whenever data augmentations are not able to exactly reproduce the original distributions or correlations, models are adjusted to preserve individual values, rather than distributions.

Preserving marginal distributions The fourth level of validity refers to preserving marginal distributions. It is required for all data augmentation applications and is the one always testable, because the empirical marginal distributions are inherent in the augmented files, as well as in the donor unit or recipient unit respectively. In practical applications, the validity of the fourth level is often assessed and overall assumptions on the validity of augmentations can be made (Rässler, 2004).

Means, relationships, and correlations can be compared in order to internally evaluate the results (Putten, 2010, p. 92). For general applications in database marketing, the first level is of major importance (Hattum & Hoijtink, 2008a). When conducting a targeted campaign, it is of primary interest to have a selection criterion on whom to choose for target groups. However, sometimes it might be interesting to apply descriptive or inductive statistical methods on a synthetical dataset. A likely case might be to investigate the distribution of income among customers. Then, it should be kept in mind that the data augmentation approach in marketing is always one where individual values are preserved, rather than overall distributions. Augmentation methods are chosen to serve this objective. Unless the results are perfectly accurate, the statistical estimates applied to the whole customer group can be biased. If a macro objective was the major focus of a data augmentation approach, different methods and decision measures would have to be chosen.

3.2.5 External evaluation of augmentation results

The external evaluation assesses the effectiveness and efficiency of data augmentation results, e.g. whether the results are able to support the targeting goal of database marketing analysts (Hattum & Hoijtink, 2008a). Baker et al. (1989) state that the result should be evaluated by how accurate decisions can be made from augmented data sources. It can be translated into the value of data augmentation for further analysis (Putten, 2010, p. 92). This directly links the evaluation of the augmentation to the respective marketing problem for which the augmentation has been set up. The improvement of company performance based on data augmentation results is more important than the accuracy of the results themselves. Data augmentation results are good, if marketing campaigns based on the results outperform campaigns without this knowledge. If this is the case, all other evaluation criteria are secondary.

Hattum and Hoijtink (2008a) stress that the levels of validity in the internal evaluation are secondary, if data augmentation results are useful in practice – for example if conversion rates or sales leads are increased by using them. An advantage of direct marketing is the exact testability of results by means of conversions (Breur, 2011). One way of evaluating data augmentation results would be to conduct a targeted campaign with a test setup. In a this test setup, a target group is selected by traditional selection criteria and another one is selected using the augmentation results. If customers are randomly attributed to one of these groups, the resulting conversion rates give a direct estimate on how effective the data augmentation results are in a practical application. Evaluating the effectiveness is equal to the question whether marketing campaigns perform better using data augmentation results than when not using them. If all other factors are equal, the marginal effect of the augmentation results can be observed.

When evaluating the efficiency, it is asked whether the cost of conducting data augmentation can be justified taking into consideration additional revenues and cost reductions triggered by the augmentation results. If the new data is definitely necessary, it can be compared to other means of acquiring the data, for example a full customer survey. One of the major external evaluation criteria has been introduced by Even et al. (2010) as data utility. Data utility assesses how useful data is in a simple comparison between the cost of acquiring and retaining (up-to-date) data and their associated value for the company. The value can be constructed from different objectives. It is not easy to calculate, as augmented data can be used for various other purposes not even known at the time of the augmentation. Thus, an evaluation of data augmentation based on one single use case would not fully embrace the efficiency of the augmentation (Putten, 2010, p. 90). There might also be different utility measures for individual marketing goals and for the overall customer database. There is usually a certain utility inequality among data categories in the database (Even et al., 2010). Nevertheless, as long as the utility is positive, it has the potential of increasing a company's profits.

The efficiency of data augmentation results can be evaluated by a variety of measures. Some of them are derived from the cost-per-effective-target market rating point, which is used in media planning (Smith et al., 2010). Some are specific to database marketing. There are fix and variable cost of advertisement. Both can be reduced by data augmentation. The relative concentration of targeted customers in the selected recipient group is a major factor in evaluating the efficiency of individual marketing campaigns. Another factor is the degree to which the targeted customers are effectively exposed to the advertisement. As the conversion probability of targeted and non-targeted customers differs, data augmentation can increase the overall conversion probability by including more target customers into the addressed groups. We do not go into detail about specific external evaluation criteria here, because the external evaluation is not in the scope of this study. It would, however, pose a valuable extension to our work.

Chapter 4

Methodological framework for data augmentation

Data augmentation in database marketing is a special field of data fusion. No general framework for this kind of data augmentation has yet been established. The scientifically regarded and practically approached augmentations in marketing are limited to representative sources, which have convenient features comparably easy to handle. The theoretical contribution of our study is to establish a methodological framework for augmenting external sources. It regards the specific case of database marketing, where the customer database is the recipient unit. Possible donor units are found within internal and external databases or are derived from surveys and other data sources. They are formally described in this chapter.

There are specifics to data augmentation in database marketing. The variables collected in surveys and other external sources are usually of categorical nature. Target variables are augmented respecting a micro validity approach, which is different to other data augmentation use cases, e.g. in official statistics. The situation of conditional associations between link and target variables and the source is described in detail.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_4

A general data augmentation model is established in this chapter, mathematically describing the populations, variables, and outcome values with respect to the specifics stated. The proposed data augmentation model is based on a univariate pattern approach. Target values are augmented separately and the most likely value is augmented. The uncertainty related to this approach is captured in dedicated probability variables.

Additionally, the source data mechanism is formally described. It is based on Rubin's (1976) and Little and Rubin's (2002) theory of the ignorability of the missing data mechanism. If there is a correlation between the source data mechanism and the target values, data augmentation results can be biased. Such sources should not be used for data augmentation purposes. The transfer and development of Rubin and Little's theory to data augmentation use cases in database marketing is part of the theoretical contribution of our study.

4.1 Data augmentation specifics in marketing

Data augmentation in database marketing has special features distinguishing it from other augmentation approaches. They influence the way data is augmented. The recipient unit is always the customer database. Possible donor units vary, but can be generally described by characteristics such as overlapping units between recipient and donor unit, number of observations contained, or being a representative sample of a bigger population or not. We have already described the customer database as recipient unit and categories of donor unit data in our previous data augmentation study (Krämer, 2010). Chapter 4.1.1 and 4.1.2 are partially resumed therefrom.

In chapters 4.1.4 and 4.1.5, the notion of the source data mechanism is introduced. It is based on Rubin's (1976) and Little and Rubin's (2002) theory of the missing data mechanism as depicted in chapter 3.1.1 and is adapted here to fit the intentionally designed situation of missing data in the data augmentation context. Its relationship to the target variables – or more precisely, its conditional relationship to the target variables, given the link variables – is explained in detail with an example in order to introduce this important concept.

4.1.1 The customer database as recipient unit

Data augmentation projects in database marketing coincide in having the same recipient unit: the customer database. A customer database is a systematically structured, physical collection of data, which is saved according to rules. The main key is a unique customer number identifying a customer (Kelly, 2007). A customer DWH is a customer database tailored to the specific purpose of providing information that is relevant to the analytical purposes of database marketing (Schmidberger & Babiuch-Schulze, 2009). All data from operational, internal and external sources are consolidated in a single, central interface. After every change in the customer contact or purchasing history, the optimal form and time for follow-up contacts can be determined and initiated. They support the design of business processes and management decisions (Wilde, 2001).

The content of customer databases can generally be divided into identification data, descriptive data, and transactional data (Huldi, 2002; Behme & Mucksch, 2001; Link & Hildebrand, 1993, p. 34ff). While identification data is mandatory to every company, the customer databases of different companies are specific in their descriptive and transactional data. They differ considerably in the number of variables saved and volume of information (Hippner, Rentzmann, & Wilde, 2002).

Identification data Identification data is used to identify customers and assure their reachability. They comprise name, address, birth date, telephone number, and other contact information such as email or mobile number. But this data is not always sufficient for the identification of individuals. Additionally, a unique identifier, e.g. a customer number, is allocated to every person during the first transaction with a company.

Descriptive data Descriptive data relates to any business relevant characteristics of customers. Demographic and psychographic information are part of the customer profile. Information on household structures, micro geography, and social networks are summarized as sociographic information. Descriptive data can either be provided by the customers themselves, e.g. in a detailed online customer profile, be collected via market research, or be acquired by external sources (Schweiger & Wilde, 1993).

Transactional data Transactional data comprises the purchase and contact history, as well as the product usage information of customers, if applicable. Transactional information is often differentiated into action data (initiated by the customers) and reaction data (initiated by the company). Transactional data is specific to the company. It is generated from the operating systems and is preprocessed for usage in the DWH. Product and brand affinity, the acceptance of communication and sales channels, as well as the post purchase behavior, are deductible from transactional data (Hippner, Leber, & Wilde, 2002).

4.1.2 Possible donor units and their characteristics

A data augmentation source can be any database with people as its main unit. It does not contain a unique identifier to the people in the customer database. Every potential data augmentation source is taken into consideration because it offers information not contained in the customer database. Relevant target variables are

- new, informative, and meaningful in order to be able to support and improve the decision making process
- observable and collectable in a way that data can conceive the informative value of the variables
- true and stable, so that the information is valid for a certain time

- available and accessible from an external source
- discriminable and augmentable, so that they are explainable by link variables in an adequate model
- efficient, so that the effort of performing data augmentation is an economical contribution to the company

Data augmentation sources can contain all kinds of information relevant to database marketing problems.

Socio-demographic data Socio-demographic data is a mixture of demographic and socio-economic information. It is contained in most external sources, especially in market research. They are easily collectable, measurable, of low complexity, and serve as the basis market segmentation criteria (Vossebein, 2000; Hofsäss & Engel, 2006, p. 107; Meffert & Bruhn, 2009, p. 133). Typical demographical information are gender, age, and city of residence (Becker, 2009, p. 250ff). Typical socio-economic variables are income, education, profession, marital status, and social class (Meffert, Burmann, & Kirchgeorg, 2008, p. 195f).

Psychographic data Psychographic variables provide information on motives, attitudes, and lifestyle. These variables cannot be observed. Relevant motives are the expected benefits ascribed to a purchase, which activate the customer and initiate his or her action (Meffert & Bruhn, 2009, p. 113). Attitudinal data can be differentiated in general personal information and in brand or product related information (Freter, 1997, p. 72ff,135ff). These are not easy to collect, but have a high relevance regarding the purchase behavior. Marketing strategies can be directly derived from them. Lifestyle variables are information on activities, interests, and opinions of consumers (Meffert & Bruhn, 2009, p. 114).

Purchase and behavioral data Purchase and behavioral data is directed at the customers' preferences regarding products and services, communication, price, and locations. They are the results of past transactions and can yield assumptions on future behavior (Freter, 1997, p. 157). Product and service information concerns the customers' demand in terms of brands, types, volume, and usage intensity. Information and communication behavior relates to interpersonal communication, as well as media usage. Price information comprises information on price classes and price elasticity. Location information concerns information on type, distance, frequency, and usage intensity of individual sellers (Freter, 1997, p. 157; Becker, 2009, p. 270ff).

Donor units are manifold and can be taken from various sources as described in chapter 2.3. Besides the classification in public, company-owned, and accumulated sources, sources can be more formally described by their overlap, their size, and whether they represent a bigger population or not.

Overlap External sources differ in terms of the elements they contain with regards to the customer database. The overlap is defined by the number of elements that donor unit and recipient unit have in common. Sources can contain exactly the same persons as the customer database (identical), completely different persons (distinct), or any other overlap (partially overlapping). If there is a common set of link variables, and if there is a correlation between the link and the target variables, a high overlap is not necessarily needed.

Size The size is defined by the total number of elements in the donor unit. If an external source has a 100% overlap to the recipient unit, it can further be differentiated in sources with exact the same number of elements and sources containing these elements among others. Although in terms of the overlap, both sources are equal, they differ in size. In the former case, the size equals the overlap. In the latter case, the size is greater.

Representation Representative surveys must be differentiated from partially overlapping sources, because although the actual overlap is low, these sources represent a bigger population. If interviewees are randomly chosen, the same information is represented as if all people in the overall population would have been asked. The number of elements representing a bigger population is referred to as random sample. The sampling rate is defined by the number of elements in the sample, divided by the number of elements in the overall population.

Donor and recipient unit can be of differing size and the overlap can vary from identical (100%) to distinct (0%). In database marketing, the customer database is usually of much bigger size than the source to be augmented. However, it can also be vice versa, e.g. if data from a website is augmented, containing information on existing customers plus other visitors. Different subgroups of customers can be represented unequally well in a source, if the donor unit is not identical to the recipient unit in terms of elements, or a representative sample of it. In some cases, it might not be possible to augment data on all customers, if they do not fit the same frame, i.e. if they are not samples from the same overall population. In that case, the recipient unit needs to be reduced to the respective frame so that at least those customers fitting the frame are able to receive new variables, as described in chapter 3.2.1. Provided the frame is met, sources can generally be used for data augmentation purposes.

4.1.3 Variable scales, values, and validity

There are certain specifics related to the variables used for data augmentation in marketing, especially in terms of scales, augmented target values, and validity. As database marketing deals with people, many variables in the data augmentation process have categorical variable scales, e.g. demographic variables like gender. Many attitudinal and behavioral characteristics are measured best on a categorical scale. Additionally, many variables available for data augmentation purposes are derived from surveys or other aggregated data sources. These variables, even if they were originally metric variables, have lower scales due to the data collection style. For example, age is a metric variable, but is usually collected on an ordinal scale as age groups. Oftentimes, the problems encountered in database marketing ask for binary answers only, e.g. whether a customer likely to buy a specific product or not.

Albeit the ability of more modern procedures and software to handle differently scaled variables, it is not desirable to mix variable scales from an interpretation perspective. Therefore, also the variables with metric scale are transformed into categorical variables during the data harmonization process step as described in chapter 3.2.2. The scale of the variables has implications on the techniques that can be used. The methods need to be able to deal with high dimensional data common in marketing (Kamakura & Wedel, 2000).

Categorical variables can be distinguished into nominal variables and ordinal variables (Agresti, 2002, p. 2). Nominal variables do not have any natural ordering, like gender, city of residence, or product categories. Ordinal variables have an order, such as age groups or preferred product quality. The two kinds can be mixed for tasks when associations between differently scaled variables are needed. If associations between ordinal variables are calculated, special methods are available accounting for the characteristics of ordinal variables.

The augmented values are realistic values. This is different to data augmentation with metric variables, where the augmented values can be an uneven mean or another artificial value. The decision for one of the real categorical values is a decision for the most likely value. A decision has to be made whether the most often occurring value, a random value, or an average value is supposed to be augmented (Bleiholder & Naumann, 2008). The random value approach does not suit the micro validity objective of data augmentation in database marketing. The average value approach would lead to a tendency to the center, thus not reflecting the variability in the data and not differentiating between classes. In database marketing, the decision for the most often occurring value is made. However, to augment a single best value would not account for the uncertainty involved in the data augmentation process. According probabilities need to be augmented with the best values.

Micro validity is desired for all data augmentation results. It is the first level of validity developed by Rässler (2004) as described in chapter 3.2.4 and refers to the ability of a data augmentation to correctly reproduce individual target values. To reproduce the values of individual customers correctly is the ultimate goal of data augmentation in database marketing, because decisions on individuals are derived from these values. If an overall distribution is reproduced correctly, while allocating the wrong values to individuals, it does not have any benefit for marketing. This is different to data augmentation approaches with a macro perspective. However, the micro validity does not really refer to individuals, but only to individuals as being distinguishable in terms of link variable classes. Because values are augmented based on the link variables only, no further criterion can separate those customers with the same link variable class. Therefore, micro validity is correctly translated to micro class validity in database marketing.

4.1.4 Conditional independence of source and target variables

If source and customer database perfectly overlap in terms of elements or if the source is a representative sample of the customers, the target variable is represented in that source in a way that its distribution and variances for the customer database are known, as well as its correlation to link variables.

If this is not the case, the reason why an element appears in the donor unit is not random, i.e. not MCAR. Therefore, an additional question must be posed: Is the source correlated with the target variable to be augmented? This shall be clarified with an example.

It is assumed that a company is interested in its customers' interest in shoes. It wants to use additional information received from a volunteer survey and augment it based in the link variable gender. In terms of overlap, the volunteer survey is a source that is partially overlapping with the company's customer group.

Number of observations		Interest in shoes					
(Row percentages)		Yes		No		Row sum	
Participation volunteer survey	Yes	55	(39%)	85	(61%)	140	(100%)
	No	15	(23%)	50	(77%)	65	(100%)
$\chi^2 = 5.18$	Column sum	70	(34%)	135	(66%)	205	(100%)

Table 4.1: Association table of participation in a volunteer survey and a special interest

In the example, the distribution of interest in shoes and participation in the volunteer survey is shown in table 4.1. It shall be noted that such a calculation is not possible in a practical application. The distribution of the interest in shoes cannot be known for persons that did not take part in the volunteer survey. The participation in the volunteer survey and the interest in shoes are not independent, as 39% of the participants of the survey have an interest in shoes (55) and only 23% of the others (15). Given the row and column sums, the expected number of persons having participated in the survey and being interested in shoes would have been 47. When testing the association with a χ^2 test, the observed χ^2 value is 5.18. The corresponding test value for a level of significance $\alpha = 5\%$ is 3.84. Since the observed value is bigger than the test value, it can be said with a 5% level of significance that the participation in the online survey and interest in shoes are dependent. If this is the case, the participants in the volunteer survey are more likely to be interested in shoes. Thus the augmentation results would be biased towards interest in shoes. The additional selectivity is not reflected in the augmentation (Sharot, 2007).

For simplicity reasons, the link variable to augment the information on is gender only. In categorical data analysis, the factors potentially influencing the relationship between two categorical variables are called covariates. If these covariates are not held constant, then the relationship between the two variables shows confounding (Agresti, 2002, p. 47). In other words, it is desirable to know whether the confounding factor, here the source, influences the relationship between the link variables and the target variable. The participation in the volunteer survey and interest in shows are conditionally independent given gender, if learning whether a person has participated in the volunteer survey does not provide any additional information on the interest in shoes, depending on the gender (Pearl, 2000, p. 11).

Women

Number of observations		Interest in shoes					
(Row percentages)		Yes		No		Row sum	
Participation volunteer survey	Yes	40	(67%)	20	(33%)	60	(100%)
	No	10	(50%)	10	(50%)	20	(100%)
$\chi^2 = 1.78$	Column sum	50	(63%)	30	(38%)	80	(100%)

Men

Number of observations		Interest in shoes					
(Row percentages)		Yes		No		Row sum	
Participation volunteer survey	Yes	15	(19%)	65	(81%)	80	(100%)
	No	5	(13%)	40	(87%)	45	(100%)
$\chi^2 = 1.25$	Column sum	20	(16%)	105	(84%)	125	(100%)

Table 4.2: Association between a volunteer survey and a special interest, given gender

This situation is shown in table 4.2. The respective participants are splitted into 80 women and 125 men in order to calculate the conditional association. When testing the association with a χ^2 test, the observed χ^2 value is 1.78 for women and 1.25 for men. With the same corresponding test value, no association can be detected between the participation in the volunteer survey and the interest in shoes for either subgroup. There is no basis anymore for claiming that the categories are dependent, given the link variable. Cases like this are referred to as Simpson's paradox (Blyth, 1972; Dawid, 1979; Simpson, 1951).

It becomes obvious that there is an association between gender (the condition) and interest in shoes (the target): Women are more interested in shoes than men. This association between the link variable and the target variable is not only comprehensible from a contextual point of view, it is also necessary in terms of data augmentation requirements. For both link variable classes, no dependence exists between the participation in the volunteer survey and the interest in shoes. In other words, if the distribution of interest in shoes and participation in the online survey is predictable by gender alone, then data augmentation based on gender is necessary and sufficient to preserve all relationships of interest (Sharot, 2007).

The idea of conditional independence in the context of data augmentation has been examined and described particularly by D'Orazio et al. (2006) and Rässler (2002), after having been fathomed by Rubin (1976). In our case, it has a slightly different meaning as the conditional independence needs to be established between the source and the target variable, given the link variables, rather than target variables that have never been jointly observed. This conditional independence is important, because information on the target variable is derived from a given source for the customers, who might not be identical to the people observed in the source.

The assumption of conditional independence is not testable, unless an auxiliary source is available with information on the conditional association. Otherwise, conditional independence may be assumed based on expert evaluations, which is common practice in data augmentation (Rässler, 2002, p. 4). This is disputable and has been argued for example by Rodgers (1984) and references therein. They claim that data augmentation results calculated under conditional dependence are not valid and lead to estimation errors. The conditional independence assumption can be a disadvantage, if it is not testable and assumed spuriously. If auxiliary information is available, the calculation itself is another challenge. Data augmentation approaches in database marketing often face the problems that variables have categorical or different scales, are differently skewed or have different numbers of possible values. This makes it difficult to find a model to efficiently prove conditional dependence or independence (Rässler, 2002, p. 4). In chapter 7, we explore whether it is permissible to assume conditional independence in the context of database marketing.

4.1.5 Source data mechanism of the donor unit

It has been stated in chapter 3.1.1 that data augmentation is a special form of missing data problem. However, unlike the original intention of missing data problems, data is missing intentionally. This means that a decision has been made at some point that data is missing, either by systematically omitting observations or variables from data collection (e.g. sub-sampling or split questionnaire survey design) or by acquiring a source from which it is known that it does not contain all observations or variables of the recipient unit. Often, the situation of missing data is only just created by regarding two data sources, where one contains more information than the other, so that a missing data situation evolves from this composition. Some of the patterns shown in figure 3.1 on page 75 in chapter 3.1.1 can have intentionally missing data: univariate (3.1b) and multivariate patterns (3.1c), split questionnaire survey design (3.1e), statistical matching (3.1f), and factor analysis (3.1g).

In the context of data augmentation, the missing data mechanism is referred to as *source data mechanism*. It focuses on the fact that data is *available* in the source (created by the source data mechanism), rather than on the fact that data is *missing* in the customer database. This is illustrated in figure 4.1. In the overall population (e.g. all people living in Germany), all link variables (blue) and target variables (gray) exist. The overall population is not observable. The external source contains values for all link and all target variables. The customer database contains the link variables only. In the induced study design, the target variables are missing for all customers. The binary source data mechanism indicator variable S marks the observations with target variables with a 1 and the observations with missing values with a 0. It is not observable. For the data augmentation problem, it is of interest whether there is a correlation between the source data mechanism and the values of the target variables.

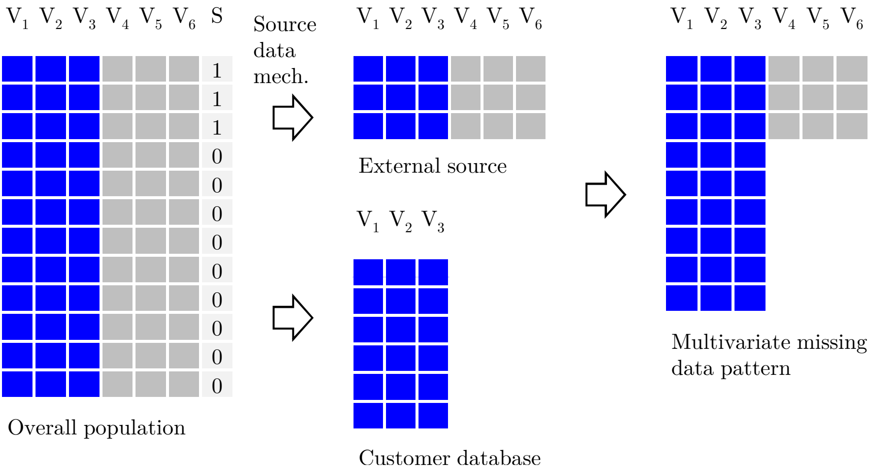


Figure 4.1: Source data mechanism in the data augmentation context

Technically speaking, there is also a *customer data mechanism* for the customer database. It is assumed that the mechanism that leads to a customer to be present in the customer database does not influence the missing values of the target variables. This is based on the initial need to differentiate customers based on target variables. If all customers were equal in terms of the target variable, it would obviate the data augmentation approach. For example, if all customers were interested in shoes, because only those interested in shoes become customers of a company (e.g. if the company is a shoe store), it would obviate a data augmentation with interest in shoes as target variable.

Supposition: It is assumed that the data mechanism of the customer database does not influence the target values.

Data augmentation is only valid without potential bias, if this supposition is true. Then, the data mechanism of the customer database is ignorable. The supposition is a requirement for the development of the theoretical data augmentation model in chapter 4.2. It can be relaxed in practice, because we are able to show in chapter 7 that the association of the data mechanism and the target variable does not compromise the data augmentation results in a categorical data augmentation context, if there is a strong relationship between link and target variables.

4.2 Data augmentation model

The following model comprises the notation and formal description of the data augmentation process in database marketing. It includes important interrelationships and statistical concepts relevant to data augmentation in database marketing. The population, all relevant samples, and their source data mechanisms are depicted from a formal point of view. The relationship between samples is outlined and implications are explained. Especially the conditional independence assumption is clarified. Categories of variables and preconditions are explained. The expected target variable values are explained and according probability and variance measures are given. The essence of this chapter is a formal definition of requirements a source must met in order to be suitable for data augmentation in database marketing.

In the context of probability notation and set theory, variables are noted in capital letters, whereas the individual values, i.e. realizations of variables (Marinell & Steckel-Berger, 1995, p. 241), and dimensions of matrices are noted in small letters. Population matrices are noted in bold letters. KPIs are generally noted in capital letters, but notations are adopted from the respective literature, where applicable. The notation of the following model is based on notations used e.g. by Rässler (2002), D'Orazio et al. (2006), Ratner (2003), and Hattum and Hoijtink (2008a), but is adjusted to fit the case of data augmentation in database marketing.

4.2.1 Populations and samples

For data augmentation in database marketing, a recipient unit is augmented with data from a donor unit in order to receive a full rectangular dataset with target variable values for all recipients. Donor and recipient unit are part of an overall population. In figure 4.2, the illustrations of figure 1.1 on page 11 in chapter 1.2.1 and figure 4.1 on page 118 in chapter 4.1.5 are combined and complemented by the abbreviations for the variables, which are explained in this chapter.

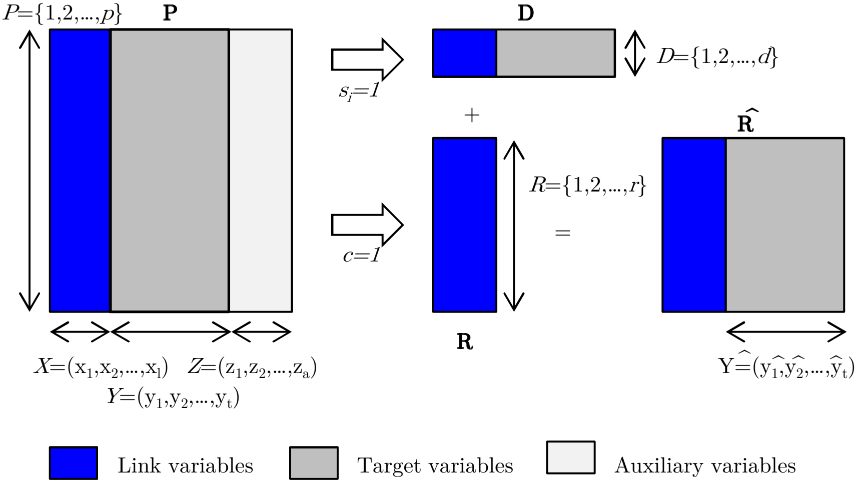


Figure 4.2: Notations used in the data augmentation model

The elements in the customer database form the recipient unit in the data augmentation model (R) with a defined number of customers or recipients (r). R is a set of customers $R = \{1, 2, ..., r\}$. The elements of the external source form the donor unit in the data augmentation model (D)

with a defined number of donors (d). D is a set of persons $D = \{1, 2, ..., d\}$. As an illustrating example, D shall be the group of people who took part in a volunteer survey. The customers are those of a publisher. Both R and D are samples of the overall population in the data augmentation model (P) with a finite overall number of elements (p). In the described example, the overall population is the German population between 14 and 69 years. R, D, and P are vectors, because they denote sets of elements, regardless of the variables describing them.

There can be an overlap between recipient unit and donor unit (O) with a calculable number of overlapping elements (o), which may vary between 0 and the smaller sample of R and D. In the example, these elements are the customers of the publisher who took part in the volunteer survey.

$$O = R \cap D \tag{4.1}$$

If an overlap rate is calculated, it is always calculated based on the number of recipients, i.e. by $\frac{o}{r}$. In real world applications, the overlap can only be calculated, if there is auxiliary information on $R \cap D$. For example, a publisher using a volunteer survey for data augmentation can calculate $\frac{o}{r}$, if one question in the online survey is "Do you have a newspaper subscription or regularly buy a newspaper from publisher XY?". From a company's perspective, R is always equal and differences in O results from different possible sources D.

All sources can be classified by their size d and their overlapping units o with the recipient unit, as shown in figure 4.3[1]. The influence of size and overlap of the sources on data augmentation results is one subject of our study. Depending on the combination of overlapping units and number of donors, different types of sources are differentiated.

[1] The notation of set theory is used, where "$\subset$" describes a subset of elements of another set, "$\cup$" describes the union of sets of elements, and "$\cap$" describes the intersection of sets of elements.

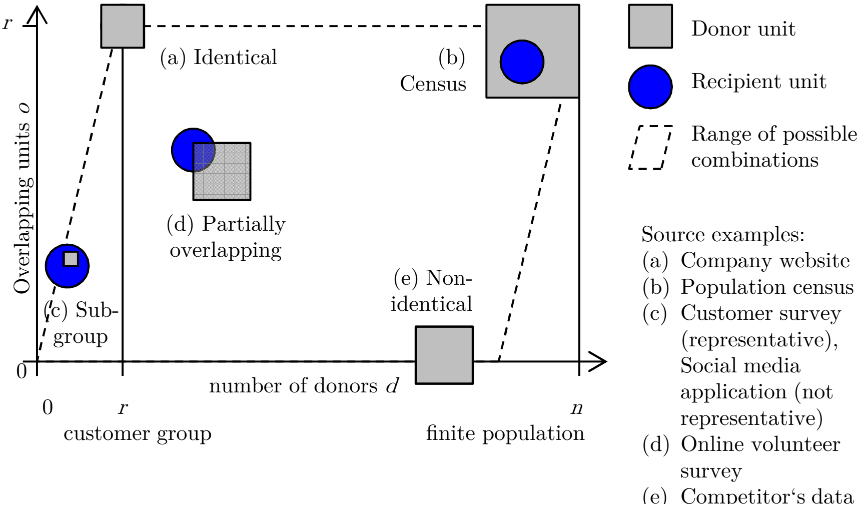


Figure 4.3: Data augmentation sources by size and overlap

(a) A source with a 100% overlap rate $D = R$ (4.3a) could be data from an internal online source, e.g. the protected area from a company's website (provided all customers use the website). It is characterized by containing the same observations, with no unique identifier to match the observations on an individual basis. It can be found in the upper left corner of the parallelogram.

(b) If the whole population P was the source, e.g. a census (4.3b), all customers would be elements of the source $D = P$. This is a rather theoretical case stated for comparison purposes, as only little information is known through official census data on the overall population, with few differentiating link variables. It can be found in the upper right corner of the parallelogram.

(c) Other sources are nonprobability samples in a way that D is not equal to R or P in terms of elements. A non-representative sample $D \subset R$ (4.3c) could be data gathered by a company's social media application or a self-initiated volunteer survey. It can be found on the line connecting the origin with in the upper left corner of the parallelogram. The overlap rate is always 100%.

(d) If the groups only partially overlap $D \supset R$ (4.3d), they could be derived from a branch survey also containing customers of other companies. They can be found anywhere within the lines of the parallelogram.

(e) If the groups do not overlap at all $D \cap R = 0$ (4.3e), they might be derived from a competitor's survey. The availability for data augmentation in database marketing is rather unlikely. Its relevance should be questioned from a rational point of view. It is included to distinguish between source types. It can be found on the x-axis between the origin and the lower right corner $(n - r)$, because the number of overlapping units always equals zero.

If the source data mechanism of a source is random, as in a representative population survey like the *Communication Networks* or the *Typologie der Wünsche* (Institut für Medien- und Konsumentenforschung, 2012a, 2012b), the information is representative for the population from which it was sampled. It is therefore denoted as $D \equiv R'$, if D is a representative sample of R, and $D \equiv P'$, if D is a representative Psample (Krämer, 2010). These sources are used for comparison purposes. They are not the focus of this study, because representative sources have been previously studied and used for data augmentation.

4.2.2 Variables

The basis for a formalization of data augmentation has already been provided by Wendt and Wendt (1983). Every element of P has certain particular characteristics relevant to the database marketing problem. These characteristics are part of a multidimensional topological space. The topological space is an abstract representation of individuals who otherwise have an unmanageable quantity of characteristics. It reduces the real world to a few variables of interest. These are defined by measurable characteristics. Every observation of P can be regarded as a vector in the topological space and the distance between observations is measurable.

Let $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{R}$ be matrices with a specific number of row vectors (observations) and column vectors (variables) (Krämer, 2010). There is a set of link variables (X) with a certain number of link variables (l) and a set of target variables (Y) with a certain number of target variables (t) available from the source. The number of target variables is of minor interest, because a univariate pattern approach is used, augmenting one target variable at a time. The link variables $X = (x_1, x_2, ..., x_l)$ are available in all matrices and the target variables $Y = (y_1, y_2, ..., y_t)$ only in $\mathbf{D}$ and $\mathbf{P}$. Supplementary, let there be a set of auxiliary variables (Z) with a certain number of auxiliary variables (a) depending on which additional variables are of interest in the data augmentation context. The auxiliary variables $Z = (z_1, z_2, ..., z_a)$ only exist in $\mathbf{P}$. In our case study, Z comprises a binary customer indicator variable (c) and a set of source data mechanism indicator variables (S). These binary variables are used for the non-random source data mechanisms $S = (s_1, s_2, ..., s_{a-1})$. This is different to data augmentation problems in a statistical matching context, where Z comprises information from the second source to be matched with the first.

$$\mathbf{P}_{p\times(l+t+a)} =$$
$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,l} & y_{1,1} & y_{1,2} & \cdots & y_{1,t} & z_{1,1} & z_{1,2} & \cdots & z_{1,a} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,l} & y_{2,1} & y_{2,2} & \cdots & y_{2,t} & z_{2,1} & z_{2,2} & \cdots & z_{2,a} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,l} & y_{p,1} & y_{p,2} & \cdots & y_{p,t} & z_{p,1} & z_{p,2} & \cdots & z_{p,a} \end{pmatrix}$$

Figure 4.4: Dimensions of the overall population

The dimension $(l + t + a)$ with $l, t, a > 0$ of all column vectors in $\mathbf{P}$ consists of X, Y, and Z. In figure 4.4, $X = (x_1, x_2, ..., x_l)$, $Y(y_1, y_2, ..., y_t)$, and $Z = (z_1, z_2, ..., z_a)$ are added to form the horizontal dimension. The size of the vertical dimension is p. While $\mathbf{P}$ is a $(p \times (l + t + a))$ matrix, $\mathbf{R}$ and $\mathbf{D}$ have dimensions of $(r \times l)$ and $(d \times (l + t))$, respectively. $\mathbf{R}$ and $\mathbf{D}$ are shown in figure 4.5.

$$\mathbf{D}_{d\times(l+t)} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,l} & y_{1,1} & y_{1,2} & \cdots & y_{1,t} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,l} & y_{2,1} & y_{2,2} & \cdots & y_{2,t} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d,1} & x_{d,2} & \cdots & x_{d,l} & y_{d,1} & y_{d,2} & \cdots & y_{d,t} \end{pmatrix}$$

(a) Donor unit **D**

$$\mathbf{R}_{r\times l} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,l} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{r,1} & x_{r,2} & \cdots & x_{r,l} \end{pmatrix}$$

(b) Recipient unit **R**

Figure 4.5: Dimensions of donor unit and recipient unit

Independent sources can be used regardless to their overlap in observations, if there is a definable set of variables appearing in both sources (Adamek, 1994). Data augmentation matches two datasets by link variables in order to receive more information for the customer database. One target variable is augmented at a time. Matches are found based on a model built for all customers in the target population (Rässler, 2002, p. 6). This is possible if there is a correlation between X and Y.

$$P(Y = y|X) = P(Y = y|x_1 \cap x_2 \cap ... \cap x_l) \tag{4.2}$$

$X = (x_1, x_2, ..., x_l)$ are the predictors for Y and need to have predictive power in order to discriminate between target variable values (Baker et al., 1989). If link and target variable were not correlated, no learning from the link variables about the target variable would be possible (Bernardo & Smith, 1994, p. 168). Even worse, link variables with no additional predictive power distort the augmentation process, lowering the quality of the overall augmentation results. It is assumed that there is a joint probability function of X and Y in the overall population. Each unit is a sample drawn from the joint probability function (Rässler, 2002, p. 20).

The link variables $X = (x_1, x_2, ..., x_l)$ need to be mutually independent. If link variables are dependent, the dependent link variables have no additional predictive power regarding the target variable. Many statistical procedures require variables to be independent. If correlations among link variables exist, some of the variables need to be deleted, or they need to be reduced by factor analysis, so that

$$P\left(\bigcap_{i=1}^{l} x_i\right) = \prod_{i=1}^{l} P(x_i) \tag{4.3}$$

In practice, where human characteristics are regarded, there is usually no case where link variables do not correlate at all, unless the set of link variables is reduced to a very low number, thereby losing the marginal predictive power of deleted link variables. It is therefore reasonable to introduce a threshold distinguishing weak correlations from strong correlations. If link variables are only weakly correlated, and the exclusion of a weakly correlated link variable would lead to a loss in predictive power, it can be decided to retain it albeit its correlation.

Data augmentation in database marketing usually uses categorical variables. Therefore, all link variables $X = (x_1, x_2, ..., x_l)$ and the target variable have a finite domain. The domain of each link variable contains a certain number of values per link variable (j). The domain of the target variable contains a certain number of values per target variable (k). Because the number of values per link variable $j(x_i)$ is limited, there is a finite number of link variable classes (v).

$$v = \prod_{i=1}^{l} j(x_i) \tag{4.4}$$

Every specific variable combination can be regarded as a class or stratum. The number of elements per class and the distribution of elements among classes can be different for **R**, **D**, and **P**. The number of classes does not have to be equal between recipient unit and donor unit.

For link variable values not represented at all in the source across classes, no statements can be made. For example, if only young people are available from a source, no values for old customers can be augmented. The augmentation frame is restricted to link variable values represented in both recipient and donor unit.

4.2.3 Univariate pattern approach

Data augmentation can have a univariate or a multivariate pattern, depending on the information available from the external source. From a managerial point of view, a source is more valuable for data augmentation, if various target variables are available. Therefore, data augmentation problems often have a multivariate pattern. There are two ways of approaching a multivariate pattern situation. If several target variables are augmented, one recipient can either receive all target values from the same donor or one at a time from potentially different donors.

If all target variables are augmented from the same donor, donor and recipient are referred to as statistical twins. They are as similar as possible in terms of link variables. Values can therefore be exchanged between them. In theory, all elements and variables are derived from the same overall distribution. Observations with the same link variables should therefore have the same target variable values.

In practice, however, different target variables are not explained equally well by the link variables. The approximation is only as good as the link variables available and their level of detail. Although link variables are chosen by their ability to discriminate between recipients, they will never fully explain the target variables. From the practical point of view, it stands to reason that every target variable should be imputed individually, taking into account the link variables explaining it best. The only reason to pursue the idea of statistical twins would be the desire to preserve the interrelations between target variables. Whether this is desirable depends on the database marketing problem.

If no inference is expected to be made between target variables after the data augmentation process, the best possible results are achieved by predicting every target variable individually. In terms of computation, univariate and multivariate models differ, because they rely on different concepts. If all target variables are modeled at the same time, the computation is more complex. We decided to deeper explore the univariate pattern approach, thereby precluding the possibility to regard correlations between augmented target variables after the augmentation.

> *Supposition:* No inference is expected to be made between the augmented target variables.

There are also auxiliary variables in $\mathbf{R}$, which are not used in the data augmentation process and are not described in the formal model. Of course, the customer database does not only consist of the link variables. After having augmented target variables to the customer database, one might be interested in making inference on these variables and the newly acquired target variables. To do so would result in the classical problem of the conditional independence assumption in a statistical matching setting. One would essentially be interested in the relationship of variables that have never been jointly observed. For solutions to this problem, the interested reader is referred to Rässler (2002), D'Orazio et al. (2006), the European Conference on Quality in Official Statistics (2012), and references therein. We exclude the problem of cross tabulation here.

> *Supposition:* No inference is expected to be made between augmented target variables and auxiliary variables present in the recipient unit.

4.2.4 Data augmentation target values and uncertainty

There are two goals that can be pursued by data augmentation in terms of enhancing direct marketing activities: selecting and segmenting customers. The output of the data augmentation process is designed to satisfy these goals. Sharot (2007) stated that an ideal data augmentation is given if donors and recipients are conditionally independent given the link variables, and if each recipient matches perfectly with one or more donors based on these link variables. Because this is never the case, especially in a categorical variable setting, the uncertainty involved in data augmentation needs to be accounted for in the output of the data augmentation process. Decision rules need to be formulated in order to decide for specific target variable values, error sources need to be explained, and measures need to be established as to how far a data augmentation diverges from this ideal.

Let $\hat{y}$ be the augmented value during the data augmentation process, whereas y is the true, but unobserved, value of Y for an individual customer. During data augmentation, the best or most likely value of the domain of the target variable is augmented to a customer's profile. This is a point estimate for Y (Little & Rubin, 2002, p. 75), as a definite value is needed for pursuing the mentioned goals. If $D = R$ and $v = r$ was true, y would be unambiguously computable for given X and $y = \hat{y}$ would always be true. But because this is not the case, there is variability and uncertainty in the results.

In order to be able to make meaningful decisions from the augmentation results, both the best value and according probability values are kept for further analysis. Especially when dealing with unidentifiable models like the conditional independence assumption, it is advisable to give more information on the augmentation results than a punctual estimate (D'Orazio, Di Zio, & Scanu, 2010). There is a probability $P(Y = y_i|X)$ with $1 < i \leq k$ for every target value $y_1, y_2, ..., y_k$, so that

$$\sum_{i=1}^{k} P(Y = y_i|X) = 1 \tag{4.5}$$

The probabilities $P(Y = y_i|X)$ convey information on how likely a specific value is to be true, given a specific link variable class X and the source D. The augmented value $\hat{y}$ is the value y_i which maximizes the probability $P(Y = y_i|X)$ of being true for a particular recipient, given the information in the source. Depending on the data augmentation goal and method, $\hat{y}$ can equal the mode, the expected value, the predicted value, or another value that is most likely.

$$P(Y = \hat{y}|X) = max\left(P(Y = y_1|X), P(Y = y_2|X), ..., P(Y = y_k|X)\right) \tag{4.6}$$

It can make sense to introduce a minimum acceptable threshold for $P(Y = \hat{y}|X)$. If all target values were equally distributed given a specific link variable class, the best value would not be specifiable. If $P(Y = \hat{y}|X) = 0.5$ for a target variable with two possible outcomes, the augmented value would be just as likely for a customer to be true as to be wrong. If uncertainty shall be reduced, an exit criterion can be introduced in a way that values are augmented only if $P(Y = \hat{y}|X)$ is greater than a certain threshold. By definition, the threshold is always greater than $\frac{1}{k}$. While from an uncertainty point of view, introducing a very high threshold might be desirable, from a marketing practice point of view, it might not. Introducing a threshold always leads to a number of customers not receiving an augmented value, because no value can be augmented with the required amount of certainty. If the group of customers without augmentation output becomes too high, it would obviate the data augmentation effort.

In order to account for the variability and the uncertainty in the data augmentation process, several new variables are matched to a recipient during data augmentation. Once a decision has been made towards a target value or not augmenting a value, more information is necessary in order to assess the likeliness of $\hat{y}$. If only a single value was augmented, it would convey no information on how likely this value was true. The probability $P(Y = \hat{y}|X)$ is added in order to give information on the likeliness. More specifically, there is a probability $P(Y = y_i|X)$ with $1 < i \leq k$ for every possible variable combination of X. But only one of the values $Y = (y_1, y_2, ..., y_k)$ is augmented for an individual customer. The probability for the augmented value to be wrong is $1 - P(Y = \hat{y}|X)$ and is called expected loss (Bernardo & Smith, 1994, p. 256) or matching noise (Paass, 1985, as cited in D'Orazio et al., 2006, p. 10).

The point estimate would be sufficient for the segmentation goal of direct marketing. Segmentations can be based on $\hat{y}$, because all customers are allocated to groups. For selection problems, not only the augmented value is of interest. Using the augmented value might not satisfy the requirements of selecting a target group depending on certain size, budget, and conversion criteria. For a campaign, it might be desirable to select 10,000 customers with high income. Even if a customer was not classified into the high income group, it is necessary to have a measure on how likely he or she would have been in that group. In order to select a specific number of customers, they are then sorted by the probability of this value and chosen accordingly. Therefore, all probabilities $P(Y = y_1|X)$, $P(Y = y_2|X)$, ..., $P(Y = y_k|X)$ are augmented for every customer. The part of the new synthetic dataset containing the probabilities for all target values is referred to as probabilistic database (Jiang et al., 2007). We promote the meaningfulness for creating a combination of a deterministic and a probabilistic database in order to have as many usage opportunities as possible.

Additionally, information on the robustness of $\hat{y}$ for given X is of interest, i.e. the variance of Y in the respective donor class. If there was no variance, and the number of donors in the class was infinite, the augmentation results would be maximally robust. Because categorical variables are used, the sampling variance is calculated by the index of qualitative variation (IQV)

developed by Wilcox (1973). It takes into account the number of target variable categories k and their according probabilities $P(Y = y_i|X)$:

$$IQV = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} P(Y = y_i|X)^2\right) \tag{4.7}$$

IQV varies between 0 and 1. It takes 0 if all observations belong to the same category and 1 if all observations are equally distributed among the categories (Wilcox, 1973). The resulting recipient unit has augmented values for the missing target variable, probabilities for every target value of the domain, and an IQV measure, as shown in figure 4.6. For simplification purposes, only one augmented target variable is shown along with its descriptive measures, as it would result from a data augmentation with a univariate pattern. After the augmentation process has been performed for several target variables, the recipient unit has a target value and according measures for every target variable, so that the recipient unit has a dimension of $r \times (l + (k + 2) \times t)$.

$$\hat{\mathbf{R}}_{r\times(l+k+2)} =$$
$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,l} & \hat{y}_1 & P(Y = y_1|X)_1 & \cdots & P(Y = y_k|X)_1 & IQV_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,l} & \hat{y}_2 & P(Y = y_1|X)_2 & \cdots & P(Y = y_k|X)_2 & IQV_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{r,1} & x_{r,2} & \cdots & x_{r,l} & \hat{y}_r & P(Y = y_1|X)_r & \cdots & P(Y = y_k|X)_r & IQV_r \end{pmatrix}$$

Figure 4.6: Augmented recipient unit

IQV is a good measure for MCAR sources. But it is influenced by how many donors represent a class, in combination with the number of elements in that class in the donor unit. If only one donor represented a class, IQV would equal 0. This seems like a desirable result. However, if the overall population had three elements in that class, with different target values, the true, but unknown IQV would be 1. Therefore, IQV always needs to be interpreted as a measure for the variance or variability of a target value in a specific link variable class *within* the source. It is a measure of how robust the augmented value is *given* the data – not in general.

4.2.5 Ignorability of the source data mechanism

During data augmentation, a decision towards a target value has to be made for every recipient. Uncertainty arises, because the target variable is not observable for the elements of the recipient unit, but has to be derived from the donor unit. Consequently, the occurrence of the event $Y = y$ has a certain probability and a decision has to be made in order to choose the most likely value. The term ignorability of the source data mechanism refers to the question whether the source data mechanism S of the source D influences the target values. The probability for a specific target value is described by $P(Y = y|X, S)$, because it is dependent on the link variable class and the source. The source data mechanism is ignorable, if

$$P(Y = y|X, S) = P(Y = y|X) \tag{4.8}$$

If $P(Y = y|X, S) = P(Y = y|X)$, beliefs about $P(Y = y|X)$ are unchanged when taking into consideration S (Bernardo & Smith, 1994, p. 45). In general, a source can only be used for augmentation purposes if the source data mechanism can be ignored. Otherwise, the source data mechanism would have to be modeled. This can be difficult, because these models face identification issues, are very complex, and rely on a number of prior information (Rässler, 2000; Schafer, 1997, p. 28).

The following theory is derived and further developed from Rubin's (1976) and Little and Rubin's (2002) theory of the ignorability of the missing data mechanism. The probability $P(S)$ gives information on whether elements of the overall population have been observed in the source. S is an indicator variable, which can assume the value 0 if data is missing and 1 otherwise. D is sampled from P, if $S = 1$. Knowledge on this probability has to be obtained from an auxiliary source or assumptions have to be made. Such assumptions are common practice in data augmentation applications (D'Orazio et al., 2006, p. 13), but have to be handled with care, because the accuracy of the model depends on their applicability. Given the probability model $P(Y|X, S)$ and the data X, values for Y can be predicted (Box & Tiao, 1973, p. 6). Additionally, let u be a random variable denoting unknown parameters. The source data mechanism is formalized as

$$P(S = s|X, Y, u) \tag{4.9}$$

In the recipient unit, X has been observed and Y has not. The weakest conditions for ignoring the source data mechanism when augmenting data are: the missing data Y is missing at random, the observed data X is observed at random, and there are no a priori ties between the data and the source data mechanism S (Rubin, 1976). This does not mean the pattern itself has to be random, but rather that the missingness does not depend on the data values (Little & Rubin, 2002, p. 12).

