Paul Fink

Contributions to Reasoning on Imprecise Data

Imprecise Classification Trees, Generalized Linear Regression on Microaggregated Data and Imprecise Imputation

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 01.03.2018

Paul Fink

Contributions to Reasoning on Imprecise Data

Imprecise Classification Trees, Generalized Linear Regression on Microaggregated Data and Imprecise Imputation

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

1. Berichterstatter: Prof. Dr. Thomas Augustin
2. Berichterstatter: Prof. Michael Smithson, Ph.D.
3. Berichterstatter: Prof. Dr. Matthias Schmid

Eingereicht am 01.03.2018
Disputation am 15.06.2018

Zusammenfassung

Diese Schrift setzt sich in vier Beiträgen für eine vorsichtige statistische Modellierung und Inferenz ein. Dieses wird erreicht, indem man Mengen von Modellen betrachtet, entweder direkt oder indirekt über die Interpretation der Daten als Menge zugrunde liegender Datensituationen. Besonderer Wert wird dabei darauf gelegt, Annahmen zu vermeiden, die zwar technisch bequem sind, aber die zugrunde liegende Unsicherheit der Daten in ungerechtfertigter Weise reduzieren.

In dieser Schrift werden verschiedene Methoden der vorsichtigen Modellierung und Inferenz vorgeschlagen, die das Potential von präzisen und unscharfen Daten ausschöpfen können, angeregt von unterschiedlichen Anwendungsbereichen, die von Politikwissenschaften bis zur amtlichen Statistik reichen.

Zuerst wird das Modell der Nonparametrischen Prädiktiven Inferenz, welches per se unscharf ist, in der vorsichtigen Auswahl von Split-Variablen bei der Erstellung von Klassifikationsbäumen verwendet, die auf Methoden der Imprecise Probabilities fußen. Diese Bäume zeichnen sich dadurch aus, dass sie sowohl eine Struktur beschreiben, als auch eine annehmbar hohe Prädiktionsgüte aufweisen.

In Abhängigkeit von der Interpretation der Unschärfe, werden dann verschiedene Strategien für den Umgang mit unscharfen Daten im Rahmen von finiten Random Sets erörtert.
Einerseits werden die zu analysierenden Daten als mengenwertige Antwort auf eine Frage in einer Fragebogen aufgefasst. Hierbei wird jede mögliche (multiple) Antwort, die eine Teilmenge des Stichprobenraumes darstellt, als eigenständige Entität betrachtet. Somit werden die finiten Random Sets auf (gewöhnliche) Zufallsvariablen reduziert, die nun in einen transformierten Raum abbilden. Im Rahmen einer Analyse von Wahlabsichten hat der vorgeschlagene Ansatz gezeigt, dass die Unentschlossenen mit ihm genauer charakterisiert werden können, als es mit den gängigen Methoden möglich ist. Obwohl die vorgestellte Analyse, betrachtet als ein erster Schritt, auf mengenwertige Daten angewendet wird, die vor dem Hintergrund der wissenschaftlichen Forschungsfrage in geeigneter Weise selbst konstruiert worden sind, zeigt diese dennoch klar, dass die Möglichkeiten dieses generellen Ansatzes nicht ausgeschöpft sind, so dass er auch in komplexeren Situationen angewendet werden kann.
Andererseits werden unscharfe Daten durch eine mengenwertige Einfachimputation (imprecise imputation) erzeugt. Hier werden die finiten Random Sets als Ergebnis einer (unspezifizierten) Vergröberung interpretiert. Der Ansatz wird im Rahmen des Statistischen Matchings vorgeschlagen, das verwendet wird, um gemeinsame Informationen über ursprünglich nicht zusammen erhobene Merkmale zur erhalten. Dieses ist insbesondere relevant bei der Datenproduktion, beispielsweise in der amtlichen Statistik, weil es erlaubt, die verschiedenartigen Informationen aus unterschiedlichen bereits vorhandenen Datensätzen zu einen neuen Datensatz zu verschmelzen, ohne dass dafür tatsächlich Daten neu erhoben werden müssen.

