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 re-

ceive 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 with-

out 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.

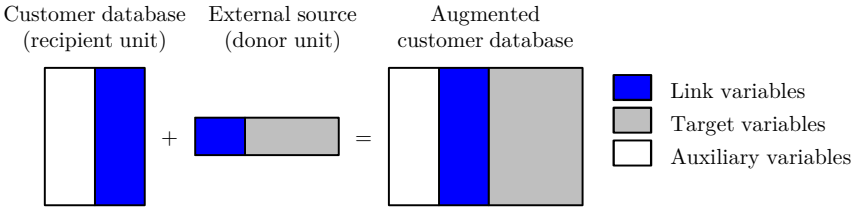

Figure 1.1: Schematic illustration of data augmentation as derived from Putten et al. (2002a, p. 2)

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 unlim-

ited 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.