There are three stages of randomness of missing data that indicate whether a source data mechanism can be ignored. If missingness does not depend on any of X or Y, the data is missing completely at random (MCAR) and can be described by

$$P(S = s|X, Y, u) = P(S = s|u) \tag{4.10}$$

In the MCAR case, $P(S)$ is referred to as reference prior. A reference prior is said to have no influence on the inferences made on X and Y. It can alternatively be denoted as non-informative prior and is applicable for large sample sizes (Bernardo & Smith, 1994, p. 298). If the source data mechanism is MCAR, it is ignorable:

$$P(S = s|X, Y, u) = P(S = s|u) \Leftrightarrow P(Y = y|X, S) = P(Y = y|X) \tag{4.11}$$

If the missingness depends on the link variables only, which have been observed for all observations, the data is missing at random (MAR) and can be described by

$$P(S = s|X, Y, u) = P(S = s|X, u) \tag{4.12}$$

If the source data mechanism is MAR, it is ignorable:

$$P(S = s|X, Y, u) = P(S = s|X, u) \tag{4.13}$$

$$\Leftrightarrow \; P(Y = y|X, S) = P(Y = y|X)$$

If the missingness depends also on the target variable, the data is missing not at random (MNAR) and the probability $P(S = s|X, Y, u)$ cannot be simplified.

In contrast to other theoretic approaches to data augmentation, amongst others described by Rässler (2002, p. 7) and D'Orazio et al. (2006, p. 6), the assumption that the source data mechanism is MAR, because it was intentionally induced by the study design, does not apply here. This is, albeit not explicitly stated by the named authors, only true for randomly sampled sources. Putten (2010, p. 91) mentioned that many concepts of data augmentation assume a representative source. But sources readily and publically available are usually not representative.

As has been shown in chapter 4.2.1, the data augmentation source D can have different relationships to P and R. If $D = P$, no data is missing, which is why it can be regarded as missing completely at random (MCAR). No missingness is present, thus it is not correlated with any variables. If $D \equiv P'$, the source has been sampled by an unknown random mechanism. Then data is MCAR as well and the source data mechanism is ignorable:

$$P(S = s|X, Y, u) = P(S = s|u) \;\; \forall \;\; D = \{P, P'\} \tag{4.14}$$

If $D = R$ or $D \equiv \; R'$, the situation can be interpreted in different ways. Actually, data is not missing completely at random from the overall population. If D and R are identical, the missing data is missing due to $c = 0$.

The fact whether an observation is a customer is observable. If the missingness only depends on observable variables, it is MAR. The same applies to the representative sample of $D \equiv R'$, whose source data mechanism is a mixture of MAR (the $D = R$ part) and MCAR (the representative sample part). In both cases, the source data mechanism is ignorable:

$$P(S = s|X, Y, u) = P(S = s|X, u) \;\forall\; D = \{R, R'\} \tag{4.15}$$

However, if $D = R$ or $D \equiv R'$, all elements relevant to such a data augmentation problem are present in R. Then, the augmentation frame does not comprise any other people than the customers. R is the overall population of interest in this case. Consequently, $R = P$ in the special cases of $D = R$ and $D \equiv R'$. Both cases are regarded as MCAR. This comprehension can be further relaxed to sources with a $R \subset D$ source data mechanism. Whenever all recipients are included in the source, all persons of interest are observed. In this case, the source data mechanism does not matter, because no bias can evolve from the nonprobability sample.

If $D \supset R$ or $D \subset R$ or $D \cap R = 0$, then data is not missing completely at random, which means that the source data mechanism is either MAR or MNAR. Whether it is MAR or MNAR depends on the association between S and Y. As stated by Rubin (1976), the source data mechanism is ignorable if S and Y are distinct and one does not provide any information on the other (Rässler, 2002, p. 77; Schafer, 1997, p. 11). This is a MAR case, because

$$P(Y = y|S) = P(Y = y) \times P(S = s) \tag{4.16}$$

$$\Leftrightarrow\; P(Y = y|X, S) = P(Y = y|X)$$

In general, the proof of independency between the source data mechanism and the target variable would be sufficient for a source to be suitable for data augmentation. However, because the new information is augmented based on link variables, the link variables need to be considered in the evaluation. In fact, the presence of unconditional independence is neither implied by, nor implies a conditionally independent relationship (Simpson, 1951). The values to be augmented are derived from X, and from X alone. Therefore, a source not satisfying all previous assumptions can be used, if, and only if, Y and S are conditionally independent given X:

$$P(Y = y, S = s|X) = P(Y = y|X) \times P(S = s|X) \tag{4.17}$$

$$\Leftrightarrow \; P(Y = y|X, S) = P(Y = y|X)$$

If Y and S are conditionally independent given X, the source data mechanism is MAR. This can shortly be written as $Y \perp S|X$ and is referred to as collapsibility in the context of categorical data analysis (Agresti, 2002, p. 358). It means that the relationship $P(Y = y|X, S)$ can be collapsed to $P(Y = y|X)$ without any loss of information. It is therefore possible to define a restricted class of conditions under which the source data mechanism can be ignored.

$$P(Y = y|X, S) = P(Y = y|X) \; \forall \left\{ \begin{array}{c} D = P, P' \\ D = R, R' \\ D : Y \perp S|X \end{array} \right\} \tag{4.18}$$

These conditions are the minimum acceptable conditions under which it is reasonable to perform data augmentation without incorporating the source data mechanism into the model. They refer to all sources

- that have a 100% overlap rate to the recipient unit ($D = R$ or $D = P$)
- that are a representative sample of a source with a 100% overlap rate to the recipient unit ($D \equiv R'$ or $D \equiv P'$)

- where the target variable and the source data mechanism are conditionally independent, given the link variables ($D : Y \perp S|X$)

The first two conditions regard MCAR sources. The source data mechanism for MCAR sources is usually known. The last condition refers to MAR sources and is of particular interest in our study.

Chapter 5

Test design for evaluating the source characteristics

From the variety of external sources available, the question arises which sources are suitable for data augmentation. There are formal requirements which need to be met, as described in the previous chapter. Beyond those requirements, it is desirable to know whether the augmented values are "good". They are supposed to be good in a way that they come close enough to the true, but unknown, values. Only then can they be used for target group selection and decision making. To have an upfront understanding of a source and its knowledge improvement capabilities is of central importance in our study. It is the managerial contribution of our work.

To answer the research question, a measure for the quality of the data augmentation results is needed. Data augmentation results are good, if many of the true values are "hit". A value is hit, if the augmented value is equal to the true, but unobserved value. Furthermore, the results need to be compared to a suitable criterion describing the ability of selecting target groups, if the data augmentation results were not available. A conversion probability lift (CPL) measure can show how much the quality of selecting target groups increases, as compared to that criterion.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_5

In this chapter, a conceptual model is built formalizing the variables affecting this KPI. Five hypotheses are derived to the research question. We describe the test set-up for the data basis given. In a business application, the quality of data augmentation cannot directly be measured. After having used the augmented data for a marketing task can the conversions be compared to a control group. We choose a case study design with simulated missing target variables and a high number of information-oriented samples, in which the conceptual model is implemented. Parameters are varied multiple times, so that the influences of changes in individual parameters can be isolated. The true target values are known, so that an internal evaluation of data augmentation results is possible.

5.1 Research question conceptualization

Our study delivers insight on the suitability of various sources for data augmentation in practice. In the following, the research question is mapped to a conceptual model. The conceptual model specifies all relevant parameters to be regarded when answering the research question, based on the theory and framework established in chapter 4. Different factors can influence the quality of data augmentation results, e.g. the source data mechanism, the augmentation methods, the characteristics of the source, and the variables used. Two major KPIs are introduced with which the quality is measured: the model lift and the conversion probability lift. After having constituted the relationships between the influencing factors and the quality KPIs, hypotheses are established regarding the properties of these influences. The hypotheses can later be tested using the case study design.

5.1.1 Conceptual model

In the previous chapters, important parameters influencing the data augmentation results have been introduced. The effectiveness of the augmentation is dependent on the predictive power of the link variables regarding the target variables (chapter 3.2.1). However, data augmentation results are seldom highly accurate and precise. Because of the categorical nature and the confined number of link variables, the degree of precision can only have a certain extent. This is due to the topological space and the simplification related to it (chapter 4.2.2). It has been shown in chapter 4.2.5 that every source can be used for data augmentation purposes, as long as the source data mechanism is ignorable. If a source includes the same people as the customer group, it means that there is definitely a match – the challenge is to find it. If the donor unit is completely different from the customer group, it seems rather unlikely to make accurate predictions, even if conditional independence applies. Other factors like overlap, number of donors, and representation might influence the quality of augmentation results (chapter 4.1.2), depending on the augmentation methods. When close neighbors are used to calculate data augmentation results, like in the nearest neighbor hot deck method, the accuracy and precision of the results depend on the closeness of similar observed elements. When logistic regressions are approximated by maximum likelihood functions, the accuracy depends on the error term, i.e. the ability of X to explain Y (chapter 3.2.3).

The conceptual model of this study summarizes the parameters mentioned and puts them into sequence. It is illustrated in figure 5.1. The conceptual model frame encircles the research question with a continuous line. The model frame defines the scope of this study. In the conceptual model, relevant relationships are shown. Each arrow symbolizes such a relationship and hypothesis tests can be used to test whether the assumed relationships are true.

A valid source data mechanism is an antecedent for data augmentation. The antecedents (1) are shown first, because their appropriateness is a precondition. MCAR sources are framed with a dashed line, because their suitability for data augmentation has previously been confirmed. MAR and

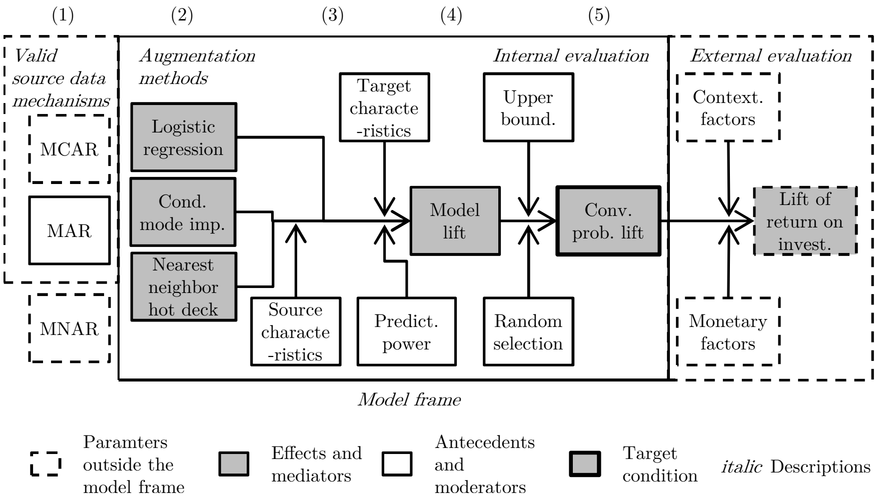


Figure 5.1: Conceptual model

MNAR sources are of primary interest to this study. The validation of the antecedents is one of our contributions to the research question.

Once a source is available for augmentation, two decisions have to be made by the database marketing analyst. The first decision is an assessment whether a source is suitable for data augmentation, i.e. whether it meets all criteria enabling a significant model lift. This assessment includes exit criteria for the source characteristics, the target variable characteristics, and the link variable characteristics. The second decision concerns the data augmentation method (2) to be used. The method is the main effect in the conceptual model, because it can be influenced by the database marketing analyst. We compare logistic regression, conditional mode imputation, and nearest neighbor hot deck, which are commonly used data augmentation methods as described in chapter 3.2.3. Giving upfront advice on the best method to be used is another contribution to the research question.

Source, link and target variables have properties relevant to the model lift and thus to the research question. They are moderators in the conceptual model (3). Moderators are immutable and cannot be changed by the database marketing analyst. The performance of methods is expected to be dependent on the source characteristics (3a). The source has several characteristics relevant to the data augmentation model. The overlap describes the number of elements identical in the donor unit and the recipient unit. The number of donors describes the overall set of elements contained in the source. Representative sources and others are differentiated. The number of link variable classes occurring in the source describes the different possible combinations of link variable values. In the illustration of the conceptual model, the relationship of methods and moderators is shown, after a method has been selected. However, if the relationship is known, the moderators observed can influence the best method to be used.

The target variable characteristics are of major importance, because they influence the boundaries of the model lift (3b). Although our goal is to make statements about data augmentation sources, it is necessary to use different target variables in order to receive results that are generalizable for various target variables. The number of target values as well as the variance as measured by the IQV limit the maximum possible model lift. This is described in detail in chapter 6.4.

The link variables are fixed in our case study. However, regarding the different target variables, their predictive power can vary (3c). The predictive power influences the model lift. The higher the predictive power, the lower the uncertainty of the augmented target values, because element in a link variable class are more homogeneous. When comparing different sources and methods, it needs to be taken into consideration.

The moderators cannot be influenced and should be examined carefully before performing data augmentation. There might be a minimum number of donors or a minimum overlap necessary for the model to be significant. Likewise, there can be a maximum number of target values and a maximum variance for enabling a significant model lift. The predictive power needs to be of some amplitude in order to produce a significant model lift. An upfront assessment of all of these characteristics is therefore just as important as the evaluation of the antecedents. The validation of the moderators is another contribution to the research question.

The goal of data augmentation is to reach a model lift as high as possible. The first parameter that is measured for every augmentation is the model lift (4). The model lift indicates the increase of correct hits resulting from augmentation as compared to the achievable hit rate by a random allocation of target values within the recipient unit. The model lift is a mediator in the conceptual model. It is a relatively robust measure, because it can be compared for differing sources. A model lift is considered significant, if the hit rate of the model is significantly higher than the hit rate that could have been achieved by chance. A more detailed description of the model lift calculation is given in chapter 6.4.

The ultimate goal of any data augmentation is a significant CPL (5). The CPL indicates the increase of correctly selected customers when using augmentation results for target group selection as compared to the number of correctly selected customers when not using augmentation results. It depends on whether the objective of a marketing campaign is to segment customers into alternative groups or to target a selected group of customers in order to use marketing potentials best. In the former case, the CPL is equal to the model lift. In the latter case, it is crucial to know how many customers are targeted. Our CPL measure regards the second case, but takes into consideration all possible target group sizes. It is moderated by the upper boundary and the capability of a random selection to hit correct values. The hit rate of a random selection is regarded as the minimum requirement for any target group selection. This is explained in detail in chapter 6.5.

Up to the CPL, the effects can be evaluated internally, as it is done in this study. The eventually desired condition is a maximum lift of return on investment of a marketing activity. For a maximal ROMI, the CPL is a mediator. It is further influenced by several other factors, including contextual and monetary factors. This can only be evaluated externally and is not in the scope of our study.

5.1.2 Hypotheses

The source data mechanism, the link and target variable characteristics, the source characteristics, and the method chosen have been identified to influence the data augmentation results. From the interdependencies, hypotheses are derived that can be tested in order to answer the research question.

Antecedents

From the theoretical consideration in chapter 4.2.5, a restricted class of minimum acceptable conditions has been stated under which the source data mechanism can be ignored in a data augmentation problem. The source data mechanism can be ignored, if the donor unit has a 100% overlap rate to the recipient unit, if it is a representative sample of a source with a 100% overlap rate to the recipient unit, or if the target variable and the source data mechanism are conditionally independent, given the link variables. While the first two conditions refer to MCAR sources, MAR and MNAR sources are distinguished by exhibiting the required conditional independence.

For practical applications, it is of particular interest, if and how the source data mechanism influences the model lift and subsequently the conversion probability lift. If a source is MNAR, the link variables and the source data mechanism can influence the augmented target values. However, it needs to be assessed in *how far* the influence of the source data mechanism is strong enough to compromise the augmentation results, when compared to MAR sources.

> *Hypothesis 1:* The model lift is significantly higher for MAR sources than for MNAR sources.

It has already been stated in chapter 3.2.5 that the theoretical soundness of the augmentation set-up is of minor interest, if the augmented information raises the conversion probability when used for segmenting or selecting customers. Thus, if the stated hypothesis was wrong, the source data mechanism would not need to be evaluated ex ante when performing a data augmentation.

Types of methods

The methods available for data augmentation are expected to influence the model lift. In the optimal case of conditional mode imputation and nearest neighbor hot deck, every recipient has exactly one donor, which is the same person as the recipient. It is uniquely identifiable by a certain link variable class. In this case, the file concatenation would have the form of record linkage, with the respective variable combination being the unique key.

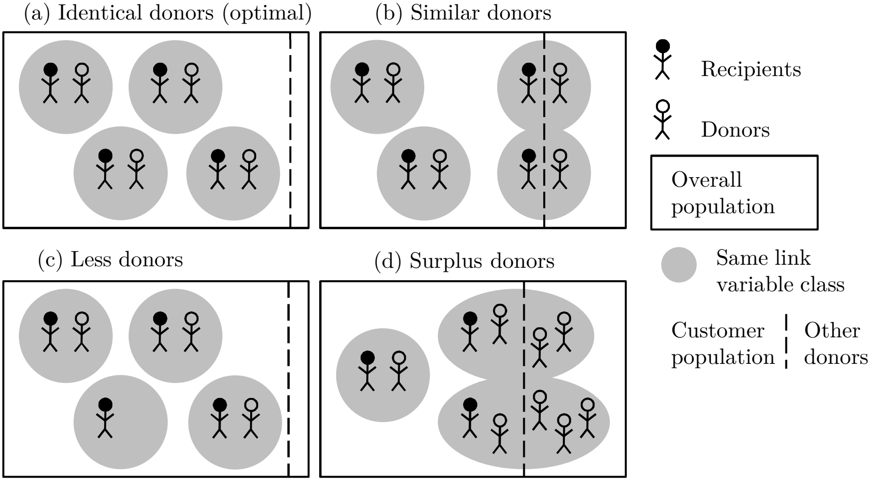


Figure 5.2: Deviations from the optimum influencing the model lift for conditional mode imputation and nearest neighbor hot deck

All parameters causing the data augmentation situation to deviate from this optimum are expected to result in a decrease of the model lift. Simplified examples of this deviation are shown in figure 5.2. The recipients are symbolized by black stickmen, the donors by white stickmen. All stickmen left to the dashed line are elements of the customer population. Thus, if white stickmen are found in this area, these customers have also been observed in the source. The number of white stickmen in this area, divided by the number of black stickmen, results in the overlap rate. The gray circle denotes donors and recipients within the same link variable class. If several stickmen of the same color are found within a class, they symbolize elements with different target values. Donors and recipients within the same or close link variable classes are paired up in the data augmentation process, and values for the recipients are augmented from the respective donors.

The optimal state of a source with a 100% overlap rate regarding the recipient unit, with every donor being distinctly relatable to a recipient, is shown in figure 5.2a. Situation 5.2b shows a case in which only two recipients have an identical donor and the other two recipients have similar donors in terms of their link variable class. The number of donors is equal to figure 5.2a, but the overlap is lower. Although having the same link variable classes, their target variable values might differ, if the link variable values are not able to discriminate perfectly between target variable values. The model lift is expected to be lower, because more uncertainty is involved.

> *Hypothesis 2a:* The model lift increases with increasing overlap of donor and recipient unit.

If there are less donors as shown in figure 5.2c, the likelihood to get an accurate estimate for every link variable class decreases. In this case, the third recipient with no matching donor either receives no value or one of the other donors which is closest in terms of a distance measure. Either way, the model lift is expected to decrease, because more uncertainty is involved.

> *Hypothesis 2b:* The model lift increases with increasing size of the donor unit.

If the source contains more donors than there are recipients, as shown in figure 5.2d, this surplus causes uncertainty. Because the "correct" donors cannot be identified, a rule needs to be introduced including all donors for every link variable class. In this case, the increase in donors does not lead to an increase in the model lift. This is different to situation 5.2c, in which the increase in donors has a positive influence on the model lift. It differs, because in figure 5.2d, the overlap remains the same as compared to the optimal state in figure 5.2a.

> *Hypothesis 2c:* The model lift decreases with increasing size of the donor unit, given a certain overlap between donor and recipient unit.

The hypotheses are tested using a case study with differing source characteristics. Because all of the hypotheses concern these source characteristics, they are subsumed in an overall hypothesis regarding conditional mode imputation and nearest neighbor hot deck methods.

> *Hypothesis 2:* The augmentation results of nearest neighbor hot deck and conditional mode imputation are influenced by the source characteristics.

The quality of the multivariate methods like logistic regression is dependent on the predictability of the link variables regarding the target variables. Variance explained by other factors cannot be captured by the models of the multivariate methods. Because human behavior can only be explained by a limited set of link variables to a certain extent, there is always uncertainty in the results. On the other hand, a cross-validated multivariate model can easily be applied to any subgroup of an overall population. It should therefore not be dependent on the source characteristics.

> *Hypothesis 3:* The augmentation results of logistic regression are not influenced by the source characteristics.

The finding that conditional mode imputation and nearest neighbor hot deck are expected to be influenced by the source characteristics, while logistic regression is not, is only relevant for database marketing analysts, if one method does not in general outperform the others. It is expected that there is an inflection point in the source characteristics, so that it is advisable to use the one kind of methods for a certain type of sources and another kind of methods for another type of sources.

> *Hypothesis 4:* There is a definable set of source characteristics for which similarity methods like conditional mode imputation and nearest neighbor hot deck perform better than logistic regression in terms of model lift, and vice versa.

Conversion probability lift through data augmentation

The final goal of data augmentation is an increase of the conversion probability when using data augmentation results for target group selections. In order to assess the CPL, a target group selected by the data augmentation results is compared to a target group selected by random selection criteria.

> *Hypothesis 5:* If all decisions are made correctly, data augmentation results are able to significantly increase conversion probabilities, when compared to randomly selected target groups.

The goal of our study is to check whether the criteria found are able to guarantee significant CPLs for all data augmentations. While MCAR sources are clearly suitable for data augmentation and have been used as such for many years, MAR sources are not yet used, because their ability to lift conversion probabilities cannot be guaranteed in general. In our case study, we validate the assumed antecedents and find rules on how to choose the best data augmentation method. All data augmentations not having been exited because of exit criteria should then, using the most appropriate method, lead to a significant CPL. This should be true independently of the target variables. If all remaining augmentation results show a significant CPL, our rules are regarded as sufficient for answering the research question.

5.2 Dataset generation and case study set up

In this chapter, the preparation process for the case study is described. The data basis available is first edited in order to fit the model frame as described in chapter 5.1.1. An overall population is designed, which contains variables relevant to data augmentation questions and from which sources are later sampled. The choice of variables is made in a way that it best fulfills the requirements described in chapter 1.3.2, so that the applicability of the case study results is maximally broad. A high number of varying sources is then sampled based on an information-oriented sampling mechanism. These samples function as cases in the case study, within, between, and across which analyses are made. Furthermore, the methods used for data augmentation are described with the specific formulas and modifications used.

5.2.1 Population characteristics and test design

A data augmentation case study needs a big number of observations to account for the many possible link variable classes. A typical data augmentation uses six to eight link variables with two to eight possible values for each link variable domain. The number of possible variable combinations is easily increased to several hundreds. In order to get a comprehensive distribution of observations among variable combinations, even when D is only a portion of P, the minimum number of observations in the overall population needs to be high.

The data used in the case study is chosen so that it resembles a real world approach as much as possible. The available dataset contains more than one million observations. However, the full dataset is not suitable for the case study purpose, because it contains many observations with incomplete or sparse data. A full and rich dataset is needed as basis for the case study. Because of that, only observations with many recorded variables are used.

Following the basic idea of data augmentation in database marketing, there is a finite market population from which a subgroup is the customer group of a certain company. The sampling frame is set to a population relevant to marketing, which is restricted by age (20-69 years), country (Germany), and language (German). Typical sources, such as surveys, use these criteria. In order to achieve a broad range of use cases from the study, the dataset available for this study is reduced and stratified to represent the German population by quotas. All samples are later taken from this finite population. With the reduction to this sampling frame, the frame of the original dataset is purposely falsified, so that from the research results, no conclusions can be drawn on the data provider.

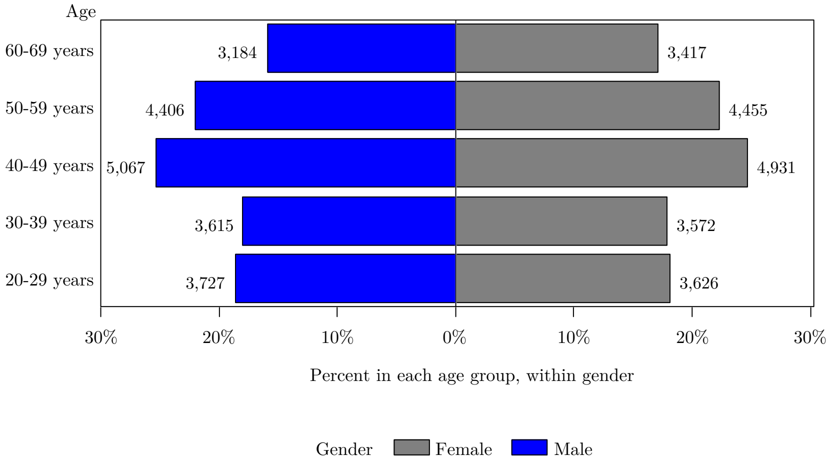


Figure 5.3: Population pyramid of the case study data

The representation of the German population is achieved by sampling and stratification of the provided data by age and gender. The strata on which the data is conditioned is taken from the German population statistics (Statistisches Bundesamt, 2013, p. 33). Figure 5.3 shows the sampled population and its respective properties. The length of the bar for each gender starting in the middle denotes the percentage of elements categorized in the respective age group. Absolute numbers are also given. The total number of observations resulting from the described process and used as overall population in our case study is 40,000. It is of sufficient size, so that even small subgroups resulting from sampling as described in chapter 5.2.3 still have a reasonable number of observations.

Figure 5.4[1] shows the design of the data augmentation case study to be carried out in order to test the stated hypotheses and to answer the research question. During the first step, suitable target variables (1) are chosen from the available data based on expert knowledge. They comprise variables usually not available from the customer database, like income, propensity to buy for products in several categories, interests, and preferences. The target variables are suitable, if the available link variables are able to discriminate between the target values, i.e. if they have predictive power. Every target variable is augmented individually by the univariate pattern approach as described in chapter 4.2.3. Various test samples are purposely sampled from an overall market population (2). The samples represent different sources thinkable in practice and are either fully, partially, or not overlapping with the customer group. This is described in more detail in chapter 5.2.3. A case is defined as a certain combination of source and target variable. In this context, *between case analysis* refers to the analysis of augmentation results with different sources, but using the same target variable. Different methods (3) are chosen according to relevant literature and data augmentation practice. *Within case analysis* refers to the analysis of augmentation results within a case of the case study, e.g. the analysis of augmentation results using different methods, but using the same target variable and source.

[1] The colored version of this figure can be found online on www.springer.com under the title of this publication.

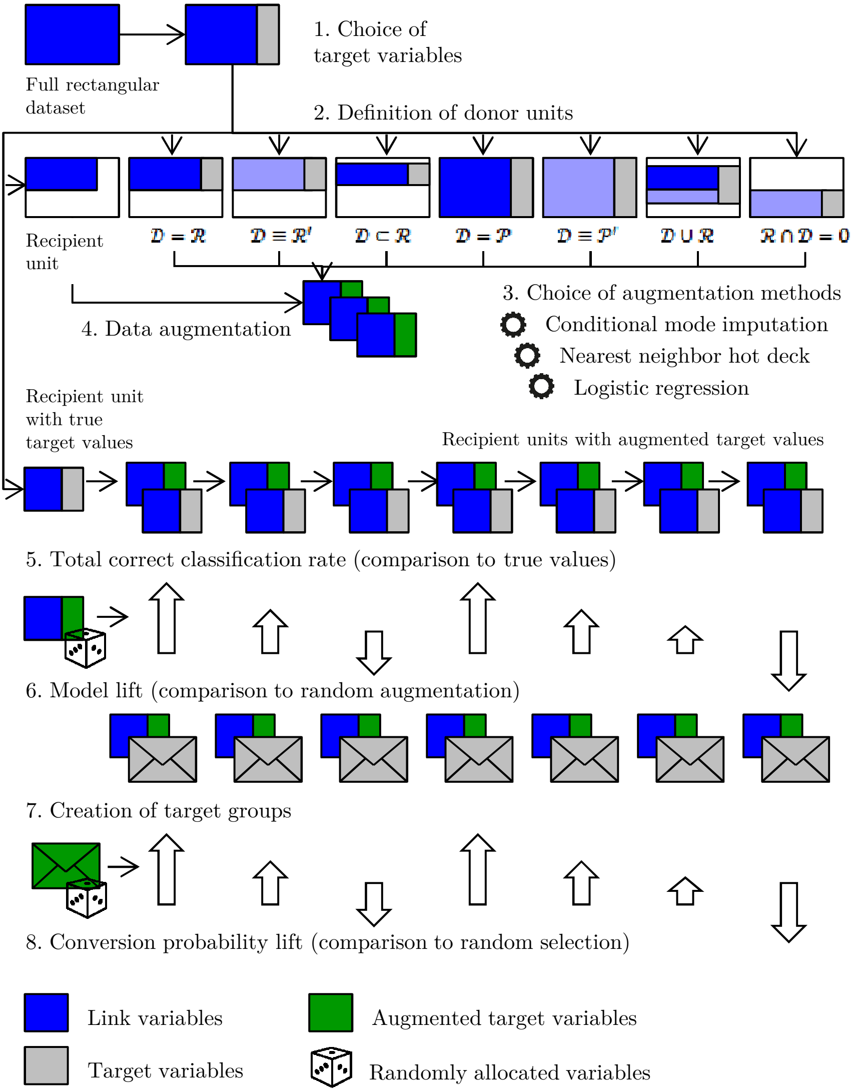

Full rectangular dataset
1. Choice of target variables
2. Definition of donor units
Recipient unit
$\mathcal{D} = \mathcal{R}$
$\mathcal{D} \equiv \mathcal{R}'$
$\mathcal{D} \subset \mathcal{R}$
$\mathcal{D} = \mathcal{P}$
$\mathcal{D} \equiv \mathcal{P}'$
$\mathcal{D} \cup \mathcal{R}$
$\mathcal{R} \cap \mathcal{D} = 0$
3. Choice of augmentation methods
Conditional mode imputation
Nearest neighbor hot deck
Logistic regression
4. Data augmentation
Recipient unit with true target values
Recipient units with augmented target values
5. Total correct classification rate (comparison to true values)
6. Model lift (comparison to random augmentation)
7. Creation of target groups
8. Conversion probability lift (comparison to random selection)
Link variables
Augmented target variables
Target variables
Randomly allocated variables

Figure 5.4: Design of the data augmentation case study

Across case analysis refers to the overall analysis of augmentation results from the case study, including all cases, target variables, and methods.

During the data augmentation step, the customer group, which has also been sampled from the overall population, is augmented with the target variables (4). The augmentation results are then compared to the true values known from the overall population (5). A total correct classification rate can be calculated, which is then compared to the hit rate achievable by a random distribution of values. From this comparison, a model lift is derived (6). Eventually, the target groups that would have been selected with random selection criteria are compared to the new target groups selected with the augmentation results (7). A CPL can be computed, indicating whether data augmentation is able to improve conversion probabilities (8). Model lifts and conversion probability lifts are analyzed across all results.

5.2.2 Choice of variables

The variables used are categorized into link variables, target variables, and auxiliary variables, as described in chapter 4.2.2. All variables are nominal or ordinal, because most marketing sources contain categorical variables only, as described in chapter 4.1.3. The distinction of the available variables into link, target, and auxiliary variables is based on information-oriented decision criteria. It is chosen to best fit the data context and to resemble a real world application as much as possible.

Link variables

A set of socio-demographic, behavioral, and preferential variables is used as link variables. It has been shown in many studies that demographic variables alone do not lead to good marketing models (Brogini, 1998, p. 113ff). Nevertheless, their utility in marketing models has recently been shown by Naseri and Elliott (2011). The best models are built by using a well-balanced mix of socio-demographic, behavioral, and preferential variables.

Figure 5.5 gives an overview of all link variables available from the dataset. There are four socio-demographics variables (Gender, age group,

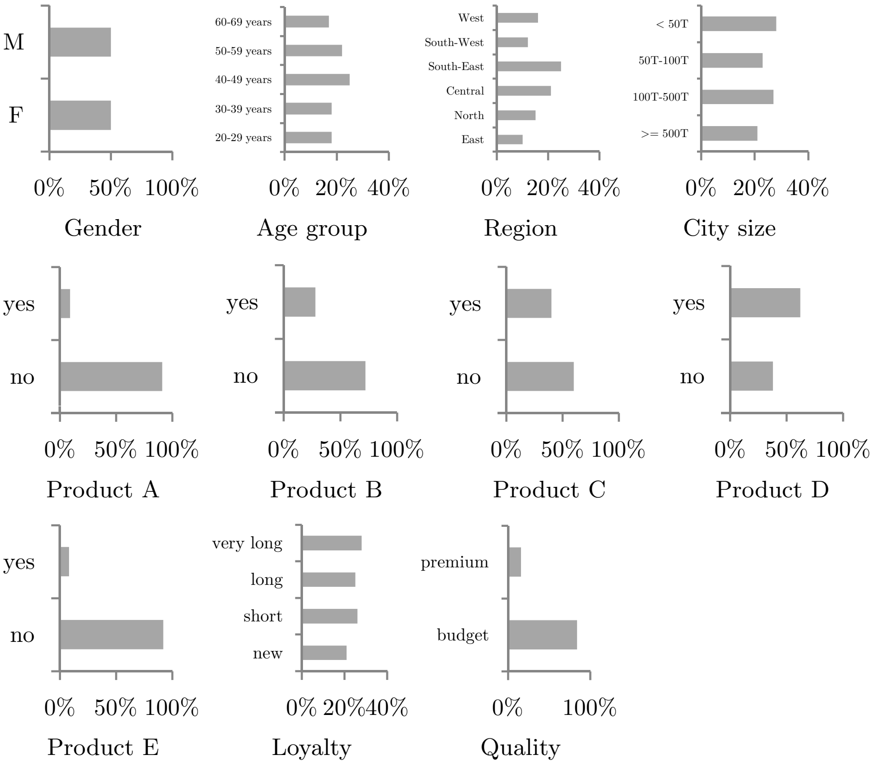


Figure 5.5: Link variables and distributions available from the dataset

region, and city size), as well as seven behavioral and preferential variables. The behavioral and preferential variables describe the purchase history of the persons with regards to five product categories and two variables describing their overall preferences and attitudes (loyalty and quality). Due to confidentiality reasons, the product categories used cannot be named, but are pseudonomized to the generic titles A to E. In the figure, a frequency diagram is given for every link variable, describing the distribution of each link variable in the overall population. Like in a real-world application, some link variables are binary and some exhibit more categories, both nominal and ordinal. Different levels of skewness are available, ranging from evenly distributed (e.g. gender) to very skew (e.g. products A and E).

There must not be any correlations between link variables. If link variables are correlated, the correlated link variables have no additional predictive power for the target variables. Therefore, the partial correlations between link variables need to be tested. The correlations between categorical variables are called associations and are evaluated using contingency analysis (Backhaus et al., 2008, p. 299). Two categorical variables are considered statistically independent, if all joint probabilities equal the product of their marginal probabilities (Agresti, 2002, p. 38). A contingency table is created by cross-tabulating the two variables and the frequency of the appearance of variable values. To test the null hypothesis of statistical independence, the observed values are compared to the expected values. This is done using the χ^2 test statistic (Backhaus et al., 2008, p. 307). If the calculated χ^2 test statistic is smaller than a critical value, chosen for a given level of significance $\alpha = 5\%$, the test is not able to show that the link variables are statistically dependent.

The χ^2 test statistic does not tell anything about the magnitude of an association. Especially, it does not perform well for big numbers of observations. The value of the χ^2 test statistic doubles when doubling the sample size, even though the strength of the association does not change (Huber, 2008, p. 5/22). $Cramer's\ V$ indicates the strength of the association. It takes into account the size and the number of possible values of the variables. For 2×2 tables, $Cramer's\ V$ has a range of -1 to 1. For larger tables, it has a range of 0 to 1. Values between -0.3 and 0.3 are considered trivial (Backhaus et al., 2008, p. 309). In this case, the detected association is a result of the large sample size (Huber, 2008, p. 5/28).

Examples for the evaluation of partial link variable correlations between two nominal link variables and a nominal and an ordinal link variable are given in table 5.1. If at least one variable is nominal, including binary variables, the measures for nominal tests need to be used. If the p-value of

Link variables (scale)	χ^2	p-value for χ^2	H_0	Cramer's V	Magnitude	Association
Age group (ord.) Product Cat. A (nom.)	95.23	1.0191E−19	rejected	0.049	none	no
Product Cat. C (nom.) Product Cat. E (nom.)	13,989.92	0	rejected	0.591	strong	yes

Table 5.1: Examples of nominal associations between link variables

the χ^2 test statistic is smaller than $\alpha = 5\%$, the null hypothesis is rejected and the association is considered significant. Depending on the absolute value (+/−), the magnitude of the association as measured by *Cramer's V* is labeled as none ($<= 0.1$), weak ($<= 0.3$), medium ($<= 0.5$), or strong (> 0.5). Only those associations with a significant and strong relationship are considered dependent, otherwise the link variables are considered independent and both variables can be used. Both of our examples show a significant χ^2 value, so that the null hypotheses are rejected. However, only *Cramer's V* of the second observation is high. This shows that the high χ^2 value of the first association is due to the large sample size, but the magnitude of the association has no strength. The first observation is considered independent, the second observation is considered dependent. The same test is performed for all other partial link variable correlations involving nominal variables.

Some of the link variables are ordinal. If the partial correlation between two ordinal variables is to be tested, the Cochran-Mantel-Haenszel (CMH) statistic is taken into consideration (Mantel & Haenszel, 1959; SAS Institute Inc., 2012). The CMH statistic is evaluated using the χ^2 test statistic. If it is smaller than a critical value, chosen for a given level of significance, the test is not able to show that the link variables are statistically dependent. The Spearman rank correlation coefficient can be used to measure the strength of the ordinal association. It takes into account rank scores of the variables (Huber, 2008, p. 5/41; Keller, 2008, p. 810). In order to evaluate the strength of the correlation, the null hypothesis that the correlation equals zero is tested. The Spearman rank correlation coefficient has a range of -1 to 1. The correlation is strong if values are close to -1 or 1. If the lower bound of the Spearman rank correlation coefficient is below zero and the upper bound is above zero, the correlation is not significant (Huber, 2008, p. 5/41).

Link variables (scale)	p-value for CMH	H_0	Spearman	Lower bound	Upper bound	Strength	Association
Age group (ord.) City size (ord.)	2.1107E−53	rejected	0.079	0.069	0.088	no	no
Age group (ord.) Customer hist. (ord.)	0	rejected	0.491	0.483	0.500	yes	yes

Table 5.2: Examples of ordinal associations between link variables

An example for the evaluation of partial link variable correlations between ordinal variables is given in table 5.2. If the p-value of the CMH statistic is smaller than $\alpha = 5\%$, the null hypothesis is rejected and the association is considered significant. The strength of the association is measured with the Spearman rank correlation coefficient. Only those associations leading to a high coefficient and with the lower and upper bounds not including 0 are considered dependent. Otherwise the link variables are independent and both variables can be used. Only the Spearman rank correlation coefficient of the second observation is high, because the value of the Spearman correlation coefficient is close to 0 for the first association shown. The first observation is considered independent, the second observation is considered dependent. The same test is performed for all other partial link variable correlations with two ordinal variables.

All possible link variables are partially tested for correlations with the χ^2 or the CMH test, as applicable. The results are shown in table 5.3. There are partial correlations between age group and length of customer history, as well as product B and product D, and product C and product

Link variable	Age group	City size	Gender	Customer hist.	Product A	Product B	Product C	Product D	Product E	Product F	Quality	Region
Age group		no	no	yes	no	no	no	no	no	no	no	no
City size	no		no	no	no	no	no	no	no	no	no	no
Gender	no	no		no	no	no	no	no	no	no	no	no
Customer hist.	yes	no	no		no	no	no	no	no	no	no	no
Product A	no	no	no	no		no	no	no	no	no	no	no
Product B	no	no	no	no	no		no	yes	no	no	no	no
Product C	no	no	no	no	no	no		no	yes	no	no	no
Product D	no	no	no	no	no	yes	no		no	no	no	no
Product E	no	no	no	no	no	no	yes	no		no	no	no
Product F	no	no	no	no	no	no	no	no	no		no	no
Quality	no	no	no	no	no	no	no	no	no	no		no
Region	no	no	no	no	no	no	no	no	no	no	no	

Table 5.3: Overview of associations between link variables

E. The association of age group and length of customer history is obvious, because they evolve together. In order to receive independent link variables, one variable of each pair needs to be deleted. This can be done based on contextual knowledge of the data, based on the stability of the variables, or based on which link variable has the best effect when being deleted; e.g. when one link variable has correlations with several other link variables. In our case, the decision is made based on contextual knowledge of the data. The variables age group, product D and product E lead to better selectivity of the overall set of link variables. They are kept for further analysis.

The combination of all possible link variable values results in the link variable classes. People are in the same link variable class, if they are equal by means of their link variable values. In our case, there are 2,805 link variable classes for the observations. There are many link variable classes only occurring once, but frequently encountered combinations appear up to 238 times. For example, people in the class with the most observations are women between 30 and 39 years old, live in a metropol city (>= 500T inhabitants) in south-eastern Germany, have a budget focus regarding quality, and are only interested in product E, but not in products A, D, and F. In contrast, only one man between 60 and 69 years old living in a medium sized city (100-500T inhabitants) in northern Germany has a budget focus regarding quality and is interested in products E and F, but not A and D[2].

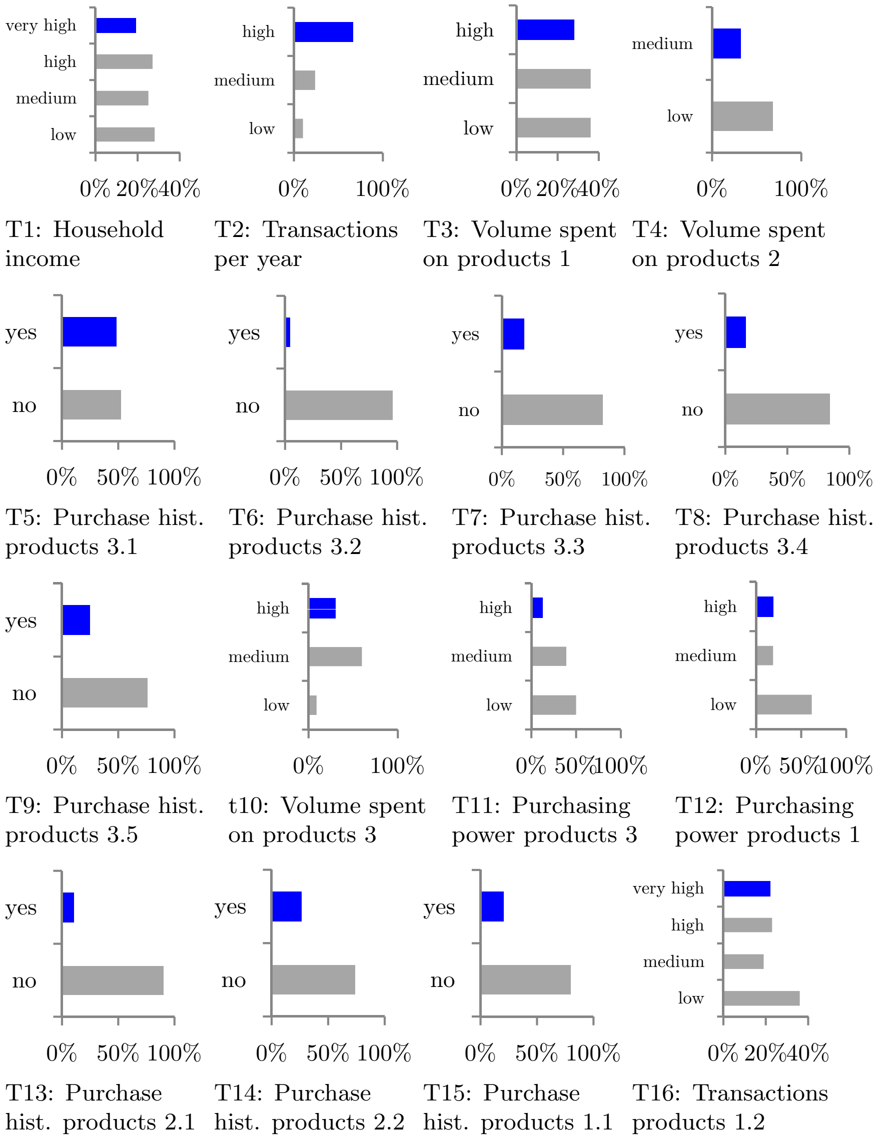
Target variables

For the target variables, we choose a mix of socio-demographic, behavioral, and preferential variables. For demonstrating purposes, some of them are binary variables with only two possible values. They have differently skewed distributions. Often, only one of the two values of a skewed distribution is of interest; e.g. whether a person is interested in a specific product (not the ones who are not). Other variables have three or four possible values with an ordinal scale. That way, different types of variables can be analyzed regarding their suitability for data augmentation.

Figure 5.6 gives an overview of all target variables available from the dataset, as they could be available for example from a survey. There are two socio-demographic variables, one indicating the net household income and one indicating the general purchasing power by number of purchase transactions per year. Furthermore, there are 14 behavioral and preferential variables describing product preferences interesting to the recipient unit. Due to confidentiality reasons, the branches from which the data is taken cannot be named, but are pseudonomized to the generic titles 1 to 3, with according product categories. There is information available on the general interest, the monetary volume persons are willing to spend on certain categories, as well as some variables indicating the purchasing power with regards to the product categories and branches.

For each target variable, the most interesting target value is highlighted. If selection is the database marketing goal, only one of the target values is

[2] An overview of the distribution of observations, i.e. persons, among the link variable classes in the population can be found in table 9.1 on page 303 in the appendix.



T1: Household income

T2: Transactions per year

T3: Volume spent on products 1

T4: Volume spent on products 2

T5: Purchase hist. products 3.1

T6: Purchase hist. products 3.2

T7: Purchase hist. products 3.3

T8: Purchase hist. products 3.4

T9: Purchase hist. products 3.5

t10: Volume spent on products 3

T11: Purchasing power products 3

T12: Purchasing power products 1

T13: Purchase hist. products 2.1

T14: Purchase hist. products 2.2

T15: Purchase hist. products 1.1

T16: Transactions products 1.2

Figure 5.6: Target variables and distributions available from the dataset

of interest. In the course of chapters 6 and 7, we use the first target variable as an example when it comes to explaining calculations and measures.

Besides analyzing the correlation between link variables, the predictive power of the link variables regarding the target variables needs to be regarded. Link variables without additional predictive power distort the augmentation process, lowering the overall augmentation results. The predictive power of link variables might decrease if there are high numbers of possible values for the link variables. Link variables can be reduced to fewer categories, if this increases their predictive power.

In a practical data augmentation, one would try to find the best set of link variables for every target variable. However, the model lift is explained to an important portion by link variables and their predictive power regarding the target variable. In order to be able to compare the model lifts in the case study, the choice and number of link variables are equal for all augmentations. It is, however, allowed for multivariate methods to omit some of the link variables in order to be more stable.

In a practical application, the predictive power of the link variables for individual target variables is analyzed within the source data. In the case study, the predictive power is tested within the overall population, because all sources are used, as long as there is a general predictive power measureable. When judging whether the link variables are able to discriminate between the target variable values, the distribution of predicted values is analyzed using logistic regression. Backward elimination is permitted, meaning that not all of the link variables have to be used for every target variable. If all target variable values have been predicted at least for some elements, the chosen link variables are able to discriminate between the target variable values.

In our dataset, eight out of 16 target variables are reproducible by the link variables using logistic regression, so that $k = \hat{k}$. If $k = \hat{k}$, all parameter values of the target variables are reproduced when predicting them with the link variables available. If the link variables are not able to predict all

Target variable	k	$\hat{k}$	Check	IQV	Minimum target class occupation
Target 6	2	1	N	0.16	4.2%
Target 13	2	1	N	0.35	9.6%
Target 8	2	2	Y	0.55	16.4%
Target 7	2	1	N	0.59	17.9%
Target 15	2	2	Y	0.65	20.5%
Target 9	2	2	Y	0.74	24.4%
Target 14	2	2	Y	0.77	25.9%
Target 4	2	2	Y	0.87	31.9%
Target 5	2	2	Y	1.00	47.8%
Target 10	3	2	N	0.80	9.2%
Target 2	3	1	N	0.75	9.8%
Target 11	3	2	N	0.89	12.0%
Target 12	3	2	N	0.82	18.7%
Target 3	3	3	Y	0.99	27.9%
Target 16	4	2	N	0.98	18.8%
Target 1	4	4	Y	0.99	19.0%

Table 5.4: Results of check for predictive power of the link variables regarding possible target variables

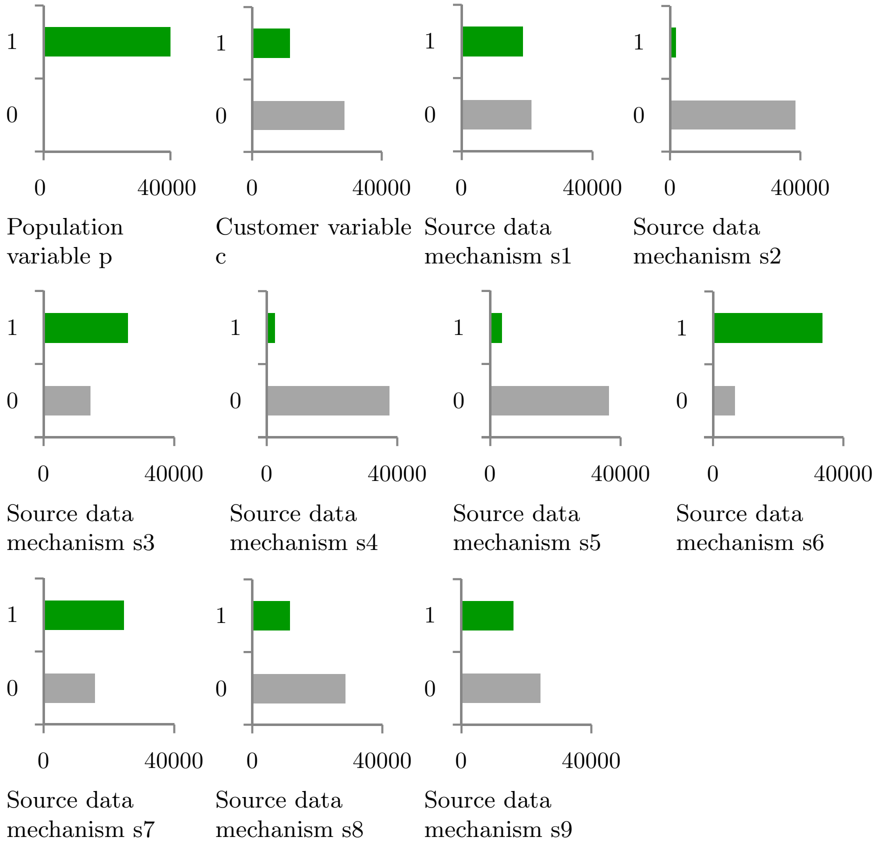
values, these target variables are not suitable for data augmentation and are not regarded further. If $k = \hat{k}$, then target variables are marked with a "Y" in the "Check" column, with a "N" otherwise. A list of the predictability check results is given in table 5.4. It can be seen from the variance in the target variable distribution as measured by IQV, that reproducible target variables have always IQV values of 0.55 or greater. Furthermore, the minimum target class occupation among the target values is shown. The minimum target class occupation refers to the percentage of the population with the least frequent target parameter value. Particular skew variables with the smallest target class occupation below 16% are never entirely reproduced. All of the target variables that did not pass the test have minimum target class occupations below 19%. For analysis purposes, we use all target variables that passed the check for predictive power (Y), plus those that have minimum occupations greater than 16%. It cannot be said that target variables with a minimum target class occupation between 16% and 19% are not suitable for data augmentation purposes, because some variables with these properties reproduced all target values. More analysis is conducted with them. From this selection, a first general rule regarding data augmentation in database marketing can be derived.