Zudem müssen die Daten für den Datenaustausch in geeigneter Weise anonymisiert sein. Für die spezielle Klasse der Anonymisierungstechnik der Mikroaggregation wird ihre Eignung im Hinblick auf die Verwendbarkeit in generalisierten linearen Regressionsmodellen geprüft. Hierfür werden die mikroaggregierten Daten als eine Menge von möglichen, unbeobachtbaren zu Grunde liegenden Datensituationen aufgefasst. Es werden zwei Herangehensweisen präsentiert: Als Erstes wird eine maximax-ähnliche Optimisierungsstrategie verfolgt, dabei werden die zu Grunde liegenden unbeobachtbaren Daten als Nuisance Parameter in das Regressionsmodell aufgenommen, was eine enge, aber auch über-optimistische Schätzung der Regressionskoeffizienten liefert. Zweitens wird ein Ansatz im Sinne der partiellen Identifikation angewendet, der per se schon vorsichtiger ist (als der vorherige), indem er nur die Menge aller möglichen Regressionskoeffizienten schätzt, die erhalten werden können, wenn die Schätzung auf jeder zu Grunde liegenden Datensituation durchgeführt wird.

Unscharfe Daten haben gegenüber präzisen Daten den Vorteil, dass sie zusätzlich die Unsicherheit der einzelnen Beobachtungseinheit umfassen. Damit besitzen sie einen höheren Informationsgehalt. Allerdings gibt es zur Zeit nur wenige glaubwürdige statistische Modelle, die mit unscharfen Daten umgehen können. Von daher wird die Erhebung solcher Daten bei der Datenproduktion vernachlässigt, was dazu führt, dass entsprechende statistische Modelle ihr volles Potential nicht ausschöpfen können. Dies verhindert eine vollumfängliche Bewertung, wodurch wiederum die (Weiter-)Entwicklung jener Modelle gehemmt wird. Dies ist eine Variante des Henne-Ei-Problems.

Diese Schrift will durch Vorschlag konkreter Methoden hinsichtlich des Umgangs mit unscharfen Daten in relevanten Anwendungssituationen Lösungswege aus der beschriebenen Situation aufzeigen und damit die entsprechende Datenproduktion anregen.

Abstract

This thesis contains four contributions which advocate cautious statistical modelling and inference. They achieve it by taking sets of models into account, either directly or indirectly by looking at compatible data situations. Special care is taken to avoid assumptions which are technically convenient, but reduce the uncertainty involved in an unjustified manner.

This thesis provides methods for cautious statistical modelling and inference, which are able to exhaust the potential of precise and vague data, motivated by different fields of application, ranging from political science to official statistics.

At first, the inherently imprecise Nonparametric Predictive Inference model is involved in the cautious selection of splitting variables in the construction of imprecise classification trees, which are able to describe a structure and allow for a reasonably high predictive power.

Dependent on the interpretation of vagueness, different strategies for vague data are then discussed in terms of finite random closed sets:
On the one hand, the data to be analysed are regarded as set-valued answers of an item in a questionnaire, where each possible answer corresponding to a subset of the sample space is interpreted as a separate entity. By this the finite random set is reduced to an (ordinary) random variable on a transformed sample space. The context of application is the analysis of voting intentions, where it is shown that the presented approach is able to characterise the undecided in a more detailed way, which common approaches are not able to. Although the presented analysis, regarded as a first step, is carried out on set-valued data, which are suitably self-constructed with respect to the scientific research question, it still clearly demonstrates that the full potential of this quite general framework is not exhausted. It is capable of dealing with more complex applications.
On the other hand, the vague data are produced by set-valued single imputation (imprecise imputation) where the finite random sets are interpreted as being the result of some (unspecified) coarsening. The approach is presented within the context of statistical matching, which is used to gain joint knowledge on features that were not jointly collected in the initial data production. This is especially relevant in data production, e.g. in official statistics, as it allows to fuse the information of already accessible data sets into a new one, without the requirement of actual data collection in the field.

Finally, in order to share data, they need to be suitably anonymised. For the specific class of anonymisation techniques of microaggregation, its ability to infer on generalised linear regression models is evaluated. Therefore, the microaggregated data are regarded as a set of compatible, unobserved underlying data situations. Two strategies to follow are proposed. At first, a maximax-like optimisation strategy is pursued, in which the underlying unobserved data are incorporated into the regression model as nuisance parameters, providing a concise yet over-optimistic estimation of the regression coefficients. Secondly, an approach in terms of partial identification, which is inherently more cautious than the previous one, is applied to estimate the set of all regression coefficients that are obtained by performing the estimation on each compatible data situation.