> *Finding:* For good data augmentation results, every target value should appear at least for 20% of the elements in the source.

In order to account for the predictive power of the link variables regarding the model lift, the R^2 of the logistic regression method is saved to function as moderator in the conceptual model. There are several ways to calculate R^2, among which McFadden's (1974) and Cox and Snell's (1989, p. 208f) measures are used most often in common statistical software (Allison, 2013). In our case study, Cox and Snell's measure is used (SAS Institute Inc., 2014a). Different measures result in different R^2 values. A high R^2 value is a good indicator for a good model fit, but can still miss important bias. It is not easy to find a universal threshold of R^2 indicating a good model fit. While in physical processes, R^2 values around 0.9 are desired, R^2 values describing human behavior are seldom greater than 0.5 (Frost, 2013). Because of these reasons, there is no exit criterion related to the predictive power in terms of R^2. However, the R^2 measure is used for analysis and moderator reasons in our case study, as changes in R^2 can give a relative indication on how much the model lift can be increased.

Auxiliary variables

The auxiliary variables comprise a binary population indicator variable, a customer indicator variable and nine binary source data mechanism indicator variables. The customer group is always the recipient unit. The source data mechanism indicator variables are used to draw the information-oriented samples, as described in chapter 5.2.3. The auxiliary variables are derived from information on media usage, contact information, and product usage information. They can have any relationship to the link and target variables. These relationships are subject to research.



Population variable p

Customer variable c

Source data mechanism s1

Source data mechanism s2

Source data mechanism s3

Source data mechanism s4

Source data mechanism s5

Source data mechanism s6

Source data mechanism s7

Source data mechanism s8

Source data mechanism s9

Figure 5.7: Auxiliary variables and distributions available from the dataset

Figure 5.7 gives an overview of all auxiliary variables available from the dataset. The parameter value used for designing the source data mechanisms is marked green. The population indicator variable has only one parameter value, because every observation in the dataset is part of the overall population. 11,560 observations are marked as customers ($c = 1$), while all other observations represent people not being customers in the overall population ($c = 0$). The nine source data mechanism variables are used for designing the information-oriented source samples. Some of them, like s_2, s_4, and s_5 denote only small groups, while greater parts of the population are represented in the sources denoted by s_3, s_6, and s_7. By choosing a variety of different source data mechanisms, a broad range of possible sources can be designed for the case study.

5.2.3 Test samples

Different data augmentation sources are created in order to test the stated hypotheses. Common sources for data augmentation are operational data, internal market research, external market research, and other external sources. Sources differ in terms of size, overlap, and representation. For every test sample, different sets of observations are chosen as donor units from the overall population. Most of them resemble realistic sources, but some are also created for comparison purposes. Figure 5.8 shows the schematic source samples from the overall population and indicates which source data mechanisms are used to create them.

Data in a source can be missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). By definition, sources with a 100% overlap rate regarding the customer database have ignorable source data mechanisms. Likewise, the source data mechanism of their representative samples can be ignored, because representative samples are drawn randomly. This has been formally argued in chapter 4.2.5. Whether a source is MCAR is usually known from the study design. Census sources and representative surveys are MCAR.

An identical subgroup ($D = R$) as shown in figure 5.8(a) is identified by $c = 1$, like the customer group (e.g. operational/behavioral data from login-protected areas of the company website). The five representative subgroups of the customer group ($D \equiv R'$) as shown in figure 5.8(b) are representative samples thereof (e.g. representative customer surveys). The representative samples are drawn randomly from the customer group. The donor unit equal to the customer group and the five randomly sampled sources with

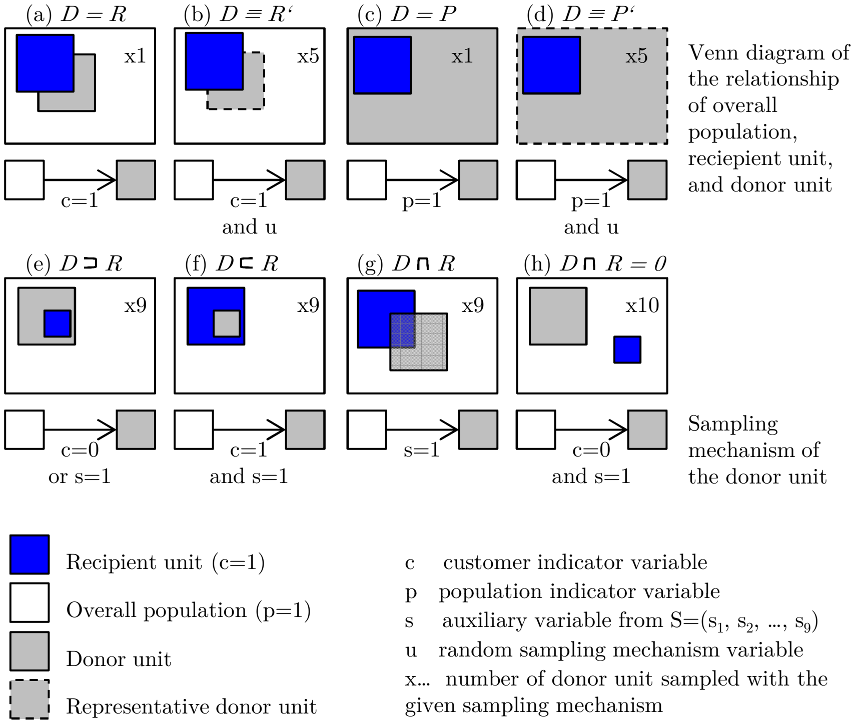

(a) D = R
x1
c=1
(b) D ≡ R'
x5
c=1
and u
(c) D = P
x1
p=1
(d) D ≡ P'
x5
p=1
and u
Venn diagram of the relationship of overall population, reciepient unit, and donor unit
(e) D ⊐ R
x9
c=0
or s=1
(f) D ⊏ R
x9
c=1
and s=1
(g) D ⊓ R
x9
s=1
(h) D ⊓ R = 0
x10
c=0
and s=1
Sampling mechanism of the donor unit
Recipient unit (c=1)
Overall population (p=1)
Donor unit
Representative donor unit
c customer indicator variable
p population indicator variable
s auxiliary variable from S=(s1, s2, ..., s9)
u random sampling mechanism variable
x... number of donor unit sampled with the given sampling mechanism

Figure 5.8: Sources created for the case study by information-oriented sampling

sampling rates between 10% and 50% are MCAR by definition. While in practice, an identical subgroup ($D = R$) is rather unlikely, a representative customer survey is common. For MCAR sources, the source data mechanism is always ignorable.

Similarly, there is a census group ($D = P$) as shown in figure 5.8(c), which exists of all people in the finite population (e.g. census data from a statistics bureau). The five representative subgroups of the population ($D \equiv P'$) as shown in figure 5.8(d) are random samples from the overall population (e.g. market media study). They are also drawn randomly from the population. The overall population and the five randomly sampled sources with sampling rates between 10% and 50% are MCAR by definition.

While in practice, a census source ($D = P$) is rather unlikely, a representative market media study is common.

In both cases, samples are randomly drawn, but from different populations. Although the identical subgroup ($D = R$) is a nonprobability sample of the overall population, both kinds are defined MCAR. This can be argued from a theoretical point of view. The overall population is defined by the marketing problem. If the problem, as well as the source, only concerns the customer population, then the customer group is defined to be the overall population. In this context, both the identical subgroup ($D = R$) and all representative samples are MCAR. Furthermore, all sources sampled by $c = 1$ or $s = 1$ are MCAR, because they have a 100% overlap rate regarding the recipient unit. We sample nine more groups of the form $D \supset R$ as shown in figure 5.8(e), defined by $c = 1$ or $s = 1$, where $S = (s_1, s_2, ..., s_9)$ is always one of the defined auxiliary variables.

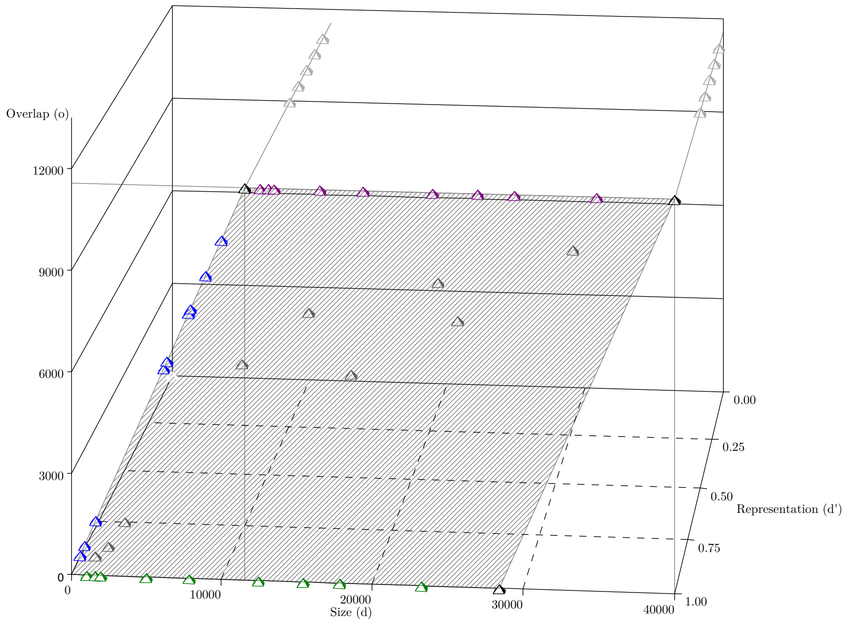
While the mentioned sources are easily classifiable, all other sources are those which database marketing analysts are particular interested in using. MCAR sources rely on strict sampling criteria. Consequently, data collection is time-consuming and costly. In contrast, MAR sources are readily available. For them, the source data mechanism cannot be defined upfront. Only if the conditional association of source and target variable, given the link variables, is independent, these sources are classified as MAR. The source data mechanisms used in the case study are binary variables, which equal 1 if a person is present in the source and 0 otherwise. MAR and MNAR data augmentation sources can have three forms. They can be partially overlapping, like a volunteer survey. They can be a subset of the customer group, like a social media source. The customer group can also be distinct from the source, although this is a less likely case.

The non-representative subgroups of the customers ($D \subset R$) as shown in figure 5.8(f) are nonprobability samples (e.g. volunteer survey on the company website). In the case study, the subgroup is identified by $c = 1$ and $s = 1$. In order to receive meaningful data to evaluate the hypothesis, nine source data mechanisms $S = (s_1, s_2, ..., s_9)$ are used to form nine different sources of the form $D \subset R$. In the course of chapters 6 and 7, we use such a source (Donors_c_s6) as an example when it comes to explaining calculations and measures.

Nine partially identical subgroups ($D \cup R$) as shown in figure 5.8(g) are chosen from the overall population by one of the source data mechanism indicator variables $S = (s_1, s_2, ..., s_9)$. The source data mechanism is a nonprobability sample (e.g. branch surveys, social media data). In the case study, such a subgroup is identified by $s = 1$. Non-overlapping subgroups ($D \cap R = 0$) as shown in figure 5.8(h) are chosen by the opposite value of the criteria as the customer group (e.g. competitor's survey). In the case study, nine of them are identified by $c = 0$ and $s = 1$ and one is the complement of the recipient unit, sampled with $c = 0$.

Not all versions mentioned are plausible sources. It is rather unlikely to find a source that has a 100% overlap rate regarding the recipient unit, but it is tested in order to evaluate the hypotheses. Likewise, the availability of census data for data augmentation is implausible in a way that information is sufficiently discriminable and interesting variables are present. The availability of donor units distinct from the recipient unit is rather unlikely, and its relevance should be questioned from a rational point of view. Nevertheless, they help to evaluate the hypotheses.

Three describing statistics are analyzed for every source: the size d, the overlap o, and the number of representing units (d'). If a source is sampled by a random sampling design, it contains all the information present in the population from which it was sampled, minus the sampling error. Size-wise, the overall population and the representative sample differ, but entropy-wise, they do not. The number of donors d is the number of *represented* units in these cases. Accordingly, a third dimension d' is introduced, being the number of *representing* units.



Color	Source data mechanism	Sampling rule
black	$D = R$	$c = 1$
	$D = P$	$p = 1$
	$D = P \setminus R$	$c = 0$
blue	$D \subset R$	$c = 1$ and $s_1, s_2, ..., s_9 = 1$
purple	$D \supset R$	$c = 1$ or $s_1, s_2, ..., s_9 = 1$
green	$D \cap R = 0$	$c = 0$ and $s_1, s_2, ..., s_9 = 1$
dark grey	$D \cup R$	$s_1, s_2, ..., s_9 = 1$
light grey	$D \equiv R'$	$c = 1$ and u
	$D \equiv P'$	$p = 1$ and u

Figure 5.9: Sources by overlap, size, and representation

Figure 5.9[3] shows the scope of parameters and where to find the sources in the three-dimensional space of overlap, size, and representation of the source. The schematic illustration from figure 4.3 on page 122 in chapter 4.2.1 is used here to plot all sources, being enhanced by a third dimension for d'. Every source represents a different combination of character-

[3] The colored version of this figure can be found online on www.springer.com under the title of this publication.

istics. All sources which are not randomly sampled can be found in the two-dimensional plane which is defined by overlap and size. All of these sources have a sampling rate of 100%, i.e. they are not randomly sampled. The vector space of possible combinations of size and overlap is shaded.

Starting from the origin, it is bordered by the $D = R$ source (upper left corner of the parallelogram, black), the $D = P$ census source (upper right corner, black) and the $D = P \setminus R$ complement source sampled with $c = 0$ (lower right corner, black). The $D = R$ source has a maximal overlap rate and the number of donors is equal to the number of overlapping units ($d = o = 11,560$). The $D = P$ source has a maximal overlap rate and a maximal number of donors ($40,000$). The $D = P\setminus R$ source has a 0% overlap rate and contains all elements of the population except for the customers, therefore the number of donors is $p - r = 28,440$.

The $D \subset R$ sources are stringed along the line connecting the origin and the upper left corner (blue). They are a subgroups of the recipient unit, so that the number of overlapping units is equal to the number of donors. All donors are also element of the recipient unit. Depending on the number of donors, the source is found on a lower or higher location along the connecting line. The $D \supset R$ sources are alined on the upper border (purple). Their overlap rate is 100%, because all recipients are also found in the source. But in contrast to the $D = R$ source, there are other donors contained in the source as well, so that the recipient unit is a subgroup of the donor unit. The $D \cap R = 0$ sources are placed on the lower border (green). They all have no overlapping units with the recipient unit. The $D \cup R$ sources are scattered in the two dimensional plane without reaching the borders (dark gray). Any combination of overlap rate and number of donors is possible.

Only those sources randomly sampled from either the overall population or the customer group can be found in the third dimension of representation. However, these are spread in the plane of size and representation only (light gray). The overlap is always equal to the number of customers, because sources representatively sampled from a MAR database are still considered MAR. The random samples of the customer group $D = R$ connect the upper left corner of the parallelogram with the opposite site of the three dimensional space. The random samples of the customer group $D = P$ connect the upper right corner with the backside.

5.2.4 Methods

Three common methods are compared during the case study: conditional mode imputation, nearest neighbor hot deck, and logistic regression. They are a choice of common and simple methods used for data augmentation in database marketing. Their influences on the model lift, interacting with the source characteristics, are analyzed in order to give advice on how to choose a suitable method.

Conditional mode imputation

As described in chapter 3.2.3, several rules are introduced in order to use the conditional mode imputation method. During conditional mode imputation, the distribution of target values is regarded for every link variable class. The most frequent value is augmented to every recipient with the respective link variable class. Conditional mode imputation is a rule-based method where several decisions have to be made. A mode value is not only augmented if the value occurs most often. The most frequent value has to occur at least for more than two donors more than the second most frequent value. This rule is supposed to decrease the uncertainty involved. For example, it means that the second value will be augmented as best value, if $d(Y = y_1|X) = 3, d(Y = y_2|X) = 10, d(Y = y_3|X) = 1$, but not if $d(Y = y_1|X) = 3, d(Y = y_2|X) = 4, d(Y = y_3|X) = 1$. It has been decided that the delta of 2 adds sufficient certainty to the augmentation.

Once the conditional mapping is finished, the recipients are split into a group where values have been augmented, because their class was unambiguous, and a group where values have not been augmented, because of the threshold. For the latter group, a second augmentation is performed omitting one link variable. By omitting a link variable, the number of classes decreases and the most frequent values can be found anew. The omitted link variable is chosen arbitrarily depending on contextual factors, such as scattering (variables with many target values lead to small classes with low occupation) and importance (as judged by the database marketing analyst). This process is iteratively continued until either all recipients have received a value or the link variables are maximally collapsed.

Nearest neighbor hot deck

For nearest neighbor hot deck, the closest donor is searched for every recipient and the according target value is augmented. In order to calculate distances, categorical information has to be transformed to binary variables. Categorical variables are decomposed into binary variables by forming one variable for every target value, denoting whether a characteristic is true or not. Ordinal variables are decomposed on an interval basis. If the observed value is greater than a threshold value, it is assigned 1, else 0 (Backhaus et al., 2008, p. 410). That way, adjacent categories are "closer" than categories with other values. Because the variables are not continuous, the Mahalanobis distance measure is used to determine proximity (Backhaus et al., 2008, p. 211). No minimum acceptable distance is specified for classification. In our case study using SAS, the nearest neighbor hot deck method is implemented with a discriminant analysis procedure. A nonparametric method option is chosen with $k_{nn} = 1$, where k_{nn} is the number of donors regarded as nearest neighbors. It means that the value is augmented from the very nearest neighbor only.

The SAS procedure produces the probabilities needed for uncertainty assessment. At the same time, the according value from the nearest neighbor is augmented. If no decision can be made, because at least two values are equally likely, no value is augmented. This is unique in the comparison of methods. The resulting customer database has a target variable distribution with an additional "unknown" target value. In the total correct classification rate and model lift measures, these "unknown" values are considered wrong, although it is not really wrong. In contrast to falsely augmented values, this target value has a meaning the database marketing analyst can interpret: For this specific customer, no statement can be made regarding this target variable. At the same time, no wrong decision can be made. However, it does not provide any new information, which is why it is treated like a falsely augmented value in the overall measures.

Logisitic regression

For logistic regression, a model is built within the source dataset and the model parameters are used to predict values for the recipient unit. In order to evaluate the significance of the influence of individual link variables, the Wald statistic is calculated. It tests the hypothesis whether the influence of a link variable is negligible. The Wald statistic has an asymptotic χ^2 distribution and can therefore be tested accordingly (Backhaus et al., 2008, p. 273). If H_0 is rejected, the respective link variable has a significant impact on the value of Y and should be used in the data augmentation model. If not, appropriate model selection methods can be used to eliminate these variables from the analysis, e.g. backward elimination.

The SAS procedure produces the probabilities needed for uncertainty assessment. The most likely value as derived from the logistic equation is augmented. Because of the logarithmic nature of the logistic regression analysis, every recipient receives a value – there is no case of a tie.

Chapter 6

Analysis of data augmentation KPIs

In order to answer the research question of which sources are suitable for increasing conversion probabilities in direct marketing, we have developed a conceptual model and identified factors influencing the conversion probability. Accordingly, we have created a test design in order to test the hypotheses derived from the conceptual model. A case study is carried out, performing multiple data augmentations, while varying defined parameters. The results of the different augmentations are used in order to give advice on how to best assess data augmentation sources ex ante.

The question of which sources are suitable for data augmentation has several components. The target variables contained must be worth knowing for the customers in the database on an individual basis. These target variables must be predictable by the link variables. A simple quality check criterion – the number of augmented target values – already gives an answer on which target variables are well predictable. From this criterion, the first managerial implications are derived in this chapter, regarding the target variable characteristics.

For every augmentation, a set of descriptive variables is kept. They include the source, target variable, and method used in the respective augmentation, as well as several characteristics describing these input parame-

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_6

ters. They all influence the data augmentation results. In order to assess the goodness of the augmentation results, we use Ratner's (2003) and Hattum & Hoijtink's (2008b) total correct classification rate and model lift measures. We extent these to other KPIs able to give insight on the augmentation performance. We add two KPIs describing the CPL, which are necessary to evaluate the quality of the results with regards to selection tasks.

The KPIs are only calculable in a simulated context, where the true values are known. By analyzing the descriptive data, it is possible to make aggregated statements on the performance of augmentations in varying settings. In this chapter, the calculation of the measures is described along with illustrating examples. Ranges and interpretation rules are given. These KPIs are used in chapter 7 to analyze the influence of different source characteristics and to answer the research question.

6.1 Preserved descriptive data and measures

For every augmentation carried out in the case study, several criteria are saved in a descriptive file, which is used to test the stated hypotheses. Every source-target combination is considered a case in the case study. Three augmentations are performed with every case using alternative methods. Variables describing the sources are preserved, such as the number of donors, the number of representing donors, the size, and the sampling rate. The relationship of the source data mechanism and the target variable is regarded for every source-target combination and according test results are saved. The results are measured by total correct classification rates, model lifts, and the parameters needed to evaluate the stated hypotheses.

The list of descriptive information and measures for every augmentation is given in table 6.1. The category, the variable name, and a description are shown along with an exemplary figure. Their meanings and functions are explained in this and the following chapter, as denoted in the last column. The augmentations are compared using this descriptive data. From the

Category	Variable name	Description	Example	Chapt.
Recipient unit	Recipients	Number of customers in the recipient unit	11,560	
Target variable characteristics	Target	Name of target variable	Target_1	
	k	Number of target values	4	
	IQV	Variance of target variable	0.963	
	n_{target}	Number of recipients with target value	4,093	
Source characteristics	Source	Name of source	Donors_c_s6	
	Donors	Number of donors in the source	10,005	
	Overlaps	Number of overlapping elements	10,005	
	VCs	Number of link variable classes	793	
	Sampl. rate	Sampl. rate for repr. sources	1	
	d_rep	Number of representing elements	10,005	
	Rsqu	Predictive power of link variables regarding target variable in the donor unit	0.024	
Method	Method	Name of method	Near. neigh.	
	k_augm	Number of augmented target values	5	6.2
	Check	Discrimination check indicator	Y	
Model lift	$TCCR_{model}$	Total correct classification rate	0.426	6.3.1
	CCR_{target}	Correct class. rate for target value	0.607	6.3.2
	ML_{chance}	Model lift as compared to a random distribution of the target variable in the recipient unit	1.534	6.4.1
	ML_{target}	Model lift for target value, as compared to a random distribution of value in the recipient unit	4.845	6.4.2
	$ML_{uniform}$	Model lift as compared to a uniform distribution of most frequent target value in recipient unit	1.203	6.4.3
	ML_{source}	Model lift as compared to a random distribution of target variable in the source	1.544	6.4.4
Conversion probability lift	CL_{global}	Conversion probability lift	1.353	6.5
	CM_{global}	Conversion probability lift magnitude	0.547	6.5
	Optimum_conv	Optimal number of recipients for a direct marketing campaign	5,799	6.5
	Uplift_max	Uplift at the point where the optimal number of recipients is highest	1.500	6.5
Source data mechanism tests	CHI Assoc.	Test result for χ^2 test with aggregated total measures	dependent	7.1.1
	CMH Ass.	Test result for CMH test	dependent	7.1.1
	Wald Ass.	Test result for influence of source data mechanism indicator variable	dependent	
	Model I Ass.	Test result for difference between inclusion and exclusion model	dependent	7.1.2
	Model II Ass.	Test result for difference between separation and exclusion model	dependent	7.1.2

Table 6.1: Descriptive information saved for every augmentation

aggregated results, overall tendencies can be detected and statistic models can be established to answer the research question.

Throughout this and the next chapter, we use the exemplary case from the fourth column in table 6.1 for illustration purposes. For all cases, the recipient unit – the customer database – is equal, while the donor units and target variables differ. The recipient unit contains 11,560 elements. Target_1, the net household income, is used as ordinal target variable in this case. It has four possible target values (k) and an index of qualitative variation of 0.963. It is thus a fairly even distributed target variable. The highest of the four values (n_{target}) is later used for target group selection. In the recipient unit, 4,093 recipients have the desired value. This is known, because the situation of missing target variables is simulated. The target variables were removed from the recipients in order to perform the augmentations.

In the case shown, Donors_c_s6 is used as external source and could be for example data gathered by the company website. It is a $D \subset R$ source and has been sampled by $c = 1$ and $s_6 = 1$ from the overall population during the data generation phase, as illustrated in figure 5.8 on page 167 in chapter 5.2.3. Donors_c_s6 has an overlap rate of $\frac{o}{r} = \frac{10{,}005}{11{,}560} = 87\%$ regarding the recipient unit. Most of the customers use the website, but not all. It is a nonprobability sample of the customer population, because not every customer has the same chance to be included in the model. No additional random sampling component is added, so that the sampling rate equals 1 and the number of representing units (d_rep) is equal to the number of donors. Nine link variables are used in the case study, of which 793 link variable classes are available from Donors_c_s6. The predictive power of these link variables regarding Target_1 in the source has a R^2 value of 0.024 in a logistic model.

In this manner, we report the results for all augmentations. Eleven target variables are augmented for each of the 49 sources described in chapter 5.2.3. For each case, three augmentations are carried out using conditional mode imputation, logistic regression, and nearest neighbor hot deck. Thus, 1,617 augmentation results are regarded in the analysis of the case study results. The remaining measures of the above example are explained at appropriate steps in the analysis.

Throughout the course of the chapter, we pursue a three-part analysis approach. First, the relevant measure is theoretically explained, including input variables, calculations, resulting KPIs and according properties. Then, an exemplary calculation is depicted for one of the augmentations in order to illustrate the procedure. Finally, the resulting KPIs for all augmentations are analyzed, compared, and interpreted.

6.2 Quality checks

Before introducing the KPIs used to describe the quality of the augmentations results, a simple and easily observable quality check criterion is introduced: the number of augmented target values. Therefrom, exit criteria are derived which allow the database marketing analyst to assess the target variables contained in the source upfront and to make a decision whether an augmentation can be successful.

6.2.1 Predictive power and discrimination

The predictive power of the link variables regarding the target variables has already been assessed during the dataset generation phase using the overall population. However, it is still possible that the link variables are not able to discriminate between target values, when using a different (smaller) source and applying it to the customer database. The reasons why not all values are reproduced are manifold.

- *Donors:* The smaller the source, the less information is contained and the less likely all values are reproduced.
- *Number of target values:* The more target values, the less likely all target values are reproduced. Target values need to have a significant

target class occupation in order to be predictable. The target class occupation is calculated from the number of elements with a specific target value, divided by all elements.

- *Skewness:* The skewer a target variable, the smaller the individual target class occupations, i.e. the less likely all target values are reproduced.
- *Predictive power:* If the predictive power between link and target variable is already weak when being observed in the source, it is more likely to be non-existing when applied to the recipient unit.
- *Method:* Certain methods reproduce values more easily than others. Multivariate methods, for example, rely more on predictive power and are less likely to reproduce values than nearest neighbor methods.
- *Recipients:* In a MAR case, it is possible that the recipients do not possess all the values. When not augmenting all values in such a case, the total correct classification rate is high, although not all target values are augmented.

In the following, we introduce a simple quality check criterion for assessing the usefulness of data augmentations results. If all customers were augmented the same target value, such an augmentation would be invalid (unless the customer database was correlated with the target value, which we precluded in chapter 3.1.1). At least two different target values need to be reproduced for the augmentation to be successful. Optimally, all target values are reproduced.

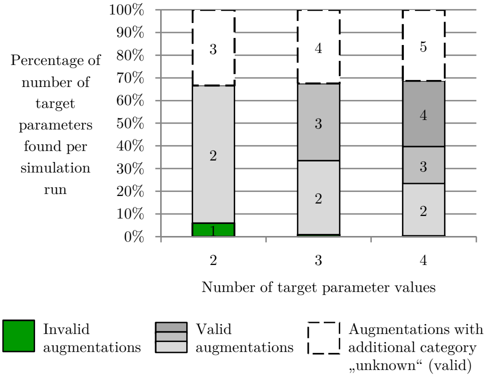
Calculation For quality check reasons, a check variable has been saved for every augmentation, indicating whether all possible target values were reproduced. Also, a variable for the number of augmented target values $\hat{k}$ has been saved. If $k = \hat{k}$, all values have been reproduced. If $k < \hat{k}$, less values have been reproduced. $\hat{k}$ can also be greater than k, when using nearest neighbor hot deck, because values can be assigned an additional category "unknown". In this case, $k = \hat{k}$ one target value has not been reproduced, but the additional category "unknown" instead.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
k (true)	4	4	4
k_augm	4	2	5
Check	Y	N	Y
Decision	valid	valid	valid

Table 6.2: Exemplary quality check in the case of Target_1 and Donors_c_s6

Example In table 6.2, the exemplary case from chapter 6.1 is resumed in order to show how this decision is made. The quality check criterion indicates whether the augmented number of values (k_augm) is equal or higher than the number of true target values($k = 4$). The results of conditional mode imputation show all four values. The logistic regression method only reproduced two of the values. Although all values have been reproduced with logistic regression when using the overall population for augmenting Target_1 in the data generation phase described in chapter 5.2.2, this is no longer true for the nonprobability sample Donors_c_s6. Nevertheless, the result is valid, in a way that some discrimination among the recipients is possible. Nearest neighbor hot deck produced five target parameter values, including the additional category "unknown".

Analysis Figure 6.1 shows the augmentations observed with the respective $\hat{k}$ (y-axis), given a certain number of target values k (x-axis). If the same target value was augmented to all recipients in the recipient unit ($\hat{k} = 1$), the data augmentation would be invalid. The percentage of invalid augmentations is shaded blue in figure 6.1. If at least two values are found, augmentations are valid ($\hat{k} > 1$). The percentage of valid augmentations is shaded gray in figure 6.1. If more values were produced by nearest neighbor hot deck than there are values ($\hat{k} > k$), the augmentation would be also valid. These cases are shaded white with a broken line as its border.



Diagram

k	2	3	4	Total
5			92	92
4		143	85	228
3	294	150	48	492
2	535	144	68	747
1	53	4	1	58
N	882	441	294	1,617

Measures

Figure 6.1: Quality check based on the number of reproduced target values for all augmentations

The number of invalid augmentations decreases with increasing k. For $k = 2$, the percentage of augmentations resulting in a uniform target value distribution is 6%. For $k = 3$ and $k = 4$, it is below 1%. At the same time, the number of augmentations with $k = \hat{k}$ decreases with increasing k. For $k = 2$, the percentage of augmentations resulting in the correct number of target values is 61%, while it drops to 34% for $k = 3$ and 29% for $k = 4$. The percentage of augmentations with valid target value numbers other than $k = \hat{k}$ increases with increasing k.

Due to invalid results, 58 augmentations are exited. They cannot further be used for our analysis. Of them, 7 were augmented using conditional mode imputation and 51 using logistic regression. Nearest neighbor hot deck augmentations did not produce invalid results. 52 of the invalid augmentations used sources with a 0% overlap rate regarding the recipient unit. This criterion is considered more thoroughly in chapter 7.2.2. For further analysis, 1,559 augmentations are considered.

In a practical application, one would switch from one method to another, if one is not able to discriminate between target values. The reproduction of at least two target variable values does not guarantee for a significant model lift, but if no discrimination can be achieved, the data augmentation definitely failed.

6.2.2 Managerial implications

With our case study, we want to give an answer on which sources are suitable for data augmentation in database marketing. But of course, the value and suitability is primarily dependent on the information contained. All variables able to describe direct marketing target groups are relevant to the database marketing problem. It implies that the information, or any similar or ample information, is not already available in the customer database or in any other system available to the company. The exact type of information is case-specific and depends on the branch, the company, the marketing goals, and the specific marketing problem.

The explanatory power of the link variables is tested within the source. Standard test methods for multivariate relationships can be used. If the general predictive power between link and target variables is already weak when being applied to the source, it is more likely to be non-existing when applied to a different group. Some methods reproduce values more easily than others. Multivariate methods, for example, rely on significant predictive power and are more likely to not reproduce target values than others. As the predictive power is essential to the data augmentation approach, any source not showing significant predictability should be exited in case of doubt. From the quality check after the augmentations in our case study, the following guideline for application and management is derived.

> **Target variable guideline:** A categorical target variable should have at most five possible target values. The distribution

of these values should be sufficiently even in order for methods to discriminate between them ($IQV > 0.8$).

There are general rules regarding suitable target variables. Target values need to have a significant target class occupation in the source in order to be predictable. The skewer the target variable, the smaller individual percentages. The more target values, the less likely all target values are reproduced. Only target variables with a moderate number of target values can be taken into consideration, in a way that every target class is sufficiently occupied. We found that a target class occupation of 20% or less is problematic. Therefrom, we concluded that only five parameter values are possible. However, the exact threshold depends on the data context, the discrimination of the link variables, and the database marketing goals in general. Our findings can only serve as an indication.

6.3 Accuracy and precision of results

In order to evaluate the accuracy and precision of the augmentation results, the augmented values of the target variable are compared to the true values of the recipients for that variable. This comparison can only be made in the case study context, where the true values are known. In a practical application, it can neither be assessed ex ante, nor ex post. Because of this disadvantage, we conducted our case study with simulated missing target variables. If rules can be established in the case study environment, where the augmented-true-comparison is possible, they can help to assess practical data augmentation projects ex ante.

There are two different measures for accuracy and precision. The accuracy measure – the total correct classification rate – was first mentioned by Ratner (2003, p. 182) and described in more detail by Hattum and Hoijtink (2008b). It regards all target values and whether they were hit correctly. For segmentation tasks, the total correct classification rate is the relevant measure. The precision measure – the correct classification rate of the target value – has been developed by Hattum and Hoijtink (2008b). It only regards one of the target values which is of particular interest. Only for this value, e.g. whether a recipient has a high income, is it measured how many recipients were assigned the right value, as compared to the true value. For selection tasks, the classification rate of the target value is the relevant measure. It reduces the regarded results to the target value of interest.

6.3.1 Total correct classification rate

The accuracy of the data augmentation results can be measured by the total correct classification rate ($TCCR$). It is a measure for how many target values were "hit", i.e. how many augmented values are equal to the true values (Ratner, 2003, p. 182). During case studies with simulated missing target variables, $TCCR_{model}$, or the hit rate, can be calculated by comparing the augmented values to the true values (Hattum & Hoijtink, 2008b). Hits only occur for discontinuous variables, because the probability of a hit for continuous variables equals zero. In data augmentation applications for database marketing, this level of validity is the most important one. The following calculation is adapted from Hattum and Hoijtink (2008b) so as to fit our notation concept.

Calculation Let there be a hit indicator variable (h), which can assume the value 1 if a target value is augmented correctly ($y = \hat{y}$) and 0 otherwise ($y \neq \hat{y}$). The number of correctly classified recipients is calculated by the sum of h_i with $0 < i <= r$. $TCCR_{model}$ is calculated by the number of correctly classified recipients divided by the number of all recipients:

$$TCCR_{model} = \frac{\sum_{i=1}^{r} h_i}{r} \; \forall \; h_i = \begin{cases} 1 & \text{if } y = \hat{y} \\ 0 & \text{if } y \neq \hat{y} \end{cases} \tag{6.1}$$

The total correct classification rate of the model has a range of $0 \leq TCCR_{model} \leq 1$. The $TCCR_{model}$ values are dependent on the distribution of the target variable and the number of target values. Target variables with a skew distribution generally have a higher $TCCR_{model}$ than those with symmetric distributions, because the target value which occurs most often is hit more easily. Equally, target variables with a low number of target values generally have a higher $TCCR_{model}$ than those with a high number of target values, because values are also hit more easily.

Example The number of correctly classified recipients can be exhibited by creating a two way frequency table with the augmented values as rows and the true values as columns. An example for such a two way table is shown in table 6.3. It shows the calculation of the total correct classification rate in the example from chapter 6.1 for the augmentation of Target_1 using Donors_c_s6 and the nearest neighbor hot deck method. In the example, $TCCR_{model} = \frac{421+467+1{,}552+2{,}486}{11{,}560} = 0.43$. $4,926$ recipients (the sum of the diagonal values), i.e. 43% of the target values, were hit by the model. All values classified as "unknown" were not really classified wrong. Their classification can be interpreted. Nevertheless, these values are not correct, so that they cannot be taken into consideration for the total correct classification rate.

Number of recipients	True y_1	True y_2	True y_3	True y_4	Row sum
Classified $\hat{y_1}$	421	89	109	84	703
Classified $\hat{y_2}$	112	467	138	122	839
Classified $\hat{y_3}$	389	448	1,552	633	3,022
Classified $\hat{y_4}$	457	523	961	2,486	4,427
Classified $\hat{y_{unknown}}$	472	516	813	768	2,569
Column sum	1,851	2,043	3,573	4,093	11,560

Table 6.3: Exemplary two-way frequency table for calculating a total correct classification rate for the augmentation of Target_1 using Donors_c_s6 and the nearest neighbor hot deck method

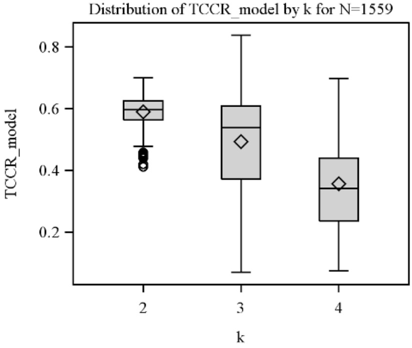
The comparison of the results from different methods given a particular case is referred to as within case analysis. The results give insight into which augmentation method to use best, given a certain setting. Table 6.4 shows the $TCCR_{model}$ values for the different methods used in the case of Target_1 and Donors_c_s6. Nearest neighbor hot deck performed best and reached a total correct classification rate of 43%. Logistic regression lead to $TCCR_{model} = 39\%$, although not all values were reproduced, as shown in table 6.2 on page 181 in chapter 6.2.1. Conditional mode imputation only yielded a total correct classification rate of 34%. In this case, nearest neighbor hot deck would be the best choice. In this manner, we have reported the $TCCR_{model}$ measures for every case in the case study.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
$TCCR_{model}$	0.34	0.39	0.42

Table 6.4: Exemplary $TCCR_{model}$ values in the case of Target_1 and Donors_c_s6

Analysis The distribution of the $TCCR_{model}$ values of all augmentations is shown in figure 6.2, distinguished by k. The number of observations (N) is 1,559 – 829 augmentations with $k = 2$, 437 with $k = 3$, and 293 with $k = 4$. A box and whiskers plot is used in figure 6.2a, where the box indicates the range of 50% of the observations (separated by a median line and marked with the mean) and the whiskers indicate the observed minimum and maximum values. The values for the whiskers, the box borders, the median, and the mean are shown in figure 6.2b.

It can be seen that $TCCR_{model}$ varies between 41% and 70% for $k = 2$, while half of the observations are found between 56% and 63%. The observations for $k = 3$ have a greater variance. They are spread between 7% and 84%, while most observations are situated between 37% and 61%. Observations with $k = 4$ have lower overall total correct classification rates, ranging from 7% to 70%, while the core of the augmentation results can be found between 24% and 44%. We stated earlier that target variables with a low number of target values generally have a higher $TCCR_{model}$ than those



Box plot

k	2	3	4
N	829	437	293
Max	0.70	0.84	0.70
Q3	0.63	0.61	0.44
Median	0.60	0.54	0.34
Mean	0.59	0.49	0.36
Q1	0.56	0.37	0.24
Min	0.41	0.07	0.08
Sh.-W.	0.97	0.96	0.96
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	369.14	<.001
Levene	224.95	<.001
Welch's ANOVA	354.27	<.001

Measures

Figure 6.2: Distribution of all $TCCR_{model}$ values by number of target values k

with a high number of target values, because values are hit more easily if the class occupation is higher. On average, the observations for $k = 2$ are better than those with $k = 3$, with a mean of 59% versus 49%. In turn, these are better than those with $k = 4$, with a mean of 36%.

In order to assess whether the differences of k are significant, a comparison between the group means is conducted using analysis of variance (ANOVA). With ANOVA, the means of two or more groups are compared, if the dependent variable ($TCCR_{model}$) is continuous and the independent variables (k) is discrete (Backhaus et al., 2008, p. 152)[1]. If the compared groups are of different size, like it is here ($k = 2$: 829, $k = 3$: 437, $k = 4$:293), it is referred to as unbalanced design. Because a single independent variable is used, this is not problematic (Milliken & Johnson, 1984, p.127). In order to conduct ANOVA, three assumptions must hold. The observations must be independent in a way that no measurement is done twice for the same combination of independent variables. This is controlled during the study design phase. The error terms of each treatment must be normally dis-

[1] k is actually an interval variable, i.e. a natural number. However, as only a small, but relevant, space of $2 \leq k \leq 4$ is regarded, it can be treated as discrete and can also be denoted as ANOVA with fixed effects (Huber, 2008, p. 2/26).

tributed – although ANOVA is relatively robust regarding deviations of the normality assumption, especially if the number of observations is sufficiently large. Furthermore, the variance of the classes must be approximately equal, which is referred to as homoscedasticity (Huber, 2008, p. 2/4).

The statistics computed for ANOVA are included in figure 6.2b. In order to assess the normality requirement, the distribution of the residuals[2] is tested using the Shapiro-Wilk test (Sh.-W.)[3]. The null hypothesis states that the data follows a normal distribution. If the calculated statistic ($0 < W \leq 1$) is high, the null hypothesis is accepted. As it is dependent on the sample size, it can only be used for samples smaller than 2,000 (SAS Institute Inc., 2014b). In figure 6.2b, the Shapiro-Wilk test yields a very high measure ($W = 0.97$ or $W = 0.96$ for the different levels of k) with a p-value below $\alpha = 5\%$ each. The normality assumption holds.

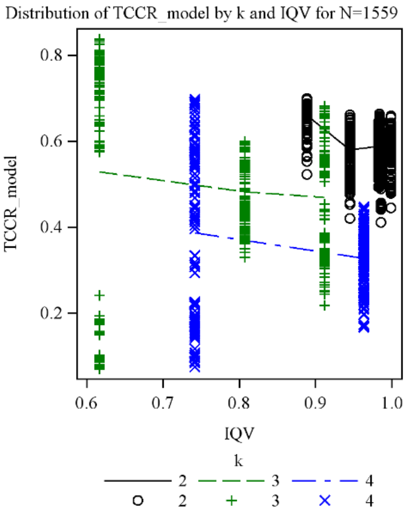
The Levene's standard homogeneity of variance test[4] is used in order to assess whether the variances of the classes are equal (SAS Institute Inc., 2014c). The null hypothesis states that all variances are equal (Backhaus et al., 2008, p. 159). If the value of the F statistic as derived from the F distribution (F) is high and the according p-value is below $\alpha = 5\%$, the null hypothesis needs to be rejected. If the variances follow heteroscedasticity, Welch's variance-weighted ANOVA can be used in order to compare the means (Huber, 2008, p. 2/38). It is robust to unequal variances. Both the ANOVA and the Welch's ANOVA value in figure 6.2b have the null hypothesis that the means are equally distributed among the classes. If the F value is larger than 1 and the according p-value is below $\alpha = 5\%$, the null hypothesis is rejected (Huber, 2008, p. 2/39). Thus, the distribution of means of $TCCR_{model}$ as measured by the group means of k is not random.

[2] The residual of each observation is calculated as the distance between the predicted mean and the actual value (Huber, 2008, p. 2/32).

[3] Other tests are also possible, including the Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test (SAS Institute Inc., 2014b). However, the Shapiro-Wilk test is robust for small sample sizes (Sen & Srivastava, 1990, p. 5) and is used here.

[4] Other tests would also be possible, including the Bartlett's and O'Brien's tests (SAS Institute Inc., 2014c). However, Levene's test is most common (Huber, 2008, p. 2/34).

The difference in means is significant, so that it can be said that $TCCR_{model}$ changes with changing k.



Scatter plot

IQV	N	Mean	Sh.-W.	p-value
k=2				
0.89	129	0.66	0.82	<.001
0.94	146	0.58	0.93	<.001
0.98	137	0.59	0.96	<.001
0.98	137	0.60	0.94	<.001
0.99	138	0.55	0.90	<.001
1.00	142	0.57	0.91	<.001
k=3				
0.62	147	0.53	0.77	<.001
0.81	143	0.48	0.94	<.001
0.91	147	0.47	0.91	<.001
k=4				
0.74	147	0.39	0.89	<.001
0.96	146	0.33	0.97	<.001

k	Test	F	p
2	ANOVA	87.80	<.001
	Levene	14.02	<.001
	Welch's A.	127.30	<.001
3	ANOVA	3.85	0.0219
	Levene	194.09	<.001
	Welch's A.	2.45	0.0885
4	ANOVA	11.09	0.0010
	Levene	272.19	<.001
	Welch's A.	11.15	0.0010

Measures

Figure 6.3: Distribution of all $TCCR_{model}$ values by number of target values k and index of qualitative variance IQV

As has been introduced in chapter 4.2.4, the index of qualitative variation developed by Wilcox (1973) is used in order to judge the variance of a categorical variable. The scatter plot in figure 6.3[5] incorporates IQV into the analysis, which varies between 0.88 and 1 for $k = 2$, between 0.62 and 0.91 for $k = 3$, and between 0.74 and 0.96 for $k = 4$. The scatter plot in figure 6.3a shows individual values marked with circles, plusses, and crosses, along with a line connecting the means of each combination indicating the overall tendency. The number of observations per k-IQV-level and the respective mean are shown in figure 6.3b. It can be seen that the

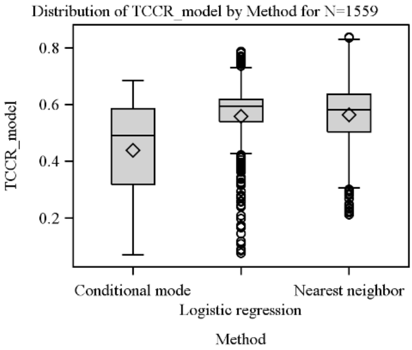
[5] The colored version of this figure can be found online on www.springer.com under the title of this publication.

variance of $TCCR_{model}$ is greater for low values of IQV. However, the average $TCCR_{model}$ decreases with increasing IQV. This has been expected, because values are hit more easily, if the distribution is skewer.

Using ANOVA, it can be assessed whether the mean differences among the IQV values per number of target values k are significant. As shown in figure 6.3b, all residuals are normally distributed for every level of k and IQV (as measures by the Shapiro-Wilk test), so thatANOVA statistic can be calculated. The distributions of the variances of IQV are heteroscedastic for every level of k (as measured by Levene's test), so that Welch's ANOVA is used. While the means differ significantly for $k = 2$ and $k = 4$, the mean differences are not significant for $k = 3$ at a $\alpha = 5\%$ level of significance. IQV cannot be regarded as a strong predictor for $TCCR_{model}$. Its interaction with other predictors is more closely examined in chapter 7.2.

Variables that did not pass the quality check have already been deleted during the data generation phase. It has become clear that target variables with a low IQV are generally not suited for data augmentation. It is difficult to differentiate between target values, if only a small portion of the donors show particular target values. From figure 6.3, it can further be seen that the variance of $TCCR_{model}$ is unacceptably high for $IQV < 0.8$. It indicates that only evenly distributed target variables are a guarantee for good data augmentation results. As found earlier, a target variable sufficiently even has at least 20% of occurrences per target value.

Figure 6.4 shows a box plot for the distribution of $TCCR_{model}$ values across all augmentations by augmentation method used. It can be seen that conditional mode imputation is generally inferior to logistic regression and nearest neighbor hot deck. The average $TCCR_{model}$ of conditional mode imputation is 44%, as compared to 56% for both logistic regression and nearest neighbor hot deck. The variation is bigger, ranging from 32% to 59% for half of the observations. The results from logistic regression and nearest neighbor reach better $TCCR_{model}$ values, where half of the augmentations reliably vary between 50% and 64%. Logistic regression results are mostly



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	0.69	0.79	0.84
Q3	0.59	0.62	0.64
Median	0.49	0.59	0.58
Mean	0.44	0.56	0.56
Q1	0.32	0.54	0.50
Min	0.07	0.08	0.21
Sh.-W.	0.91	0.85	0.97
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	124.31	<.001
Levene	53.27	<.001
Welch's ANOVA	99.87	<.001

Measures

Figure 6.4: Distribution of all $TCCR_{model}$ values by applied augmentation method

found between 54% and 62%. The total correct classification rate cannot be judged in absolute terms, as it is highly dependent on the distribution of the target variable. However, because all augmentations are performed for the same source-target combinations, the relative differences are notable.

An ANOVA is calculated in order to judge whether the mean differences of the methods are significant. From the Shapiro-Wilk test, it can be concluded that the residuals are normally distributed. The value of the W statistic for Shapiro-Wilk's test for normally distributed residuals (W) as shown in figure 6.4b is close to 1 for all augmentation methods. The variances are not equal, which is confirmed by Levene's test yielding a high F value, so that the null hypothesis of equal variances needs to be rejected. Because of that, Welch's variance-weighted ANOVA is used to assess the variance of the means. The F value is greater than 1, so that the null hypothesis of equally distributed means is rejected. The difference is significant with $\alpha = 5\%$: $TCCR_{model}$ does not have equal means for the methods used. From the box plot, it can be seen that the means of logistic regression and nearest neighbor hot deck are equal. The performance of these two methods is explored in more detail in chapter 6.4.

6.3.2 Correct classification rate of the target value

The precision is a measure of how well a single desired target value is augmented. The desired value is referred to as positive predictive value, because it is derived from medical tests and the proportion of positively tested patients being correctly diagnosed (Altmann & Bland, 1994). This is especially important for selection tasks, where one specific output value is of interest. It differs from the previous measure in that it only regards the proportion of the correctly augmented target values to the true elements carrying that value. The following calculation is adapted from Hattum and Hoijtink (2008b) so as to fit our notation concept.

Calculation Let y_{target} be a target value relevant to a database marketing problem. The number of recipients that have the target value are denoted as $r(y_{target})$. The correct classification rate of the target value (CCR) is calculated by the number of hits regarding this value y_{target}, divided by all recipients that have the positive value $r(y_{target})$.

$$CCR_{target} = \frac{\sum_{i=1}^{r(y_{target})} h_i}{r(y_{target})} \;\forall\; h_i = \begin{cases} 1 & \text{if } y_{target} = \hat{y} \\ 0 & \text{if } y_{target} \neq \hat{y} \end{cases} \tag{6.2}$$

CCR_{target} ranges between 0 and 1. While $TCCR_{model} = 100\%$ is a perfect augmentation, $CCR_{target} = 100\%$ does not necessarily lead to a good selection criterion. It provides no information on the false positives. In a data augmentation where no differentiation is possible, CCR_{target} is 100%, if the desired target value is the most frequent value, and 0%, if it is a less frequent value. No selection is possible, because not only the recipient with the desired target value are augmented this value, but all others, too.

Example y_4 is the desired value of Target_1 in the previous example stated in table 6.3 on page 186. The number of true positives is 2,486, whereas 4,093 recipients actually have that value. The precision in this case is $CCR_{target}(y_4) = \frac{2{,}486}{4{,}093} = 0.61$. 61% of the target values are hit by the model. $TCCR_{model}$ and CCR_{target} are not dependent. Especially, if the desired value occurs seldom in a source, CCR_{target} can be low, although $TCCR_{model}$ is high.

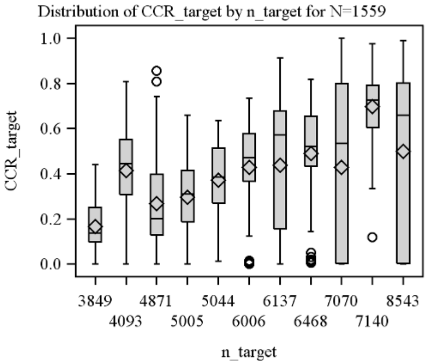
Method	Conditional Mode	Logistic Regression	Nearest Neighbor
$CCR_{target}(y_4)$	0.53	0.73	0.61

Table 6.5: Exemplary $CCR_{target}(y_4)$ values in the case of Target_1 and Donors_c_s6

A within case analysis is possible, when comparing the CCR_{target} values for the same source-target combination, but different augmentation methods. In our example from chapter 6.1, nearest neighbor hot deck reached a correct classification rate of 61% for y_4. However, logistic regression performed even better and lead to $CCR_{target}(y_4) = 73\%$. It would be the best choice in this case. Conditional mode imputation resulted in a 53% correct classification rate for y_4. In this manner, we have reported the CCR_{target} measures for every case, with a value saved for each method used.

Analysis Figure 6.5 shows the distribution of CCR_{target} among the augmentations for different levels of n_{target}. In order to calculate CCR_{target}, one of the values of each target variable has been assigned to be of particular interest. In figure 5.6 on page 161 in chapter 5.2.2, the respective value is shaded blue. It is the highest target value each, e.g. the highest income class for all levels of the household income. Only this value is regarded when calculating CCR_{target}.

Pertaining to CCR_{target}, neither k nor IQV are relevant parameters to be analyzed. k is not of interest, because only one target value is regarded. In this case, the target value is virtually reduced to two values: the target value versus all others. The same is true for IQV, because IQV relates to the distribution of all observations among the target values. Because of that, n_{target} is regarded. It relates to the number of recipients in the customer database actually having the target value sought. CCR_{target} is expected to increase with increasing n_{target}, because the chance of hitting the target value is higher if more recipients have this value.



Box plot

n_target	3,849	6,006	8,543
N	129	142	147
Max	0.44	0.74	0.99
Q3	0.25	0.58	0.80
Median	0.14	0.47	0.66
Mean	0.17	0.43	0.50
Q1	0.10	0.37	0.00
Min	0.00	0.00	0.00
Sh.-W.	0.94	0.89	0.81
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	46.76	<.001
Levene	93.30	<.001
Welch's ANOVA	128.46	<.001

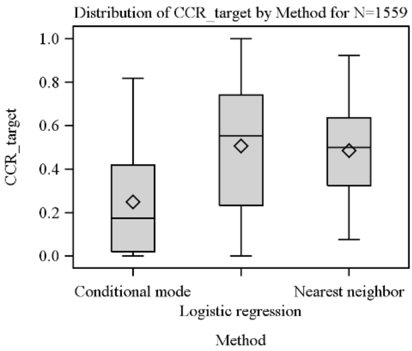
Measures

Figure 6.5: Distribution of all CCR_{target} values by number of observations with the target value n_{target}

The assumption is confirmed when regarding the means in the boxplot in figure 6.5[6]. The means roughly follow a positive line. This is confirmed by ANOVA. A calculation of ANOVA is possible, because the residuals are normally distributed as confirmed by the results of the Shapiro-Wilk test in figure 6.5b. The variances are not equal, so that Levene's test yields a significant F value. The F value of Welch's ANOVA is above 1 and hence significant. The variation can be high, because it depends on the respective target variable used. The box of the box plot occupies almost the whole range of possible CCR_{target} values for $n_{target} = 7,070$ and $n_{target} = 8,543$. The mean CCR_{target} values for $n_{target} = 4,093$ and $n_{target} = 7,140$ are noticeably higher than the general trend. It becomes clear that other factors also influence CCR_{target}, e.g. the predictive power of the link variables.

The results for CCR_{target} by augmentation method for all augmenta-

[6] Table 6.5b only contains selected measures for illustration purposes. The full table can be found in table 9.2 on page 304 in the appendix.



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	0.82	1.00	0.92
Q3	0.42	0.74	0.64
Median	0.17	0.55	0.50
Mean	0.25	0.51	0.49
Q1	0.02	0.23	0.32
Min	0.00	0.00	0.08
Sh.-W.	0.89	0.95	0.98
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	176.49	<.001
Levene	67.11	<.001
Welch's ANOVA	189.08	<.001

Measures

Figure 6.6: Distribution of all CCR_{target} values by applied augmentation method

tions are shown in figure 6.6 using box plots. Again, conditional mode imputation is inferior to logistic regression and nearest neighbor hot deck. CCR_{target} is 25% on average for conditional mode imputation, versus 51% for logistic regression and 49% for nearest neighbor hot deck. This time, the variation for logistic regression is greater than for nearest neighbor hot deck. The main part of logistic regression results varies between 23% and 74%. For nearest neighbor hot deck, it varies only between 32% and 64%. This is confirmed by ANOVA. It can be conducted, because the residuals of all classes are normally distributed (all values of the Shapiro-Wilk statistic are significant), but Welch's ANOVA has to be used, because the variances are not equally distributed ($F = 67.11$ for Levene's test). From the p-value of Welch's ANOVA, it can be seen that there is a significant difference between the means of the methods. While conditional mode imputation does not seem to be a relevant method, the decision between logistic regression and nearest neighbor hot deck is not as obvious. More influencing parameters are examined in chapter 7.2 to give a substantiated recommendation.

6.4 Model lifts assessing the results

The model lift (ML) is a measure of how much the model increased the data augmentation results, when compared to results that could have been achieved without using augmentation results. To that end, $TCCR_{model}$ is compared to a total correct classification rate that could be achieved if the data augmentation results were not known. By means of $TCCR$ and ML, data augmentation results can be evaluated internally. Different conditions for data augmentation can be compared using appropriate tests.

In our study, we assess and compare four different model lift measures. They have different use cases and relate to different states of information. For example, the state of information can differ in terms of the true target variable distribution in the recipient unit. If the overall distribution was known, the state of information is better than if it was not known. This is explained in more detail in this chapter. Furthermore, like for the accuracy measures, different measures are available depending on whether segmentation or selection is the database marketing goal for the augmentation. In the following, each of the four measures is explained and its respective use case is explicated. The chapter is finished with a comparison of the four model lift measures and their respective advantages and disadvantages.

Of the four measures, only the first is derived from Ratner's (2003) and Hattum and Hoijtink's (2008b) literature. We extend the idea of the model lift to three other measures, which respect other use cases and states of information each. Ratner's (2003) and Hattum and Hoijtink's (2008b) model lift is used for the internal evaluation of segmentations. We examine their model lift measure more thoroughly and add some general rules and ranges to the discussion. Their model lift measure implicitly requires knowledge on the overall distribution of the target variable in the recipient unit. We introduce their model lift measure first, before extending it to other use cases, where other intentions and requirements apply.

6.4.1 Model lift (chance)

In order to calculate a model lift, i.e. a measure of how much the augmentation results improved the knowledge on the recipients, the total correct classification rate of the model is compared to a random allocation of target values. One way of calculating a total correct classification rate that could be achieved if the data augmentation results were not known is to establish a total correct classification rate that is achieved when randomly distributing target values among the recipients (Hattum & Hoijtink, 2008b).

Calculation Such a total correct classification rate $TCCR_{chance}$ is calculated by the sum of squared expected percentages of each target value, given the number of recipients $r(y_i)$ that have the respective target value (Ratner, 2003, p. 182).

$$TCCR_{chance} = \sum_{i=1}^{k} \left(\frac{r(y_i)}{r} \right)^2 \tag{6.3}$$

The model lift ML_{chance} is calculated as the total correct classification rate of the model, divided by the total correct classification rate that would have been achieved by chance (Ratner, 2003, p. 221f). It is an index showing how much the model increased (or decreased) $TCCR_{model}$, compared to $TCCR_{chance}$.

$$ML_{chance} = \frac{TCCR_{model}}{TCCR_{chance}} \tag{6.4}$$

ML_{chance} cannot be used without limitations, if different target variables and their augmentation results are compared. ML_{chance} has different ranges depending on the skewness and the number of target values in the domain. Both $TCCR_{chance}$ and $TCCR_{model}$ values are dependent on this distribution. Target variables with a skew distribution generally have higher $TCCR_{chance}$ values than those with symmetric distributions, because the target value which occurs most often is hit more easily. Equally, target variables with a low number of values generally have a higher $TCCR_{model}$ than those with a high number of target values, because there is not so much variability among the target values, so that values are hit more easily. Neither Ratner (2003) nor Hattum and Hoijtink (2008b) described these properties in more detail. However, if different model lift measures shall be compared, like in our case study, it is important to respect these ranges.

The range of the total correct classification rate achievable by chance is $0 < TCCR_{chance} < 1$. In contrast to $TCCR_{model}$, it can never be equal 0 or 1. Every target value needs to appear at least once in the donor unit. No general advice can be given on the maximum applicable skewness of a distribution. It depends on the size of the donor unit, the size of the recipient unit, and the information value of the target variable to a specific database marketing problem. We therefore only require a target value to occur at least once. In theory, the highest possible $TCCR_{chance}$ is given by the following equation, where $r(y_1) = r(y_2) = ... = r(y_{k-1}) = 1$ and $r(y_k) = r - k - 1$.

$$max(TCCR_{chance}) = (k-1) \times \left(\frac{1}{r}\right)^2 + \left(\frac{r-k-1}{r}\right)^2 \tag{6.5}$$

In the skewest possible distribution, every but one target value has only one element in the recipient unit and the other target value has all other elements. In this case, $max(TCCR_{chance})$ converges to 1 with increasing k and r. $max(TCCR_{chance})$ does not vary much, as even in the smallest case $k = 2$ and with a low number of recipient (e.g. $r = 100$), the maximum total classification rate achieved by chance would be 0.98. In contrast to $max(TCCR_{chance})$, $min(TCCR_{chance})$ depends on k. The lowest $TCCR_{chance}$ value possible is that of a symmetrically distributed target variable, where $r(y_1) = r(y_2) = ... = r(y_k) = \frac{r}{k}$.

$$min(TCCR_{chance}) = k \times \left(\frac{\frac{r}{k}}{r}\right)^2 = \frac{1}{k} \tag{6.6}$$

The minimal total correct classification rate that can be achieved by chance is highest, if $k = 2$. With raising k, it converges to 0. Accordingly, $max(TCCR_{chance})$ is different for differing target variables. This limits the possibility to compare the model lift of augmentations with differing target variables. While the smallest possible model lift $min(ML_{chance})$ is always 0 (if $TCCR_{model} = 0$), the highest possible model lift $max(ML_{chance})$ is defined by $TCCR_{model} = 1$ and $min(TCCR_{chance})$.

For example, if a recipient unit had 10,000 recipients and a maximally skew binary target variable with $r(y = 1) = 9,999$ and $r(y = 0) = 1$, the highest possible value of $TCCR_{chance}$ would be 0.9998. If the augmentation was perfect with $TCCR_{model} = 1$, the highest possible value of ML_{chance} would be 1.0002. Thus, even for a perfect augmentation, only a minimal model lift would be possible. Contrarily, if a binary variable was equally distributed with $r(y = 1) = 5,000$ and $r(y = 0) = 5,000$, the highest possible value of ML_{chance} would be 2, because $min(TCCR_{chance})$ would equal 0.5. In that case, the data augmentation model provides 100% more correct hits for all target values than would have been achieved by chance. This is even more obvious for a target variable with four target values, like in our example. For four target values, the maximal model lift varies between $1.0004 \leq max(ML_{chance}) \leq 4$. A model lift of 1.53 for a binary target variable is therefore better than the same model lift for a target variable with four target values. Table 6.6 shows the different ranges for $min(TCCR_{chance})$, $max(TCCR_{chance})$, and resulting $max(ML_{chance})$ values for different target values.

Accordingly, ML_{chance} values cannot be compared for different target variables. Only if ML_{chance} is regarded pertaining to the variance and the number of target values, it is possible to compare different target variables. While the boundaries of the model lift are defined by the properties of the target variable, the value itself is also influenced by the link variables. The number of link variables and the explanatory power of these link variables regarding the target variable explain much of the variability of the model

	r=10,000			
k	$min(TCCR_{chance})$	$max(TCCR_{chance})$	$min(ML_{chance})$	$max(ML_{chance})$
2	0.500	0.9998	0	2.00
3	0.333	0.9996	0	3.00
4	0.250	0.9994	0	4.00
5	0.200	0.9992	0	5.00
6	0.167	0.9990	0	6.00
7	0.143	0.9988	0	7.00
8	0.125	0.9986	0	8.00

Table 6.6: Examples for upper and lower boundaries of total correct classification rate and model lift for random allocation of values given different numbers of target values

lift. The quality of any data augmentation project highly depends on these properties (Kamakura & Wedel, 1997). When building a model for varying target variables, the number of target values k, the index of qualitative variation IQV, and a measure for the predictive power of the link variables need to be incorporated as moderators into the model.

The model lift relativizes the total correct classification rate by dividing it by $TCCR_{chance}$, the total correct classification rate achieved by a random allocation of target values. By ML_{chance}, it is possible to compare different data augmentations using the same target variable. It is an artificial measure, because in a practical application, the true distribution of target values in the recipient unit is not known. Because it is not known, $TCCR_{chance}$ cannot be achieved in a practical application. Thus, it would be wrong to state that the knowledge has increased by ML_{chance} after having used data augmentation.

Example In the example shown in table 6.3 on page 186 in chapter 6.3.1, the total correct classification rate achieved by a random allocation of target values is calculated as follows.

$$TCCR_{chance} = (\frac{1,851}{11,560})^2 + (\frac{2,043}{11,560})^2 + (\frac{3,573}{11,560})^2 + (\frac{4,093}{11,560})^2 = 0.28 \quad (6.7)$$

It means: If all values were allocated randomly, knowing that the distribution of Target_1 is $(r(y_1) = 1,851, r(y_2) = 2,043, r(y_3) = 3,573, r(y_4) = 4,093)$ in the recipient unit, 28% of the values would also be hit by chance. This comparison measure requires an upfront knowledge on the overall distribution of the target variable in the recipient unit.

In the same example, $ML_{chance} = \frac{TCCR_{model}}{TCCR_{chance}} = \frac{0.43}{0.28} = 1.53$. The data augmentation model provides 53% more correct hits for all target values than would have been hit by chance. The effects of variance and number of target values of the target variables are balanced out, because not only the dividend increases with raising variance and number of target values, but also the divisor. Thus, the ML_{chance} measure is a more comparable measure than $TCCR_{model}$. ML_{chance} is a good measure for comparing results between cases, e.g. for different sources.

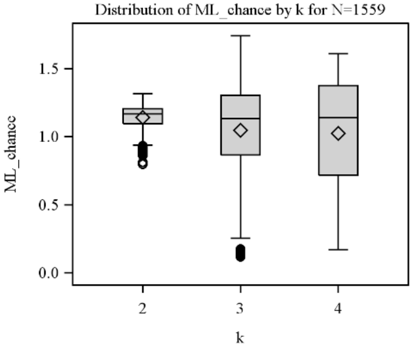
Method	Conditional Mode	Logistic Regression	Nearest Neighbor
ML_{chance}	1.23	1.40	1.53

Table 6.7: Exemplary ML_{chance} values in the case of Target_1 and Donors_c_s6

When augmenting values for Target_1 using Donors_c_s6 as a source in our example from chapter 6.1, conditional mode imputation leads to a model lift of 1.23, logistic regression to a model lift of 1.41, and nearest neighbor hot deck to a model lift of 1.53. In this case, nearest neighbor hot deck would be the best choice for a data augmentation method. In this manner, we have reported the ML_{chance} measures for every case, with a value saved for each method used.

Analysis The results of ML_{chance} observed in our case study differentiated by k are shown in figure 6.7. The variation increases with increasing k.

50% of the augmentations vary between 1.10 and 1.20 for $k = 2$, as opposed to 0.87 and 1.30 for $k = 3$, and 0.71 and 1.37 for $k = 4$. Levene's test for equal variances yields a significant F value, indicating heteroscedasticity. It can be expected that target variables with more target values would have an even bigger variance, with a core range (Q1) starting well below and averaging (mean) below $ML_{chance} = 1$.



k	2	3	4
N	829	437	293
Max	1.31	1.74	1.61
Q3	1.20	1.30	1.37
Median	1.17	1.13	1.14
Mean	1.14	1.05	1.02
Q1	1.10	0.87	0.72
Min	0.80	0.12	0.17
Sh.-W.	0.92	0.92	0.91
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	28.11	<.001
Levene	195.63	<.001
Welch's ANOVA	24.63	<.001

Box plot

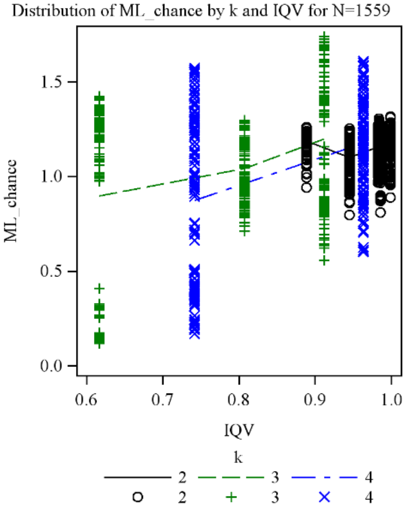
Measures

Figure 6.7: Distribution of all ML_{chance} values by number of target values k

The means decrease from 1.14 for $k = 2$ to 1.05 for $k = 3$ and 1.02 for $k = 4$. These differences are significant, as confirmed by Welch's ANOVA. It is used, because the variances are not equal. The normality prerequisite is met, because the Shapiro-Wilk statistics are significant for all levels of k. This is similar to the distribution of $TCCR_{model}$ by k as shown in figure 6.2 on page 188 in chapter 6.3.1. However, the differences are not as strong. Because $TCCR_{model}$ and $TCCR_{chance}$ are positively correlated, the effect of $TCCR_{model}$ is relativized by $TCCR_{chance}$.

In terms of IQV, $TCCR_{model}$ and ML_{chance} differ, as shown in figure 6.8[7]. While the $TCCR_{model}$ values were higher for lower values of IQV, this is no longer true for ML_{chance}. The target values are hit more easily for

[7] The colored version of this figure can be found online on www.springer.com under the title of this publication.



Scatter plot

IQV	N	Mean	Sh.-W.	p-value
k=2				
0.89	129	1.19	0.82	<.001
0.94	146	1.10	0.93	<.001
0.98	137	1.16	0.96	<.001
0.98	137	1.17	0.94	<.001
0.99	138	1.09	0.90	<.001
1.00	142	1.14	0.91	<.001
k=3				
0.62	147	0.90	0.77	<.001
0.81	143	1.04	0.94	<.001
0.91	147	1.20	0.91	<.001
k=4				
0.74	147	0.87	0.89	<.001
0.96	146	1.18	0.97	<.001

k	Test	F	p
2	ANOVA	26.93	<.001
	Levene	15.74	<.001
	Welch's A.	28.90	<.001
3	ANOVA	26.04	<.001
	Levene	133.38	<.001
	Welch's A.	21.56	<.001
4	ANOVA	49.18	<.001
	Levene	151.65	<.001
	Welch's A.	49.38	<.001

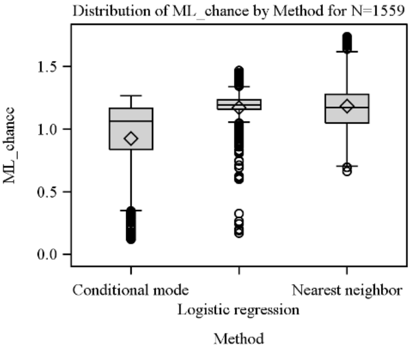
Measures

Figure 6.8: Distribution of all ML_{chance} values by number of target values k and index of qualitative variance IQV

skew variables in a data augmentation, but the same applies to the random distribution to which it is compared for the model lift. After having been corrected for $TCCR_{chance}$, the results are better for high IQV values. For $k = 3$ and $k = 4$, the means rise for increasing IQV. This is confirmed by Welch's ANOVA, which is used because the variances are not equal (as measure by Levene's test). The means of ML_{chance} given IQV significantly differ for $k = 2$, too. However, the direction is not as clear. This could be due to the low range of IQV for $k = 2$ (all observations have high IQV values). It could also be that the association of IQV and ML_{chance} is more complex. This is further explored in chapter 7.2.

In our case study, the average model lift for $IQV < 0.8$ is below 1. This confirms the earlier finding. The characteristics of the target variable give insight into possible exit criteria. Suitable target variables can be used for data augmentation purposes, while one should refrain from data augmentation, if these criteria are not met.

> *Finding:* Only evenly distributed target variables are suitable for data augmentation, in a way that every value occurs at least for 20% of the donors. Therefore, only target variables with at most 5 target values can be taken into consideration. The IQV value should be at least 0.8.



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	1.27	1.47	1.74
Q3	1.17	1.23	1.28
Median	1.07	1.19	1.17
Mean	0.93	1.17	1.18
Q1	0.84	1.16	1.05
Min	0.12	0.17	0.66
Sh.-W.	0.81	0.74	0.98
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	177.57	<.001
Levene	52.75	<.001
Welch's ANOVA	132.15	<.001

Measures

Figure 6.9: Distribution of all ML_{chance} values by applied augmentation method

If the available target variables are suitable, the main question for the database marketing analyst is which method should be used for data augmentation. Figure 6.9 shows the differences in ML_{chance} pertaining to the applied augmentation method. Again, conditional mode imputation performs worse than the two other methods. It has a mean of 0.93 and varies mainly between 0.84 and 1.17. In contrast, logistic regression and nearest neighbor hot deck have means of 1.17 and 1.18, respectively. Welch's ANOVA confirms that the means of all three methods are not equal (it is used, because the F value of Levene's test is significant, indicating unequal variances). However, the difference between the mean of logistic regression and nearest neighbor hot deck is not significant[8]. It can thus be said that logistic regression and nearest neighbor hot deck results are significantly higher than those of conditional mode imputation. The results of conditional mode imputation are even below $ML_{chance} = 1$ on average, so that they are not better than those of a random allocation of target values. Performing data augmentation with conditional mode imputation is not effective.

> *Finding:* Nearest neighbor hot deck and logistic regression are superior to conditional mode imputation.

There are also augmentations using logistic regression or nearest neighbor hot deck which do not satisfy the $ML_{chance} = 1$ precondition. However, most of the observations can be found between 1.16 and 1.23 for logistic regression and between 1.05 and 1.28 for nearest neighbor hot deck. Nearest neighbor hot deck even reaches ML_{chance} values of 1.74, which means that true values are hit 74% more likely than when having no other knowledge than the overall distribution.

> *Finding:* Nearest neighbor hot deck can lead to higher model lifts than logistic regression. However, the variance of logistic regression results is lower.

Although this is a central finding, it is not yet helpful. Database marketing analysts would not know which method to use, i.e. how to achieve the best augmentation results. Nearest neighbor hot deck is more effective, because results are slightly better on average and because there are less negative outliers. But logistic regression results are less variable, so that

[8] The parameter estimates of conditional mode imputation and logistic regression are contrasted to nearest neighbor hot deck by solving the normal equations of the ANOVA model (SAS Institute Inc., 2014d). The table of parameter estimates for this ANOVA model solution can be found in table 9.3 on page 304 in the appendix.

this certainty could be preferable in a marketing context. Some more information is needed in order to make the decision. This is regarded in detail in chapter 7.2.

6.4.2 Model lift (target)

If only a certain target variable parameter y_{target} is of interest, a model lift ML_{target} is calculable by comparing the correct classification rate of the target value CCR_{model} to the correct classification rate that can be achieved for this value by chance CCR_{chance}.

Calculation CCR_{chance} is calculated by the squared expected percentages of the target value, given the number of recipients $r(y_{target})$ that have the respective target value. The model lift ML_{target} is calculated as the correct classification rate of the target value, divided by the correct classification rate that would have been achieved by chance. It is an index showing how much the model increased (or decreased) CCR_{model}, compared to CCR_{target}.

$$ML_{target} = \frac{CCR_{model}}{CCR_{chance}} = \frac{\frac{h_i}{r}}{\left(\frac{r(y_{target})}{r}\right)^2} \tag{6.8}$$

As for ML_{chance}, the range of ML_{target} differs depending on the number of target values k and the distribution of the target variable as measured by IQV. For an evenly distributed target variable with $k = 2$, the maximum ML_{target} is 4, because the minimum $CCR_{target} = \left(\frac{1}{2}\right)^2 = 0.25$. In contrast to ML_{chance}, however, the minimum CCR_{target} is not necessarily smallest for an evenly distributed variable. Because only the positive part of the target values is of interest, the minimum CCR_{target} converges to 0 with increasing IQV, if y_{target} is a value occurring less often than other values. It converges to 1 with increasing IQV, if y_{target} is a value occurring more often than other values. ML_{target} theoretically varies between 0 and infinity.

It is, like ML_{chance}, not comparable among target variables with different target values and different distributions.

ML_{target} regards the model lift regarding a specific target value. If the publisher from our previous example would like to contact customers interested in shoes, he does not care about how many customers not interested in shoes were hit correctly. His only measure is how many positive values were hit. If this number is significantly higher than what could have been achieved by chance, the augmentation is successful. If segmentation is the database marketing goal, ML_{chance} is a meaningful measure. If selection is the database marketing goal, ML_{target} is more meaningful. It isolates the model lift of the relevant value without taking into account hit rates of other values. ML_{target} can be high, if ML_{chance} is low, and vice versa.

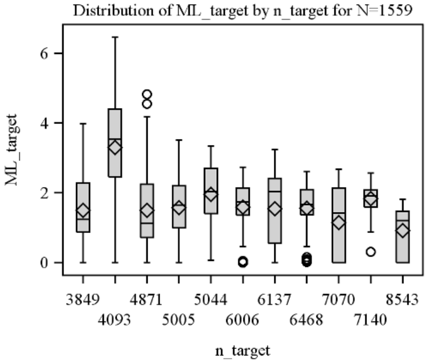
Example In the example shown in table 6.3 on page 186, $CCR_{chance} = \left(\frac{4{,}093}{11{,}560}\right)^2 = 0.13$ for y_4. If all values are allocated randomly, knowing that 4,093 of 11,560 recipients have value y_4, 13% of the values are also hit by chance. This comparison measure requires an upfront knowledge on the overall distribution of the target variable in the recipient unit. Consequently, $ML_{target} = \frac{CCR_{model}}{CCR_{chance}} = \frac{0.61}{0.13} = 4.85$. The data augmentation model provides 384% more correct hits for the desired target value than would have been hit by chance.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
ML_{target}	4.22	5.85	4.85

Table 6.8: Exemplary ML_{target} values in the case of Target_1 and Donors_c_s6

In our example from chapter 6.1, logistic regression performed best and reached a ML_{target} value of 5.85. Nearest neighbor hot deck and conditional mode imputation yielded ML_{target} values of 4.85 and 4.22, respectively.

Analysis The distribution of all CCR_{target} values by number of target values n_{target} can be observed in figure 6.10[9]. In order to calculate ML_{target}, one of the values of each target variable has been assigned to be of particular interest, as described for the CCR_{target} in chapter 6.3.2. Only this value is regarded when calculating CCR_{chance} and CCR_{target}.



Box plot

k	3,849	6,006	8,543
n_target	3,849	6,006	8,543
N	129	142	147
Max	3.97	2.73	1.81
Q3	2.27	2.14	1.47
Median	1.24	1.74	1.21
Mean	1.50	1.59	0.91
Q1	0.88	1.36	0.00
Min	0.01	0.00	0.00
Sh.-W.	0.94	0.89	0.81
p-value	<.001	<.001	<.001

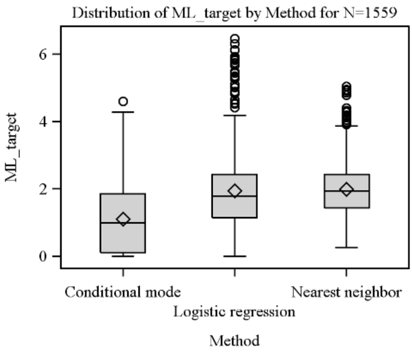
Test	F	p
ANOVA	58.78	<.001
Levene	44.82	<.001
Welch's ANOVA	38.70	<.001

Measures

Figure 6.10: Distribution of all ML_{target} values by number of observations with the target value n_{target}

Like ML_{chance}, ML_{target} is relativized in comparison to CCR_{target}, because target values with a high n_{target} values are hit more easily in the augmentation, but also more easily in a random allocation as represented by CCR_{chance}. The differences of the means are confirmed by Welch's ANOVA, which is used due to the unequal variances. However, some means depart more noticeably from the overall mean of $ML_{target} = 1.67$ than others. It becomes obvious that those having already stood out in the boxplot of CCR_{target} (figure 6.5 on page 195 in chapter 6.5) even more clearly stand out here. The mean ML_{target} values for $n_{target} = 4,093$ and $n_{target} = 7,140$ are well above average. This confirms our earlier finding that other factors also influence ML_{target}. n_{target} only seems to be a weak predictor.

[9] Table 6.10b only contains selected measures for illustration purposes. The full table can be found in table 9.4 on page 305 in the appendix.



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	4.59	6.45	5.03
Q3	1.84	2.43	2.42
Median	1.00	1.78	1.94
Mean	1.11	1.94	1.99
Q1	0.11	1.15	1.44
Min	0.00	0.00	0.27
Sh.-W.	0.91	0.91	0.96
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	118.63	<.001
Levene	19.77	<.001
Welch's ANOVA	131.35	<.001

Measures

Figure 6.11: Distribution of all ML_{target} values by applied augmentation method

Figure 6.11 shows the box plots for the three augmentation methods used measured by ML_{target} for all augmentation results. Again, the inferiority of conditional mode imputation can be observed. Results can be found on the whole range between 0 and 4, while only outliers exceed 4. Logistic regression and nearest neighbor hot deck reach ML_{target} values up to 6.5 and 5, respectively. The mean difference is also shown by the high F value of Welch's ANOVA, which is used because Levene's homoscedasticity test yields a significant F value. The upper border of the box as well as the whisker indicating the maximum value observed for logistic regression slightly exceed the nearest neighbor hot deck results in terms of ML_{target}, but not significantly[10]. Like for ML_{chance}, logistic regression and nearest neighbor hot deck are superior to conditional mode imputation when regarding the distribution of ML_{target} by method.

[10] The table of parameter estimates for this ANOVA model solution can be found in table 9.5 on page 305 in the appendix.

6.4.3 Model lift (uniform)

If all target values are equal, it is referred to as a uniform distribution. When comparing $TCCR_{model}$ to a distribution where all recipients are assigned the same, most frequent value, another model lift can be calculated. To assign the same value to all recipients does not enable segmentation or selection based on the augmented variable. All the more, the comparison to such a uniform distribution is reasonable. Albeit being undesirable, it can happen in practice that the link variables are too weak to discriminate between the target values, so that the same (usually the most frequent) value is assigned to all recipients. In this case, ML_{chance} would result in a value above 1, diluting the database marketing analyst into having successfully augmented the target variable. Such a prediction error is easily detectable after the augmentation, because the resulting distribution of target values is observable. However, in the case study context, it makes sense to introduce and regard this model lift measure for quality check purposes.

Calculation The total correct classification rate $TCCR_{uniform}$ that would be achieved if all recipients were augmented the most frequent value in the recipient unit y_{max}, $max(r(y_1), r(y_2), ..., r(y_k))$ is calculated by the number of correctly classified recipients divided by the number of all recipients:

$$TCCR_{uniform} = \frac{\sum_{i=1}^{r} h_i}{r} \; \forall \; h_i = \begin{cases} 1 & \text{if } y = y_{max} \\ 0 & \text{if } y \neq y_{max} \end{cases} \tag{6.9}$$

$TCCR_{uniform}$ can never equal one, because only target variables with at least two target values are taken into consideration for data augmentation purposes.

The respective model lift $ML_{uniform}$ is calculated as the total correct classification rate of the model, divided by the total correct classification rate that would be achieved if all recipients were augmented the same, most frequent, value. It is an index showing how much the model increased (or decreased) $TCCR_{model}$, compared to $TCCR_{uniform}$.

$$ML_{uniform} = \frac{TCCR_{model}}{TCCR_{uniform}} \tag{6.10}$$

Inherently, $ML_{uniform}$ is more strict than ML_{chance}. While $TCCR_{model}$ does not change, $TCCR_{uniform} > TCCR_{chance}$. As for $TCCR_{chance}$, $TCCR_{uniform}$ is influenced by the target variable characteristics. The skewest possible distribution is that in which every but one target value has only one element and the other target value has all other elements. The highest possible $TCCR_{uniform}$ converges to 1 with increasing r and a limited k. It is given by $max(TCCR_{uniform})$, with $r(y_1) = r(y_2) = ... = r(y_{k-1}) = 1$ and $r(y_k) = r - k$.

$$max(TCCR_{uniform}) = \frac{r-k-1}{r} \tag{6.11}$$

$min(TCCR_{uniform})$ depends on k. The lowest possible $TCCR_{uniform}$ is that of a perfectly equally distributed target variable with $r(y_1) = r(y_2) = ... = r(y_k) = \frac{r}{k}$.

$$min(TCCR_{uniform}) = \frac{1}{k} \tag{6.12}$$

The range of possible $TCCR_{uniform}$ values is similar to the range of possible $TCCR_{chance}$ values. It also depends on the number of target values of the target variable k and the index of qualitative variation IQV. Hence, when building a model on $ML_{uniform}$ for varying target variables, these moderators need to be incorporated into the model.

Example The number of correctly classified recipients can be obtained by creating a two way frequency table, only this time all but one column contain zeros only, as shown in table 6.9. In the example, $TCCR_{uniform} = \frac{0+0+0+4{,}093}{11{,}560} = 0.35$.

	True y_1	True y_2	True y_3	True y_4	Row sum
Classified $\hat{y_1}$	0	0	0	0	0
Classified $\hat{y_2}$	0	0	0	0	0
Classified $\hat{y_3}$	0	0	0	0	0
Classified $\hat{y_4}$	1,851	2,043	3,573	4,093	11,560
Column sum	1,851	2,043	3,573	4,093	11,560

Table 6.9: Two-way frequency table for a uniformly distributed target variable based on Target_1

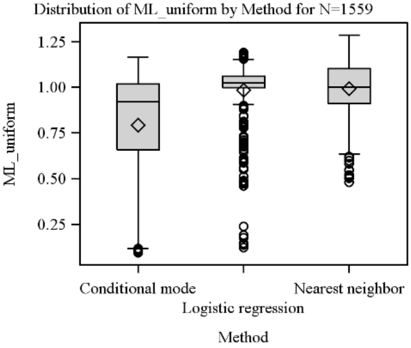
In the example shown in table 6.3 on page 186, $ML_{uniform} = \frac{0.43}{0.35} = 1.20$ for the nearest neighbor hot deck method. The data augmentation model provides 20% more correct hits for all target values than if all recipients were augmented the same most frequent value.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
$ML_{uniform}$	0.97	1.11	1.20

Table 6.10: Exemplary $ML_{uniform}$ values in the case of Target_1 and Donors_c_s6

With $ML_{uniform} = 1$, it is very likely that the method has not been able to discriminate between the target values at all. Cases where no discrimination was possible have already been removed during the quality check described in chapter 6.2. Nevertheless, all augmentations should at least result in $ML_{uniform} = 1$ in order to have a significant informative value for the customer database. The $ML_{uniform}$ reached for our example from chapter 6.1 as shown in table 6.10 are slightly lower than the ML_{chance} values, with $ML_{uniform} = 0.97$ for conditional mode imputation, $ML_{uniform} = 1.11$ for logistic regression, and $ML_{uniform} = 1.20$ for nearest neighbor hot deck. In this case, conditional mode imputation does not reach the minimum criterion for a good augmentation of $ML_{uniform} = 1$. However, it is not equal to 1 either, which would be an indicator for a uniform distribution.

Figure 6.12 shows the distribution of $ML_{uniform}$ by augmentation method for all augmentations. For $ML_{uniform}$, the mean of conditional mode imputation is 0.79, 0.98 for logistic regression, and 0.99 for nearest



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	1.17	1.19	1.29
Q3	1.02	1.06	1.10
Median	0.92	1.02	1.00
Mean	0.79	0.98	0.99
Q1	0.66	1.00	0.91
Min	0.09	0.12	0.48
Sh.-W.	0.85	0.68	0.97
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	134.90	<.001
Levene	92.55	<.001
Welch's ANOVA	94.79	<.001

Measures

Figure 6.12: Distribution of all $ML_{uniform}$ values by applied augmentation method

neighbor hot deck. Again, the comparison of means shows that the means of the methods are significantly different. However, when regarding the parameter estimates of the ANOVA model, only conditional mode imputation differs significantly from nearest neighbor hot deck, logistic regression does not[11]. On average, data augmentation results are therefore not better in terms of hits than if all customers were assigned the same, true but unknown most frequent value. However, data augmentation results enable data selection, as long as at least two target values have been augmented. A uniform distribution does not enable data selection, because all recipients receive the same value. The uniform distribution is therefore not an alternative to which the augmentation results are compared. It serves as a quality check criterion. We come back to the average $ML_{uniform}$ values in chapter 7.3.3, where all KPIs are compared for the sources and methods generally suitable for data augmentation as derived from our criteria yet to establish in chapter 7.

[11] The table of parameter estimates for this ANOVA model solution can be found in table 9.6 on page 305 in the appendix.

6.4.4 Model lift (source)

All previous measures assume that knowledge is available on the true, but unobservable distribution of the target variable in the recipient unit. ML_{chance} compares the total correct classification rate to a random distribution, allocated based on the *true* distribution of target values. $ML_{uniform}$ compares the total correct classification rate to a uniform distribution, which assigns all recipients the *true* most frequent target value. For MCAR sources, this is sufficient in order to answer the question how much the knowledge on individual customers increased after an augmentation, because the target value distribution is equal in donor and recipient unit. However, we explicitly address MAR sources, for which this condition is not given. Therefore, we developed an additional model lift measure which respects this difference.

If the distributions of the target value in donor and recipient unit differ, the most meaningful measure is to compare the augmentation results to a random distribution based on the target variable distribution of the source. The distribution of the target variable in the source would be the reference, if data augmentation was not performed, because it shows the state of knowledge *before* the augmentation. This distribution will be close to the distribution within the customers, if the overlap is high, but can differ significantly, if the overlap is low.

Calculation The comparison measure for $TCCR_{model}$ in this context is derived from the distribution of the target variable in the donor unit. If this distribution is applied to the recipient unit, $TCCR_{source}$ can be calculated.

Figure 6.13 shows an exemplary schematic illustration of a target variable with $k = 3$, which has a distribution of 1:1:2 in the donor unit, and 1:2:2 in the recipient unit. If the distribution of the donor unit is applied to the recipient unit in order to calculate the total correct classification rate of the source that can be achieved by chance, $TCCR_{source}$, a random hit rate is calculated for the parts that are similar in both units. In the example,

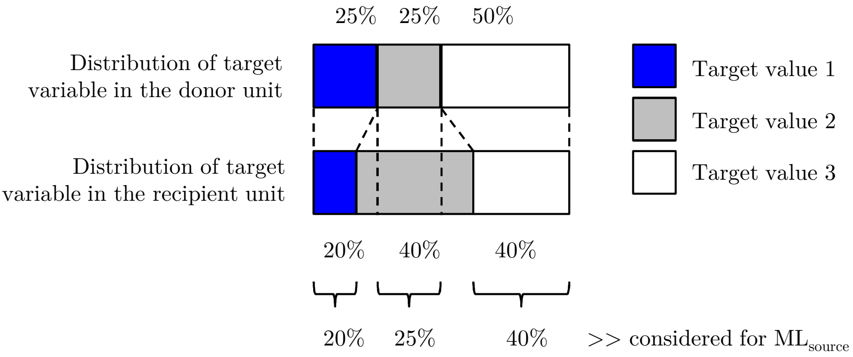

25%
25%
50%
Distribution of target variable in the donor unit
Distribution of target variable in the recipient unit
Target value 1
Target value 2
Target value 3
20%
40%
40%
20%
25%
40%
>> considered for ML_source

Figure 6.13: Exemplary schematic illustration of a ML_{source} calculation

the random hit rate is calculated for 20% of target value 1, 25% of target value 2, and 40% of target value 3. The remaining 15% would definitely not lead to a hit when applied to the recipient unit. As compared to the recipient unit, target value 2 is underrepresented in the donor unit. These 15% cannot be found in any of the numerators in the calculation of $TCCR_{source}$.

$$TCCR_{source} = \sum_{i=1}^{k} \left(min\left(\frac{r(y=i)}{r}, \frac{d(y=i)}{d} \right) \right)^2 \tag{6.13}$$

The according model lift measure is calculated as follows. It is an index showing how much the model increased (or decreased) $TCCR_{model}$, compared to $TCCR_{source}$.

$$ML_{source} = \frac{TCCR_{model}}{TCCR_{source}} \tag{6.14}$$

ML_{source} is equal to ML_{chance}, if donor unit and recipient unit are identical in terms of elements. If the distributions differ, the chance that true values are hit by chance when using a random distribution decreases with decreasing overlap. Accordingly, ML_{source} can reach much higher values than ML_{chance}. The range of ML_{source} is from 0 to infinity, as for ML_{target}. ML_{source} expresses the knowledge profit, if no other information than the source and its aggregated distribution is available. However, the regarded total classification rate does not change. Only the comparison measure changes. Thus, the augmentation does not become better. A high ML_{source} value cannot be interpreted to deliver better augmentation results than that of an augmentation resulting in a lower ML_{source} value. It is possible that the initial state of information before using an unsuitable source is so low, that data augmentation significantly increases this state of information. Possibly, the new state of information is still unsatisfactory, not leading to a CPL. Therefore, ML_{source} is not a valid measure to compare data augmentation results between or across cases. However, *given* a certain source and target variable, ML_{source} indeed gives an indication of how much the state of information improved and can be compared for different applied methods. This is referred to as *within* case analysis. It mirrors the true information increase for a company, given the supposition that the true distribution of target values in the recipient unit is not known.

Target_1	y_1 (%)	y_2 (%)	y_3 (%)	y_4 (%)	Row sum
Recipient unit	1,851	2,043	3,573	4,093	11,560
	(16.0%)	(17.7%)	(30.9%)	(35.4%)	(100%)
Donors_c_s6	1,581	1,768	3,073	3,583	10,005
	(15.8%)	(17.7%)	(30.7%)	(35.8%)	(100%)
Percentage point delta	0.2 Pp.	0.0 Pp.	0.2 Pp.	−0.4 Pp.	

Table 6.11: Distributions of the values of Target_1 in the recipient unit and the donor unit Donors_c_s6

Example The observed frequencies for Target_1 in Donors_c_s6 are shown in table 6.11. The values in figure 6.3 on page 186 were augmented therefrom. As the overall overlap is 87%, the distributions of donor and recipient unit do not differ greatly. If all values were allocated randomly, assuming that 15.8% of the elements have value y_1, 17.7% have value y_2, 30.7% have value y_3, and 35.8% have value y_4, 13% of the values would also be hit by chance. In this case, $TCCR_{source}$ is calculated as follows.

$$TCCR_{source} = \left(min\left(\frac{1,851}{11,560}, \frac{1,581}{10,005}\right)\right)^2 + \left(min\left(\frac{2,043}{11,560}, \frac{1,768}{10,005}\right)\right)^2 +$$

$$\left(min\left(\frac{3,573}{11,560}, \frac{3,073}{10,005}\right)\right)^2 + \left(min\left(\frac{4,093}{11,560}, \frac{3,583}{10,005}\right)\right)^2 = 0.13$$

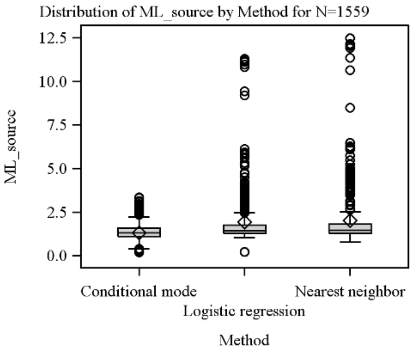
Consequently, $ML_{source} = \frac{TCCR_{model}}{TCCR_{source}} = \frac{0.43}{0.13} = 1.54$. The data augmentation model provides 54% more correct hits for the desired target value than would be hit by chance, if only knowing the donor unit distribution.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
ML_{source}	1.24	1.42	1.54

Table 6.12: Exemplary ML_{source} values in the case of Target_1 and Donors_c_s6

The ML_{source} values for the three methods used to augment Target_1 from Donors_c_s6 are shown in table 6.12. All of the three values are very similar to the ML_{chance} values in table 6.7 on page 202 in chapter 6.4.1. If the overlap is high, like for Donors_c_s6, the ML_{chance} and ML_{source} values are not significantly different. The knowledge available from the source is already close to the true, but unobserved distribution in the recipient unit.

Analysis If the state of information is already high *before* the augmentation, the model lift can only have a certain extent. However, we also regarded sources with other properties in our case study. This can be seen when regarding the ML_{source} ranges in figure 6.14. While most of the ML_{source} values are well below 2.5, the maximum ML_{source} values of logistic regression and nearest neighbor augmentations reach model lifts of 11.3 and 12.5, respectively. It means that these augmentations are 12 times as good, or more than 1,000% better, than what could be achieved without the data augmentation knowledge – taking into account the knowledge derivable from the overall source distribution.



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	3.35	11.3	12.5
Q3	1.58	1.75	1.82
Median	1.30	1.43	1.47
Mean	1.32	1.94	2.02
Q1	1.10	1.29	1.28
Min	0.18	0.21	0.76
Sh.-W.	0.95	0.49	0.53
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	44.01	<.001
Levene	10.81	<.001
Welch's ANOVA	74.32	<.001

Measures

Figure 6.14: Distribution of all ML_{source} values by applied augmentation method

Data augmentation results can be arbitrary high, if $TCCR_{source}$ is low. It can be seen from the boxes in the box and whiskers plot in figure 6.14 that the core of the augmentations ranges between 1.10 and 1.82 for all methods, which is largely similar to the previous established measures. But sources with very different distributions in donor and recipient unit cause the outliers. The inferiority of conditional mode imputation versus logistic regression and nearest neighbor hot deck can again be proven here when regarding the parameter estimates of the ANOVA model[12].

6.4.5 Comparison of the KPIs

The stated model lift measures have different informative values. One is not superior to the other. Rather, their different meanings can be used to draw different conclusions from the augmentation results. They can be

[12] The table of parameter estimates for this ANOVA model solution can be found in table 9.7 on page 305 in the appendix.

used to make within, between, and across case analysis in the case study. A summary of the described measures is given in table 6.13.

Model lift measure (range)	Assumption	Comparability	Function
ML_{chance} $(0 - k)$	The target variable distribution among the recipients is known.	*Between* cases with different sources and methods, but for the same target variable	Standardized measure for the overall quality of the augmentation results
ML_{target} $(0 - \infty)$	The number of recipients with a desired target value is known.	*Between* cases with different sources and methods, but for the same target variable	Measure for the quality of augmentation results, if selection is the targeting goal
$ML_{uniform}$ $(0 - k)$	The most frequent target value in the recipient unit is known.	*Between* cases with different sources and methods, but for the same target variable	Quality check criterion
ML_{source} $(0 - \infty)$	The target variable distribution is known among the donors only.	*Within* cases with the same source and target variable, using different methods	Measure for the increase in information achieved by data augmentation

Table 6.13: Comparison of different model lift measures

Within case analysis is possible, if the comparison measure is not influenced by the choice of the method. As none of the model lift measures are influences by the method, all model lift measures can used for within case analysis. The absolute value of ML_{source} cannot be used for any other type of analysis. It highly depends on the volatile comparison measure, which results from the different source properties. A high ML_{source} value does not implicate a high $TCCR_{model}$, and vice versa.