Vague data are deemed favourable to precise data as they additionally encompass the uncertainty of the individual observation, and therefore they have a higher informational value. However, to the present day, there are few (credible) statistical models that are able to deal with vague or set-valued data. For this reason, the collection of such data is neglected in data production, disallowing such models to exhaust their full potential. This in turn prevents a throughout evaluation, negatively affecting the (further) development of such models. This situation is a variant of the chicken or egg dilemma.

The ambition of this thesis is to break this cycle by providing actual methods for dealing with vague data in relevant situations in practice, to stimulate the required data production.

Acknowledgements

Without the various types of aid of several people, the present thesis would have never surfaced.

First of all I am utmost grateful to my supervisor Thomas Augustin, who provided me with the opportunity to write this thesis. He stimulated my research by providing me with freedom and faith in my research activities and was always willing to discuss openly all topics. Furthermore, he also encouraged me to get involved into teaching allowing me to get a deeper understanding of several statistical concepts[1], as well as into several administrative tasks, which were a helpful break permitting new perspectives.

I also like to thank Michael Smithson and Matthias Schmid who kindly agreed to be part of my examination committee and to review this thesis. I am thankful to Michael for the chats during the breaks of the ISIPTA conferences in 2013 and 2017, which gave rise to new ideas, and to Matthias Schmid for his interest into the findings of my bachelor's thesis.

I am thankful to Christian Heumann and Helmut Küchenhoff for their willingness to steer the examination committee.

Furthermore, I am grateful to all present and past members of the working group, for all the discussions, chats and joint lunch breaks, stimulating further research ideas, not just for myself but also for scientific cooperation, and also for the smooth cooperation in all of the administrative projects we worked on together: Thomas Augustin, Johanna Brandt, Marco Cattaneo, Eva Endres, Cornelia Fütterer, Christoph Jansen, Aziz Omar, Julia Plass, Georg Schollmeyer, Patrick Schwaferts and Andrea Wiencierz.

I am grateful to Christina Schneider for the joint seminar (attempts), joint teaching of courses and many inspiring discussions on the foundations of probability, both from a philosophical and measure theoretical perspective.

Special thanks are also to Micha Schneider for discussions during regular lunch breaks and out-of-research activities, providing me with a refreshing outside perspective.

I like to thank Elke Höfner and Brigitte Maxa with their dogs Betty and Luna, respectively, for the daily encouraging 'Good morning!' and close cooperation in all the administrative tasks they were involved.

I like to express my gratitude to the participants of the Fifth Workshop on Principles and Methods of Statistical Inference with Interval Probability (WPMSIIP 2012) who, directly after finishing my master's thesis, stimulated my interest in scientific research with their inspiring talks and discussions.

I am also thankful to all the people of the department of statistics for creating a positive atmosphere which made working there a pleasure.

On a personal note, I like to express my gratitude to my parents, my sister and my late grandparents, for their continuous support, backing and trust in me.

Finally, I like to thank all people, who gave me inspiration to follow my path and made this thesis possible.

[1] One only knows that one has understood a topic, if one is able to explain it to another person who lacks knowledge about the topic in question.

Author's contributions

This page details the author's contributions to the individual papers, given in chronological order. Each contribution is publicly available, accessible at the URL provided in the listing of the contributions on page 3.

For Fink and Crossman (2013), Paul Fink developed the minimum entropy algorithm for credal sets generated by the Nonparametric Predictive Inference model. He also solely carried out the simulation and wrote a R-package specifically for this task. Richard Crossman aided in writing the introduction and in the justification of the minimal entropy algorithm. Furthermore, the proposed ad-hoc criterion for the splitting was developed by the authors in close cooperation. They both contributed to revising and proof-reading the paper.

The basic ideas in Plass et al. (2015b) were developed by Julia Plass, Thomas Augustin and Paul Fink in close cooperation. Norbert Schöning was consulted for his background knowledge on political theories. Paul Fink drafted Section 2.3 and 5.3, which contained an introduction to imprecise classification trees and the analysis of the data by means of the former. The respective data analyses of this part were also carried out by him, with the help of the aforementioned R-package. He also contributed to the minute revisions of the entire technical report. The mentioned contributions apply in the same way and amount to the underlying proceedings paper as well.