ML_{source} is the only measure that truly respects the state of information before an augmentation. The target values are not known for any of the recipients. Likewise, the overall distribution of target values in the recipient unit is not known. Given this supposition, the only knowledge on the target variable is given by the source. If so, without data augmentation, only the overall distribution is known. ML_{source} compares the total correct classification rate to this state of information. Therefore, ML_{source} is a relevant measure of how the state of information improved by data augmentation.

ML_{chance} and $ML_{uniform}$ are more stable than the other two measures, because they have defined ranges, while ML_{target} and ML_{source} range between 1 and infinity. Although the ranges differ for varying k and IQV, ML_{chance}, ML_{target}, and $ML_{uniform}$ can be compared between cases for differing sources. Between case analysis is possible, if the properties of the source do not influence the model lift measure. Because the three KPIs mentioned relate to a comparison measure which is derived from the true target value distribution in the recipient unit, it is not influenced by varying donor units.

While $TCCR_{chance}$ is a measure of how well target groups could be selected when information on the overall distribution of the target variable is present, $TCCR_{uniform}$ is a measure that does not differentiate target groups at all. Assigning the target values uniformly would not solve the marketing problem and therefore cannot serve as a benchmark. However, it can be seen from a managerial point of view that it is desirable to have better data augmentation results than what could have been achieved in terms of hits by assigning the most frequent value to all customers. Furthermore, ML_{chance} can be intriguing, because it is positive even if all recipients were assigned the same target value, which is undesirable. $ML_{uniform}$ serves as a quality check for these cases. It is the sternest measure. Only if $ML_{uniform}$ is positive, the data augmentation is a real success.

As argued earlier, ML_{chance} gives a good indication for the results of segmentation. If only a single value is of interest, ML_{target} measures how well exactly this value has been augmented. The two measures can differ. The suitable measure should be used for selection and segmentation goals, respectively.

Across case analysis would be possible, if the measures were not influenced by the target variable properties. However, all of the KPIs are influenced by them. Hence, across case analysis is only possible when taking into account the target variable properties. This is done in the more detailed examination in chapter 7.

6.5 Conversion probability lift

The final goal of data augmentation is not the augmentation results themselves, but the increase of conversion probabilities when using the augmented variables for target group selections or segmentations. Essentially, database marketing analysts are interested in how much the model improves the conversion probabilities, as compared to not using the model (Ratner, 2003, p. 234). The specific use case of target group selection after data augmentation has not yet been regarded in the existing literature. Hattum and Hoijtink (2008b) only regarded the segmentation case. Consequently, no measure yet exists describing the CPL for varying selections as derived from the data augmentation results. We establish two KPIs describing the data utility in terms of the CPL of the data augmentation results.

Calculation Let $P(conv)$ be the conversion probability for an individual customer and a given target variable value $\hat{y}$. Without loss of generality, let $P(conv)$ be an absolute value (b) if $\hat{y}$ is true for a customer, and 0 otherwise.

$$P(conv) = \begin{cases} b & \text{if } y = \hat{y} \\ 0 & \text{if } y \neq \hat{y} \end{cases} \tag{6.15}$$

In the formula, $P(conv)$ is directly dependent on the similarity of y and $\hat{y}$. The model can be relaxed to a non-linear relationship and to cases where $P(conv) = a$ if $y \neq \hat{y}$, instead of 0. In practice, the relationship can be more complex. It has to be evaluated whether the available target variables Y are good predictors for $P(conv)$. This can only be evaluated externally and cannot be shown in this study.

The conversion probability for a customer as derived from data augmentation is given by the general conversion probability $P(conv)$ and the probability of $\hat{y}$ to be true, given the corresponding link variable class:

$$P(conv|X) = P(conv) \times P(Y = \hat{y}|X) \tag{6.16}$$

When segmentation is the database marketing goal, the CPL is equal to the model lift, given that Y is a predictor of $P(conv)$. When selection is the database marketing goal, the CPL also depends on the number of customers to be selected for a specific marketing campaign. In order to calculate the conversion probability lift (CL) for a specific number of selected customers for a direct marketing campaign (n) and for a desired target value y_{target}, the data augmentation results are sorted by $P(Y = \hat{y}_{target}|X)$. If a certain number of customers is selected to be contacted, e.g. $n = 1,000$, the 1,000 customers with the highest $P(Y = \hat{y}_{target}|X)$ values are selected. In our case study, the number of hits can be observed for the customers. The number of customers selected with the correct value, if recipients were drawn by chance from the customer group, n_{chance}, can be calculated by

$$n_{chance} = \frac{r(y_{target})}{r} \times n \tag{6.17}$$

The conversion probability lift $CL(N = n)$ is calculated as the conversion probability of the selected target group with n recipients by means of data augmentation results, divided by the conversion probability of a randomly selected target group, where n_{hit} is the number of selected customers that were hit.

$$CL(N = n) = \frac{\frac{P(conv) \times n_{hit}}{n}}{\frac{P(conv) \times n_{chance}}{n}} = \frac{n_{hit}}{n_{chance}} \tag{6.18}$$

The CPL shows how much the new selection criteria obtained by augmentation increase (or decrease) the conversion probability for given n. It is a ratio of how many customers with the desired values were hit using the augmented data, versus using a random distribution. The CPL does not depend on $P(conv)$. Consequently, $P(conv)$ does not have to be predicted in our case study. $P(conv)$ should be assessed during external evaluation, where its direct implications for conversion rates and sales can be observed.

It is desirable to know the best number of recipients, i.e. for which n the CPL is highest. The CPL is dependent on the number of customers selected for a target group. If the target group is large, the lift can only have a certain extent, because the number of customers having a desired value is limited. The uplift is easily high on a relative scale, if only a very small number of customers is selected. However, both the small absolute value and the small target group size are undesirable in a practical setting.

Figure 6.15 shows the distribution of correctly selected customers and customers selected in total. In a perfect selection (a), the hit rate is 100%. It forms the upper boundary for data selection possibilities. The line has a slope of 1 until $n = n_{target}$ and is then flat with a slope of 0, because no additional customers with the correct value can be added to the selection. The hit rate achieved with the data augmentation results (b) is desired to be closer to the perfect selection as the hit rate achieved by a random selection (c). The random selection always has a $\frac{r_{target}}{r}$ slope. It forms the lower boundary for data selection possibilities, but it is possible that selections cross the lower boundary.

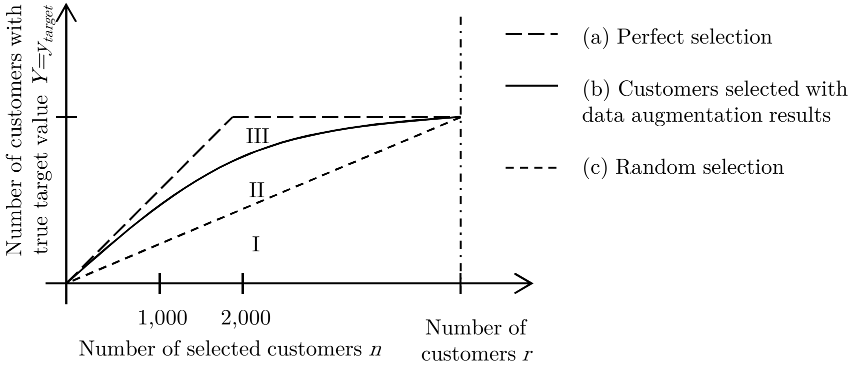


Figure 6.15: Conversion probabilities by data augmentation, random selection, and perfect selection

The vertical distance between the augmentation line and the random selection line quantifies by how much the augmentation selection is better than a random selection for given n. The y-value of the augmentation line is divided by the y-value of the random selection line for the same x-value in order to calculate a conversion uplift. From figure 6.15, it can be seen that the CPL can be highest at $n = r(y_{target})$, because the distance between the random selection line and the perfect selection line is greatest, in comparison to the distance of the random selection line to the x-axis. From the case study results, insights on the best number of customers to select are derived. Therefore, the n with the greatest distance between the random selection line and the data augmentation line is calculated.

It is desirable to have a global measure in order to make a robust statement on how well the data augmentation improved the conversion probability – independent of how many customers are to be selected. In order to give an overall indication of the conversion probability lift in general, the areas under and between the conversion probability curves (A) are regarded. In order to calculate the areas under the curve, the space between each curve and the x-axis is regarded for the range between 0 and r. A_a describes the area under the curve of the perfect selection line (I + II + III). A_b describes the area under the curve of the data augmentation line (I + II). There is a function $f(n)$ describing the number of correct target values hit (h) by the data augmentation results: $f(n) = \sum_{i=1}^{n} h$. Finally, A_c describes the area under the curve of the random selection line (I).

$$A_a = \frac{r(y_{target})^2}{2} + r(y_{target}) \times (r - r(y_{target})) \tag{6.19}$$

$$A_b = \sum_{1}^{n} \left(f(n-1) + \frac{f(n) - f(n-1)}{2} \right) \tag{6.20}$$

$$A_c = \frac{r(y_{target}) \times r}{2} \tag{6.21}$$

The line of the data augmentation selection is not a curve, but a scatter plot with one y-value per x-value. It is continuous, so that no x-value has a lower y-value than the previous x-value. The formula reflects the fact that the areas are polygons. From the figure, a global CPL measure can be derived. The area below the data augmentation selection line (A_b) and the area below the random target group line (A_c) are compared in order to calculate a CPL measure.

$$CL_{global} = \frac{A_b}{A_c} \tag{6.22}$$

The global CPL shows how much the new selection criteria obtained by data augmentation increase (or decrease) the global conversion probability, independent of n. CL_{global} is limited by the perfect selection line, so that the maximum possible CL_{global} is

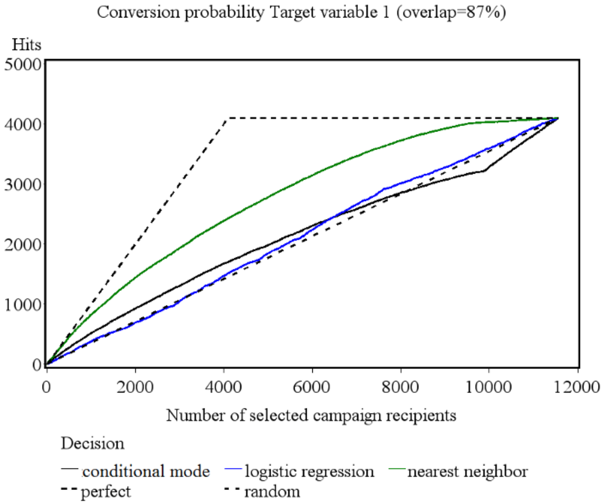
$$max(CL_{global}) = \frac{A_a}{A_c} \tag{6.23}$$

CL_{global} varies between 1 and 2, depending on the number of target values k and the variance as measured by IQV. The limit of $max(CL_{global})$ is 2 for $n_{target} \rightarrow 0$, but the actual maximum possible CPL can be smaller, if many customers have the desired target value. To better benchmark the measure, the area below the data augmentation selection line is compared to the upper boundary of the perfect selection line with the conversion probability lift magnitude (CM).

$$CM_{global} = \frac{A_b - A_c}{A_a - A_c} \tag{6.24}$$

The CPL magnitude shows how close the results using the new selection criteria are to a perfect selection. It varies between 0 and 1, if $CL_{global} >= 1$ and between -0 and -1, if $CL_{global} < 1$. A small CPL does therefore not always imply a "bad" data augmentation. The second measure enhances the global CPL. Data augmentation results can have a low CL_{global}, while having a high CM_{global}, and vice versa.

Examples Two examples of selected augmentations are shown in figures 6.16 and 6.17. The plots show a random selection line, a perfect selection line, and a data augmentation selection line for each conditional mode imputation, logistic regression augmentation, and nearest neighbor hot deck augmentation.



Plot

Perfect selection
$A_a = 38,938,755.5$

Augmentation results
nearest neighbor
$A_{b1} = 32,019,942.5$

Augmentation results
logistic regression
$A_{b2} = 24,382,556.5$

Augmentation results
conditional mode
$A_{b3} = 24,609,766.5$

Random selection
$A_c = 23,657,540$

Measures

Figure 6.16: Conversion probability plot for a source with a 87% overlap rate

Figure 6.16[13] shows the plot for Target_1 and source Donors_c_s6, which has a 87% overlap rate regarding the recipient unit. The number of recipients with the desired target value of $n_{target} = 4,093$ is about a third of the customers. The maximum possible CPL as measured by the perfect selection line is $max(CL_{global}) = \frac{A_a}{A_c} = 1.65$. The result of the nearest neighbor

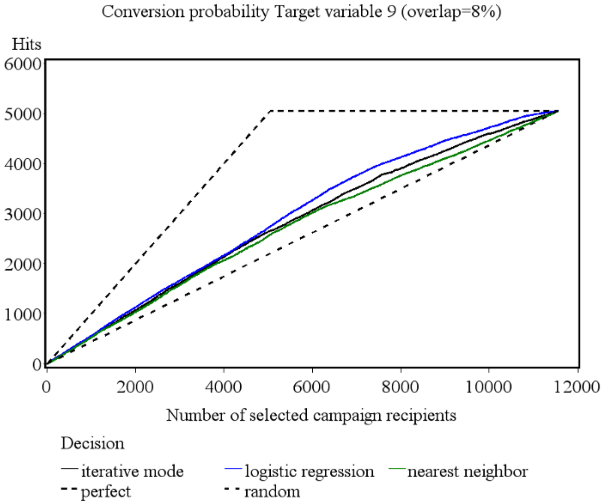
[13] The colored version of this figure can be found online on www.springer.com under the title of this publication.

hot deck augmentation approaches the perfect selection line well. The global CPL is high for nearest neighbor hot deck ($CL_{global} = \frac{A_{b1}}{A_c} = 1.35$), as well as the magnitude ($CM_{global} = \frac{A_{b1}-A_c}{A_a-A_c} = 55\%$). Conditional mode imputation and logistic regression do not perform well in this setting. Only a few more customers are hit, compared to a random selection. The global CPL is low ($CL_{global} = 1.04$ for conditional mode and $CL_{global} = 1.03$ for logistic regression), as well as the magnitude ($CM_{global} = 4\%$ and $CM_{global} = 3\%$, respectively). The conditional mode imputation line even drops below the random selection line at times. Nearest neighbor hot deck is clearly superior to the other methods in this setting.

Method	Conditional Mode	Logistic Regression	Nearest Neighbor
CL_{global}	1.04	1.03	1.35
CM_{global}	0.06	0.05	0.55
Optimum_conv	3,965	7,619	5,799
Uplift_max	1.19	1.08	1.50

Table 6.14: Exemplary CL_{global} and CM_{global} values in the case of Target_1 and Donors_c_s6

For our example from chapter 6.1, the CPL KPIs are given in table 6.14. It lists the CL_{global} and CM_{global} measures for the three methods in this case, along with a measure for the optimal number of customers to be selected (Optimum_conv), i.e. the location on the x-axis where the distance between the data augmentation line and the random selection line is greatest. For this location, or number of customers to select, the maximum conversion probability uplift (Uplift_max) is given. It can be seen that the uplift is greatest when using nearest neighbor hot deck as an augmentation method. If 5,799 customers were selected, after having been sorted by $P(Y = y_4|X)$, 50% more customers would be correctly selected than if a random selection would be used. The results of conditional mode imputation would lead to an uplift of 1.19 when selecting $n = 3,965$ customers, while with the logistic regression results, a maximal uplift of 1.08 can be reached for a target group with $n = 7,619$ customers.



Plot

Perfect selection
$A_a = 45,587,672$

Augmentation results
nearest neighbor
$A_{b1} = 31,631,083$

Augmentation results
logistic regression
$A_{b2} = 33,892,132$

Augmentation results
conditional mode
$A_{b3} = 32,581,384$

Random selection
$A_c = 29,154,320$

Measures

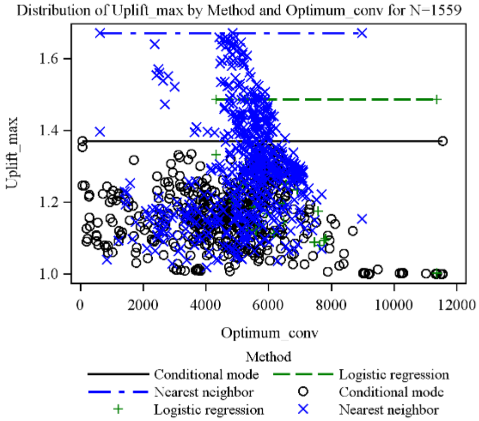
Figure 6.17: Conversion probability plot for a source with a 8% overlap rate

Figure 6.17[14] shows the plot for a source with a 8% overlap rate. The number of recipients with the desired target value Target_9 is higher than in the previous example ($n_{target} = 5,044$), therefore the maximum possible CPL as measured by the perfect selection line is lower ($max(CL_{global}) = 1.56$). This time, the global CPL is low for nearest neighbor hot deck ($CL_{global} = 1.08$), as well as the magnitude ($CM_{global} = 15\%$). The logistic regression augmentation performs better in this setting. Although the global CPL is only moderately greater to the previous example ($CL_{global} = 1.16$), the magnitude is better ($CM_{global} = 29\%$). This is due to the lower maximum possible CPL as measured by the perfect selection line. Because it is not possible to reach a CL_{global} value greater than 1.56, the magnitude is stronger. Nearest neighbor hot deck is inferior to logistic regression in

[14] The colored version of this figure can be found online on www.springer.com under the title of this publication.

this context. Conditional mode imputation can be found between the two methods, with $CL_{global} = 1.11$ and $CM_{global} = 21\%$.

Analysis The ranges of conversion uplifts for all augmentations are shown in figure 6.18[15]. The symbols mark the uplift in relation to the optimal number of selected recipients. The lines mark the maximum uplift reached by the respective method.



Scatter plot

	Conditional mode	Log. regression	Nearest neighbor
N	532	488	539
Optimum_conv			
Max	11,553	11,368	8,987
Mean	5,042	6,230	5,244
Min	60	4,329	618
Uplift_max			
Max	1.37	1.49	1.67
Mean	1.15	1.18	1.30
Min	1.00	1.00	1.02

Measures

Figure 6.18: Scatter plot for conversion uplift by augmentation method used and optimal number of customers to be selected

The uplift is always greater than 1 for the best number of customers to be selected. It means that all source-target-combinations with all methods lead to results better than random, when regarding the optimal number of customers selected for a target group. The maximum uplift reaches values up to 1.37 for conditional mode imputation, 1.49 for logistic regression, and 1.67 for nearest neighbor hot deck. Nearest neighbor hot deck results reach significantly better values than the other methods. On average, the uplift

[15] The colored version of this figure can be found online on www.springer.com under the title of this publication.

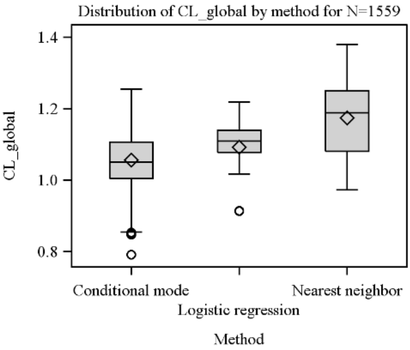
has values of 1.15 (conditional mode imputation), 1.18 (logistic regression), and 1.30 (nearest neighbor hot deck).

The number of customer to be selected at the best point is around the number of customers that have the desired target value $r(y_{target})$. It varies mainly between 3,696 and 6,038 for conditional mode imputation, between 5,193 and 6,557 for logistic regression, and between 4,777 and 5,972 for nearest neighbor hot deck. It can be seen that the uplift is never high at high levels of customers to be selected.

The distribution of the CL_{global} measures calculated for all augmentations in the case study is shown in figure 6.19. From the box plot, a hierarchy between the methods can be observed. Conditional mode imputation performs worst, with an average CPL of 1.05. The average CPL of logistic regression is a better ($CL_{global} = 1.09$). The CPL values vary between 1.07 and 1.14 for 50% of the observations, while the lower border of the box of conditional mode imputation drops to 1.00. Nearest neighbor hot deck has the highest average CPL ($CL_{global} = 1.17$) and a variance of 1.08 to 1.25. The difference between the means is significant as confirmed by Welch's ANOVA. Welch's F statistic is used for assessing the mean differences, because Levene's test yields a significant F value, showing that the variances of the observations are not equal. This confirms the overall finding that nearest neighbor hot deck is in general superior to logistic regression and conditional mode imputation, also for selection tasks. Again, logistic regression has the lowest variance.

The nearest neighbor hot deck results are significantly better than those of logistic regression. This is different to the model lift measures, which are relevant to the decision for segmentation tasks. The parameter estimates for this ANOVA model solution show that the difference between the nearest neighbor hot deck mean and the logistic regression mean is significant[16].

[16] The table of parameter estimates for this ANOVA model solution can be found in table 9.8 on page 306 in the appendix.



Box plot

Method	Cond. mode	Log. reg.	Near. neighbor
N	532	488	539
Max	1.25	1.22	1.38
Q3	1.11	1.14	1.25
Median	1.05	1.11	1.19
Mean	1.06	1.09	1.17
Q1	1.00	1.08	1.08
Min	0.79	0.91	0.97
Sh.-W.	0.97	0.84	0.97
p-value	<.001	<.001	<.001

Test	F	p
ANOVA	249.14	<.001
Levene	41.56	<.001
Welch's ANOVA	209.53	<.001

Measures

Figure 6.19: Distribution of all CL_{global} values by applied augmentation method

On average, the conversion probability lift reachable is higher when using the nearest neighbor hot deck method.

The CPL and the conversion magnitude do not necessarily show how much the knowledge of a company increased after data augmentation. They only show the increase in knowledge, if no other information is available to select target groups. In fact, the CPL depends on the ability of a database marketing department to predict conversion probabilities *before* applying data augmentation. If this ability is high already, CPLs can only have a certain extent. Companies are able to select target groups before augmentation projects. Data augmentation results increase the precision in these selections. In an external evaluation in practice, the data augmentation results have to live up to the traditional selection criteria used before. Furthermore, the database marketing analyst combines the traditional selection criteria with the newly acquired target variables in order to construct a meaningful mix of selection criteria which best target the target group[17].

[17] It would have been possible to include a traditional selection line in figure 6.15 on page 224. This line might be situated somewhere in the triangle between random selection line and perfect selection line. However, such a line depends on the respective company and its abilities to select target groups before a data augmentation approach. It is not generalizable. Therefore, no such line is included in our figure.

Chapter 7

Analysis of results and test of hypotheses

The KPIs established in chapter 6 are used in this chapter to test the hypotheses stated in chapter 5.1.2. All of the tests are only possible in the case study setting, where the true target values, as well as the source data mechanisms, are known. We validate the source data mechanism antecedents (hypothesis 1). We examine whether MNAR sources significantly bias the augmentation results. First, the conditional association of source and target variable given the link variables is tested for categorizing the source data mechanism type. Then, the augmentation results are compared to the results that would have been achieved by a MCAR source. We show that calculating the conditional association is complex in a categorical setting and does not lead to unambiguous results.

Subsequently, we test the individual influences of the source characteristics representation (hypothesis 2a), overlap (hypothesis 2b), number of donors (hypothesis 2c), and surplus (hypothesis 2d) on the model lift. The overall hypothesis 2 and hypothesis 3 regarding the general influenceability of the methods by the source characteristics are analyzed by a regression

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_7

model explaining the model lift with all characteristics. That way, interactions and the marginal effects of the characteristics are observable.

Eventually, we show that in certain settings, logistic regression is more suitable for reaching the best augmentation results than nearest neighbor hot deck, and vice versa (hypothesis 4). We find that the overlap between donor and recipient unit is significant in this decision. Conditional mode imputation turns out to be generally inferior to the other two methods, as expected from the findings in chapter 6. We are able to show that if all decisions regarding target variable exit criteria, source variable exit criteria, and choice of the best method are made correctly, data augmentations are able to lead to significant CPLS for MAR and MNAR sources. The differentiation between the latter thus becomes unessential, when respecting the rules regarding the observable parameters found.

7.1 Validating the source data mechanism

The source data mechanism can be MCAR, MAR, and MNAR. With the first two forms, the source data mechanism is ignorable. An ignorable source data mechanism is an antecedent for data augmentation in the conceptual model established in chapter 5.1.1. For all sources that are not MCAR, the ignorability is manifested in the conditional independence between the source data mechanism and the target variable, given the link variables. It means that for every link variable class, i.e. every combination of $x_1, x_2, ..., x_n$, the values of the target variable Y cannot change when taking the source data mechanism S into consideration. Otherwise, augmentation results can be biased. This has already been theoretically described in chapter 4.2.5. In the categorical setting, it is a complex relation. Different tests can be applied and a decision has to be made towards the most meaningful test result. They can only be applied in a case study, where the true target values as well as the source data mechanism are known.

We apply two existing test statistics for conditional associations, the χ^2 test with aggregated total measures and the CMH test. The results give an answer on whether the augmentations can be carried out without the risk of biased results. However, the theoretical risk of biased results does not necessarily imply that the results are actually biased. From a managerial point of view, it is desirable to know whether an existing conditional dependence is strong enough to systematically falsify the augmentation results. Therefore, we add two more tests assessing whether the data augmentations results significantly differ when augmenting data from a source as compared to when augmenting it from the overall population.

7.1.1 Test based conditional dependency calculation

If a source is not MCAR, it needs to be at least conditionally independent from the target variables, given the link variables, in order to be suitable for data augmentation. A conditional association is calculated by determining the association between the source data mechanism variable and the target variable for every link variable class. There are 539 cases, i.e. source-target combinations, in our case study, of which 231 are MCAR by definition, because the sources either have a 100% overlap rate regarding the recipient unit or are a representative sample thereof. There are 28 source data mechanisms that are not MCAR, which result from the different forms of $D \subset R$, $D \cup R$, and $D \cap R = 0$. We have illustrated these sources in figure 5.8 on page 167 in chapter 5.2.3. These 308 source-target combinations need to be evaluated using appropriate tests.

The contingency table for conditional independence has a $v \times 2 \times k$ form, where v indicates the number of link variable classes and can take various values, the source data mechanism variable has always two target values and the target variable can take several target values. The unconditional association between source data mechanism indicator variables and target variables is calculated for every possible link variable class. By building partial two-way cross-sectional tables, the relationship between the source data mechanism and the target variable can be controlled for the link variables. These tables contain no information on the link variables, but they show the relationship between source data mechanism indicator variable and target variable at the same location of the overall distribution (Agresti, 2002, p. 48). The observed associations are conditional associations. Thus, more than a thousand individual statistical tests (as many as link variable classes) might be necessary for calculating conditional dependencies.

Two established tests are applied for testing the conditional association. The first is a combination of individual χ^2 tests with aggregated total measures. The second is the CMH test for conditional associations. The tests have different properties and different tendencies and are shortly delineated. Conditional association tests are always performed for the whole population, where the source indicator variable indicates whether an element has been observed in a certain source. For that reason, the proposed test can only be performed in a case study context.

Calculation For the χ^2 test with aggregated total measures, a χ^2 test is performed for every link variable class. In the simple case of a $2 \times 2 \times k$ form, the χ^2 tests is performed twice on the $2 \times k$ table for each of the two link variable values. If there are several link variables with several target values, the number of tests multiplies accordingly. For every test, it is decided whether there is an association between source data mechanism indicator variable and target variable. In the narrowest sense, source data mechanism indicator variable and target variable are conditionally independent given the link variables, if they are independent for any given variable combination of the link variables (Agresti, 2002, p. 48). In practice, however, the partial tables rarely uniformly point into one direction (Agresti, 2002, p. 236). In order to find an aggregated measure, the χ^2 test statistics of the individual tests are added and the degrees of freedom are calculated as $df = v \times (k-1) \times (2-1)$, where the first factor is derived from the condition, the second factor is derived from the target variable, and the third factor is derived from the source data mechanism indicator variable (Pennsylvania State University, 2013a). In this context, it always equals 1. The aggregated total measure for the χ^2 tests is calculated as follows.

$$\chi^2_{df} = \sum_{i=1}^{v} \chi^2(Y, S|X) \tag{7.1}$$

The χ^2 test statistic with according degrees of freedom has a corresponding p-value, which indicates the probability of the test statistic to take a value at least as extreme as the observed value, assuming that the null hypothesis is true. The null hypothesis is rejected, if the p-value is smaller than a defined level of significance.

The CMH test verifies the null hypothesis that all conditional odds are equal to 1 (Agresti, 2002, p. 232). The calculation of test statistics differs, depending on the scales of the variables. If both variables are nominal, the number of degrees of freedom is calculated as $df = (v-1) \times (k-1)$. If one of the variables is nominal and one is ordinal, the number of degrees of freedoms is calculated as $df = (v-1)$. If both variables are ordinal, the number of degrees of freedoms is calculated as $df = 1$ (Pennsylvania State University, 2013b). The CMH test statistic with according degrees of freedom has a corresponding p-value, which indicates the probability of the test statistic to take a value at least as extreme as the observed value, assuming that the null hypothesis is true. The null hypothesis is rejected, if the p-value is smaller than a defined level of significance. Only if both tests agree, it can be said with certainty that the conditional association is dependent or independent.

A frequently encountered problem when calculating the individual dependencies for link variable classes occurs when no statistics can be calculated, because all observations have the same source data mechanism indicator or target variable value. These partial tables are referred to as sparse (Agresti, 2002, p. 233). They pose a problem to the calculation of the overall conditional association. In the case of the χ^2 test with aggregated total measures, the strata with missing statistical tests are interpreted as independent components. They add to the sum of the overall χ^2 value with a value of 0 each, but are taken into account for the degrees of freedom. Thus, they shift the overall tendency of the test result towards independence. This approach can be argued from a heuristic point of view. If there are less than two non-missing levels, changes in the one variable do not influence the value of the other variable. Thus, they are totally independent. Sparse cross-sectional tables are integrated in the aggregated total measure.

A different approach is taken for the CMH test. It is implemented in SAS so that only statistics calculated for non-sparse tables are taken into account, shifting the overall tendency of the test result towards dependence. Consequently, the results of the CMH test can differ from those of the χ^2 test with aggregated total measures, if there are a lot of variable combinations with at least two non-missing levels.

Example An example for the results of test based conditional independence calculations is shown in table 7.1. A different example than the one stated in chapter 6.1 is used for illustration purposes.

Test	p-value	Association	Estimates (from v)
χ^2 test with aggregated total measures	0.99997	independent	959 (1,468)
CMH test	6.15E-20	dependent	

Table 7.1: Example for test based conditional independence calculation in the case of Donors_c0_s8 and target variable 5

The χ^2 test with aggregated total measures has a p-value of 1 and hence it cannot be proven that there is an association between the source data mechanism indicator and the target variable. Contrarily, the CMH test has a p-value close to 0, so that it can be assumed with a 5% level of significance that there is an association between the source data mechanism indicator and the target variable. From the χ^2 test with aggregated total measures, it can be seen that only $\frac{959}{1{,}468} = 65\%$ of the link variable classes delivered estimates for the respective class. Because the two tests do not point into the same direction and because the sparse data problem is severe, no decision can be made on the conditional association between the source and the target variable with this test based calculation.

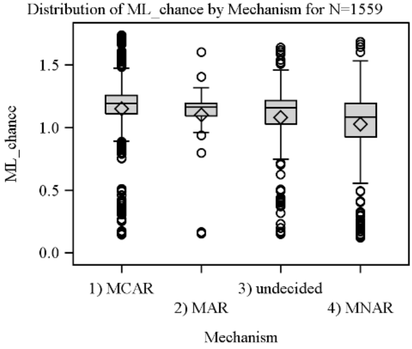
Analysis The two tests agreed in 76% of the augmentations, as shown in table 7.2. It shows the number of cases and their classification by both of the tests. Most associations are considered dependent by both tests.

Cases (not MCAR)		CMH test		
		dependent	independent	Total
χ^2 test with aggregated total measures	dependent	215	10	225
	independent	64	19	83
	Total	279	29	308

Table 7.2: Agreement of test based calculations of conditional dependencies

For most of the discordant pairs, the χ^2 test with aggregated total measures delivers an independent association and the CMH test delivers a dependent association. For these cases, no final decision can be made on the conditional association. In order to analyze the consequences of different source data mechanisms, the MCAR sources are included in the analysis. From the test results, the case study yields three more categories of source data mechanisms to be differentiated: MAR, MNAR, and "undecided" due to differences in the two tests. Their influence on the model lift is shown in figure 7.1. Because there are up to three ML_{chance} measures per case, resulting from the application of different methods, the number of observations multiplies accordingly.

All of the source data mechanisms have outliers, but the main part of the observations can be found between $ML_{chance} = 0.93$ and $ML_{chance} = 1.26$, although there is variance among the source data mechanisms. Only MNAR sources drop below the mark of $ML_{chance} = 1$ when regarding the first quartile. The MCAR category with an overlap rate of 100% and no sampling performs bests with an average model lift of 1.15. MNAR sources



Box plot

	MCAR	MAR	und.	MNAR
N	688	57	216	598
Max	1.74	1.61	1.64	1.69
Q3	1.26	1.19	1.22	1.20
Median	1.19	1.17	1.16	1.09
Mean	1.15	1.10	1.08	1.03
Q1	1.11	1.10	1.03	0.93
Min	0.15	0.15	0.15	0.12
Sh.-W.	0.81	0.65	0.85	0.88
p-val.	<.01	<.01	<.01	<.01

Test	F	p
ANOVA	21.78	<.001
Levene	0.94	0.419
Welch's ANOVA	22.03	<.001

Measures

Figure 7.1: Distribution of all ML_{chance} values by the source data mechanism as derived from the test based calculation of conditional dependencies

perform worst with an average model lift of 1.03. MAR sources average around 1.10 and the source-target combinations that have not been assigned a source data mechanism due to conflicting test results reach an average model lift of 1.08. Because the residuals are normally distributed, which can be seen from the Shapiro-Wilk statistics, and because the variances are equally distributed as confirmed by Levene's test, it is possible to compare the means using ANOVA. The F statistic is significant, proving that the means are not equal. When solving the ANOVA model, it is possible to assess whether ML_{chance} results of MNAR sources are significantly worse than those with valid source data mechanism, as well as than those classified as "undecided". The difference is significant for every pairwise comparison, so that it can be said with certainty that MNAR sources perform worse than other sources as measures by ML_{chance}[1]. Our first hypothesis is proven.

> *Proof of hypothesis 1:* The model lift is significantly higher for MAR sources than for MNAR sources.

[1] The table of parameter estimates for this ANOVA model solution can be found in table 9.9 on page 306 in the appendix.

Whenever possible and assessable upfront, MNAR sources should be excluded from data augmentation. However, such an upfront assessment is usually not possible. Furthermore, because of the "undecided" sources and because the ultimate decision criterion is whether the detected dependencies influence the data augmentation results, two model based calculations of conditional dependencies are added.

7.1.2 Model based conditional dependency calculation

The knowledge on the conditional association is not the main concern of the database marketing analyst. Rather, it is desirable to know whether the selection of the donors based on the source data mechanism significantly influences the data augmentation results. This can be translated into a model where Y is explained both by X and S. If the predicted values of Y significantly differ when including S into or excluding S from the model, the target variable and the source data mechanism are not conditionally independent given the link variables. The hypothesis has to be tested in the overall population. It can only be performed in the case study, because such information is not available in a practical application.

Calculation We developed two models that are tested using logistic regression. The first model includes the source data mechanism:

$$Y = \beta_0 + \beta_1 \times x_1 + ... + \beta_l \times x_l + \beta_{l+1} \times S \tag{7.2}$$

In order to evaluate the significance of the influence of the source data mechanism indicator variable, the Wald statistic is calculated. It tests the hypothesis whether the influence of the variable is negligible, i.e. if its regression coefficient β_{l+1} is zero (Backhaus et al., 2008, p. 273). The Wald statistic, however, is only of minor interest. The augmentation results achieved with the model including the source data mechanism indicator variable are compared to the Y values augmented without inclusion of the source data mechanism indicator variable. A comparison procedure pairs up the results from including the source data mechanism indicator variable versus excluding it. It observes whether a significant portion of the recipients is assigned different target values by the models tested. The test can show whether a significant portion of the observed results is different for the two models.

The second model separates two groups as defined by $S = 1$ and $S = 0$. Target values are augmented separately for these groups in order to compare the results with the augmentation where the groups are not separated.

$$Y(S = 1) = \beta_0 + \beta_1 \times x_1 + ... + \beta_l \times x_l + \beta_l \times x_l \tag{7.3}$$

$$Y(S = 0) = \beta_0 + \beta_1 \times x_1 + ... + \beta_l \times x_l + \beta_l \times x_l \tag{7.4}$$

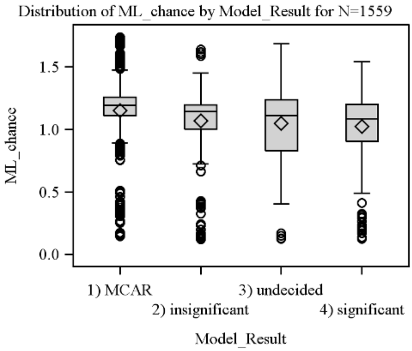
The augmentation results achieved with the model separating the groups by the source data mechanism indicator variable are compared with the Y values augmented without separation. The model based calculations test whether including a source data mechanism indicator variable into a model significantly changes the values of the target variable. Any test based proven conditional dependence is obviated, if data augmentation results do not change when taking into consideration a nonprobability sample. In fact, model based calculations of conditional dependencies do not actually test conditional associations. They inspect the influence of the source on the target variable. Because data is augmented based on link variables, the influence of the source data mechanism needs to be of sufficient strength in order to change the values of the augmented target variable. It can therefore be assumed that the stronger the predictive power of the link variables, the less important a conditional independence test.

Analysis The first model includes the source data mechanism indicator as an explaining variable along with the link variables. In our setting, 88% of the source-target combinations showed significant Wald measures, meaning that the source mechanism indicator variable has an influence on the model lift. The question whether the allocation of target values significantly changes when including or excluding the source data mechanism indicator variable into or from the model can be assessed when comparing identical elements after both augmentations. While the inclusion-exclusion model gives a fairly good insight into the properties of the influence of the source data mechanism indicator variable, it does not regard the conditional association of target variable and the source given the link variables. Rather, it analyzes the joint predictability of the target variable with both link variables and the source as explaining variables. Therefore, a separation test is added, building a model for both $S = 0$ and $S = 1$. After data augmentation, the target values of identical elements are compared to an augmentation where the groups were not separated in order to check whether a significant number of units received a different value. In both models, the association between target and source data mechanism indicator variable is considered insignificant, if at least 95% of the elements in the overall population receive the same target value as when not considering S.

Cases (not MCAR)		Inclusion-exclusion model		
		significant	insignificant	Total
Separation model	significant	144	2	146
	insignificant	24	138	162
	Total	168	140	308

Table 7.3: Agreement of model based calculations of conditional dependencies

It can be seen in table 7.3 that the two model based calculations agree in 92% of the augmentations. 144 cases have been classified as dependent by both models and 138 cases as independent. For most of the discordant pairs, the inclusion-exclusion model delivers an independent association and the separation model a dependent association. If both the inclusion test and the separation test do not show significant differences to an augmentation where the source was ignored, the source data mechanism can safely be ignored for data augmentation. In order to analyze the consequences of different source data mechanisms, the MCAR sources are included in the analysis. From the comparison results, the case study yields three more categories of source data mechanisms to be differentiated: "insignificant" (where the results were not different in both models), "significant" (where the results significantly differed in both models), and "undecided" due to differences in the two models. Their influence on the model lift is shown in figure 7.2.



Box plot

	MCAR	insign.	und.	sign.
N	688	57	216	598
Max	1.74	1.64	1.69	1.54
Q3	1.26	1.20	1.24	1.20
Median	1.19	1.14	1.11	1.09
Mean	1.15	1.07	1.05	1.03
Q1	1.11	1.01	0.83	0.91
Min	0.15	0.12	0.12	0.13
Sh.-W.	0.81	0.78	0.92	0.90
p-val.	<.01	<.01	<.01	<.01

Test	F	p
ANOVA	20.22	<.001
Levene	2.93	0.033
Welch's ANOVA	19.89	<.001

Measures

Figure 7.2: Distribution of all ML_{chance} values by the categories as derived from the model based calculation of conditional dependencies

From figure 7.2, it can be seen that the ML_{chance} mean of MCAR sources is higher than those of the other sources. It has a mean of 1.15, while the sources classified "insignificant" have a mean of 1.07, the sources classified as "undecided" have a mean of 1.05 and the sources classified as "significant" have a mean of 1.03. The mean difference is significant as confirmed by Welch's ANOVA. Welch's statistic is used, because the variances are not equal (Levene's F statistic is greater than 1). However, when regarding the parameter estimates for the solution of the ANOVA model, it becomes obvious that the difference is not significant between "undecided" and "significant" sources. The difference between "significant" and "insignificant" sources is only slightly significant with a p-value of 2.7%[2].

7.1.3 Discussion

It has been shown with the test based calculations of conditional dependencies that the type of the source data mechanism influences the augmentation results. This confirms our hypothesis as derived from the conceptual model in chapter 5.1.1. However, the results are not entirely satisfying for several reasons. Firstly, the two tests used for estimating the conditional association between target variable and source, given the link variables, did not agree in 24% of the cases. For these, no usage recommendation can be given. Secondly, the two models estimating the conditional association in order to assess the influence of a nonprobability sampling mechanism on the results did not agree in 8% of the cases. A significant part of these undefined cases lead to unacceptable results ($ML_{chance} < 1$). Thirdly, we have not yet shown whether the MNAR sources (as classified by the test based calculations) are really those sources leading to significant differences in the augmentation results (as defined by the model based calculations). Finally, we have not resolved the problem that the source data mechanism is not observable in practice and thus that none of the performed tests are implementable in order to separate MAR from MNAR sources.

In this chapter, we discuss the correlation of the test and model based calculation results in order to evaluate whether it is really the source data mechanism that influences the corruption of results. We examine the portion of unacceptable results (those that lead to $ML_{chance} < 1$) and whether this is correlated to the source data mechanism or result bias. From a managerial point of view, even biased results can be acceptable, if they have a positive impact on the model lift. If it is possible to quality check sources based on

[2] The table of parameter estimates for this ANOVA model solution can be found in table 9.10 on page 307 in the appendix.

observable parameters, so that biased or unacceptable results are precluded, it can obviate an ex ante assessment of the source data mechanism type.

Cases (not MCAR)		Model based calculations			
(Row percentages)		insignificant	undecided	significant	Row sum
Test based calculations	MAR	17 (89%)	0 (0%)	2 (11%)	19 (100%)
	undecided	44 (60%)	6 (8%)	24 (32%)	74 (100%)
	MNAR	77 (36%)	20 (9%)	118 (55%)	215 (100%)
$\chi^2 = 29.3$	Column sum	138 (45%)	26 (8%)	144 (47%)	308 (100%)
$Cramer's\ V = 0.22$ (medium)					

Table 7.4: Comparison of results from test and model based calculations of conditional dependencies for sources other than MCAR

Table 7.4 shows the classification of the source data mechanism as derived from the test based calculations of conditional dependencies cross tabulated with the results from the model based calculations. From the χ^2 measure, which is significant with $\alpha = 5\%$, it can be seen that the results from test and model based calculations of conditional dependencies for sources other than MCAR are correlated. However, the correlation is not very strong, as indicated by $Cramer's\ V$. Of those augmentations with sources classified as MAR, 89% lead to insignificant differences in the augmentation results, which means that the values of the target variables are not changed when augmenting data from the source, rather than from the overall population. However, 11% of the augmentations also lead to significant differences, as compared to augmentations using the overall population. It cannot be said with certainty that MAR sources do not influence the augmentation results. Of those augmentations with sources classified as MNAR, 35% lead to insignificant differences in the augmentation results and 55% lead to significant differences. It cannot be said with certainty that MNAR sources bias the augmentation results. Of those augmentations with sources classified as "undecided", 60% do not lead to differences in the augmentation results, but 32% lead to differences. It can be seen that MAR sources largely do not lead to biased results, while the risk is much higher for MNAR sources. However, there seem to be other parameters influencing the results other than the classification of sources as MAR or MNAR.

It has been shown that the assignation of source data mechanisms for categorical variables is not easy in the conditional association setting. Before drawing a final conclusion on the source data mechanism, it is examined how the source data mechanism influences the data augmentation results. For this purpose, the model lifts of the augmentations are analyzed, where the result is classified as ”acceptable” if $ML_{chance} > 1$ and ”unacceptable” else.

Augmentations (not MCAR)			Result	
(Row percentages)		acceptable	unacceptable	Row sum
Test based calculations	MAR	48 (84%)	9 (16%)	57 (100%)
	undecided	168 (78%)	48 (22%)	216 (100%)
	MNAR	375 (63%)	223 (37%)	598 (100%)
$\chi^2 = 24.0$	Column sum	591 (68%)	280 (32%)	871 (100%)
$Cramer's\ V = 0.17$ (weak)				

Table 7.5: Comparison of source data mechanisms and augmentations results for sources other than MCAR

Table 7.5 shows a comparison of the source data mechanism classification as derived from the test based calculations of conditional dependencies with the augmentation results. From the χ^2 measure (significant with $\alpha = 5\%$), it can be seen that the source data mechanism and the acceptability of the results are correlated. However, the strength of the association is weak, as indicated by $Cramer's\ V$. Although most of the MAR sources lead to acceptable results (84%), there is also a significant portion leading to unacceptable results. The majority of MNAR sources lead to acceptable results (63%), while 37% lead to unacceptable results. It seems like the source data mechanism is a factor, but there are other, possibly more important, parameters influencing whether augmentations results are acceptable. We will return to the influence of the source data mechanism when concurrently regarding other influencing parameters in chapter 7.3.3.

7.2 Influences of source characteristics

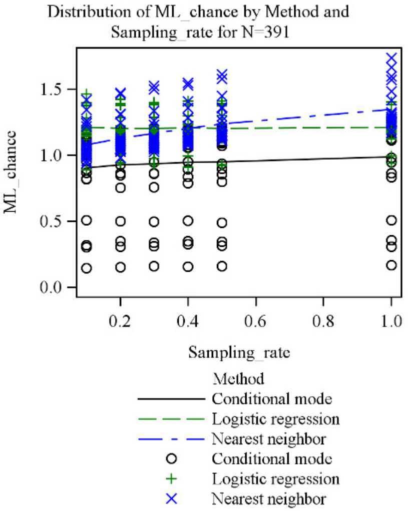
It has been clearly seen that the three methods do not lead to the same quality of results. Next it is examined for which source-target combinations which method works best (as measured by ML_{chance}). With nearest neighbor hot deck, 53% of the cases reached the highest results. 34% of the cases reached the highest results with logistic regression. But for 12% of the cases, conditional mode imputation would be the best choice. Even if these augmentations did not have a high model lift, the model lift was highest in the within case analysis. It is therefore desirable to know which method should be used in which situation.

7.2.1 Representation

The representative MCAR sources are regarded separately due to their special features. Figure 7.3[3] shows the model lift depending on different sampling rates for different methods in contrast to the model lift reached, if the full population is used (sampling rate= 1). The sampling rate is equal to the number of donors representatively sampled from the overall population, divided by the overall population. 11 sources are regarded ($D = R$, 5 representative sources with varying sampling rates of the form $D \equiv R'$, $D = P$, and 5 representative sources of the form $D \equiv P'$). From every source, the eleven target variables were augmented using conditional mode imputation, logistic regression, and nearest neighbor hot deck each. Because some of the results have already been deleted during the quality check phase described in chapter 6.2, 391 augmentations are examined.

In figure 7.3, the symbols mark the values observed. The lines connect the means of each level of overlap for conditional mode imputation, logistic regression, and nearest neighbor hot deck, respectively. From them, an overall tendency can be derived. The table of measures contains the number

[3] The colored version of this figure can be found online on www.springer.com under the title of this publication.



Scatter plot

Sampl. rate	N	Mean	Sh.-W.	p-val.
Conditional mode				
0.10	22	0.91	0.80	<.001
0.50	22	0.95	0.76	<.001
1.00	22	0.99	0.76	<.001
Logistic regression				
0.10	20	1.21	0.88	0.02
0.50	21	1.21	0.87	0.01
1.00	22	0.21	0.89	<.001
Nearest neighbor				
0.10	22	1.08	0.79	<.001
0.50	22	1.24	0.78	<.001
1.00	22	1.35	0.80	<.001

Method	Test	*F*	p
Cond. mode	ANOVA	0.16	0.976
	Levene	0.04	0.999
Log. reg.	ANOVA	0.01	1.000
	Levene	0.03	1.000
Near. neigh.	ANOVA	9.41	<.001
	Levene	0.14	0.984

Measures

Figure 7.3: Distribution of all ML_{chance} values by overlap and applied augmentation method with representative MCAR sources

of observations, the mean, the Shapiro-Wilk statistic for the distribution of the residuals, and the respective p-value for each method and selected sampling rate levels for illustration purposes.

It can be seen that conditional mode imputation is generally inferior to the other methods, while the mean lines of logistic regression and nearest neighbor hot deck intersect at about 40% sampling rate. With ANOVA, it can be assessed whether the mean differences per method are significant. It is permissible to use ANOVA, because the residuals are normally distributed, as measured by the Shapiro-Wilk statistic. Levene's test yields F values below 1 for every method, indicating equal variances, i.e. homoscedasticity. If the F value of the ANOVA test statistic is above 1, the mean differences are significant. It can be seen from figure 7.3b that the means significantly differ for different overlaps using the nearest neighbor method, but not for conditional mode imputation and logistic regression. A common representative source has a sampling rate well below 40%. Therefore, logistic regression is the best method to be used for representative MCAR sources.

> *Finding:* Representative MCAR sources with a small sampling rate ($< 40\%$) are best augmented using logistic regression.