The contribution Fink and Augustin (2017) was mainly written and developed by Paul Fink. Thomas Augustin aided with structuring of the paper, suggested the partial identification approach and hinted further literature. The simulation was performed by Paul Fink. A fruitful discussion with Georg Schollmeyer on the involved optimisation tasks was of further assistance. Both authors contributed to the revisions of the paper.

The main idea in Endres et al. (2018), namely imprecise imputation is due to Eva Endres. Paul Fink developed the embedding of the approach into the theory of finite random sets, including the estimation of probability statements on the synthetic data, in Section 4. He also introduced the tuple notation for a concise representation of the synthetic data set. Additionally, he aided with the simulation by rewriting some previously heavily time- and memory-consuming code for a notable performance gain, and by compiling the analysis of the simulation. All authors contributed to revisions of the paper.

Contents

1 Introduction

The key aspect of statistical analyses is the description of phenomena in the real world by utilising a model in the language of mathematics. The model in itself typically provides an abstraction layer on the 'real world', i.e. the world we live in comprising all observable and unobservable features, as by modelling several aspects of the complex 'real world' are neglected or diminished, while others, particularly those of interest, are (hopefully) strengthened. Thereby, a general assumption of any model is that in reality there exists a true underlying, yet unobservable, mechanism that affects or generates the aspect of interest. An ideal model, i.e. one in which all aspects of the real world are captured, is able to tell the result of this mechanism, the so-called *data generating process* directly and in an error-free way. However, in practice, the observable output of the data generating process may be blurred. Therefore, one aims at finding a suitable approximation of the true but unobservable data generating process. A schematic representation of the statistical learning process is depicted in Figure 1.1, which in the following is further explained.

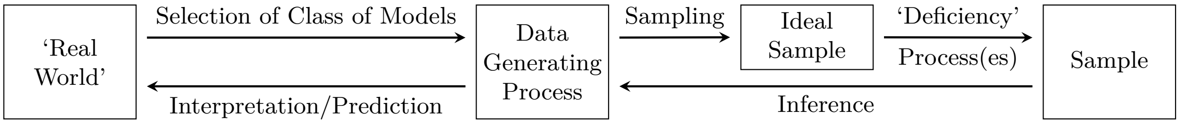

Figure 1.1: Schematic representation of the statistical learning process: The arrows indicate directed linkage between the aspects and are not interpretable in a causal way.

In order to be able to reasonably infer about properties of the data generating process, one needs to take the sampling and the deficiency process[1] into account. In Figure 1.1 they are depicted as two different steps, but this does not necessarily mean that they are unrelated (in a statistical sense). Therefore, it can be seen as a single blurring (process), on which assumptions need to be made. Those can be specified either jointly or separately for the sampling and deficiency process.

In any case, the strength of the assumptions on the blurring process maintained does matter for the statistical inference, and moreover for the interpretation of its results with respect to the 'real world':
On the one hand, as Manski stated in his *Law of Decreasing Credibility*, '[the] credibility of inferences decreases with the strength of the assumptions maintained' (Manski 2007, p.3), which essentially means that if there are too strong assumptions on the blurring process, the result of the analysis is devalued, as it is only generalisable to fewer situations, namely those with similar assumptions. On the other hand, assuming virtually nothing about the blurring, which in the terminology of imprecise probability is called being *vacuous*, one obtains results which have a higher credibility as they cover more situations, but may be practically irrelevant at all, as they might state only obvious facts. Hence one should carefully decide which assumptions to keep to derive credible and still meaningful results.

[1] The term 'deficiency process' itself may be misleading. This statement will become evident in Chapter 4.

The previous discussion concerned primarily the issue of selection of a suitable class of models (*Selection of Class of Models* in Figure 1.1) satisfying the needs to appropriately describe the data generating process. However, if settled for a specific (set of) model(s) – and hence the underlying data generating process(es) – another complication becomes evident: The model usually has to be estimated on the basis of finitely many observations, the *sample*. This layer of blurring is the so-called *sampling error*, i.e. the inference error made from looking only at a comparably small finite sample instead of the entire population, which in case of a very large size is typically modelled as if it was infinite. This error is usually not precisely measurable, yet if settled for a model, a bound for this error depending on the applied inference procedure may be derived, for instance in form of confidence regions or credible intervals, in case of frequentist or Bayesian analysis, respectively. This naturally requires that at least one model within the class of models is estimable, in case of multiplicity a single favourable model is usually selected by looking at the performance of the models in question.

As seen in Figure 1.1, only by means of a sample inference on the underlying data generating process(es) is achievable in an objective manner. It is commonly assumed that the data within the sample are precisely measured (with or without measurement error). However, there are situations in which the data collection might not be able to provide the precisely measured data the model expects. Yet, this does not necessarily mean a fault or shortcoming of the data collection in itself, even though the usually used term 'deficient data' might strengthen this impression. A more neutral term would be *vague*, *multi-valued* or *set-valued data*. Amongst this class of data are interval-valued data (for variables on metric scales), which might occur for instance as a result of heaping, but also coarse-valued data[2], which might arise when multiple precise categories are united to a set-valued category. This might be steered by a specific deficiency process, the so-called *coarsening process*, but could also be inherent to the data itself.
In the data collection by surveys deficient data (could) naturally arise, amongst others either by missingness of the respondents' answers, or by the questions asked themselves, e.g. providing an interval containing the precise value in question instead of the value itself directly, or by allowing for multiple categorical answers, e.g. as discussed herein in the context of voting intentions, or by post-processing the initially precisely collected data with anonymisation techniques to allow for data sharing in a legally appropriate manner. In general, in the data collection step the deficiency is tried to be controlled as much as possible by using appropriate techniques for the task in question. This means that, in the setting of Figure 1.1, the actual sample at hand has already been subject to this blurring process, possibly deviating from an *ideal sample*, which would have been obtainable if there had not been a 'deficiency process'.

Typically, traditional statistic methodology is unable to cope in a reliable way with vague data directly, but instead resort to enforcing strong assumptions to either be able to handle them or to produce precise results. In the light of cautious inference, such a strategy is questionable. The methodology of *partial identification* (e.g. Manski and Tamer 2002; Manski 2003; Manski 2007), tries to overcome this by looking at the identifiability of models and only excluding those from the class of models which are not compatible with the provided data, typically resulting in a set of models. The extreme cases obtainable by this methodology are the complete class of models and a single model, when it is identifiable. By providing an intermediate identification level, the methodology of partial identification tears down the usual focus on identifiable models.

A different approach provides the methodology of *imprecise probabilities*[3], which allows to relax the assumption of having only precise probability measures to directly allow sets of those

[2] The term *coarse data* is used synonymously.

[3] e.g. Augustin et al. (2014) for a broad introduction

in the inference. As such the effects of the blurring process are also taken into account cautiously. The fundamental difference between the approaches is that for partial identification the (possible) multiplicity of models arises because of multiple structural assessments inducing different probability models, while for imprecise probability the multiplicity of models, specified via a set of precise probability measures, is directly assumed.

In general, when dealing with multi-valued observations, as stressed by Couso et al. (2014), one needs to distinguish between two different natures of the data. On the one hand there is the *ontic view* in which the multi-valued observation is regarded as an entity of its own, which means that it represents total knowledge, implying that the multi-valued observation cannot be reduced to a subset by means of additional knowledge. On the other hand the *epistemic view* assumes that there is indeed a true precise value contained within the multi-valued observation, corresponding to the true value, which would have been precisely observable, if there had been sufficient [Author's note: in the colloquial sense] information available.

In this setting the present thesis consists of four contributions[4] which address different aspects and layers for uncertainty within the statistical learning process:

P. Fink and R. J. Crossman (2013). 'Entropy based classification trees'. In: *ISIPTA '13: Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*. Edited by F. Cozman, T. Denœux, S. Destercke and T. Seidfenfeld. Manno: SIPTA, pages 139–147. URL: `http://www.sipta.org/isipta13/index.php?id=paper&paper=014.html`.

J. Plass, P. Fink, N. Schöning and T. Augustin (2015a). *Statistical modelling in surveys without neglecting "The Undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision – extended version*. Technical report 179. Department of Statistics, LMU Munich. URL: `https://epub.ub.uni-muenchen.de/23816/`.
Extended version of:

> J. Plass, P. Fink, N. Schöning and T. Augustin (2015b). 'Statistical modelling in surveys without neglecting "The Undecided": Multinomial logistic regression models and imprecise classification trees under ontic data imprecision'. In: *ISIPTA '15: Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*. Edited by T. Augustin, S. Doria, E. Miranda and E. Quaeghebeur. Rome: Aracne, pages 257–266. URL: `http://www.sipta.org/isipta15/data/paper/19.pdf`.