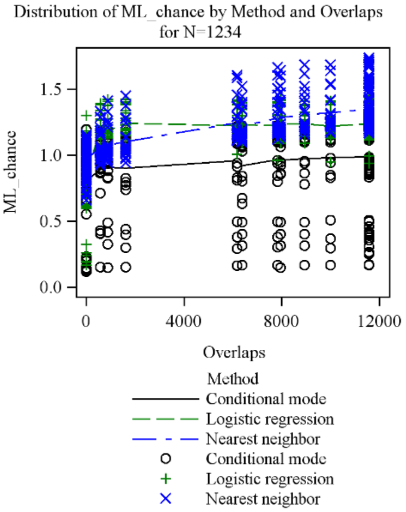
For MAR sources, the coherence between source characteristics and method to be chosen is expected to be more complex. It is examined using the criteria overlap, size, and surplus.

7.2.2 Overlap

The first source characteristic to be examined regarding its influence on the model lift using MAR sources is the overlap between the recipient unit and the donor unit. In practice, most external sources partially overlap with the recipient unit ($D \cup R$) or are a subset thereof ($D \subset R$). In the case study, we also added 100% overlapping sources ($D = R$, $D = P$, and $D \supset R$) and source with no overlap ($D = P \setminus R$ and $D \cap R = 0$) in order to illustrate the relative differences pertaining to the overlap. For each source, as shown in figure 5.9 on page 170 in chapter 5.2.3, eleven target values have been augmented using the three methods. After the quality check described in chapter 6.2, 1,234 observations remained for analysis purposes.

The scatter plot in figure 7.4[4] shows the distribution of ML_{chance} for MAR sources and MCAR sources without representative components, given the overlap, by the applied augmentation methods. Again, the line connecting the means has a higher incline for nearest neighbor hot deck than for logistic regression and conditional mode imputation. When regarding the ANOVA statistic, it can be assessed whether the mean differences are significant for each method. The differences are significant for all methods.

[4] The colored version of this figure can be found online on www.springer.com under the title of this publication. Table 7.4b only contains selected measures for illustration purposes.



Scatter plot

Overlap	N	Mean	Sh.-W.	p-val.
Conditional mode				
0	103	0.81	0.80	<.001
6,387	22	0.92	0.82	<.001
11,560	121	0.99	0.74	<.001
Logistic regression				
0	65	0.83	0.87	<.001
6,387	22	1.23	0.88	0.01
11,560	121	1.24	0.87	<.001
Nearest neighbor				
0	110	0.97	0.95	<.001
6,387	22	1.22	0.82	<.001
11,560	121	1.35	0.81	<.001

Method	Test	F	p
Cond. mode	ANOVA	2.04	0.028
	Levene	0.17	0.998
Log. reg.	ANOVA	34.62	<.001
	Levene	12.87	<.001
	Welch's	10.49	<.001
Near. neigh.	ANOVA	54.63	<.001
	Levene	0.67	0.756

Measures

Figure 7.4: Distribution of all ML_{chance} values by overlap and applied augmentation method with non-representative sources

For logistic regression, Welch's ANOVA statistic is regarded, because the variances are not equal. The parameter estimates for the solution of the ANOVA model for nearest neighbor hot deck show that all levels of overlap are significantly different from the highest mean at an overlap level of 100%[5]. The correlation between overlap and model lift tested for nearest neighbor hot deck is positive and significant at a 1% level of significance ($\rho = 0.578$). Our hypothesis established in chapter 5.1.2 can be confirmed for this method.

> *Partial proof of hypothesis 2a:* For nearest neighbor hot deck methods, the model lift increases with raising overlap of donor and recipient unit.

[5] The table of parameter estimates for the ANOVA model solution for nearest neighbor hot deck, conditional mode imputation, and logistic regression can be found in table 9.11 on page 308 in the appendix.

For conditional mode imputation, it cannot be confirmed. Although the overall ANOVA statistic confirms that the means are not equal for conditional mode imputation and logistic regression, the solutions of their ANOVA models reveal a different picture. The means do not significantly differ from the base mean for an overlap level of 100%, except for $o = 100\%$. Ergo, apart from the $D \cap R = 0$ sources, there is actually no difference in the average ML_{chance} values. The heteroscedasticity as shown by Levene's test for logistic regression is also due to the high variance of results for $o = 0$. This can be observed when excluding sources with $o = 0\%$ from ANOVA. Then, 956 augmentations are still regarded. While the F value is still significant for nearest neighbor hot deck ($F = 17.64$), it is insignificant for conditional mode imputation ($F = 0.41$) and logistic regression ($F = 0.85$) in this examination[6]. The correlation between overlap and model lift tested for conditional mode imputation ($\rho = 0.105$) and logistic regression ($\rho = 0.103$) are not significant at a $\alpha = 5\%$ level of significance.

The critical value of ML_{chance} is 1. If ML_{chance} is below 1, the data augmentation results are worse than allocating all target values randomly to the recipients. It can be seen from figure 7.4 that some of the augmentations do not fulfill this requirement. The percentage of these "unacceptable" augmentations ($ML_{chance} < 1$) is 66% for sources with $o = 0$ and only 15%, if the overlap is greater than 0. This is true for all of the applied methods. Because a high portion of augmentations with $o = 0$ are not able to produce good data augmentation results, this is introduced as an exit criterion.

> *Finding:* Sources should not be used for data augmentation, if the overlap with the recipient unit is zero.

It has been stated earlier that these sources have only been included in the case study to show this difference. From a conceptual point of view, it is

[6] The statistics for the ANOVA model excluding sources with $o = 0\%$ can be found in table 9.12 on page 308 in the appendix.

very unlikely to acquire a source for data augmentation that has no overlap to the customers at all. Since the objective of data augmentation is to get more information on the customers, it is more than reasonable to search for sources that have at least a small overlap in order to make statements on the customers.

While conditional mode imputation is inferior to the other methods as found in chapter 6, there is an intersection between the two regression lines of nearest neighbor hot deck and logistic regression. It shows that nearest neighbor hot deck is not in general superior to logistic regression. This is only true for high number of overlapping units. For low number of overlapping units, logistic regression is the better choice.

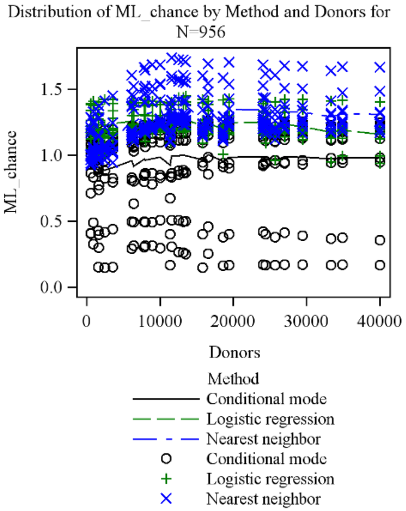
> *Finding:* Nearest neighbor hot deck is better suited for high levels of overlap. Logistic regression is better suited for low levels of overlap.

7.2.3 Number of donors and surplus

The second source characteristic to be examined is the size of the source in terms of donors. The surplus denotes the delta between the number of overlapping units o and the number of donors d.

Figure 7.5[7] shows the distribution of ML_{chance} values by number of donors and applied augmentation method with non-representative sources. As decided before, only the augmentations with $o > 0$ are taken into consideration for further analysis. The influence of the number of donors on the model lift is not as clear as for the overlap. Not all residuals are normally distributed as confirmed by the Shapiro-Wilk statistic, so that the ANOVA results should be analyzed with care – although ANOVA is relatively robust regarding deviations from the normality requirement, if the number of observations is high (Huber, 2008, p. 2/4). The homoscedasticity

[7] The colored version of this figure can be found online on www.springer.com under the title of this publication. Table 7.5b only contains selected measures for illustration purposes.



Scatter plot

Donors	N	Mean	Sh.-W.	p-val.
Conditional mode				
585	11	0.90	0.78	0.01
11,560	11	1.00	0.79	0.01
40,000	11	0.98	0.72	<.001
Logistic regression				
585	11	1.23	0.95	0.63
11,560	11	1.27	0.80	0.01
40,000	11	1.16	0.87	0.07
Nearest neighbor				
585	11	1.06	0.86	0.06
11,560	11	1.39	0.78	0.01
40,000	11	1.31	0.78	0.01

Method	Test	F	p
Cond. mode	ANOVA	0.13	1.000
	Levene	0.10	1.000
Log. reg.	ANOVA	1.44	0.076
	Levene	0.19	1.000
Near. neigh.	ANOVA	5.74	<.001
	Levene	0.25	1.000

Measures

Figure 7.5: Distribution of all ML_{chance} values by number of donors and applied augmentation method with non-representative sources ($o > 0$)

requirement is met for all methods. Regarding ANOVA's F statistic alone, the means are equal for conditional mode imputation, but not for logistic regression and nearest neighbor hot deck. When regarding the parameter estimates for the solutions of the ANOVA models, the relationships turn out to be more complex[8]. For logistic regression, the mean ML_{chance} is lowest at the highest level of overlap ($d = 40,000$). It increases unsteadily with decreasing number of donors until $d = 12,579$, where it averages at $ML_{chance} = 1.27$. At lower levels of number of donors, it does not decrease significantly. For nearest neighbor hot deck, the means do not vary significantly between $d = 6,168$ and $d = 40,000$. The means of $d = 3,583$ and less are significantly lower than the average model lift of the $d = 40,000$

[8] The table of parameter estimates for the ANOVA model solution for conditional mode imputation, logistic regression, and nearest neighbor hot deck can be found in tables 9.13, 9.14, and 9.15 on page 309 in the appendix.

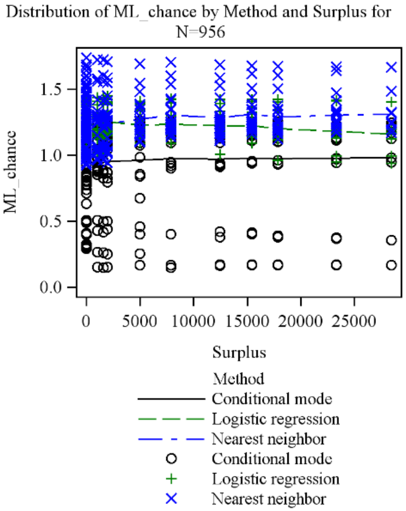
sources. Because the model lift does not increase significantly for sources where the number of donors is high, the stated hypothesis regarding the number of donors can only partially be proven.

> *Partial proof of hypothesis 2b:* For nearest neighbor hot deck, the model lift increases with increasing size of the donor unit until a certain level of donors. Then, additional donors do not add to the model lift.

Because both the overlap o and the number of donors d have a positive influence on the model lift for nearest neighbor hot deck, the positive influence of d has to be partially attributed to o. The interaction with the overlap still has to be examined in order to detect the true predictive relationship of the source characteristics regarding the model lift. However, the finding above is correct in itself. If the overlap of a source is not known, the size can be a relevant indicator when judging the suitability of a source for data augmentation.

In figure 7.6[9], the overlap has been subtracted from the number of donors in order to examine the influence of the surplus on the augmentation results. In that examination, the surplus does not have any influence on the model lift for any of the methods. All ANOVA statistics, or Welch's ANOVA alternatively used, if $F > 1$ for Levene's test, are insignificant. It becomes clear that most of the variability of the model lift regarding the number of donors is attributed to the overlap. The third hypothesis that the model lift decreases with increasing size of the donor unit, given a certain overlap between donor and recipient unit for nearest neighbor hot deck and conditional mode imputation cannot be proven. From the individual examination of influencing parameters, no overall advice on the reactiveness of the model lift pertaining to changes in the parameter settings can be given. A joint

[9] The colored version of this figure can be found online on www.springer.com under the title of this publication. Table 7.6b only contains selected measures for illustration purposes.



Scatter plot

Surplus	N	Mean	Sh.-W.	p-val.
Conditional mode				
0	110	0.95	0.82	<.001
7,870	22	0.98	0.73	<.001
28,440	11	0.98	0.72	<.001
Logistic regression				
0	109	1.25	0.89	<.001
7,870	22	1.24	0.84	<.001
28,440	11	1.16	0.87	0.07
Nearest neighbor				
0	110	1.25	0.96	<.001
7,870	22	1.31	0.85	<.001
28,440	11	1.31	0.78	0.01

Method	Test	F	p
Cond. mode	ANOVA	0.05	1.000
	Levene	0.23	0.993
Log. reg.	ANOVA	0.05	1.000
	Levene	2.07	0.026
	Welch's	0.37	0.960
Near. neigh.	ANOVA	0.94	0.497
	Levene	1.42	0.170
	Welch's	0.95	0.492

Measures

Figure 7.6: Distribution of all ML_{chance} values by surplus and applied augmentation method with non-representative sources ($o > 0$)

analysis of the parameters is necessary in order to detect interactions and to extract the marginal influences of the parameters on the model lift.

7.3 Overall estimation of influences

It has already been stated that the best method for each augmentation differs. After having abandoned the sources with zero overlap, the best methods are regarded more thoroughly. From the previous plots, it has become clear that conditional mode imputation is generally inferior to the other methods. In figure 7.7[10], the best method for the remaining source-target combinations is shown as measured by ML_{chance}. If the two best methods are very close to each other, the case is marked as ”tie”. A case

[10] The colored version of this figure can be found online on www.springer.com under the title of this publication.

is considered a tie, if the model lift for the method performing best is less than 1% better than the model lift of the second best method. The respective square is blue, if nearest neighbor hot deck performs best for this combination. It is green for logistic regression, black for conditional mode imputation, and gray otherwise.

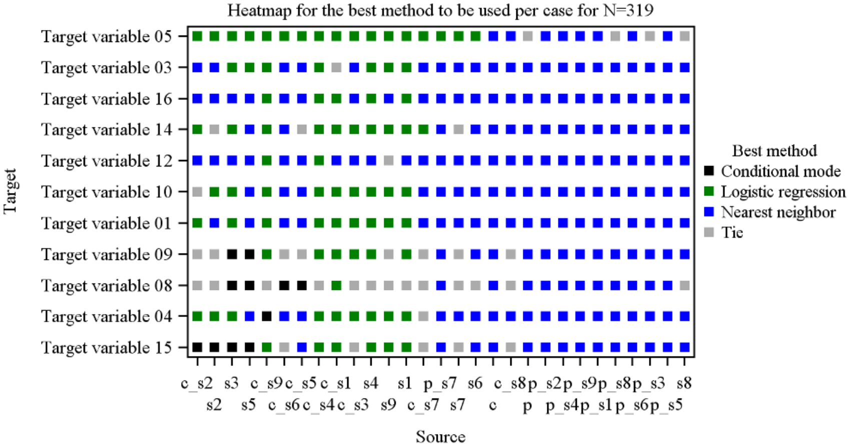


Figure 7.7: Best method per source-target combination as measured by ML_{chance} for all cases using sources with a 100% sampling rate and $o > 0$

60% of the source-target combinations perform best with nearest neighbor hot deck and 25% perform best with logistic regression. 12% have a tie and only 3% perform best using conditional mode imputation. There are sources that are best augmented using nearest neighbor hot deck, independent of the augmented target variable, e.g. $D = R$ (c), $D = P$ (p), and most sources of the form $D \supset R$ (p_s1, ..., p_s9, except for p_s7). Other sources largely perform best using logistic regression, e.g. $D \subset R$ (c_s1, c_s4, c_s9) and $D \cup R$ (s1, s4, s9). As 3% is not a relevant portion and the distance of the model lift of these data augmentations is on average only 2% better than the second best method, it can safely be argued that conditional mode imputation is not a relevant method in the data augmentation context.

> *Finding:* Conditional mode imputation is generally inferior to nearest neighbor hot deck and logistic regression. It does not need to be considered in the method question.

The database marketing analyst's decision before approaching a data augmentation is therefore reduced to a decision between nearest neighbor hot deck and logistic regression. If only regarding these two methods, 65% of the source-target combinations perform best using nearest neighbor hot deck and 35% perform best using logistic regression.

7.3.1 Regression models estimating the model lift

The source characteristics and other influencing factors can have interactions that are not captured when regarding them in an isolated way and plotting only one dimension pertaining the model lift, as it has been done in chapter 7.2. Therefore, a regression model is built, which includes the source characteristics and the target variable characteristics. The overlap o, the number of donors d, and the number of link variable classes v are considered, as well as the number of target values k, a measure for the skewness of the target variable IQV, and a measure for the predictive power of the link variables regarding the target variable R^2. By including the target variable characteristics, an across case analysis is possible.

$$ML_{chance} = \beta_0 + \beta_1 \times o + \beta_2 \times d + \beta_3 \times v + \beta_4 \times R^2 + \beta_5 \times k + \beta_6 \times IQV \quad (7.5)$$

A multiple linear regression model is built separately for every method in order to assess the relevant influencing factors for each. Besides the parameter estimate and the t-value, a standardized estimate is calculated in order to assess the strength of the influence of relevant parameters in explaining the model lift. Even if the t-value is significant, the influence of the beta value can be low, if the standardized estimate is much smaller than that of other predictive variables.

Dependent variable: ML_{chance}, N=319

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	6	7.47478	1.24580	165.94	<.0001
Error	312	2.34235	0.00751		
Corrected Total	318	9.81714			

Root MSE	0.08665	R-Square	0.7614
Dependent Mean	1.26545	Adj R-Sq	0.7568
Coeff Var	6.84704		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t-Value	Pr> |t|	Standardized Estimate
Intercept	1	−0.20868	0.08438	−2.47	0.0139	0
Overlaps	1	0.00003269	0.00000183	17.89	<.0001	0.73679
Donors	1	−0.00000238	6.487285E−7	−3.67	0.0003	−0.14778
VCs	1	−0.00001871	0.00000633	−2.96	0.0033	−0.10497
Rsqu	1	0.85589	0.06315	13.55	<.0001	0.57154
k	1	0.17438	0.00816	21.37	<.0001	0.76679
IQV	1	0.78132	0.06964	11.22	<.0001	0.52095

Table 7.6: Regression parameters for nearest neighbor hot deck for all cases using sources with a 100% sampling rate and $o > 0$

Table 7.6 shows the regression parameters for nearest neighbor hot deck[11]. 319 cases using nearest neighbor hot deck have been used to inspect the influence of the explaining parameters regarding the model lift (all non-representative sources with $o > 0$). The influence can be derived from the table of parameter estimates. The sign of the parameter estimate shows whether the parameter has a positive or negative influence on the model lift. If the p-value of the parameter estimate is smaller than 0.05, the influence is significant. The standardized parameter estimate shows the relative weight of parameters in explaining the model lift. The target

[11] The applicability of linear regression regarding the required assumptions for this model is discussed and confirmed with figure 9.1 and table 9.16 on page 312 in the appendix.

variable characteristics and the predictive power have a significantly positive influence on the model lift, as expected. Their influence demonstrates the necessity of including the target variable characteristics in a model for across case analysis. The range of the model lift increases with increasing k and IQV, so that the positive influence is obvious. The predictive power is a condition precedent to data augmentation and has a positive influence, too.

In the nearest neighbor hot deck model, the overlap has a significantly positive influence on the model lift, which reinforces the proof of hypothesis 2a. It is the second most important parameter as measured by the standardized estimate. The donors have a significantly negative influence on the model lift. Its influence is not as big as the overlap's. However, by including both the overlap and the number of donors in the regression model, the isolated influence of the number of donors given the overlap is observable. Now, the stated hypothesis can be confirmed.

> *Proof of hypothesis 2c:* For nearest neighbor hot deck, the model lift decreases with increasing size of the donor unit, given a certain overlap between donor and recipient unit.

The number of link variable classes has a significantly negative influence on the model lift. However, for practical applications, the influence of the number of donors and of the number of link variable classes should not be overestimated. As shown by the standardized estimate, their influence is relatively small in comparison to the overlap. All in all, the overall hypothesis can be proven for the nearest neighbor hot deck method.

> *Proof of hypothesis 2:* The augmentation results of nearest neighbor hot deck are influenced by the source characteristics.

The overall fit of the nearest neighbor hot deck regression model is good ($R^2 = 76\%$). We have been able to isolate the influencing factors of the model lift for nearest neighbor hot deck to a relevant portion. By focusing on sources with a high overlap and low surplus, database marketing analysts will reach good model lifts for data augmentations using nearest neighbor hot deck methods.

Dependent variable: ML_{chance}, N=318

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	6	1.51104	0.25184	49.35	<.0001
Error	311	1.58721	0.00510		
Corrected Total	317	3.09825			

Root MSE	0.07144	R-Square	0.4877
Dependent Mean	1.22972	Adj R-Sq	0.4778
Coeff Var	5.80938		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t-Value	Pr> |t|	Standardized Estimate
Intercept	1	0.37085	0.06957	5.33	<.0001	0
Overlaps	1	0.00001160	0.00000151	7.69	<.0001	0.46324
Donors	1	−0.00000339	5.348742E−7	−6.33	<.0001	−0.37279
VCs	1	−0.00001855	0.00000522	−3.56	0.0004	−0.18490
Rsqu	1	0.58353	0.05207	11.21	<.0001	0.69350
k	1	0.09471	0.00673	14.07	<.0001	0.74055
IQV	1	0.57043	0.05743	9.93	<.0001	0.67635

Table 7.7: Regression parameters for logistic regression for all cases using sources with a 100% sampling rate and $o > 0$

Table 7.7 shows the regression parameters for logistic regression[12]. 318 augmentations have been used to build the model for the logistic regression results[13]. The number of target values, IQV, and the predictive power have a significant positive influence on the model lift. They account for the greatest part of the variance as measured by the standardized estimates.

The overlap has a significantly positive influence on the model lift. However, it is not as high as in the nearest neighbor model. Likewise, the number

[12] The applicability of linear regression regarding the required assumptions for this model is discussed and confirmed with figure 9.2 and table 9.17 on page 313 in the appendix.

[13] The number of observations is smaller than for the nearest neighbor hot deck method due to the quality check selection described in chapter 6.2

of donors has a negative influence, but not as strong as for nearest neighbor hot deck. The number of link variable classes has a weak, but significantly negative influence on the model lift. The model lift is to the greatest part explained by R^2, k and IQV. The overall hypothesis that logistic regression is not influenced by the source characteristics needs to be rejected.

> *Rejection of hypothesis 3:* The augmentation results of logistic regression *are* influenced by the source characteristics. However, the influence is not as strong as for nearest neighbor hot deck.

The remark that the model lift is not as strongly influenced by the source characteristics for logistic regression as for nearest neighbor hot deck is also confirmed by the low fit of the regression model. Only 49% of the variance of the model lift is explained by the source and target variable characteristics. There seem to be other influences, which have not been captured by the conceptual model.

7.3.2 Estimating the best method to be chosen

In order to find the influencing factors regarding the choice between logistic regression and the nearest neighbor hot deck methods, a logit model is built with the best method choice for each source-target combination as its output variable. The model uses the same predictors as the previous models (o, d, v, k, and IQV). Only R^2 is omitted, because it is not directly observable and because its calculation involves certain obstacles, making it a rather unstable predictor. In the following equation, NN stands for nearest neighbor to be the best method to be used and LR for logistic regression.

$$Logit(Y_{NN/LR}|X) = \beta_0 + \beta_1 \times o + \beta_2 \times d + \beta_3 \times v + \beta_4 \times k + \beta_5 \times IQV \quad (7.6)$$

If this model is able to discriminate well between the two methods, the influencing factors can be found for an optimal method choice. The model is built upon a 50% random test sample of the augmentations. It is then validated against the other 50% control sample in order to have a robust test for the third hypothesis.

Model Information

Data Set	TESTSAMPLE
Response Variable	Method
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Method	Total Frequency
1	Logistic regression	52
2	Nearest neighbor hot deck	108

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr>ChiSq
1	VCs	1	4	0.3590	0.5490
2	Donors	1	3	1.3970	0.2372
3	IQV	1	2	2.8191	0.0931

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Stand. Error	Wald ChiSq	Pr>ChiSq	Stand. Estimate
Intercept	1	12.6321	2.7664	20.8513	<.0001	
Overlaps	1	−0.00100	0.000191	27.6046	<.0001	−2.1516
k	1	−2.4502	0.6827	12.8825	0.0003	−1.0878

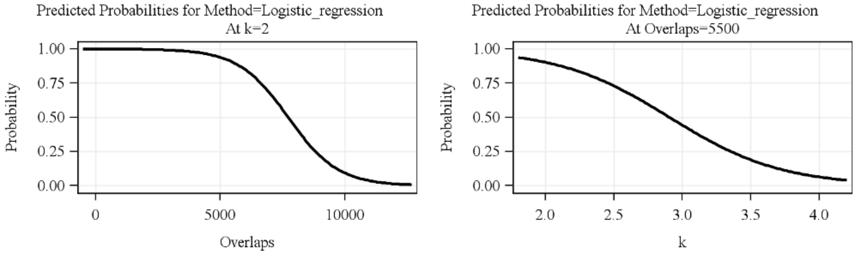
Fit Statistics for SCORE Data

Data Set	Total Frequency	Log Likelihood	Misclassification Rate
TESTSAMPLE	160	−34.3125	0.0875
CONTROLSAMPLE	159	−53.2837	0.1132

Table 7.8: Logistic regression parameters for the best method to be used for all cases using sources with a 100% sampling rate and $o > 0$

The model is built for a test sample of 160 observations, of which 52 have performed best with logistic regression and 108 have performed best with nearest neighbor hot deck. The probability modeled is method='Logistic regression' per case, i.e. per source-target combination. A backward elimination procedure is used in order to keep those parameters that have a significant influence on the best performing method. During backward elimination, the number of donors, the number of link variable classes, and IQV are removed. Their p-values are shown in the summary of the backward elimination table. Although they influence the model lift, they do not have a share in predicting the best method to be used. Only the overlap and the number of target values k have a significant effect on the method performing best. They both have a negative influence in the model, which means that nearest neighbor hot deck should be chosen when the both overlap and number of target values are high. Logistic regression should be chosen for low levels of overlap and k. Of these effects, the influence of the overlap is twice as high as the influence of k (as measured by the standardized estimate).

The overall model fit is good: the misclassification rate in the test sample is 9%. The control sample is scored using the determined parameters. The misclassification rate of 11% is only slightly higher. The model is therefore very stable, delivering accurate decisions with a 11% level of significance. Moreover, for the cases where observations have been misclassified, the average difference between the model lift for logistic regression and nearest neighbor hot deck is 0.03. This shows that using the method performing worse in these cases does not significantly lower the augmentation results. On the contrary, the difference between the model lift for logistic regression and nearest neighbor hot deck is 0.11 for correctly classified observations.



Effects plot for the overlap

Effects plot for k

Figure 7.8: Effect plots for the best method to be used regarding the overlap and the number of target values for all cases using sources with a 100% sampling rate and $o > 0$

The effect plots in figure 7.8 show the interactions of the two effects. In figure 7.8a, the decision of the overlap is shown when $k = 2$. Given this number of target values, the inflection point of the overlap is at roughly 70% overlap rate. Sources with an overlap rate of less than 70% should be augmented using logistic regression, sources with an overlap rate of more than 70% should be augmented using nearest neighbor hot deck. The inflection point changes when changing k. For example, it would be around 50% overlap rate for $k = 3$. A similar plot is shown in figure 7.8b. For an overlap rate of almost 50%, logistic regression is the best choice for $k = 2$ and nearest neighbor hot deck for $k = 4$. For $k = 3$, there is a tie. This can be seen in figure 7.8b, where the plotted line has a probability of approximately 50% for each method, given $k = 3$. It becomes clear that the effects interact. There is no single best solution to the method decision problem. However, several findings can be derived that are helpful in practice.

> *Proof of hypothesis 4:* Logistic regression is more suitable for low levels of overlap. For high levels of overlap, nearest neighbor hot deck is the more powerful method. Thus, there is a definable set of source characteristics for which nearest neighbor hot deck performs better than logistic regression in terms of model lift, and vice versa.

This has already been shown earlier. The overlap has been extracted to be the only relevant source characteristic in the decision between logistic regression and nearest neighbor hot deck for data augmentation. The other source characteristics influence the model lift, but they do not influence the decision for the method. Sources with a high overlap are close to exact matching problems. For these, a nearest neighbor hot deck method is most meaningful, because it attempts to find an exact match for every recipient. In contrast, logistic regression detects overall correlations and thus is suitable for sources with a low overlap to the recipient unit. For data augmentation applications outside our case study, the exact inflection point can vary. It depends on the interactions with the other parameters and the general data augmentation context.

> *Finding:* Logistic regression is more suitable for low numbers of target values. For high numbers of target values, nearest neighbor hot deck is the more powerful method.

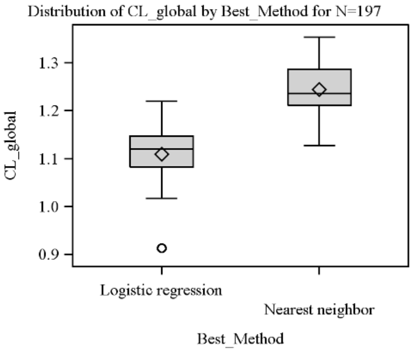
Polynomial logistic regression is able to handle various values for the explained variable. However, it works best for binary output and becomes less stable with an increasing number of target values. This is also confirmed by our model results. Nearest neighbor hot deck can handle more target values better. At the same time, it should be kept in mind that the overlap has a stronger influence on the model for the best method than k. Consequently, a source with a very low overlap should always be augmented using logistic regression, while a source with a very high overlap should always be augmented using nearest neighbor hot deck. Only for sources with an intermediate overlap, the target value number is relevant.

7.3.3 Conversion probability and conclusion

The model lift is a valuable measure if segmentation is the database marketing goal. If a selection is supposed to be based on data augmentation results, the ability of the augmentation results to narrow down the target group is of interest. In the conceptual model, the model lift is a mediator for the CPL. There is a strong correlation between ML_{chance} and both CL_{global} and CM_{global}, even if no other influencing factors are taken into account. The Pearson correlation coefficients are 0.53 for ML_{chance} and CL_{global}, and 0.59 for ML_{chance} and CM_{global}. The respective p-value is below 0.0001 for each of the correlations.

It is of interest for the research question whether the data augmentation results lead to a significant CPL after applying all rules stated. The quality check must yield at least two target parameter values after augmentation, otherwise they are no longer regarded, as described in chapter

6.2.1. Sources with $o = 0$ are not suitable for data augmentation, thus all augmentations with such a source have been exited, as described in chapter 7.2.2. Conditional mode imputation methods are generally inferior to logistic regression and nearest neighbor hot deck, thus all augmentations using conditional mode imputation have been exited, as described in chapter 7.3. Of the remaining augmentations, the methods having been classified performing best with either nearest neighbor hot deck and logistic regression are kept, as described in chapter 7.3.2. Because the decision making process is retraced here, the method to be chosen based on the logistic regression model is used, as opposed to the augmentation actually having performed best (they differ in 10% of the cases). Because the 197 MAR sources are of primary interest, the MCAR sources are not used to validate the final hypothesis. The 122 remaining MCAR sources all have a significant CPL, with a mean of 1.28.



Box plot

Method	Log. reg.	Near. neighbor
N	89	108
Max	1.22	1.35
Q3	1.15	1.29
Median	1.12	1.24
Mean	1.11	1.24
Q1	1.08	1.21
Min	0.91	1.13
Sh.-W.	0.85	0.98
p-value	<.001	0.08

Test	*F*	p
ANOVA	270.28	<.001
Levene	2.71	0.101
Welch's ANOVA	258.87	<.001

Measures

Figure 7.9: Distribution of all CL_{global} values given the best method for sources other than MCAR with $o > 0$

Figure 7.9 shows the distribution of the CPL among the chosen augmentations. The augmentations for which the previous model suggested logistic regression vary between 1.08 and 1.15 for half the observations. For nearest neighbor hot deck, it varies between 1.21 and 1.29. The nearest neighbor hot deck results are far better than the logistic regression results – the mean difference as confirmed by Welch's ANOVA is obvious. However, after the allocation of the best method, nearest neighbor hot deck is used for sources with a high overlap and logistic regression is used for sources with a low overlap. The low CL_{global} are not due to the applied method, but due to the low overlap. In these cases, nearest neighbor hot deck would have performed even worse.

All CPLs greater than 1 confirm the stated hypothesis: If all decisions are made correctly, data augmentation results are able to significantly increase conversion probabilities, compared to randomly selected target groups. This is true for 98% of the cases. Only five augmentations using logistic regression did not reach this goal. They are marked as outliers in the box plot. But also for logistic regression, the error rate is below 5%. With a level of significance of 5%, the stated hypothesis can be confirmed.

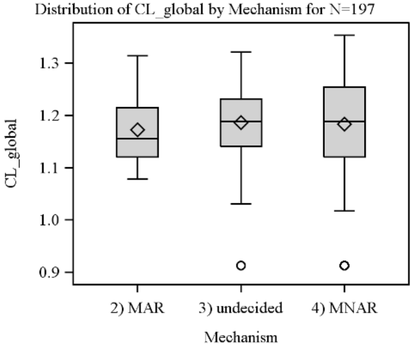
> *Proof of hypothesis 5:* If all decisions are made correctly –
>
> - at least two augmented target parameter values
> - no sources with $o = 0$
> - no use of conditional mode imputation
> - choice of the method based on overlap and number of target values
>
> – data augmentation results are able to significantly increase conversion probabilities, when compared to randomly selected target groups.

When closer analyzing the five augmentations that did not reach the goal, it can be observed that three of these augmentations did not discriminate perfectly between the target values. Of the three or four possible target values, only two or three were augmented. This is still a valid data augmentation as defined in chapter 6.2. However, it is evidence of an imperfect augmentation. From the 192 significant CPLs, 22 also fall into that category (valid augmentations with not all target values reproduced). They show a good CPL, nevertheless. Consequently, the failure to reproduce all target values for target variables with more than two target values cannot be introduced as an exit criterion in general. It should, however, be approached carefully, when already in doubt whether the data augmentation will perform well.

> *Finding:* If other criteria in the decision process are already close to exit criteria (e.g. very low overlap or low IQV), sources should not be used for data augmentation, if not all parameters are reproduced by the augmentation.

In chapter 7.1, we did not finally answer the question on whether the source data mechanism significantly compromises the data augmentation results. In figure 7.10, the distribution of all CL_{global} values is shown pertaining to the source data mechanism as defined by the results of the χ^2 and CMH tests applied in chapter 7.1.1. After having established the rules stated, 197 data augmentations remained of which 18 were formerly categorized as MAR, 62 as "undecided", and 117 as MNAR.

The means of CL_{global} do not differ significantly for the three source data mechanism categorizations. Welch's ANOVA is used to assess the mean differences, because the variances are not equal as measured by Levene's test. The F value of Welch's ANOVA is smaller than 1, so that a significant difference among the means cannot be observed. This means that if all decisions are made correctly, the associations between source data mechanism and target variables, given the link variables, is secondary. This is not only noticeable from a theoretical point of view. It is also of practical interest, because the source data mechanism is not assessable outside of a simulated case study context. But if the source data mechanism has no practical meaning, respecting the rules found, it does not have to be assessed in practice. This is a central finding of our study, because it conveniently solves the problem of the unfeasible conditional independence calculation.



Box plot

Mech.	MAR	undec.	MNAR
N	18	62	117
Max	1.31	1.32	1.35
Q3	1.22	1.23	1.25
Median	1.16	1.19	1.19
Mean	1.17	1.19	1.18
Q1	1.12	1.14	1.12
Min	1.08	0.91	0.91
Sh.-W.	0.93	0.96	0.97
p-value	0.19	0.06	<.001

Test	F	p
ANOVA	0.17	0.843
Levene	1.66	0.192
Welch's ANOVA	0.24	0.788

Measures

Figure 7.10: Distribution of all CL_{global} values by assigned source data mechanism based on the χ^2 and CMH tests for sources other than MCAR, with $o > 0$, using the best augmentation method

Finding: If all established decisions are made correctly, data augmentation results are able to significantly increase conversion probabilities, even if the source data mechanism is MNAR.

This can also be argued from a contextual point of view. In marketing applications, where humans are the elements of interest, correlations between two variables are seldom very strong. The reasons why persons take part in a survey, for instance, are manifold and cannot be explained by a single observable variable. Likewise, there will not be a clear conditional dependence between this source data mechanism and a target variable in the survey, given a wide range of socio-demographic and other variables. This lack of correlation becomes an advantage in data augmentation. If the source data mechanism can be ignored for any source, it does not need to be assessed upfront before approaching a data augmentation project. This is convenient, as the source data mechanism is not observable and would only be able to be roughly estimated with an auxiliary source.

The same is true for the "customer data mechanism" as described in chapter 3.1.1. Theoretically, it is possible that there is a correlation between the target variable values and the fact that a person is a customer of a company, e.g. interest in shoes is related with a person being a customer of a shoe store. Just as for the source data mechanism, it can be assumed that a strong correlation between link variables and the target variable balances a potential correlation between source data mechanism and target variable.

With the stated findings, a good upfront evaluation of data augmentation sources is possible. It is decidable which augmentation method should be used. Even if no total correct classification rates, model lifts, and CPLs can be calculated, the database marketing analyst can be sure that the data augmentation significantly improves the data basis for segmentation and data selection tasks.

KPI	Minimum	5th Pctl	Q1	Median	Q3	Maximum
ML_chance	1.003	1.123	1.180	1.210	1.323	1.689
ML_target	0.374	1.102	1.679	2.226	2.613	6.453
ML_uniform	0.727	0.929	1.007	1.052	1.105	1.248
ML_source	1.168	1.207	1.291	1.417	1.666	6.259
CL_global	0.913	1.031	1.120	1.187	1.248	1.354
CM_global	-0.185	0.047	0.250	0.407	0.510	0.848

Table 7.9: Final measures for data augmentation results complying with the established rules for $N = 197$ sources other than MCAR, with $o > 0$, using the best augmentation method

Table 7.9 summarizes the KPIs of all augmentations complying with our established guidelines. The minimum and maximum values observed are given, as well as the significance border of the 5% percentile. Lower quartile, median, and upper quartile border the main part of the observations. It can be seen from table 7.9, that all ML_{chance} and ML_{source} values are greater than 1 for all observations complying with the rules. ML_{target} has some outliers that lead to values below 1. However, it can be said with a 5% level of significance that our rules lead to significant ML_{target} values, too. The strictest KPI, $ML_{uniform}$, is greater than 1 for 75% of the observations as bordered by the lower quartile. 25% of the observations did not pass the quality check. This illustrates the fact that data augmentations usually only lead to a small information increase for companies. It is often not much better in terms of hits than allocating the most frequent value to all customers. However, the data augmentation results enable target group selection, while a uniform allocation of target values does not.

As stated earlier, CL_{global} is greater than 1 for more than 95% of the observations. It can be said with a 5% level of significance that our rules lead to significant CPL, which answers the research question. The CM_{global} values are positive for 95% of the observations, which is equally desirable. Half of the observations lead to CM_{global} values between 25% and 51%. It means that the augmentation results are able to come as close as 25% to 51% to a perfect selection, as compared to a random selection. Half of the information increase possible is reached by data augmentation based on the established rules.

7.4 Managerial implications

The findings of the case study are used to state guidelines regarding an ex ante suitability check for potential data augmentation sources. We were able to prove that if all decisions are made correctly, data augmentation results are able to significantly increase conversion probabilities, when compared to randomly selected target groups. Our findings regarding the source data mechanism, the overlap, and the best method to choose are summarized here and illustrated in a one page graphical overview on page 276.

Source characteristics

The source characteristics overlap, size, and number of link variable classes influence the augmentation results. The size is countable from the donors contained in the source. The overlap can directly be observed, if there is a customer indicator variable available in the source. In a survey, this would equal the question "Have you ever bought something from this company or brand?" or similar. Sources which are not a subgroup of the customer group and which do not contain a customer indicator variable are more challenging. In this case, like for the source data mechanism, auxiliary information is needed if to estimate the overlap e.g. by means of expert knowledge or otherwise.

The results of data augmentation improve with increasing overlap. For a given overlap, however, the results can decrease with increasing size. In this case, and if the overlapping units are identifiable, it can be reasonable to omit the surplus donors. Both measures can be used to assess the expected quality of the data augmentation. The smaller the source, the less information is contained and the less likely all values are reproduced. It has been shown in our case study that an important portion of augmentations with a zero overlap did not yield significant model lifts.

> **Overlap guideline:** The higher the overlap between the donor unit and the recipient unit, the better the data augmentation results. If there is no overlap, the hit rate of the results does not outperform a random distribution of target values. In this case, sources should not be used for data augmentation.

If not all target values are reproduced after data augmentation, the reasons should be reviewed carefully. If only one target value is reproduced for all recipients, the link variables were not able to discriminate between target values. These augmentations are invalid. They do not lead to an increase in information. If two out of three or three out of four values are reproduced, the results can still be reasonable. However, if already in doubt whether the data augmentation will perform well because of other criteria, sources should not be used for data augmentation, given not all parameters are reproduced by the augmentation.

Methods

We made certain suppositions in order to motivate the general choice of applied data augmentation methods. The following guidelines are only valid in this context and cannot be transferred, if the data augmentation set up is different. No inference is sought to be made between the augmented target variables. Accordingly, a univariate pattern approach has been chosen. Otherwise, the values from all target variables for one recipient would have to be augmented from the same donor in order to preserve correlations. No inference is supposed to be made between augmented target variables and auxiliary variables present in the customer database. Consequently, we do not need to regard the case in which variables have never been jointly observed. Otherwise, the augmented target values would have to be different in order to contribute to the four levels of validity stated earlier. Furthermore, the target values are not known for any of the recipients. Likewise, the overall distribution of target values in the recipient unit is not known. If either one of the two suppositions was true, different methods would be possible and more meaningful in the respective context.

We compared conditional mode imputation, logistic regression, and nearest neighbor hot deck regarding their general suitability for data augmentation and regarding their performance under certain circumstances. Our goal was to make suggestions in which situation to use which method. We found that nearest neighbor hot deck and logistic regression are generally superior to conditional mode imputation. This is because conditional mode imputation only uses very general rules augmenting the same value for a broad class of recipients. This is often not sufficient concerning the data augmentation needs. The achieved model lifts were significantly worse than those of the other two methods applied in the same context.

MCAR sources with a small sampling rate are best augmented using logistic regression. Logistic regression detects overall relationships among the variables, which are stable even if only a small donor unit is available. For MAR sources and MCAR sources with high sampling rates or complete populations, nearest neighbor hot deck can lead to higher model lifts than logistic regression. However, the variance of logistic regression results is smaller. If no overlap can be observed, logistic regression should be chosen for its smaller general risk.

For nearest neighbor hot deck, the size has a positive influence on the model lift. It does, however, have a negative influence on the model lift, given a certain level of overlap. The number of link variable classes has a negative influence on the model lift. All in all, nearest neighbor hot deck is significantly influenced by the source characteristics. Multivariate methods are also influenced by the source characteristics, but the influence is not as strong as for nearest neighbor hot deck.

> **Method guideline:** Logistic regression is best suited for sources with a low overlap or sampling rate. Nearest neighbor hot deck is best suited for sources with a high overlap. If the overlap is not known, logistic regression should be used, because it performs equally well for all levels of overlap.

There is a definable set of source characteristics for which nearest neighbor methods perform better than multivariate methods, and vice versa. It has been shown that the overlap is the only strong indicator for choosing the best augmentation method, regarding the source characteristics. In our case, the inflection point was generally at an overlap rate of 50%. For a lower overlap, logistic regression is superior. For a higher overlap, nearest neighbor hot deck should be chosen. This can also be argued from a theoretical point of view. The smaller the distance between observations, the better the data augmentation results from nearest neighbor hot deck. The overall distance decreases with increasing overlap. The perfect state for a nearest neighbor hot deck approach is that of record linkage. Contrarily, multivariate methods detect general relationships between variables and can be applied to any subpopulation. They do not rely on the overlap.

Additionally, logistic regression is more suitable for low number of target values. For a high number of target values, nearest neighbor hot deck is the more powerful method. Logistic regression is the stronger, the more

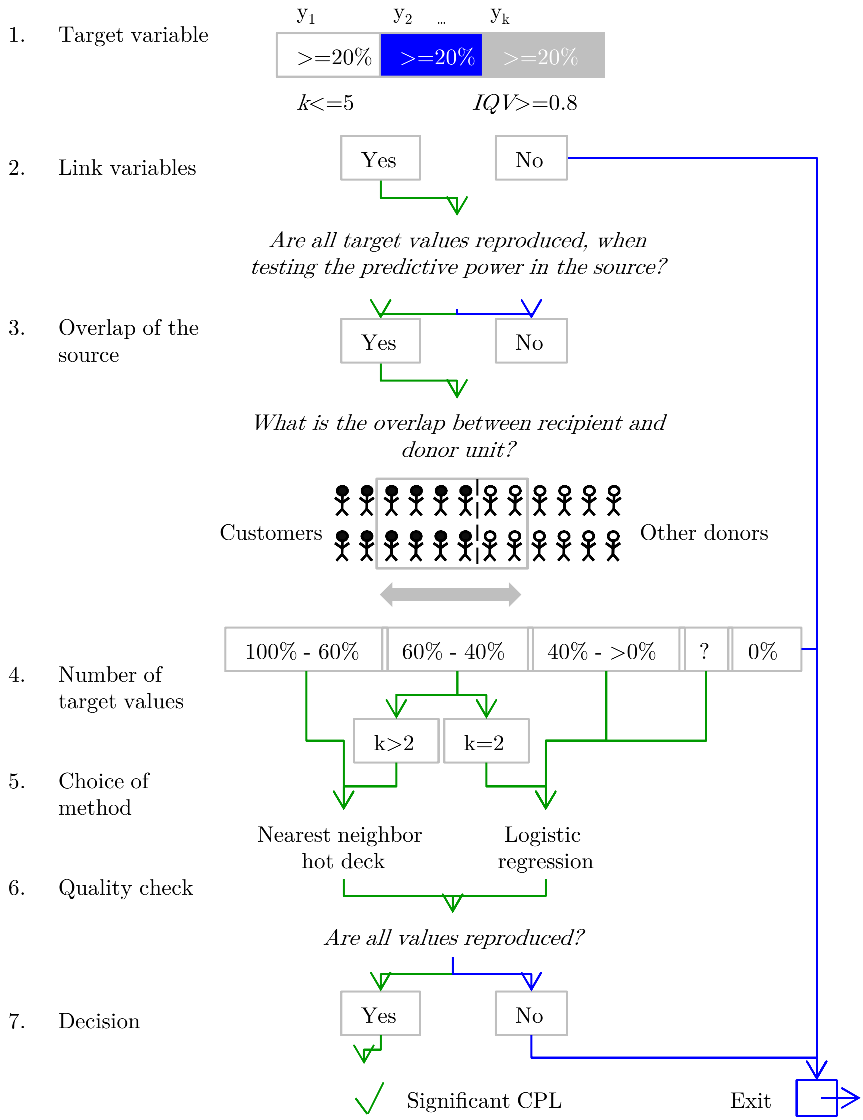


Figure 7.11: Practical guide for evaluating external sources

observations can be found in each link value class. Target class occupations of individual target values are higher, if the overall number of target values is lower. However, this rule only applies, if the overlap rate is around 50%. Otherwise, the overlap criterion is the stronger effect. Again, the exact threshold depends on the data augmentation context. It can vary for other applications and needs to be verified for other use cases. A summary of the practical guide is shown in figure 7.11 on page 276.

Source data mechanism

There are two types of source data mechanisms valid for performing data augmentation: sources with data missing completely at random (MCAR) and data missing at random (MAR). If data is missing not at random (MNAR), the fact whether a donor has been observed in a source can bias the augmentation results. The source data mechanism of MCAR sources is usually known from the study design of the source. Volunteer surveys, web data, social media sources, and others are MAR or MNAR sources. These sources usually partially overlap with the recipient unit, the source is a subgroup of the customer database, or they do not overlap at all. A source is MAR, if it is known from the source profile that selection was only made based on observable variables, or if the source and the target variable are conditionally independent given the link variables. It has been shown during the case study, how difficult the calculation of conditional associations is in a categorical setting.

In practice, an additional challenge is the fact that the mechanism that leads to donors being present in a source is not observable. Moreover, a situation in which the source mechanism indicator, the target variable, and the link variables are present, so that the conditional association can be calculated, is hardly ever available.

> **Source data mechanism guideline:** An external source does not need to be representatively sampled. If all variables are

categorical and if the link variables are able to predict the target variables well, and if all other rules are respected, the source data mechanism is negligible.

Because of these findings, we recommend to assume conditional independence, as long as there are no other substantiated reasons, e.g. expert knowledge. Moreover, we were able to show that model lifts do not significantly differ for different source data mechanisms, if all rules are respected. We found that in a categorical data augmentation context, the association of the source data mechanism and the target variable does not compromise the data augmentation results. The source data mechanism is important in theory, but not relevant in practice.

Chapter 8

Limitations of data augmentation and outlook

Data augmentation sources are manifold and oftentimes easily available. The main question arising from these sources is whether using them for augmentation purposes will lead to an increase in information and a better basis for decision making. The decision for or against data augmentation is simplified with our guidelines. However, they do not substitute a thorough case specific examination and pretesting phase.

In order to correctly use data augmentation in database marketing, it is necessary to consider what data augmentation is capable of and what it is not. In this chapter, our findings are enhanced with hints on how to manage expectations regarding data augmentation results. Data augmentation is a tool for providing information in an area of database marketing where existing customer data is sparse. It should always be regarded as such. We contemplate the limitations of data augmentations and the successional decision whether data augmentation is the right tool to answer questions in database marketing. To understand these limitations is an important factor in the appreciation of data augmentation results.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_8

We were able to answer the research question and to provide comprehensive findings regarding external sources in database marketing. However, our study has also raised new questions. We have started to examine the source data mechanism. More research should be dedicated to this topic and how different source data mechanisms can be approached in practice. Certain steps in our proposed data augmentation process can be enhanced by deeper exploration. Our case study is a comprehensive example of data augmentation settings. However, some parameters have been fixed in order to establish comparability. These parameters vary in practice. More data augmentation use cases should be regarded in order to support our findings. Eventually, we point out the difference between predictive and uplift models, the latter focusing on how customers are motivated and activated through direct marketing communication.