P. Fink and T. Augustin (2017). '(Generalized) linear regression on microaggregated data – From nuisance parameter optimization to partial identification'. In: *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*. Edited by A. Antonucci, G. Corani, I. Couso and S. Destercke. Volume 62. Proceedings of Machine Learning Research. PMLR, pages 157–168. URL: `http://proceedings.mlr.press/v62/fink17a.html`.

E. Endres, P. Fink and T. Augustin (2018). *Imprecise imputation: A nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data*. Technical report 214. Department of Statistics, LMU Munich. URL: `https://epub.ub.uni-muenchen.de/42423/`.

In the following, the contributions will be briefly summarised. A more detailed description of each contribution, including the embedding into its respective literature, is to be found in the accordingly dedicated chapters.

[4] The breakdown of the author's share to each contribution is to be found on p. ii.

The first contribution (Fink and Crossman 2013) generalises[5], by addition of a user-adjustable weighting scheme, a proposed method of building so-called *imprecise classification trees* based on entropy ranges. As input this approach expects data of precise observations and involves them in each node of the tree for the estimation of a *credal set*, i.e. a (convex) set of probability measures, by means of a Nonparametric Predictive Inference (NPI) model. Furthermore, an exact algorithm for deriving distributions with minimal entropy from those estimated credal sets is developed. In context of the previous outline of statistical learning, the application of local imprecise models allows to take more potentially underlying data generating processes into account. This in turn makes the tree more flexible with respect to classification of new observations, which have no similar counterparts in the data used for growing the tree. In this sense a precise (inference) model is enriched by building the statistical inference on a set of models.

An application of those imprecise classification trees is presented in the second contribution (Plass et al. 2015a)[6]. In the setting of a pre-election study the distinction between different notions of imprecise data is made, and detailed that in this case the ontic view is suitable. It is demonstrated therein that those imprecise data with an ontic perspective are interpretable as precise data on a transformed sample space, and are then analysed by means of a multinomial regression model and the aforementioned imprecise classification tree. It is concluded that in principle, after embedding it into this setting by transforming the sample space, any appropriate method for precise data is suitable. Within the outline of statistical learning this contribution fits into dealing with vague data, which are seemingly deficient but in this setting turn out to be – in a literal meaning – not deficient at all.

As previously mentioned, anonymised data could be regarded as deficient data as well. The third contribution (Fink and Augustin 2017) develops two main approaches on how to derive meaningful parameter estimates for a (generalised) linear regression model, where the independent variables (covariates) are anonymised by microaggregation and the dependent variable (response) is left unchanged. Both approaches try to estimate the structural parameters of the regression model of the underlying true data. The first employs a likelihood-based approach, in which the differences from the microaggregated data to the true underlying ones are treated as additional nuisance parameters in the model. The parameters of interest then are estimated in a manner similar to an EM algorithm. The second approach relies on the methodology on partial identification by re-interpreting the microaggregated data as a set of possible underlying data situations, and estimates then the set of maximum likelihood estimators compatible with the set, the so-called *collection region*. In the paper an outer approximation of the collection region was actually estimated. This contribution directly shows an application in which the strategy of adding modelling assumptions – herein the nuisance parameters – could lead to more precise but also questionable results, circumvented elegantly by the partial identification approach.

The fourth contribution (Endres et al. 2018) presents a micro approach for statistical matching for categorical data. In order to obtain a meaningful fused data set in the end, the task of statistical matching is treated as a missing data problem, for which three different strategies on imputing in a set-valued manner are proposed:

i. A domain imputation, where for any missing value the entire domain of the variable in question, i.e. the set of possible values of the variable, is imputed,
ii. a variable wise imputation scheme based on so-called *donation classes*, i.e sets of observations which are similar with respect to the matching variables, where each missing

[5] Author's note: In order to simplify the attribution to the respective contribution this word is spelled differently in the chapter heading.

[6] It is based on Plass et al. (2015b) as extended version.

value within a donation class is replaced by the set of all actually observed values within the same donation class of the same variable, and finally

iii. a case-wise imputation scheme, also based on donation classes, but this time the missing values of a single observation within a donation class are jointly replaced by the set of already observed values in the sample space of the joint.