8.1 Limitations and alternatives

When conducting a data augmentation, managing expectations is an important part of the project. Any analyst making secondary use of the data should have a clear picture on the meaning of the available data. The augmented values are directly saved in the customer database and become an analysis basis themselves. But using it like a primary data source is dangerous, because it dilutes the analyst in knowing something not actually known (Dempster & Rubin, 1983), and data might be used for conclusions not possible from augmented data. Data augmentation can enrich the existing data in a way not possible otherwise and its results can serve as important decision criteria. However, data augmentation results are only approximations, and decisions based on the results should be treated accordingly.

Essentially, data augmentation is only an option if no complete dataset is available, or if it is impractical to collect all data from a single customer (Adamek, 1994). Sometimes, there are better possibilities to reach a particular marketing goal. Much information can already be found by mining the donor or recipient unit only. It is not always the best solution to mix up sources in order to receive valuable insights (Putten et al., 2002b).

First and foremost, all additional information is valuable to database marketing analysts. In practice, targeting formulas are oftentimes not very sophisticated. Knowledge on the majority of customers is limited and predictions of conversion probabilities are associated with great uncertainty. The benefits of data augmentation oftentimes outweigh the risks, so that data augmentation has become a popular tool (Adamek, 1994). There is a trade-off between the economical contribution of the new information and the cost related to the augmentation project. Because marketing is usually afflicted with not knowing very much on most of the customers, any additional information is valuable and money is attributed to this cause. Whenever target group decisions can be facilitated and direct marketing campaigns become more efficient, a data augmentation project is worth the effort. This should be kept in mind when comparing the usefulness of data augmentation to the challenges and limitations to be stated below.

The knowledge achievable from data augmentation is limited. Because of the categorical nature of variables, data augmentation information is usually not very precise. One could even argue that it is misleading to say that data is really augmented. It combines the already existing link variables in a way that conversion probabilities for certain target variables can be predicted. Not the target values are the new information, but the model combining the link variables pointing to the target value with a certain probability. The information has thus already inherently been present in the data. This combination of variables is the value of the data augmentation results.

The augmented data is always only as good as the data in the source. Data describing human behavior and preferences is volatile and may age quickly. These facts should be considered before approaching a data augmentation project and should be evaluated carefully for any source. Obstacles related to the quality and usefulness of the data are:

- *Data quality:* The results depend directly on the data quality of both recipient and donor unit. Some values might be deficient or missing and some values might not convey the intention of the person from whom it was collected.

- *Availability of link variables in the recipient unit:* In order to perform data augmentation, some existing knowledge on the customers is necessary. In fact, the more information is already known, the better more information can be augmented. This is not necessarily in accordance with the marketing strategy for data augmentation, which often involves receiving a better picture of customers on whom not so much information is available already.

- *Comparability of link variables:* Every data source contains variables collected according to specified concepts and definitions, saved in a specific format and scale. Link variables can only be used, if concepts and definitions are similar, and if formats and scales can be adjusted to be the same.

- *Correlation between link and target variables:* If the link variables are not able to predict the target variables, data augmentation is not possible.

- *Meaning and interpretation of variables:* Some variables might not have the meaning that the database marketing analyst ascribes to it. The donor unit is often a source whose data was collected for a different purpose. There are hard facts like age or gender, which are universally understood. When talking about interests, the situation is not as clear. If a woman records to be interested in shoes by adding this interest in her Facebook profile, it might mean a lot of things to her. The database marketing analyst would interpret the interest in shoes as the intention to buy shoes. While there certainly is a correlation between being interested in shoes and buying them, the two notions have a slightly different meaning. "Borrowing" someone else's data for a different project may lead to bias because of differences in interpretation one is not aware of (Ozimek, 2010).

- *Usefulness of target variables:* Even if the meaning of a variable is captured correctly, the variable might still not be a direct hint for the conversion probability of an offer. Targeting is the application of a mixture of variables, of which every variable is expected to have predictive power regarding the conversion. However, this predictive power cannot be tested upfront. It can only be observed after having conducted a marketing campaign. If a person is interested in shoes, and also likes to buy shoes, it might still not mean that this intention is influenced by a campaign. The conversion probability concept is more complex and thus manipulating it is not trivial.

It cannot be expected from data augmentation to exactly reproduce every single value. It has been shown in the past that many statistics judging data augmentation results were within acceptable limits of the real values (Baker et al., 1989). Sometimes, it makes sense to illustrate the uncertainty associated with data augmentation by introducing uncertainty bounds (D'Orazio et al., 2006, p. 97ff). Additionally, data augmentation is only able to identify groups of customers that can be targeted by different marketing mix strategies. It is still far from a one-to-one marketing solution (Hattum & Hoijtink, 2008a).

The data augmentation approach as described here is a micro approach. The most likely target value is augmented to the customers. Only the target value's probability for people with the according link variable class in the source is available. Data augmentation results are not useful for aggregated statements, because the best value is sought for every customer, which does not necessarily adhere to the overall macro validity. Statements like "half of our customers are interested in this product" are not valid. It would always have to be put into the context of the source. The correct interpretation of the data might be difficult to explain to external parties, e.g. general management. Likewise, no correlations should be made between the augmented variables and other existing variables. The augmentation results are a tool for enabling better segmentations and target group selections. If aggregated statements shall be made, market research is a better alternative.

Data augmentations rely on some assumptions that cannot be tested in practice. For example, if the customer database is assumed to be MAR (as described in chapter 3.1.2) and no auxiliary source is available to confirm or object this assumption, the results rely on the correctness of this assumption. Several theoretical, empirical, and simulation studies have shown that there are risks associated with data augmentation (Adamek, 1994; Rodgers, 1984). As in every research model, there are several steps in the data augmentation process where decisions have to be made by the researcher. Every poor decision can compromise data augmentation results.

The data augmentation results should always be used in combination with existing data. The uncertainty inherent in the augmentation results is too strong for decisions to be based solely on these results. Data augmentation results are only meaningful, if the current information available for decision making is sparse. For example, when introducing new products or cross selling other product categories, data augmentation results can enhance the decision basis. For categories or products for which much is already known, data augmentation results derived from an external source might not have an additional informative value. But whenever other, more substantiated information is available, data augmentation results can be used to improve decisions and to have more variables to base a decision on. They can also be used to validate preliminary decisions.

8.2 Further research opportunities

This study is a starting point in the exploration of data augmentation with external sources in database marketing. More questions arise from our findings. These are delineated below. Furthermore, other use cases relevant to the practice are given, broadening the field of data augmentation research in database marketing.

8.2.1 Ignorability of the source data mechanism

MCAR source data mechanisms are easily differentiable, while MAR and MNAR source data mechanisms are not observable. Whether a source is MAR or MNAR depends on the conditional association of source and target variable given the link variables.

It has previously been suspected that a clear distinction between conditionally dependent and independent sources is difficult, because of the categorical nature of the variables and the complexity of the association resulting from a big number of link variables and the various classes related to them. In a conditional independence test, partial two-way cross-sectional tables are built and independence is tested for source and target variable for every class. A major problem encountered has been the problem of sparse data in many of these classes. Another problem has been the frequent disagreement of the χ^2 test with aggregated total measures and the CMH test used. Because a MNAR source data mechanism did not always compromise the augmentation results, we eventually concluded that conditional independence can be assumed. It should be validated whether the assumption of conditional independence, as suggested in this study, is admissible when no auxiliary data is available. It was sufficient in our research context, but could prove differently when validated in a different context.

The process of assessing the ignorability of the source data mechanism is complex and hardly performable in practice. Even if we had been able to definitely prove or disprove conditional independence, the same test cannot be performed in practical applications, where the source data mechanisms, as well as the elements not observed are not known. More research is necessary in this field.

8.2.2 Deeper exploration of the augmentation process

In order to get a deeper understanding of data augmentation in marketing, further research is suggested at three points of the data augmentation process. The first enhancement concerns other and more complex data augmentation methods. The second enhancement is related to the uncertainty assessment of the target values, once a data augmentation has been carried out. Eventually, a study researching on the external evaluation of the augmentation results would be valuable in order to complement our study.

The methods presented here do not reflect the state of the art in statistical matching. They are hands-on ad hoc approaches to data augmentation, as it is frequently found in companies. This choice is not only for simplification purposes, but also because the practical application of these methods is more likely in database marketing than other, more complex methods. Nevertheless, the use of more complex methods is possible and their benefits regarding effectiveness and efficiency should be evaluated in order to get a more comprehensive picture of possible data augmentation methods. Examples are likelihood-based inference methods, as for example the EM algorithm, or Markov Chain Monte Carlo methods, as for example Gibbs sampling or the Metropolis-Hastings algorithm (Schafer, 1997, p. 2). These methods involve complex simulations of posterior distributions (Schafer, 1997, p. 4). It has not been feasible for us in our case study context to apply methods which would require a simulation each. Methods of these fields could also be useful in gaining more insight into the conditional association assessment. To contrast the effectiveness of such methods against our approaches would be a valuable extension to our work.

The probabilities augmented in our study are the direct observable probabilities as derived from the source. Its interpretation always takes into consideration the source derivation. These probabilities do not include any kind of uncertainty assessment related to the fact that MAR sources are not identical to or representative of the customer population. If possible, it would be desirable to get a more comprehensive measure, including an uncertainty part as derived from the source characteristics. It might prove difficult, because probabilities reflecting all uncertainty might easily decrease to very low numbers not catching differences between target values anymore. Further research is necessary in this area before such a comprehensive measure can be developed.

Eventually, an external validation of this study in terms of actual conversions and return on marketing investment is desirable to complement our conceptual model. Many factors influence the ROMI and it is not easy to conduct a comprehensive study including various sources, like it has been done here. To establish comparability when real sources are used is rather difficult. The external validation is much more focused on practical applications, and several case studies can be combined in order to perform an overall external validation of our conceptual model.

8.2.3 Further augmentation opportunities and use cases

In this study, we have collected, specified, and structured the features of data augmentation in database marketing for the special case of external data. The data augmentation approach as proposed in this study is built on the use case of optimal target group selection in direct marketing. Although being highly relevant for database marketing practice, it has never been comprehensively assessed in a scientific context. In order to give guidance to database marketing analysts, basic rules, relevant aspects, and cruxes of data augmentations are stated. These are verified with a suitable case study. More examples are needed in order to build a comprehensive picture for marketing in general. Other use cases are thinkable and many of them have different properties in terms of variable scales, inference requirements, and recency expectations. The augmentation opportunities comprise, but are not limited to, the following use cases.

Our conceptual model and all our methods are based on the assumption that the typical sources for data augmentation applications contain categorical variable scales. This might be different, if branches like the financial sector were regarded, where many variables are metric. In this case, it is possible to reduce these variables to categories. However, this would result in a loss of information in terms of accuracy. In these cases, it might be reasonable to develop a conceptual model for metric variables in order to receive more meaningful data augmentation results. Also, many sources contain a mixture of variable scales. In our approach, we propose to harmonize all variables to the same scale. A comparison of such methods to the proposed standard methods would be a valuable extension to our work.

In our data augmentation proposition, we chose a univariate pattern approach. It means that all target variables are augmented individually, so that the most accurate results are achievable for every variable. However, by doing so, no inference can be made between different target variables augmented from the same source. It means that there have to be made as many augmentations as there are target variables. This can prove time consuming, because every model has to be established and adjusted to best fit the respective target variable. Hence, marketing problems exist where the advantages of a multivariate pattern approach outbalance the accuracy advantages of the univariate pattern approach. Such multivariate pattern approaches are more complex than univariate pattern approaches, because different target variables have to be explained at the same time. Only methods using statistical twins, such as the nearest neighbor method, can guarantee that all target values are taken from the same donor. Interrelations existing between these target variables are preserved and it is possible to attempt to make inferences between these variables after having augmented the data. Establishing such approaches and comparing them to univariate pattern approaches would be a meaningful enhancement to our work.

Our approach is based on a classical database marketing structure, where unique customers are the elements the analyses are based on. While this is true for most channels in the offline world and in email marketing, other channels have different identifiers, such as cookies, online accounts being used by multiple persons, or other structures not directly relatable to unique persons. If it is not possible to convert these structures to a unique-person-element structure, classical data augmentation sources cannot be properly used. At the same time, there are possible sources in the internet world consisting of these non-classical structures. Likewise, these sources cannot be used for data augmentation, if no transformation is possible. Both the identification of unique persons in these structures and the usage of other units for data augmentation purposes (both as recipient units and donor units) are interesting and seminal fields which can be regarded.

Eventually, all data augmentation structures should be transferred to automatic processes in which new information is generated in real-time. One time approaches, especially as derived from dedicated surveys, have little information value, because the information is not valid for a long time period. Most marketing problems as described here, regarding target group selections in direct marketing, return frequently. An automated augmentation process has several advantages. Firstly, both the augmented information and the link variables information is augmented on are always up-to-date. Secondly, the ROMI increases with every reuse of a source that has once been connected to the customer database with data augmentation. Of course, the automation process has afflicted costs itself that need to be evaluated in the decision process. But the maintenance effort is low when compared to a whole new data augmentation. Finally, the data augmentation automation infrastructure can be used for new sources, again leading to economies of scale.

8.2.4 Uplift models

Our models assume that the fact that someone has a certain characteristic, i.e. a certain target value, is a good predictor for that customer to convert after having received a marketing communication. Uplift models reach one step further. They try to predict for which customers the probability of conversion is maximally *increased* when being contacted by a marketing campaign. It does not necessarily prefer those customers with a high affinity to an offer, if these customers would have bought an offered product anyways. Rather, it tries to identify those customers whose propensity significantly increases through a marketing campaign.

Marketing campaigns using uplift models might not reach as high conversion rates as those based on our approach of data augmentation. Nevertheless, the delta between a potential control group and a target group chosen by an uplift model is higher. When adding up general sales not motivated by marketing actions, the sales leads generated by the marketing action, and the costs of the marketing campaign, uplift models yield higher revenues than simple conversion probability models.

Uplift models are more complex than simple conversion models. The general conversion probability needs to be separated from the uplift in conversion probability induced by a marketing campaign. It needs to respect many volatile factors. However, if mastered, uplift models have a high profit, because the marketing budgets can be directed at exactly the customers that respond best and the overall earnings can be maximized.

Data augmentation results can also be used to specify uplift models. This is a logical extension to our work. Uplift models need to be trained well in order to make the separation between the general conversion probability and the uplift in conversion probability induced by a marketing campaign. Such information cannot be found in an external source. This makes it difficult to *augment* the probability for a significant uplift. But the augmented information can be used to build an uplift model *within* the customer database.

Chapter 9

Summary

External sources like web tracking, volunteer surveys, and social media are rarely used for data augmentation in database marketing yet, even though such information can lead to significant conversion probability lifts when used for targeting and personalization. Many external sources are suspected to lead to biased augmentation results, because the available data is not representative for the customer group to which it is augmented. With our study, we deliver insights on which sources can be used beneficially for data augmentation and which augmentation methods should be used for different kinds of sources.

Introduction

An increasing amount of data is needed in order to *segment* and *select* customers for personalized and targeted marketing activities. But the data basis for targeting customers is often sparse, so that the customers with the highest propensity cannot be identified. At the same time, data has never been easier available externally, e.g. from website click behavior, volunteer surveys, or social media. Data augmentation is a beneficial tool for harnessing this information. But it is used hesitantly, because no validation of the augmentation results is possible. Moreover, it cannot be assessed up-

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4_9

front whether the augmentation is effective and efficient in a way that using the results increases the conversion rates of direct marketing campaigns. With our study, we establish guidelines for evaluating external augmentation sources.

External sources are beneficial, because they contain information types not available from the customer database. It is usually not possible to acquire the same information in an easier, cheaper, and more up-to-date way. External sources are defined by not being technically connected to the customer database, with a unique customer identifier missing. In order to acquire the same information otherwise, companies would have to ask the customers – which is unfeasible in most cases – or entrust external agencies with market research. The objective of data augmentation is to augment the external information to the customer database instead based on similarities.

In order to perform data augmentation, no unique identifier is needed in the external source. As long as there is a set of *link variables* present both in the customer database (recipient unit) and the source (donor unit), the *target variables* can be augmented. Link and target variables are categorical in a marketing approach. The most likely value is augmented for every single customer, given a certain link variable class, along with a probability for this value to be accurate. These values can be used to manage direct marketing campaigns and to improve the basis for decision making. We favor a univariate pattern approach, in which one target variable is augmented at a time. Thereby, the link variables are combined differently for every target variable in a way that each target variable is predicted best.

It is desirable to be able to evaluate the external sources ex ante and to make a decision whether they will increase conversion probabilities. The main question regarding external sources is: Are the target values observed in the source transferable to the customers? For example, let there be an internet volunteer survey that asked for computer literacy. If computer literacy is to be augmented, it is possible that the results will be biased, because most people who take part in an internet volunteer survey are computer literate. The results would indicate that there are more people computer literate than there really are in the customer database. Such a source is not feasible for data augmentation. In a feasible source, the target variable is not correlated to the fact whether somebody has been observed in the source or not.

Literature review

The potential of data augmentation using external sources in marketing has already been realized in contemporary literature (Breur, 2011; Jiang et al., 2007; Ratner, 2001b). Data augmentations in marketing have been performed by Kamakura and Wedel (1997, 2000), Putten et al. (2002a, 2002b), and Gilula et al. (2006). Our paper contributes to Hattum and Hoijtink's (2008a, 2008b) idea of data fusion for direct marketing and extends it to a broader range of possible applications. The methods used for data augmentation have been developed in the contexts of missing data imputation (Little & Rubin, 2002; Rubin, 1976; Schafer, 1997). Some have also been developed for statistical matching in official statistics (D'Orazio et al., 2006; Okner, 1972) and media planning (Rässler, 2002; Wendt, 1977).

Related data augmentation approaches are also referred to as data fusion (Breur, 2011; Gilula et al., 2006; Hattum & Hoijtink, 2008b), statistical matching (D'Orazio et al., 2006; Rässler, 2002), or deterministic data integration (Jiang et al., 2007). Data augmentation suits the use of external information in database marketing best, because it focuses on supplementing information already available. It is different from record linkage, where recipient unit and donor unit contain identical elements, with a missing unique identifier. It also differs from exact matching, where a unique identifier is present. In contrast to exact matching, the statistical augmentation of external data is always legally sound (Plath & Frey, 2009).

Available external sources are seldom properly sampled like representative surveys. Previous data augmentation studies in marketing have only regarded sources with a random sampling mechanism. External sources usually partially overlap with the customer database. The reasons why a person is observed in a source are manifold. We refer to the mechanism that leads to a person being observed in a source as *source data mechanism*, derived from the missing data mechanism in missing data theory (Little & Rubin, 2002; Rubin, 1976). The source data mechanism is a nonprobability sample (Powell, 1997, p. 67), if the observed persons are not randomly chosen.

Little and Rubin (2002) differentiate three stages of sampling types for missing data. These describe the relationship between which elements of the overall population have been observed and the values of the target variable. Data is missing completely at random (MCAR), if the availability or missingness of data is not dependent on any factors. A randomly sampled survey is MCAR. Data is missing at random (MAR), if the availability or missingness of data is only dependent on observable variables. If the information on credit card ownership is particularly missing for older people, but the age has been observed, then data is MAR. Data is missing not at random (MNAR), if the availability or missingness of data depends also on variables not observed. If information on the income volume is particularly missing for people with a high income, then data is MNAR. If data is MCAR or MAR, the missing data mechanism can be ignored when augmenting data from these sources.

The augmentation process consists of data screening, preparation, augmentation, and evaluation steps. It has been described by Putten et al. (2002b) and D'Orazio et al. (2006, chapter 6). The link variables must be present in both donor and recipient unit. Relevant target variables are chosen, given that the link variables are able to predict these target variables. A suitable augmentation method needs to be chosen. In the existing literature, there are no general rules on which method to use best in marketing. We deliver insight into how to choose the best method. Eventually, the augmentation results are evaluated by their ability to increase conversion rates in targeting applications.

Theory and model

The three stages of missing data can be transferred to the source data mechanisms of external sources. It refers to whether a person has been observed in a source. The target variable can have each of the three relationships to the source data mechanism. The source can be randomly sampled, so that it is MCAR. If a source is MAR, the target values depend only on the variables observed for the customers, i.e. the link variables. Tthe source data mechanism can be ignored, because data is augmented based on the link variables. If a source is MNAR, the fact whether a person has been observed in a source is correlated with the target values. Such a source should not be used for data augmentation, because the results can be biased.

In our study, we are particularly interested in the quality of augmentation results of sources with a MAR and MNAR source data mechanism. A source is MAR, if the source data mechanism is conditionally independent from the augmented target variable, given the link variables. In our previous example, the participants in the volunteer survey are particularly computer literate. Augmentation results are suspected to be biased towards high computer literacy levels. However, maybe target variables are augmented based on the link variable age. If a high number of young people have taken part in the survey, having a higher share of computer literates than older segments, the tendency to being computer literate can be attributed to the age. In that case, the source is MAR, because the age is observable.

The conversion probability lift (CPL) is an indicator showing how much the ability to select prospective customers for a direct marketing campaign increases when using augmentation results. If the true target values are known for the customers, a total correct classification rate (TCCR) can be calculated describing how many true target values were hit by the results. The model lift of the augmentation is calculated by comparing the TCCR of the model to the TCCR achievable, if target values were allocated randomly among the recipients (Hattum & Hoijtink, 2008b; Ratner, 2003, p. 182).

The database marketing analyst has to choose an appropriate augmentation method, which influences the model lift. In our conceptual model as shown in figure 5.1 on page 142 in chapter 5.1.1, the main effect of the model lift is the chosen method. All other parameters are moderators and the appropriate source data mechanism is an antecedent. The model lift shows how many customers more would be correctly assigned to a target group, if segmentation is the targeting goal. If selection is the targeting goal, the CPL is influenced by the number of customers to be selected. In that case, the model lift is a mediator for the CPL. Only the internal evaluation is in the scope of this study. In an external evaluation in a business setting, a lift of return on investment is calculated to finally assess the utility of the augmented data.

Three common data augmentation methods are compared: conditional mode imputation, nearest neighbor hot deck, and logistic regression. For conditional mode imputation, the most frequent value of a link variable class is augmented. Classes are iteratively collapsed to broader ranges, if the mode cannot explicitly be determined. The nearest neighbor hot deck method uses the target value from the donor closest to the recipient in terms of a distance measure regarding all link variables. In logistic regression, the most likely target value is predicted based on a maximum likelihood function using the link variables as predictors. The choice of the best method is expected to be dependent on the moderators. Conditional mode imputation and nearest neighbor hot deck are expected to be influenced by the source characteristics. They work best, if recipient unit and donor unit contain identical elements. Every deviation from this optimum results in higher uncertainty and lower model lifts. Contrarily, logistic regression detects general correlations between link and target variables. These can be transferred to different subgroups.

The model lift is expected to be influenced by the source characteristics. Sources are characterized by its size (number of donors), the overlap between recipient unit and donor unit (in terms of identical elements), and whether the donors were representatively sampled or not. The model lift is dependent on the target variable characteristics. The range of the model lift differs, depending on the number of target values and the skewness of the target variable distribution as measured by the index of qualitative variation. Finally, it depends on the predictive power of the link variables regarding the target variables.

We test the hypothesis that there is a definable set of source characteristics influencing which method to use best for data augmentation. A logistic regression model describes the parameters that lead to the method with the best model lift. The output variable is a categorical variable indicating which method yielded the best augmentation results in the respective setting. The models give insight into factors leading to good model lifts and methods to choose in particular settings. Our goal is to provide a set of rules leading to significant CPLs. If all decisions are made correctly based on our findings, augmentation results are expected to be significantly better than randomly selected target groups.

Methodology and results

A model lift can only be calculated in a simulated setting, where target values are augmented, although the true values are known. This is not feasible in a business setting, because the true values are not known for the customers. For this purpose, an existing customer dataset from a renowned German company is used as a case study object. It already contains target values for all elements and serves as overall population. A subgroup of the elements in the dataset is chosen to be the customer group. For this subgroup, all target values are removed. After the augmentations, the augmented target values can be compared to the true values. In order to compare the influencing factors, different sources are simulated and used for augmentation. The sources are sampled based on an information-oriented sampling approach. We have varied the following factors in the case study:

- *49 Sources:* size, overlap, and representation
- *11 Target variables:* number of target values in the domain and skewness
- *3 Methods:* conditional mode imputation, nearest neighbor hot deck, and logistic regression

1,617 augmentations have been performed, each producing a set of measurement data describing the input parameters and results. For example, a source is sampled based on a feasible variable which functions as source data mechanism. The source contains twice as many donors as the recipient unit, but has only 50% overlap to the recipient unit. 33 augmentations are performed with this source: 11 target variables are augmented individually, each using one of the three methods.

A case study approach is necessary, because augmentation results can only be evaluated in a simulated setting where the true target values are known. Likewise, the source data mechanism is not observable in a practical application. A high number of different sources with different levels of source characteristics must be given in order to give general advice on which source characteristics have an influence of the augmentation results and which should not be used. This is not possible in a real-world setting. However, the results of a case study always need to be regarded in the context of the study and are limited to this frame. Some parameters like the link variables to be used and the recipient unit have been fixed. The observed results give insight on the relationships of influencing parameters in this setting. Therefore, the frame is broadened in our study as much as possible. In order to overcome the problem of a lack of generalizability inherent in case study approaches, the data is especially *realistic*, so that correlations observable with humans relating to the purchase behavior are available, *qualitatively rich*, so that many different variables and categories are represented, and *quantitatively rich*, so that many different permutations can be performed with each resulting in an reasonably sized subset of the data. From the case study, insights for an ex-ante assessment of data augmentation sources in practice can be derived not otherwise attainable.

If the source data mechanism is MNAR, using it can lead to biased augmentation results. In the original dataset, a source data mechanism indicator variable marks whether a person has been observed ($S = 1$) in a source or not ($S = 0$). Therefrom, the conditional association between the target variable and the source data mechanism indicator variable, given the link variables, is calculated. We performed a χ^2 test with aggregated total measures and a Cochran-Mantel-Haenszel test in order to assess the conditional association. We then assessed whether the augmentation results of these sources are biased by comparing the data augmentation results with or without incorporating the source data mechanism in the model. Although MNAR sources have a higher tendency to lead to biased results than MAR sources, a decisive recommendation could not be given. Particularly, if there is a strong correlation between the link variables and the target variable, the source data mechanism cannot be so strong that it compromises the data augmentation results. We later found that if all other rules in our practical guide based on observable source characteristics are respected, data augmentations do not lead to biased results, even if the sources are MNAR.

Target variables need to be sufficiently even distributed and can only have a moderate number of target values. Then the methods are able to discriminate between them. The discriminability depends on the predictive power of the link variables regarding the target variables. A quality check after each augmentation shows whether all target values have been reproduced. Each target value should appear sufficiently often in the donor unit, with a class occupation of at least 20%. The average model lift decreases with an increasing number of target values and increasing skewness. Likewise, the variance of the model lift increases, leading to less stable results.

In a good augmentation, all target values are reproduced. Augmentations can also lead to significant model lifts, if only some values are reproduced (in cases with target variables with at least three target values). However, augmentations in which not all values have been reproduced do not reliably lead to significant CPLs. If other criteria in the decision process are already close to exit criteria (e.g. low overlap), sources should not be used for data augmentation, if not all parameter values are reproduced by the augmentation.

The best method to be used depends on the overlap. Nearest neighbor hot deck is better suited for high levels of overlap. It works best at a 100% overlap rate, where there is a match for every recipient. Logistic regression is better suited for low levels of overlap, because the model lift is not influenced by the overlap. Logistic regression is able to detect general relationships in the data and to transfer them to partially different subgroups of the overall population. Conditional mode imputation is inferior to the other two methods and does not lead to significant model lifts.

Sources should only be used, if there is a significant overlap to the recipient unit. The uncertainty involved in an augmentation with zero overlap is high. On average, no significant model lift can be observed. In practice, it is rather unlikely to acquire a source that has nothing in common with the customer database. We have only included such sources in order to show that distinct sources cannot be used, if the source data mechanism is MAR or MNAR.

The overlap is the only relevant source characteristic influencing the best method to be used. We have used a backward elimination procedure in the method model to isolate the parameters with a significant influence. Only the overlap and the number of target values have remained. Of these, the effect of the overlap has been twice as strong as the number of target values. The number of target values is only relevant at medium levels of overlap. In that case, logistic regression should be used for binary target variables and nearest neighbor hot deck for variables with more target values. After having chosen the best method based on the logistic regression results, 98% of the augmentations show a significant CPL.

Conclusions

The hesitation regarding data augmentation in database marketing is causeless. The source data mechanism is negligible in a categorical setting, where the predictability of link variables regarding the target variables is strong. Sources need to comply with certain minimum criteria, so that the benefits of data augmentation outweigh the risks associated with nonprobability sampling mechanisms of the source. From our study, the following guidelines are derived regarding data augmentation in database marketing.

- *Target variable:* A categorical target variable should have at most five possible target values. The distribution of these values should be sufficiently even in order for methods to discriminate between them ($IQV > 0.8$).

- *Link variables:* The link variables need to have predictive power regarding the target variable. They should be able to discriminate between all possible values of the target variable domain. Link variables have to be independent.

- *Source data mechanism:* An external source does not need to be representatively sampled. If all variables are categorical and if the link variables are able to predict the target variables well, the source data mechanism is negligible.

- *Overlap:* The higher the overlap between the donor unit and the recipient unit, the better the data augmentation results. If there is no overlap, the hit rate of the results does not outperform a random distribution of target values.

- *Method:* Logistic regression is best suited for sources with a low overlap or sampling rate. Nearest neighbor hot deck is best suited for sources with a high overlap. If the overlap is not known, logistic regression should be used, because it performs equally well for all levels of overlap.

Simple methods like nearest neighbor hot deck and logistic regression lead to significant model lifts. But data augmentation is not limited to these methods. The literature on missing data imputation comprises more complex methods, including multiple imputation and Bayes approaches. To contrast the effectiveness of such methods against our approach would be a valuable extension to our work.

More questions arise from our study regarding the augmentation process. The ignorability of the source data mechanism requires more inspection. For the application in business, a measure describing the uncertainty related to a data augmentation project is desirable. Our approach is a univariate pattern augmentation, augmenting one target variable at a time. A multivariate pattern augmentation is also thinkable, preserving the interrelations of the augmented variables. While our case study is an application in a specific context, the theory is generalizable and can be applied in other contexts. We encourage its extension to more use cases. Furthermore, the results of our study need to be evaluated externally in order to assess the financial contribution of augmentation results on marketing profits.

It is important to manage the expectations regarding data augmentation results. Data augmentation is a tool for acquiring information that is not available in any other (better) way. Data augmentation results are never definite, but always only probabilities pointing to a certain value. The results should be used in combination with existing data, because the uncertainty inherent in the augmentation results is too strong for decisions to be based solely on these values. Stakeholders and analysts making use of the augmentation results should never use the data like a primary source. Before conducting an augmentation, the possible benefit needs to be contrasted against the effort of the augmentation. We support this process with our findings. Database marketing analysts are now able to evaluate external augmentation sources upfront. We have been able to show that feasible sources lead to significant CPLs and encourage their use for data augmentation in database marketing.

Appendix

Table complementing chapter 5

Table 9.1 shows the distribution of observations, i.e. persons, among the link variable classes in the population, as described in chapter 5.2.2.

Obs.	> 200	150<=200	100<=150	50<=100	5<=10	10<=50	<=5
Classes	5	22	62	133	550	484	1,549

Table 9.1: Number of link variable classes by numbers of observations

Figures and tables complementing chapter 6

Table 9.2 contains the full list of measures for the distribution of all CCR_{target} values by number of target values k, from which the selected measures in table 6.5 on page 195 in chapter 6.3.2 were taken.

© Springer Fachmedien Wiesbaden 2016
B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4

n_{target}	N	Max	Q3	Median	Mean	Q1	Min	Sh.-W.	p-value
3,849	129	0.44	0.25	0.14	0.17	0.10	0.00	0.94	<.001
4,093	146	0.81	0.55	0.44	0.41	0.31	0.00	0.95	<.001
4,871	143	0.86	0.40	0.20	0.27	0.13	0.00	0.93	<.001
5,005	137	0.66	0.41	0.31	0.29	0.19	0.00	0.98	0.020
5,044	137	0.64	0.51	0.39	0.37	0.27	0.01	0.95	<.001
6,006	142	0.74	0.58	0.47	0.43	0.37	0.00	0.89	<.001
6,137	147	0.91	0.68	0.57	0.44	0.16	0.00	0.88	<.001
6,468	138	0.82	0.65	0.52	0.49	0.43	0.00	0.90	<.001
7,070	147	1.00	0.80	0.53	0.43	0.00	0.00	0.83	<.001
7,140	146	0.98	0.79	0.73	0.70	0.61	0.12	0.96	<.001
8,543	147	0.99	0.80	0.66	0.50	0.00	0.00	0.81	<.001

Table 9.2: Full table of measures for the distribution of all CCR_{target} values by number of observations with the target value n_{target}

Table 9.3 contains the parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by method, as shown in figure 6.9 on page 205 in chapter 6.4.1. The parameter estimates of conditional mode imputation and logistic regression are contrasted to nearest neighbor hot deck by solving the normal equations of the ANOVA model (SAS Institute Inc., 2014d).

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	1.184	0.011	109.36	<.001
Method Conditional mode	−0.258	0.015	−16.82	<.001
Method Logistic regression	−0.0115	0.016	−0.73	0.463
Method Nearest neighbor	0.000			

Table 9.3: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by method

Table 9.4 contains the full list of measures for the distribution of all ML_{target} values by number of target values k, from which the selected measures in table 6.10 on page 209 in chapter 6.4.2 were taken.

Table 9.5 contains the parameter estimates for the ANOVA model solution describing the distribution of ML_{target} by method, as shown in figure 6.11 on page 210 in chapter 6.4.2.

Table 9.6 contains the parameter estimates for the ANOVA model solution describing the distribution of $ML_{uniform}$ by method, as shown in figure 6.12 on page 214 in chapter 6.4.3.

n_{target}	N	Max	Q3	Median	Mean	Q1	Min	Sh.-W.	p-value
3,849	129	3.97	2.27	1.24	1.50	0.88	0.01	0.94	<.001
4,093	146	6.45	4.40	3.54	3.30	2.46	0.00	0.95	<.001
4,871	143	4.82	2.24	1.14	1.50	0.72	0.01	0.93	<.001
5,005	137	3.52	2.20	1.65	1.57	1.00	0.00	0.98	0.020
5,044	137	3.34	2.70	2.03	1.95	1.41	0.07	0.95	<.001
6,006	142	2.73	2.14	1.74	1.59	1.36	0.00	0.89	<.001
6,137	147	3.24	2.41	2.03	1.55	0.56	0.00	0.88	<.001
6,468	138	2.61	2.09	1.66	1.56	1.38	0.01	0.90	<.001
7,070	147	2.67	2.14	1.43	1.14	0.01	0.00	0.83	<.001
7,140	146	2.56	2.08	1.90	1.83	1.59	0.31	0.96	<.001
8,543	147	1.81	1.47	1.21	0.91	0.00	0.00	0.81	<.001

Table 9.4: Full table of measures for the distribution of all ML_{target} values by number of observations with the target value n_{target}

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	1.990	0.045	44.23	<.001
Method Conditional mode	−0.881	0.064	−13.81	<.001
Method Logistic regression	−0.048	0.065	−0.74	0.461
Method Nearest neighbor	0.000			

Table 9.5: Parameter estimates for the ANOVA model solution describing the distribution of ML_{target} by method

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	0.993	0.010	103.69	<.001
Method Conditional mode	−0.199	0.014	−14.66	<.001
Method Logistic regression	−0.009	0.014	−0.64	0.524
Method Nearest neighbor	0.000			

Table 9.6: Parameter estimates for the ANOVA model solution describing the distribution of $ML_{uniform}$ by method

Table 9.7 contains the parameter estimates for the ANOVA model solution describing the distribution of ML_{source} by method, as shown in figure 6.14 on page 210 in chapter 6.4.4.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	2.016	0.057	35.32	<.001
Method Conditional mode	−0.699	0.081	−8.63	<.001
Method Logistic regression	−0.081	0.083	−0.97	0.330
Method Nearest neighbor	0.000			

Table 9.7: Parameter estimates for the ANOVA model solution describing the distribution of ML_{source} by method

Table 9.8 contains the parameter estimates for the ANOVA model solution describing the distribution of CL_{global} by method, as shown in figure 6.19 on page 232 in chapter 6.5.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	1.173	0.004	309.12	<.001
Method Conditional mode	−0.117	0.005	−21.80	<.001
Method Logistic regression	−0.081	0.006	−14.79	<.001
Method Nearest neighbor	0.000			

Table 9.8: Parameter estimates for the ANOVA model solution describing the distribution of CL_{global} by method

Figures and tables complementing chapter 7

Table 9.9 contains the parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by the source data mechanism as derived from the test based calculation of conditional dependencies, as shown in figure 7.1 on page 240 in chapter 7.1.1. The parameter estimates of MCAR, MAR, and undecided sources are contrasted to MNAR sources by solving the normal equations of the ANOVA model.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	1.027 B	0.011	92.02	<.001
Mechanism 1) MCAR	0.123	0.015	8.07	<.001
Mechanism 2) MAR	0.074	0.038	1.96	0.050
Mechanism 3) undecided	0.057	0.022	2.64	0.008
Mechanism 4) MNAR	0.000			

Table 9.9: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by the source data mechanism as derived from the test based calculation of conditional dependencies

Table 9.10 contains the parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by the source data mechanism as derived from the model based calculation of conditional dependencies, as shown in figure 7.2 on page 244 in chapter 7.1.2. The parameter estimates of MCAR, insignificant, and undecided sources are contrasted to significant sources by solving the normal equations of the ANOVA model.

Parameter	Estimate	Standard Error	t Value	p-value
Intercept	1.025	0.014	75.36	<.001
Mechanism 1) MCAR	0.125	0.017	7.32	<.001
Mechanism 2) insignificant	0.043	0.019	2.22	0.027
Mechanism 3) undecided	0.023	0.034	0.67	0.503
Mechanism 4) significant	0.000			

Table 9.10: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by the categorization as derived from the model based calculation of conditional dependencies

Table 9.11 contains the parameter estimates for the ANOVA model solutions describing the distribution of ML_{chance} by overlap for every method used, as shown in figure 7.4 on page 251 in chapter 7.2.2. The parameter estimates of every overlap level observed are contrasted to the highest overlap level possible ($o = 11,560$) by solving the normal equations of the ANOVA model.

Table 9.12 contains the table ANOVA statistics for the distribution of all ML_{chance} values by overlap and applied augmentation method with non-representative sources ($o > 0$), as described in chapter 7.2.2.

Tables 9.13, 9.14, and 9.15 contain the parameter estimates for the ANOVA model solutions describing the distribution of ML_{chance} by number of donors for every method used, as shown in figure 7.5 on page 254 in chapter 7.2.3. The parameter estimates of every donor level observed are contrasted to the highest donor level possible ($d = 40,000$) by solving the normal equations of the ANOVA model.

Figure 9.1 contains the fit diagnostics for the regression model applied for nearest neighbor hot deck augmentations as described in chapter 7.3.1. In a linear regression model, both the independent and dependent variables are continuous. Some assumptions must be met in order to apply a linear regression model to the descriptive data. In particular, the predicted value must follow linearity and the errors terms need to be independent, have

Parameter	Estimate	Standard Error	t Value	p-value
Conditional mode imputation				
Intercept	0.992	0.030	33.11	<.001
Overlaps 0	−.184	0.044	−4.18	<.001
Overlaps 585	−.099	0.076	−1.30	0.196
Overlaps 882	−.082	0.076	−1.08	0.281
Overlaps 1,624	−.088	0.076	−1.15	0.249
Overlaps 6,168	−.034	0.076	−0.44	0.659
Overlaps 6,387	−.070	0.076	−0.92	0.359
Overlaps 7,829	−.026	0.076	−0.34	0.737
Overlaps 7,956	−.027	0.076	−0.36	0.720
Overlaps 8,944	−.018	0.076	−0.23	0.816
Overlaps 10,005	−.011	0.076	−0.14	0.887
Overlaps 11,560	0.000			
Logistic regression				
Intercept	1.239	0.014	86.63	<.001
Overlaps 0	−.406	0.024	−16.78	<.001
Overlaps 585	−.036	0.036	−1.00	0.318
Overlaps 882	−.051	0.037	−1.37	0.171
Overlaps 1,624	0.003	0.036	0.10	0.924
Overlaps 6,168	−.012	0.036	−0.34	0.735
Overlaps 6,387	−.007	0.036	−0.20	0.843
Overlaps 7,829	−.015	0.036	−0.41	0.679
Overlaps 7,956	0.002	0.036	0.05	0.964
Overlaps 8,944	0.001	0.036	0.04	0.970
Overlaps 10,005	−.019	0.036	−0.52	0.602
Overlaps 11,560	0.000			
Nearest neighbor hot deck				
Intercept	1.352	0.013	106.39	<.001
Overlaps 0	−.386	0.018	−20.94	<.001
Overlaps 585	−.290	0.032	−8.97	<.001
Overlaps 882	−.273	0.032	−8.43	<.001
Overlaps 1,624	−.247	0.032	−7.61	<.001
Overlaps 6,168	−.104	0.032	−3.22	0.001
Overlaps 6,387	−.130	0.032	−4.01	<.001
Overlaps 7,829	−.069	0.032	−2.13	0.034
Overlaps 7,956	−.065	0.032	−2.01	0.045
Overlaps 8,944	−.042	0.032	−1.29	0.197
Overlaps 10,005	−.028	0.032	−0.87	0.386
Overlaps 11,560	0.000			

Table 9.11: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by overlap and method

	Conditional mode			Logistic regression			Nearest neighbor		
Test	N	F	p-value	N	F	p-value	N	F	p-value
ANOVA	319	0.41	0.930	318	0.85	0.572	319	17.64	<.001
Levene	319	0.16	0.997	318	0.23	0.990	319	0.38	0.942

Table 9.12: Full table of ANOVA statistics for the distribution of ML_{chance} by o and method excluding sources with $o > 0$

constant variances, and be normally distributed with a mean of 0 (Backhaus et al., 2008, p. 79ff).

Parameter	Estimate	Standard Error	t Value	p-value
Conditional mode imputation				
Intercept	0.981	0.102	9.62	<.001
Donors 585	−.078	0.144	−0.54	0.590
Donors 882	−.051	0.144	−0.36	0.723
Donors 1,604	−.098	0.144	−0.68	0.498
Donors 1,624	−.078	0.144	−0.54	0.587
Donors 2,469	−.091	0.144	−0.63	0.526
Donors 3,583	−.076	0.144	−0.52	0.600
Donors 6,168	−.020	0.144	−0.14	0.892
Donors 6,387	−.061	0.144	−0.42	0.671
Donors 7,829	−.011	0.144	−0.08	0.939
Donors 7,956	−.018	0.144	−0.12	0.902
Donors 8,944	−.004	0.144	−0.03	0.977
Donors 10,005	0.007	0.144	0.05	0.962
Donors 11,388	−.057	0.144	−0.39	0.694
Donors 11,560	0.018	0.144	0.13	0.899
Donors 12,579	0.017	0.144	0.12	0.907
Donors 13,147	0.019	0.144	0.13	0.894
Donors 13,519	0.016	0.144	0.11	0.914
Donors 15,826	−.015	0.144	−0.10	0.918
Donors 16,561	0.024	0.144	0.16	0.869
Donors 18,633	−.026	0.144	−0.18	0.858
Donors 19,430	0.007	0.144	0.05	0.964
Donors 24,025	0.008	0.144	0.05	0.957
Donors 24,387	−.009	0.144	−0.06	0.950
Donors 25,693	−.018	0.144	−0.13	0.900
Donors 27,003	0.008	0.144	0.05	0.956
Donors 29,424	0.005	0.144	0.04	0.970
Donors 33,303	−.006	0.144	−0.04	0.964
Donors 34,858	0.000	0.144	0.00	0.999
Donors 40,000	0.000			

Table 9.13: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by number of donors for conditional mode imputation

The linearity assumption can be evaluated by the R^2 measure. With 0.7614, ML_{chance} is not perfectly linear with regards to the predictors, but close. The equality of variances can be assed using the White test of first and second moment specification (UCLA, 2014). It yields a significant χ^2 value, indicating heteroscedasticity. However, the first plot for the residuals by predicted values of ML_{chance} shows no apparent patterns in the scattering around the 0 line, so that the residuals are assumed to be randomly distributed. The heteroscedasticity does not seem to be very strong. The errors follow normality, as the points in the residual by quantile plot largely follow the diagonal line. This is also confirmed by the histogram of the resid-

Parameter	Estimate	Standard Error	t Value	p-value
Logistic regression				
Intercept	1.159	0.029	39.63	<.001
Donors 585	0.067	0.041	1.63	0.105
Donors 882	0.060	0.042	1.42	0.156
Donors 1,604	0.020	0.041	0.47	0.637
Donors 1,624	0.086	0.041	2.07	0.040
Donors 2,469	0.000	0.041	0.01	0.993
Donors 3,583	0.081	0.041	1.96	0.051
Donors 6,168	0.100	0.041	2.43	0.016
Donors 6,387	0.096	0.041	2.33	0.020
Donors 7,829	0.102	0.041	2.46	0.015
Donors 7,956	0.096	0.041	2.31	0.022
Donors 8,944	0.106	0.041	2.56	0.011
Donors 10,005	0.106	0.041	2.56	0.011
Donors 11,388	0.049	0.041	1.18	0.239
Donors 11,560	0.108	0.041	2.60	0.010
Donors 12,579	0.112	0.041	2.70	0.007
Donors 13,147	0.109	0.041	2.65	0.009
Donors 13,519	0.110	0.041	2.65	0.008
Donors 15,826	0.067	0.041	1.63	0.104
Donors 16,561	0.089	0.041	2.14	0.033
Donors 18,633	0.035	0.041	0.84	0.404
Donors 19,430	0.089	0.041	2.16	0.032
Donors 24,025	0.094	0.041	2.26	0.024
Donors 24,387	0.057	0.041	1.37	0.172
Donors 25,693	0.028	0.041	0.67	0.502
Donors 27,003	0.078	0.041	1.88	0.061
Donors 29,424	0.059	0.041	1.44	0.152
Donors 33,303	0.016	0.041	0.38	0.705
Donors 34,858	0.031	0.041	0.76	0.450
Donors 40,000	0.000			

Table 9.14: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by number of donors for logistic regression

ual in the lower left corner. In a polynomial regression, the predictors must not be correlated (collinearity). If the variance inflation factor as shown in table 9.16 is lower than 10 for all variables, no variables are as strongly correlated as to disturb the regression results. This applies for all variables in the nearest neighbor hot deck model. All in all, the regression model can be applied with caution. The results should not be over-interpreted, but general findings can be derived, as it is done in chapter 7.3.1.

Figure 9.2 contains the fit diagnostics for the regression model applied for logistic regression augmentations as described in chapter 7.3.1. The $R^2 = 0.488$ reveals that only half of the variance of ML_{chance} is explained by the predictors, indicating that the regression model is only a mediocre

Parameter	Estimate	Standard Error	t Value	p-value
Nearest neighbor hot deck				
Intercept	1.312	0.044	29.48	<.001
Donors 585	−.249	0.063	−3.96	<.001
Donors 882	−.230	0.063	−3.65	<.001
Donors 1,604	−.252	0.063	−4.01	<.001
Donors 1,624	−.201	0.063	−3.20	0.002
Donors 2,469	−.236	0.063	−3.76	<.001
Donors 3,583	−.212	0.063	−3.37	0.001
Donors 6,168	−.047	0.063	−0.75	0.457
Donors 6,387	−.066	0.063	−1.04	0.299
Donors 7,829	−.002	0.063	−0.03	0.978
Donors 7,956	−.001	0.063	−0.02	0.985
Donors 8,944	0.017	0.063	0.28	0.783
Donors 10,005	0.042	0.063	0.67	0.506
Donors 11,388	−.115	0.063	−1.82	0.070
Donors 11,560	0.077	0.063	1.22	0.223
Donors 12,579	0.073	0.063	1.16	0.248
Donors 13,147	0.071	0.063	1.13	0.261
Donors 13,519	0.070	0.063	1.11	0.267
Donors 15,826	−.049	0.063	−0.78	0.435
Donors 16,561	0.041	0.063	0.65	0.516
Donors 18,633	−.082	0.063	−1.31	0.192
Donors 19,430	0.038	0.063	0.60	0.550
Donors 24,025	0.030	0.063	0.47	0.638
Donors 24,387	−.021	0.063	−0.34	0.735
Donors 25,693	−.057	0.063	−0.90	0.368
Donors 27,003	0.019	0.063	0.30	0.761
Donors 29,424	0.015	0.063	0.23	0.816
Donors 33,303	−.018	0.063	−0.29	0.770
Donors 34,858	0.005	0.063	0.08	0.934
Donors 40,000	0.000			

Table 9.15: Parameter estimates for the ANOVA model solution describing the distribution of ML_{chance} by number of donors for nearest neighbor hot deck

Variable	DF	Parameter Estimate	Stand. Err.	t-Value	Pr> |t|	Tolerance	Variance Inflation
Intercept	1	−0.20868	0.08438	−2.47	0.0139		
Overlaps	1	0.000033	0.000002	17.89	<.0001	0.45100	2.21729
Donors	1	−0.000002	6.487E−7	−3.67	0.0003	0.47291	2.11459
VCs	1	−0.000019	0.000006	−2.96	0.0033	0.60719	1.64694
Rsqu	1	0.85589	0.06315	13.55	<.0001	0.43007	2.32522
k	1	0.17438	0.00816	21.37	<.0001	0.59408	1.68328
IQV	1	0.78132	0.06964	11.22	<.0001	0.35470	2.81929

Table 9.16: Parameter estimates for the regression model explaining ML_{chance} by the descriptive data for nearest neighbor hot deck augmentations, including tolerance and variance inflation

choice. Like for the nearest neighbor hot deck model, the White test yields a significant χ^2 value, indicating heteroscedasticity, while the scattered marks

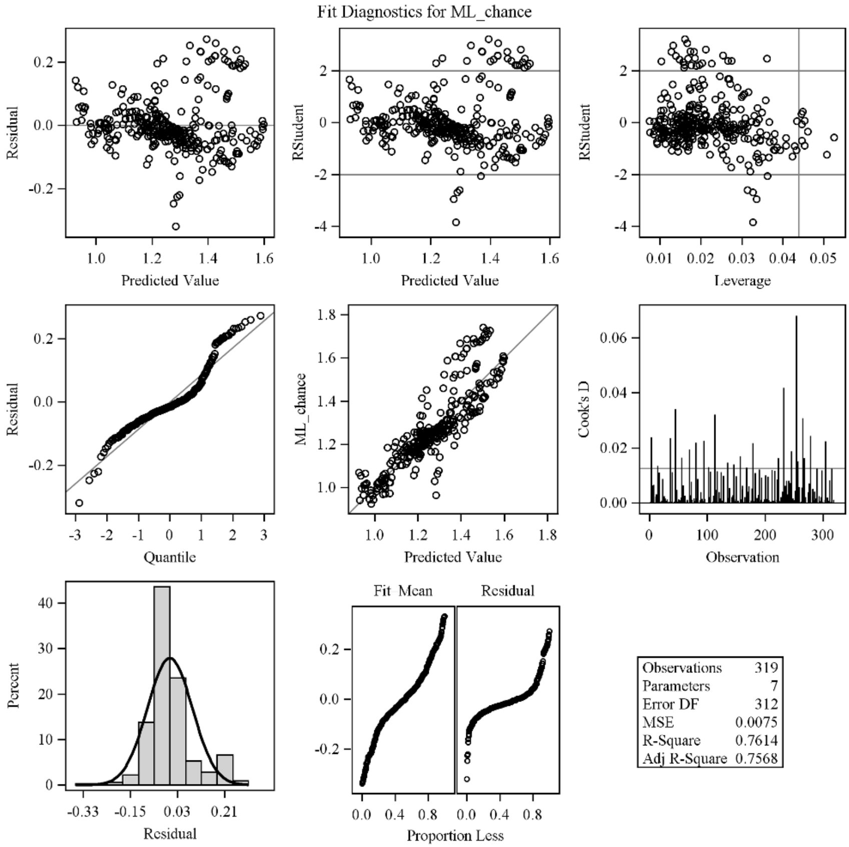


Figure 9.1: Fit diagnostics for the regression model applied for nearest neighbor hot deck augmentations

in the plot for the residuals by predicted values of ML_{chance} has no particular pattern. The points in the residual by quantile plot largely follow the diagonal line and the histogram of the residual shows an approximate normal distribution, so that it can be assumed that the residuals are largely normally distributed. The variance inflation factor as shown in table 9.17 is lower than 10 for all variables, so that no collinearity is found. All in all, to apply a regression model to this data is disputable, which is one of the findings in chapter 7.3.1. However, because minimum criteria apply, general findings can be derived for comparison purposes.

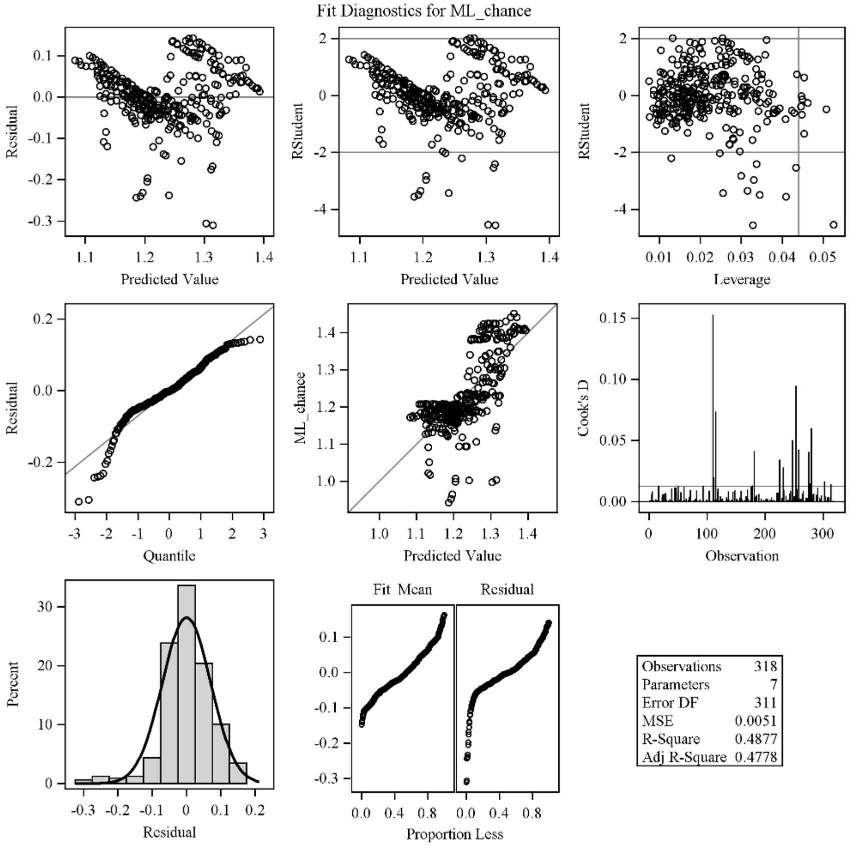


Figure 9.2: Fit diagnostics for the regression model applied for logistic regression augmentations

Variable	DF	Parameter Estimate	Stand. Err.	t-Value	Pr> |t|	Tolerance	Variance Inflation
Intercept	1	0.37085	0.06957	5.33	<.0001		
Overlaps	1	0.000012	0.000002	7.69	<.0001	0.45362	2.20448
Donors	1	−0.000003	5.349E−7	−6.33	<.0001	0.47536	2.10369
VCs	1	−0.000019	0.000005	−3.56	0.0004	0.60939	1.64098
Rsqu	1	0.58353	0.05207	11.21	<.0001	0.43022	2.32440
k	1	0.09471	0.00673	14.07	<.0001	0.59500	1.68066
IQV	1	0.57043	0.05743	9.93	<.0001	0.35529	2.81458

Table 9.17: Parameter estimates for the regression model explaining ML_{chance} by the descriptive data for logistic regression augmentations, including tolerance and variance inflation

Glossary

Across case analysis the overall analysis of augmentation results from the case study, including all cases, target variables, and methods

Augmentation frame describes all possible sets of elements, link variables, and target variables relevant to the augmentation process. All samples are drawn therefrom.

Below the line media all channels through which customers can directly and individually be reached and where access is limited, e.g. email, letter, SMS, or promoters at the point of sale

Between case analysis the analysis of augmentation results with different sources, but using the same target variable

Conversion the reaction to a marketing communication, e.g. a sale, a response, the participation in a raffle, or any other specified desired customer activity

Conversion probability lift indicates the increase of correctly selected customers when using augmentation results for target group selection as compared to the number of correctly selected customers when not using augmentation results

Conversion rate the number of conversions, divided by the number of recipients of a marketing campaign

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4

Customer a person who has already had a contact with a company, e.g. bought a product or service, has taken interest in doing so, or subscribed for promotional communication

Data augmentation adding supplemental data to the customer database based on similarities of elements instead of a unique identifier

Data utility assesses how useful data is in a simple comparison between the cost of acquiring and retaining (up-to-date) data and their associated value for the company

Donor unit contains customers and possibly other persons available from the source

Elements rows in a database table. They are used synonymously here with customers, persons, or observations.

Even missing data pattern observations and variables can be grouped or sorted in a way that a rule can be established describing the pattern

External evaluation assessing the utility of the results in terms of return on marketing investment

(External) source every data source that is not directly collected and saved on a personal level with the existing customer data. It is defined by not having unique identifiers for the customers.

Haphazard missing data pattern data is randomly missing for observations or variables, so that no rule can be established as to which data is missing

Hit the augmented value is equal to the true, but unobserved value

Intentionally missing data a decision has been made at some point that data is missing, either by systematically omitting observations or variables from data collection (e.g. sub-sampling or split questionnaire survey design) or by acquiring a source from which it is known that it does not contain all observations or variables of the recipient unit

Internal evaluation validating the augmentation results and derived KPIs in comparison to the true values

Link variables variables present in both the customer database and the external source

Link variable class a unique combination of link variable values similar for all elements contained in this class

Model lift indicates the increase of correct hits resulting from augmentation as compared to the achievable hit rate by a random allocation of target values within the recipient unit

Overlap the number of elements that donor unit and recipient unit have in common

Population a unit of elements, i.e. people, conforming to a set of defined criteria

Recipient unit contains all customers relevant to the augmentation frame

Relevance attracting the positive attention of the recipient to the content or offer

Segmenting customers allocating customers to several target segments in order to distinguish them, usually in terms of offers, prices, or creative appeal

Selecting customers choosing the best target group for a given offer or communication, often given certain constraints like target group size, budget, or required conversion rate

Size the total number of elements in the donor unit

Source data mechanism refers to the mechanism describing whether a person has been observed in an external source or not

Targeting the ability to differentiate customers based on data in order to provide them individual and personalized offers

Target group a selected group of prospective customers for which the cost-benefit ratio of marketing is highest

Target variables variables of interest in the external source, which are not available from the customer database

Unit (case study) the unit that is measured and analyzed

Within case analysis the analysis of augmentation results within a case of the case study, e.g. the analysis of augmentation results using different methods, but using the same target variable and source

References

Abel, R. B. (2011). Zulässigkeit von Online-Marketing auf der Basis von Erkenntnissen aus sozialen Netzwerken. In C. Bauer, G. Greve, & G. Hopf (Eds.), *Online Targeting und Controlling: Grundlagen, Anwendungsbeispiele, Praxisbeispiele* (p. 124-136). Wiesbaden: Gabler / Springer Fachmedien.

Acker, O., Gröne, F., Blockus, A., & Bange, C. (2011). In-memory analytics: Strategies for real-time CRM. *Database Marketing & Customer Strategy Management*, *18*(2), 129-136.

Adamek, J. C. (1994). Fusion: Combining data from separate sources. *Marketing Research*, *6*(3), 48-50.

Adriaans, P., & Zantinge, D. (1998). *Data mining.* Harlow: Addison-Wesley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Albright, S. C., Winston, W. L., & Zappe, C. J. (2011). *Data analysis, optimization, and simulation modeling* (4th ed.). Mason, Ohio: South-Western, Cengage Learning.

Allison, P. (2012). *Handling missing data by maximum likelihood.* Statistical Horizons. Retrieved September 21, 2012, from `http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf`

Allison, P. (2013). *What's the best R-squared for logistic regression?*

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4

Statistical Horizons. Retrieved September 6, 2013, from `http://www.statisticalhorizons.com/r2logistic`

Altmann, D. G., & Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *BMJ*, 102.

Ang, L., & Buttle, F. (2009). Customer development strategies for exceeding expectations: An exploratory study. *Journal of Database Marketing & Customer Strategy Management*, *16*, 267-275.

Arndt, D., & Koch, D. (2002). Datenschutz im Web Mining: Rechtliche Aspekte des Umgangs mit Nutzerdaten. In H. H., M. M., & K. D. Wilde (Eds.), *Handbuch Web Mining im Marketing: Konzepte, Systeme, Fallstudien* (p. 77-103). Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Arnold, S. (2011). Data fusion, discovery, and the next big thing in research. *Information Today*, *April*, 20-21.

Axel Springer AG. (2012). *Über die VerbraucherAnalyse.* VerbraucherAnalyse. Retrieved August 9, 2013, from `http://www.verbraucheranalyse.de/home`

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, *6*(4), 355-385.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2008). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (12th ed.). Berlin: Springer.

Baker, K., Harris, P., & O'Brien, J. (1989). Data fusion: An appraisal and experimental evaluation. *Journal of the Market Research Society*, *31*(2), 153-212.

Bakker, B. F. M. (2012). Micro-integration: State of the art. *Report on WP1: State of the Art on Statistical Methodologies for Data Integration*, 77-108.

Bankhofer, U., & Praxmarer, S. (1998). Angewandte Marktforschung und das Problem fehlender Daten. *Planung & Analyse*, *6*, 46-48.

Barney, J. (1991). Firm resources and sustained competitive advantage.

Journal of Management, *17*(1), 99-120.

Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. *The Qualitative Report*, *13*(4), 544-559.

Becker, J. (2009). *Marketing-Konzeption: Grundlagen des zielstrategischen und operativen Marketing-Managements.* München: Franz Vahlen.

Behme, W., & Mucksch, H. (2001). Anwendungsgebiete einer Data Warehouse-gestützten Informationsversorgung. In W. Behme & H. Mucksch (Eds.), *Data Warehouse-gestützte Anwendungen: Theorie und Praxiserfahrungen in verschiedenen Branchen* (p. 3-32). Wiesbaden: Gabler.

Bensberg, F. (2002). Segmentierung im Online-Marketing. In H. H., M. M., & K. D. Wilde (Eds.), *Handbuch Web Mining im Marketing: Konzepte, Systeme, Fallstudien* (p. 163-190). Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Berekoven, L., Eckert, W., & Ellenrieder, P. (2009). *Marktforschung: Methodische Grundlagen und praktische Anwendung* (12th ed.). Wiesbaden: Gabler / GWV Fachverlage.

Bernardo, J. M., & Smith, A. F. (1994). *Bayesian theory.* Chichester: John Wiley & Sons.

Berry, J. (2009). How should the goals for 'contact optimisation' be set, and how should contact optimisation be managed in a multi-channel inbound and outbound environment? *Journal of Database Marketing & Customer Strategy Management*, *16*, 241-245.

Berry, M., & Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed.). Indianapolis: Wiley Publishing.

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). *Database marketing: Analyzing and managing customers* (Vol. 29). New York: Springer Science + Business Media.

Bleiholder, J., & Naumann, F. (2008). Data fusion. *ACM Computing*

Surveys, *41*(1), 1-41.

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364-366.

Box, G. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis* (Vol. 107). Reading, MA: Addison-Wesley.

Breur, T. (2011). Data analysis across various media: Data fusion, direct marketing, clickstream data and social media. *Journal of Direct, Data and Digital Marketing Practice*, *13*(October/December 2011), 95-105.

Brogini, M. (1998). *Über Kundengruppen zur Marktstruktur: Das Modell der Segmentintensität* (Vol. 18). Bern: Paul Haupt.

Bruhn, M. (2009). *Kommunikationspolitik: Systematischer Einsatz der Kommunikation für Unternehmen* (5th ed.). München: Franz Vahlen.

Bull, N., & Passewitz, G. (1994). *Finding customers: Market segmentation.* Ohio State University. Retrieved November 13, 2012, from `http://ohioline.osu.edu/cd-fact/1253.html`

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378-394.

Bundesdatenschutzgesetz. (BDSG). *Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), zuletzt geändert durch Artikel 15 Abs. 53 des Gesetzes vom 5. Februar 2009 (BGBI. I S. 160).*

Carlson, B. L., Cox, B. G., & Bandeh, L. S. (2012). *SAS macros useful in imputing missing survey data.* Mathematica Policy Research. Retrieved October 9, 2012, from `http://mathematica-mpr.org/publications/PDFs/misssurdata.pdf`

Casteleyn, J., Mottart, A., & Rutten, K. (2009). How to use facebook in your market research. *International Journal of Market Research*, *51*(4), 439-447.

Chen, Y., & Iyer, G. (2002). Research note: Consumer addressability and customized pricing. *Marketing Science*, *21*(2), 197-208.

Chen, Y., Narasimhan, C., & Zhang, Z. J. (2001). Individual marketing with imperfect targetability. *Marketing Science*, *20*(1), 23-41.

Cibella, N., Guigó, M., Scanu, M., & Tuoto, T. (2012). Literature review update on record linkage: Introduction. In *Report on WP1: State of the art on statistical methodologies for data integration* (p. 8-11). Athens: European Conference on Quality in Official Statistics.

Cochran, W. G. (1983). Historical perspective. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys: Volume 2, theory and bibliography* (p. 11-25). New York: Academic Press.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). London: Chapman & Hall.

Cuthbertson, R., & Messenger, S. (2013). *Marrying market research and customer relationship marketing.* Ipsos. Retrieved February 26, 2013, from `http://www.ipsos.com/loyalty/sites/ipsos.com.loyalty/files/Marrying_MR_2013`

Dallmer, H. (2002). Das System des Direct Marketing: Entwicklungsfaktoren und Trends. In H. Dallmer (Ed.), *Das Handbuch Direct Marketing & More* (8th ed., p. 3-32). Wiesbaden: Gabler.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(1), 1-31.

Dekker, L. (1993). Simulation of systems. *Simulation Practice and Theory*, *1*, 1-3.

Dempster, A. P., & Rubin, D. B. (1983). Introduction. In W. G. Madow, I. Okin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys: Volume 2, theory and bibliographies* (p. 3-10). New York: Academic Press.

Dey, D., Sarkar, S., & De, P. (1998). A probabilistic decision model for entity matching in heterogeneous databases. *Management Science*, *44*(10), 1379-1395.

Dialog Marketing Monitor. (2009). *Dialogmarketing Deutschland 2009: Dialog Marketing Monitor Studie 21.* Deutsche Post Dialogmarketing.

Dialog Marketing Monitor. (2012). *Dialogmarketing Deutschland 2012: Dialog Marketing Monitor Studie 24.* Deutsche Post Dialogmarketing.

Dong, X., Manchanda, P., & Chintagunta, P. K. (2009). Quantifying the benefits of individual-level targeting in the presence of firm strategic behavior. *Journal of Marketing Research*, *XLVI*, 207-221.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* Chichester, West Sussex: John Wiley & Sons.

D'Orazio, M., Di Zio, M., & Scanu, M. (2010). *Old and new approaches in statistical matching when samples are drawn with complex survey designs.* Department of Statistical Sciences University of Padua. Retrieved November 4, 2013, from `http://homes.stat.unipd.it/mgri/SIS2010/Program/18-SSXVIII_Luzi/872-1502-1-DR.pdf`

Däubler, W., Klebe, T., Wedde, P., & Weichert, T. (2010). *Bundesdatenschutzgesetz: Kompaktkommentar zum BDSG. 3. Aufl.* Frankfurt a. M.: Bund-Verlag.

Dull, S. F., Stephens, T., & Wolfe, M. T. (2001). *How much are customer relationship management capabilities really worth?: What every CEO should know: Executive summary.* Accenture. Retrieved March 4, 2010, from `http://www.accenture.com/NR/rdonlyres/86DB4186-2474-47AC-8FA6-9EFB183CF881/0/crm_komp_value.pdf`

Dyk, D. A. v., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1-50.

Eickmeier, F., & Hansmersmann, P. (2011). Datenschutzkonforme Nutzerprofile im Internet. In C. Bauer, G. Greve, & G. Hopf (Eds.), *Online Targeting und Controlling: Grundlagen, Anwendungsfelder, Praxisbeispiele* (p. 96-121). Wiesbaden: Gabler / Springer Fachmedien.

Eimeren, B. v., & Frees, B. (2012). Ergebnisse der ARD / ZDF-Onlinestudie 2012: 76 Prozent der Deutschen online – Neue Nutzungssituationen durch mobile Endgeräte. *Media Perspektiven*, *7-8*, 362-379.

Eimeren, B. v., Gerhard, H., & Frees, B. (2002). ARD / ZDF-Online-Studie 2002: Entwicklung der Onlinenutzung in Deutschland – Mehr

Routine, weniger Entdeckerfreude. *Media Perspektiven*, *8*, 346-362.

Eimeren, B. v., Oehmichen, E., & Schröter, C. (1997). *ARD-Online-Studie 1997: Onlinenutzung in Deutschland – Nutzung und Bewertung der Onlineangebote von Radio- und Fernsehsendern.* ard-zdf-onlinestudie.de. Retrieved February 6, 2013, from `http://www.ard-zdf-onlinestudie.de/fileadmin/Online97_98/Online97.pdf`

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532-550.

European Conference on Quality in Official Statistics. (2012). *Report on wp1: State of the art on statistical methodologies for data integration.* Athens.

Even, A., Shankaranarayanan, G., & Berger, P. D. (2010). Inequality in the utility of customer data: Implications for data management and usage. *Journal of Database Marketing & Customer Strategy Management*, *17*, 19-35.

Feinberg, F. M., Krishna, A., & Zhang, Z. J. (2002). Do we care what others get?: A behaviorist approach to targeted promotions. *Journal of Marketing Research*, *XXXIX*, 277-291.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, *12*(2), 219-245.

Fogarty, K. (2008). How to get real-time analytics from a data warehouse. In *CIO Insight, Supplement* (p. 7-9).

Ford, B. L. (1983). An overview of hot-deck procedures. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys: Volume 2, theory and bibliographies* (p. 185-207). New York: Academic Press.

Freter, H. (1997). *Markt- und Kundensegmentierung: Kundenorientierte Markterfassung und -bearbeitung* (2nd ed.). Stuttgart: W. Kohlhammer.

Frost, J. (2013, May). *Regression analysis: How do I interpret R-squared and assess the goodness-of-fit?* The Minitab Blog.

Retrieved September 6, 2013, from `http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit`

Garg, R., Rahman, Z., & Kumar, I. (2010). Evaluating a model for analyzing methods used for measuring customer experience. *Journal of Database Marketing & Customer Strategy Management*, *17*, 78-90.

Gesetz gegen den unlauteren Wettbewerb. (UWG). *Gesetz gegen den unlauteren Wettbewerb in der Fassung der Bekanntmachung vom 3. März 2010 (BGBl. I S. 254), zuletzt geändert durch Artikel 2 G vom 29. Juli 2009 I 2413.*

Gilula, Z., McCulloch, R. E., & Rossi, P. E. (2006). A direct approach to data fusion. *Journal of Marketing Research*, *XLIII*(9), 73-83.

Greve, G., Hopf, G., & Bauer, C. (2011). Einführung in das Online Targeting. In C. Bauer, G. Greve, & G. Hopf (Eds.), *Online Targeting und Controlling: Grundlagen, Anwendungsbeispiele, Praxisbeispiele* (p. 4-21). Wiesbaden: Gabler / Springer Fachmedien.

Guigó, M. (2012). Preface. In *Report on WP1: State of the art on statistical methodologies for data integration* (p. 4-6). Athens: European Conference on Quality in Official Statistics.

Hagedorn, J., Bissantz, N., & Mertens, P. (1997). Data Mining (Datenmustererkennung): Stand der Forschung und Entwicklung. *Wirtschaftsinformatik*, *39*(6), 601-612.

Hattum, P., & Hoijtink, H. (2008a). Improving your sales with data fusion. *Journal of Database Marketing & Customer Strategy Management*, *16*(81), 7-14.

Hattum, P., & Hoijtink, H. (2008b). The proof of the pudding is in the eating. *Journal of Database Marketing & Customer Strategy Management*, *15*(80), 267-284.

Hattum, P., & Hoijtink, H. (2010). Reducing the optimal to a useful number of clusters for model-based clustering. *Journal of Targeting, Measurement and Analysis for Marketing*, *18*(2), 139-154.

Hebestreit, S. (2009, April). *Das Gold des 21. Jahrhunderts.* Frankfurter Rundschau. Retrieved February 12, 2013, from http://www.fr-online.de/datenschutz/private-informationen-das-gold-des-21--jahrhunderts,1472644,2746314.html

Heinrich, C. (2011). *Forschen mit Facebook: Wissenschaftler nutzen Soziale Netzwerke als Labor für ihre soziologischen Studien.* ZEIT Online. Retrieved February 8, 2013, from http://www.zeit.de/2011/23/T-Facebook

Helm, S., & Günter, B. (2006). Kundenwert: Eine Einführung in die theoretischen und praktischen Herausforderungen der Bewertung von Kundenbeziehungen. In B. Günter (Ed.), *Kundenwert: Grundlagen, Innovative Konzepte, Praktische Umsetzungen* (3rd ed., p. 4-38). Wiesbaden: Gabler / GWV Fachverlage.

Herzog, T. N., & Rubin, D. B. (1983). Using multiple imputations to handle nonresponse in sample surveys. In W. G. Madow, I. Olkin, & D. B. Rubin (Eds.), *Incomplete data in sample surveys: Volume 2, theory and bibliographies* (p. 209-254). New York: Academic Press.

Heun, F. W. (2002). Verlorene Kunden zurückholen. *Sales Business*, *11*(12), 20-21.

Höhl, M. (1999). One-to-One-Marketing. *Wirtschaftsinformatik*, *41*(1), 74-76.

Hipperson, T. (2010). The changing face of data insight and its relationship to brand marketing. *Journal of Database Marketing & Customer Strategy Management*, *17*, 262-266.

Hippner, H., Leber, M., & Wilde, K. D. (2002). Kundendatenbanken als strategischer Erfolgsfaktor. In K. D. Wilde (Ed.), *Marketing Information Provider: Professionelle Qualifizierung Ihrer Kundendaten* (p. 9-31). Düsseldorf: Verlagsgruppe Handelsblatt.

Hippner, H., Rentzmann, R., & Wilde, K. D. (2002). Marketing Information Providing: der gekaufte Erfolg. In K. D. Wilde, H. Hippner, & P. Hanser (Eds.), *Marketing Information Provider: Professionelle*

Qualifizierung Ihrer Kundendaten (p. 73-82). Düsseldorf: Verlagsgruppe Handelsblatt.

Hippner, H., & Wilde, K. D. (2001). Der Prozess des Data Mining im Marketing. In H. Hippner, U. Küsters, M. Meyer, & K. Wilde (Eds.), *Handbuch Data Mining im Marketing: Knowledge Discovery in Marketing Databases* (p. 21-91). Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Hofsäss, M., & Engel, D. (2006). *Praxishandbuch Mediaplanung: Forschung, Studien und Werbewirkung – Mediaagenturen und Planungsprozess – Mediagattungen und Werbeträger* (4th ed.). Berlin: Cornelsen.

Homburg, C., & Sieben, F. G. (2005). Customer Relationship Management (CRM): Strategische Ausrichtung statt IT-getriebenem Aktivismus. In M. Bruhn & C. Homburg (Eds.), *Handbuch Kundenbindungsmanagement* (5th ed., p. 501-528). Wiesbaden: Gabler.

Hopf, G. (2011). In vier Schritten zum Online Geschäftsmodell. In C. Bauer, G. Greve, & G. Hopf (Eds.), *Online Targeting und Controlling: Grundlagen, Anwendungsbeispiele, Praxisbeispiele* (p. 24-42). Wiesbaden: Gabler / Springer Fachmedien.

Huber, M. (2008). *Statistics I: Introduction to ANOVA, regression, and logistic regression course notes.* Cary, NC, USA: SAS Institute Inc.

Hudson, D., Seah, L.-H., Hite, D., & Haab, T. (2004). Telephone presurveys, self-selection, and non-response bias to mail and internet surveys in economic research. *Applied Economics Letters*, *11*, 237-240.

Huldi, C. (2002). Nutzenpotenziale von CRM im Unternehmen. In H. Dallmer (Ed.), *Das Handbuch Direct Marketing & More* (8th ed., p. 1103-1119). Wiesbaden: Gabler.

Insitut für Demoskopie Allensbach. (2009). *Zu wenig Datenschutz?: Die meisten sind mit persönlichen Daten vorsichtiger geworden.* Allensbacher Berichte. Retrieved February 6, 2013, from http://www.ifd-allensbach.de/uploads/tx_reportsndocs/prd_0906.pdf

Institut für Medien- und Konsumentenforschung. (2012a). *Communica-*

tion Networks (CN) 15.0. Institut für Medien- und Konsumentenforschung. Retrieved August 9, 2013, from `http://www.imuk.de/cn.html`

Institut für Medien- und Konsumentenforschung. (2012b). *Typologie der Wünsche.* Institut für Medien- und Konsumentenforschung. Retrieved August 9, 2013, from `http://www.imuk.de/tdw.html`

Iyer, G., Soberman, D., & Villas-Boas, J. M. (2005). The targeting of advertising. *Marketing Science*, *24*(3), 461-476.

Jiang, Z., Sarka, S., De, P., & Dey, D. (2007). A framework for reconciling attribute values from multiple data sources. *Marketing Science*, *53*(12), 1946-1963.

Kamakura, W. A., & Wedel, M. (1997). Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, *34*(1), 485-498.

Kamakura, W. A., & Wedel, M. (2000). Factor analysis and missing data. *Journal of Marketing Research*, *XXXVII*(November), 490-498.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G. M., Paxson, V., & Savage, S. (2008). *Spamalytics: An empirical analysis of spam marketing conversion.* University of California School of Engineering. Retrieved January 15, 2013, from `http://cseweb.ucsd.edu/~klevchen/kklevps-ccs08.pdf`

Keller, G. (2008). *Statistics for management and economics: Identify, compute, interpret.* USA: South-Western College Pub.

Kelly, S. (2007). Total recall: Transforming the possibilities of customer intelligence in an age of intelligent commerce. *International Journal of Market Research*, *49*(4 Data Integration Special Issue), 425-433.

Kim, J., Baek, S., & Cho, S. (2004). A preliminary study on common variable selection strategy in data fusion. *Advances in Consumer Research*, *31*, 716-720.

Küppers, B. (1999). *Data Mining in der Praxis: Ein Ansatz zur Nutzung der Potenziale von Data Mining im betrieblichen Umfeld* (Vol. 2373). Frankfurt a. M.: Peter Lang.

Krämer, B. (2010). *Mehr über den Kunden wissen: Die Nutzung von Ergebnissen der Markt-Media-Forschung im Database Marketing – Ein Verfahren zur Datenanreicherung am Beispiel der Kundendatenbank von Lufthansa Miles & More und der Markt-Media-Studie Communication Networks 13.0.* Mainz: Hausarbeit zur Erlangung des akademischen Grades Diplom-Medienwirtin.

Kroeber-Riel, W. (1988). Kommunikation im Zeitalter der Informationsüberlastung. *Marketing: Zeitschrift für Forschung und Praxis*, *10*(3), 182-189.

Küsters, U. (2001). Data Mining Methoden: Einordnung und Überblick. In H. Hippner, U. Küsters, M. Meyer, & K. Wilde (Eds.), *Handbuch Data Mining im Marketing: Knowledge Discovery in Marketing Databases* (p. 95-130). Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Kuhner, C. (2013). *How to enrich customer data for efficient marketing analytics.* The Cross-Channel Conversation. Retrieved August 13, 2013, from `http://blog.neolane.com/marketing-analytics/marketing-analytics-enrich-customer-data/`

Kutter, I. (2013). Facebook macht dick, blöd und wirkt wie Sex: Ist das tatsächlich zu beweisen? *Die Zeit*, 36.

Laase, C. M. (2011). Neue Wege im Online Targeting. In C. Bauer, G. Greve, & G. Hopf (Eds.), *Online Targeting und Controlling: Grundlagen, Anwendungsbeispiele, Praxisbeispiele* (p. 198-210). Wiesbaden: Gabler / Springer Fachmedien.

Lambertz, P. (2009). *Was erlaubt ist und was nicht: Die werbliche Ansprache nach dem neuen BDSG.* Datenschutz Praxis. Retrieved February 8, 2013, from `http://www.datenschutz-praxis.de/fachwissen/fachartikel/die-werbliche-ansprache-nach-dem-neuen-bdsg/`

Leitzmann, C.-J. (2002). Kampagnenmanagement zur Steuerung des Multi-Channel-Marketing: Eine Einführung mit Fokus E-Mail-Marketing. In

H. Dallmer (Ed.), *Das Handbuch Direct Marketing & More* (8th ed., p. 371-397). Wiesbaden: Gabler.

Leser, U., & Naumann, F. (2007). *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen.* Heidelberg: dpunkt.verlag.

Liehr, T. (1999). Data Warehouse + Data Mining = Wissen auf Knopfdruck?: Eine Perspektive für Marktforschung im Informationszeitalter. *Planung & Analyse*, *1*, 44-49.

Liehr, T. (2001). 'Data Matching' bei Finanzdienstleistungen: Steigerung des Share of Wallet bei Top-Kunden. In H. Hippner, U. Küsters, M. Meyer, & K. Wilde (Eds.), *Handbuch Data Mining im Marketing: Knowledge Discovery in Marketing Databases* (p. 725-740). Braunschweig: Friedr. Vieweg & Sohn Verlagsgesellschaft.

Link, J. (2000). Zur zukünftigen Entwicklung des Online Marketing. In J. Link (Ed.), *Wettbewerbsvorteile durch Online Marketing: Die strategischen Perspektiven elektronischer Märkte* (2nd ed.). Berlin: Springer.

Link, J., & Hildebrand, V. (1993). *Database Marketing und Computer Aided Selling: Strategische Wettbewerbsvorteile durch neue informationstechnologische Systemkonzeptionen.* München: Franz Vahlen.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (10th ed.). Hoboken, NJ: John Wiley & Sons.

Lugtig, P., Lensvelt-Mulders, G. J., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(5), 669-686.

Madow, W. G., Olkin, I., & Rubin, D. B. (1983). *Incomplete data in sample surveys: Volume 2, theory and bibliographies.* New York: Academic Press.

Mahrdt, N. (2009). *Crossmedia: Werbekampagnen erfolgreich planen und umsetzen.* Wiesbaden: Gabler / GWV Fachverlage.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719-748.

Marinell, G., & Steckel-Berger, G. (1995). *Einführung in die Bayes-Statistik: Optimaler Stichprobenumfang.* München: Oldenbourg Verlag.

marketinghub. (2009). *Marketing efficiency survey 2009.* www.marketinghub.ch. Retrieved February 2, 2013, from `http://www.marketinghub.ch/fileadmin/user_upload/DAM_09/M_Efficiency_09/Mefficiency_09__Leseprobe.pdf`

McCrary, M. (2009). Enhanced customer targeting with multi-stage models: Predicting customer sales and profit in the retail industry. *Journal of Targeting, Measurement and Analysis for Marketing*, *17*, 273-295.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (p. 105-142). New York: Academic Press.

McKay, L. (2009). Information overload. *CRM Magazine*, *13*(12), 30-35.

Meffert, H. (2000). *Marketing: Grundlagen marktorientierter Unternehmensführung – Konzepte, Instrumente, Praxisbeispiele* (9th ed.). Wiesbaden: Gabler.

Meffert, H., & Bruhn, M. (2009). *Dienstleistungsmarketing: Grundlagen, Konzepte, Methoden* (6th ed.). Wiesbaden: Gabler / GWV Fachverlage.

Meffert, H., Burmann, C., & Kirchgeorg, M. (2008). *Marketing: Grundlagen marktorientierter Unternehmensführung: Konzepte, Instrumente, Praxisbeispiele* (10th ed.). Wiesbaden: Gabler / GWV Fachverlage.

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data: 1. designated experiments.* Van Nostrand Rienhold.

Moe, W. W., & Fader, P. S. (2004). Dynamic conversion behavior at e-commerce sites. *Management Science*, *50*(3), 326-335.

Naseri, M. B., & Elliott, G. (2011). Role of demographics, social connectedness and prior internet experience in adoption of online shopping: Applications for direct marketing. *Journal of Targeting, Measurement* *and Analysis for Marketing*, *19*, 69-84.

Noller, S. (2009). Intelligentes Zielgruppenmanagement Online. In S. Duttenhöfer, B. Keller, & S. Vomhoff (Eds.), *Handbuch Zielgruppenmanagement* (p. 155-166). Frankfurt am Main: Fritz Knapp.

Nordholt, E. S. (1998). Imputation: Methods, simulation experiments and practical examples. *International Statistical Review / Revue Internationale de Statistique*, *66*(2), 157-180.

Okner, B. A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, *1*(3), 325-342.

Ozimek, J. (2010). Holding customer data: Why organisations need to start to consider their customers. *Journal of Database Marketing & Customer Strategy Management*, *17*, 1-5.

Palgrave Macmillan. (2013). *Journal of marketing analytics.* www.palgrave.com. Retrieved August 12, 2013, from `http://www.palgrave-journals.com/dbm/index.html`

Pauli, D. (2009). Direktmarketing und die Gewinnung von Kundendaten: Ist die Veranstaltung eines Gewinnspiels ein geeigneter Weg? *Wettbewerb in Recht und Praxis*, *10*, 245-249.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Pennsylvania State University. (2013a). *Conditional independence.* STAT 504: Analysis of Discrete Data. Retrieved February 12, 2013, from `https://onlinecourses.science.psu.edu/stat504/node/112`

Pennsylvania State University. (2013b). *Cochran-Mantel-Haenszel test.* STAT 504: Analysis of Discrete Data. Retrieved February 12, 2013, from `https://onlinecourses.science.psu.edu/stat504/node/113`

Perry, D. E. (2000). *Case studies.* 382C Empirical Studies in Software Engineering: Lecture 6. Retrieved September 11, 2014, from `http://users.ece.utexas.edu/~perry/education/382c/L06.pdf`

Petras, A. (2007). Die Befindlichkeit der Konsumenten erforschen. In J. Kalka & F. Allgayer (Eds.), *Zielgruppen: Wie sie leben, was sie kaufen, woran sie glauben* (2nd ed., p. 90-91). Landsberg am Lech: mi-Fachverlage/Redline.

Piller, F. (2006). *Mass Customization: Ein wettbewerbsstrategisches Konzept im Informationszeitalter* (4th ed.). Wiesbaden: Deutscher Universitäts-Verlag / GWV Fachverlage.

Pineau, V., & Slotwiner, D. (2003). *Probability samples vs. volunteer respondents in internet research: Defining potential effects on data and decision-making in marketing applications.* Knowledge Networks, Inc. Retrieved January 22, 2013, from `http://www.knowledgenetworks.com/insights/docs/Volunteer`

Plath, K.-U., & Frey, A.-M. (2009). Direktmarketing nach der BDSG-Novelle: Grenzen erkennen, Spielräume optimal nutzen. *Betriebs-Berater*, *34*, 1762-1768.

Powell, R. R. (1997). *Basic research methods for librarians* (3rd ed.). Westport, CT: Ablex Publishing Corporation.

Putten, P. (2010). *On data mining in context: Cases, fusion and evaluation* (No. 28). Amsterdam: Proefschrift ter verkrijging van de graad van Doctor aan de Universiteit Leiden.

Putten, P., Kok, J. N., & Gupta, A. (2002a). *Working paper 4342-02: Data fusion through statistical matching* (No. 4). Cambridge: MIT Sloan School of Management.

Putten, P., Ramaekers, M., Uyl, M. d., & Kok, J. (2002b). *A process model for a data fusion factory* (No. 14). Leiden.

Radner, D. (1980). *An example of the use of statistical matching in the estimation and analysis of the size distribution of income.* Social Security Administration, Office of Policy, Office of Research and Statistics, Division of Economic Research.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429, Applications and Case Studies), 54-63.

Rapp, R. (2002a). *Customer Relationship Managment: Das neue Konzept zur Revolutionierung der Kundenbeziehungen.* Frankfurt a. M.: Campus Verlag.

Rapp, R. (2002b). Die Rolle des Direct Marketing im CRM. In H. Dallmer (Ed.), *Das Handbuch Direct Marketing & More* (8th ed., p. 73-86). Wiesbaden: Gabler.

Ratner, B. (2001a). Finding the best variables for direct marketing models. *Journal of Targeting, Measurement and Analysis for Marketing*, *9*(3), 270-296.

Ratner, B. (2001b). Identifying the best customers: Descriptive, predictive and look-alike profiling. *Journal of Targeting, Measurement and Analysis for Marketing*, *10*(1), 66-78.

Ratner, B. (2003). *Statistical modeling and analysis for database marketing: Effective techniques for mining big data.* Boca Raton, FL: Chapman & Hall/CRC.

Reichheld, F., & Schefter, P. (2001). Warum Kundentreue auch im Internet zählt. *Harvard Business Manager*, *23*(1), 70-80.

Rhee, E. (2010). Multi-channel management in direct marketing retailing: Traditional call center versus internet channel. *Journal of Database Marketing & Customer Strategy Management*, *17*, 70-77.

Roberts, M. L., & Berger, P. D. (1999). *Direct marketing management* (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Robinson, S. (2004). *Simulation: The practice of model development and use.* Chichester, West Sussex: John Wiley & Sons.

Rodgers, W. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, *2*(1), 91-102.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, *15*(4), 321-340.

Rässler, S. (2000). Imputation of missing data in surveys. *Jahrbücher für*

Nationalökonomie und Statistik, *220*(1), 64-94.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications and alternative bayesian approaches* (Vol. 168). New York: Springer.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, *33*(1&2), 153-171.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-192.

Rubin, D. B. (1986). Statistical matching using file concatenationwith adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87-94.

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

SAS Institute Inc. (2012). *Cochran-Mantel-Haenszel statistics* (2nd ed.). SAS/STAT(R) 9.2 User's Guide. Retrieved October 18, 2012, from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_freq_a0000000648.htm

SAS Institute Inc. (2014a). *Generalized coefficient of determination* (2nd ed.). SAS/STAT(R) 9.2 User's Guide. Retrieved January 7, 2014, from http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect031.htm

SAS Institute Inc. (2014b). *Goodness-of-fit tests* (3rd ed.). SAS/STAT(R) 9.2 User's Guide. Retrieved September 22, 2014, from http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect037.htm

SAS Institute Inc. (2014c). *Homogeneity of Variance in One-Way Models.* SAS/STAT(R) 9.22 User's Guide. Retrieved September 22, 2014, from http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glm_a0000000869.htm

SAS Institute Inc. (2014d). *Usage Note 38384: Interpreting the results of the SOLUTION option in the MODEL statement of PROC GLM.* Knowledge Base: Samples & SAS Notes. Retrieved September 24, 2014, from http://support.sas.com/kb/38/384.html

Schaar, P. (2011). *Bundesdatenschutzgesetz: Text und Erläuterung.* Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit. Retrieved February 8, 2013, from http://www.bfdi.bund.de/SharedDocs/Publikationen/Infobroschueren/INFO1_Januar_2011.pdf?__blob=publicationFile

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schiff, M. (2010). *Data augmentation and enhancement: Thinking outside the organization.* www.tdwi.org. Retrieved August 13, 2013, from http://tdwi.org/articles/2010/03/11/data-augmentation.aspx

Schmidberger, M., & Babiuch-Schulze, A. (2009). Optimierung von Direktmarketing durch systematische Kundendaten-Analyse. In S. Duttenhöfer, B. Keller, & S. Vomhoff (Eds.), *Handbuch Zielgruppenmanagement* (p. 85-94). Frankfurt a. M.: Fritz Knapp.

Schweiger, A., & Wilde, K. D. (1993). Database Marketing: Aufbau und Management. In W. Hilke (Ed.), *Direkt-Marketing* (Vol. 47, p. 89-125). Wiesbaden: Gabler.

Scovotti, C., & Spiller, L. D. (2006). Revisiting the conceptual definition of direct marketing: Perspectives from practioners and scholars. *The Marketing Management Journal*, *16*(2), 188-202.

Sen, A., & Srivastava, M. (1990). *Regression analysis: Theory, methods, and applications.* New York: Springer.

Sharot, T. (2007). The design and precision of data-fusion studies. *International Journal of Market Research*, *49*(4: Data Integration Special Issue), 449-470.

Shaw, R., & Stone, M. (1988). *Database marketing: Strategy and implementation.* United States of America: John Wiley & Sons.

Simon, P. (2013). *Too big to ignore: The business case for big data.* New York: Wiley.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238-241.

Singh, A. (2013). *Is big data the new black gold?* Wired © 2012 Condé Nast Digital. Retrieved February 13, 2013, from `http://insights.wired.com/profiles/blogs/is-big-data-the-new-black-gold#axzz2NJ74HGtA`

Smith, J. A., Boyle, B. A., & Cannon, H. M. (2010). Survey-based targeting fine-tunes television media planning: A case for accuracy and cost efficiency. *Journal of Advertising Research*, 428-439.

Statistisches Bundesamt. (2013). *Statistisches Jahrbuch Deutschland und Internationales 2013.* Wiesbaden: Statistisches Bundesamt.

Tanner, M. A., & Wong, W. H. (1997). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528-540.

Telemediengesetz. (TMG). *Telemediengesetz vom 26. Februar 2007 (BGBl. I S. 179), das zuletzt durch Artikel 1 des Gesetzes vom 31. Mai 2010 (BGBl. I S. 692) geändert worden ist.*

The SmartAgent Company. (2013). *Smartagent's activities.* The SmartAgent Company. Retrieved July 30, 2013, from `http://89.146.41.141/smartagents-activities/activities.html`

Tiedtke, D. (2000). Bedeutung des Online Marketing für die Kommunikationspolitik. In J. Link (Ed.), *Wettbewerbsvorteile durch Online Marketing: Die strategischen Perspektiven elektronischer Märkte* (2nd ed., p. 77-119). Berlin: Springer-Verlag.

tns Infratest. (2012). *(N)ONLINER Atlas 2012: Basiszahlen für Deutschland – Eine Topgrafie des digitalen Grabens durch Deutschland – Nutzung und Nichtnutzung des Internets, Strukturen und regionale Verteilung.* Initiative D21. Retrieved February 6, 2013, from `http://www.initiatived21.de/wp-content/uploads/2012/06/NONLINER-Atlas-2012-Basiszahlen-f%C3%BCr-Deutschland.pdf` tns infratest. (2012). *TNS EX-A-MINE: Data Fusion* (No. 124). tns infratest. Retrieved July 20, 2012, from `http://www.tns-infratest.com/marketing_tools/pdf/EXAMINE/EXAMINE_DataFusion_engl.pdf`

UCLA. (2014). *Regression with sas: Chapter 2 – regression diagnostics.* UCLA Institut for digital research and education. Retrieved October 13, 2014, from `http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter2/sasreg2.htm`

Vossebein, U. (2000). Grundlegende Bedeutung der Marktsegmentierung für das Marketing. In W. Pepels (Ed.), *Marktsegmentierung: Marktnischen finden und besetzen* (p. 19-46). Heidelberg: Sauer.

Wagner, E. (2010). Datenschutz bei Kundenkarten. *Datenschutz und Datensicherheit*, *34*(1), 30-33.

Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining: A practical guide.* San Francisco: Morgan Kaufmann Publishers.

Wendt, F. (1977). *Beschreibung einer Fusion: Methodenbereicht der Frauen-Typologie 3 fusioniert in MA 76* (Vol. 21). Hamburg: Gruner + Jahr.

Wendt, F., & Wendt, I. (1983). *Vom Leser pro Nummer zur Nutzungswahrscheinlichkeit: Teil 2* (Vol. 5). Frankfurt a. M.: Media Micro-Census.

Wilcox, A. R. (1973). Indices of qualitative variation and political measurement. *Western Political Quarterly*, *26*(2), 325-343.

Wilde, K. D. (2001). Data Warehouse, OLAP und Data Mining im Marketing: Moderne Informationstechnologien im Zusammenspiel. In H. Hippner, U. Küsters, M. Meyer, & K. Wilde (Eds.), *Handbuch Data Mining im Marketing: Knowledge Discovery in Marketing Databases* (p. 1-19). Braunschweig: Friedr. Vieweg & Sohn.

Winand, U., & Pohl, W. (2000). Die Vertrauensproblematik in elektronischen Netzwerken. In J. Link (Ed.), *Wettbewerbsvorteile durch Online Marketing: Die strategischen Perspektiven elektronischer Märkte* (2nd ed., p. 261-277). Berlin: Springer.

Wirtz, B. W. (2009). *Medien- und Internetmanagement* (6th ed.). Wiesbaden: Gabler / GWV Fachverlage.

Woo, K., & Fock, H. K. (2004). Retaining and divesting customers: An exploratory study of right customers, at-risk right customers, and wrong customers. *The Journal of Services Marketing*, *18*(2/3), 187-197.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: Sage.

Yin, R. K. (2014). *Case study research: Design and methods* (5th ed.). Thousand Oaks, CA: Sage.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neederlandica*, *66*(1), 41-63.

Curriculum Vitae

Bettina Hüttenrauch (geb. Krämer) has been born on April 5th, 1986 in Mannheim, Germany. She completed her high school graduation (Abitur) in 2005 at Prälat-Diehl-Schule, Groß-Gerau. She studied Media Management at Johannes Gutenberg-University, Mainz, where she graduated with a Diploma on June 15th, 2010. In collaboration with Miles & More International GmbH, a subsidiary of Deutsche Lufthansa AG, she wrote her Diploma thesis *"Mehr über den Kunden wissen: Die Nutzung von Ergebnissen der Markt-Media-Forschung im Database Marketing – Ein Verfahren zur Datenanreicherung am Beispiel der Kundendatenbank von Lufthansa Miles & More und der Markt-Media-Studie Communication Networks 13.0"* – "Knowing more about the customers: Using the results of marketmedia research in database marketing – an approach to data augmentation based on the example of Lufthansa Miles & More's customer database and the market media study Communication Networks 13.0". Upon completion of her thesis, she started working at Miles & More International GmbH as a database marketing analyst. In 2011, she decided to write a doctoral thesis in part-time as an external PHD student at the Johannes Gutenberg-University, Mainz, supervised by Prof. Dr. Heinz-Werner Nienstedt. The dissertation was completed with the oral exam on October 12th, 2015. She is currently working as a project manager at Deutsche Lufthansa AG and is responsible for building up the "Analytics Factory" for an advanced analytics program.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4

Abstract

An increasing amount of data is needed in order to segment and select customers for personalized and targeted marketing activities. But the internal data basis of companies is often not sufficient for targeting purposes. At the same time, data has never been easier available externally, e.g. from website click behavior, surveys, or social media. Data augmentation is a beneficial tool for harnessing this information. The results can be used to manage direct marketing campaigns. But external sources are used hesitantly, because no validation of augmentation results is possible. Moreover, it cannot be assessed upfront whether the augmentation results lead to an increase of conversion rates. We conduct a case study approach in order to test the suitability of different external sources for data augmentation. The result of our case studies is a set of guidelines for database marketing analysts approaching data augmentation projects with external sources. It gives guidance on how to identify relevant variables. It comprises a checklist of suitable source characteristics and advice on which augmentation method to use. We further analyze the fact that external sources of real world marketing data are not representatively sampled. In this context we could show, that the source data mechanism is negligible in a categorical setting, where the predictability of link variables regarding target variables is strong. In general we have been able to show that feasible sources lead to significant conversion probability lifts. As a result, the hesitation regarding data augmentation in database marketing is unfounded.

B. Hüttenrauch, *Targeting Using Augmented Data in Database Marketing*,
DOI 10.1007/978-3-658-14577-4