The second approach might be seen as the hull of all possible (random) hot deck imputations on a per-variable basis, while the last one could be regarded as a collection of possible random hot deck imputations obtained in a multivariate way by utilising the block-wise missing pattern typical for statistical matching. Furthermore, it is also demonstrated how imprecise imputation is embedded into the theory of random sets in order to obtain meaningful estimates on the imputed data. In this sense the contribution is involved in the cautious estimation approaches under deficient data. Moreover, treating statistical matching as a missing data problem is a cautious approach as well for dealing with the blurring occurring in the step from the data generating process to the actually observed data.

The remainder of this thesis is structured as follows: The next chapter provides a brief sketch of important concepts in order to familiarise the reader. Each of the then following chapters contains the specific theoretical background(s) and embedding of one contribution into the respective literature. And at the end of each chapter remarks and perspectives are provided. Chapter 3 deals with Fink and Crossman (2013), while in Chapter 4 Plass et al. (2015b) is discussed. The contribution on imprecise imputation for statistical matching by Endres et al. (2018) is detailed in Chapter 5, and the (generalised) linear regression on microaggregated data by Fink and Augustin (2017) is presented in Chapter 6. The thesis concludes with Chapter 7 giving final remarks.

2 General Theoretical Background

In this chapter an overview of the theoretical background of those concepts are presented, which are essential for the understanding of the contributions[1]. At first, in Section 2.1 an introduction into generalised linear regression is given, before in Section 2.2 the setting of linear regression is used to exemplify the concept of partial identification, which may lead to set-valued results. The chapter ends with Section 2.3 on random set theory, which provides a (multi-valued) generalisation of random variables, which is used herein for the modelling of 'deficient data'.

2.1 Generalised Linear Regression

In the embedding of the statistical learning process, generalised linear regression models provide a class of statistical models which aim at providing a structural assessment how the dependent (response) variable Y is linked to $p \geq 1$ independent variables (covariates) $X = (X_1, \ldots, X_p)$. The basis for the inference is a random sample of n observations, stemming from the whole population under investigation.

In the classical setting of ordinary linear regression the response is linked to a linear combination of covariates, where the deviation of the individual observation from this global sum is absorbed by an individual (random) error $\mathcal{E} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$. The structural parameters of interest are the intercept β_0 and weights $\beta_1, \ldots, \beta_p$. In formula, the realisation of the i-th observation takes the form:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \, , \quad i \in \{1, \ldots, n\} \, .$$

In order to be meaningful, assumptions on the error and its relation to the covariates are posed, namely that the errors are identical distributed with expectation $\mathbb{E}(\mathcal{E}_i) = 0$ and variance $\mathrm{Var}(\mathcal{E}_i) = \sigma^2$, uncorrelated amongst each other and to the covariates. A beneficial but not required assumption is that the errors are normally distributed. The parameters of interest, as well as the error variance, are estimable from the sample by the method of ordinary least squares, which corresponds to a maximum likelihood approach when assuming normality of the errors. The estimators are then analytically obtained by solving a linear equation system.

Maybe because of its compelling simplicity, the model is popular in applications, may it be appropriate or not for the specific situation and data at hand. Some deviations from the assumptions appear to be less critical than others, while the underlying continuous scale of the response is certainly limiting: A linear regression in the above setting with binary (or even multinomial) response seems most awkward, even though few argue in favour of it in specific situations (e.g. Hellevik 2009).

A substantial extension as introduced by Nelder and Wedderburn (1972) are generalised linear regression models, (e.g. Smithson and Merkle 2013; Fahrmeir et al. 2013), in which

[1] The following sections are written in a way to aid the reader in understanding the contributions, but are by no means exhaustive or even complete with respect to the theory described.

certain assumptions are modified, such that the ordinary linear regression is a special case. For generalised linear regression the form of the conditional distribution $Y|X$ is specified by a certain representative of the exponential family. Then the conditional expectation is modelled by the transformed linear predictor η_i:

$$\mathbb{E}(Y_i|X_i = x_i) = h(\eta_i), \quad \text{with} \quad \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

and h as the so called response function[2], which is required to be bijective and two times continuously differentiable. Typically, the parameters to be estimated are then the coefficients $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ and – depending on the choice of the distribution – a dispersion parameter ϕ, which provides a scaling for the variance, independent of the observations. The parameters are estimated in a maximum likelihood approach, which typically involves solving an equation system with no analytical solution. Therefore, the estimates are calculated by means of Fisher-Scoring, which corresponds to an iterative reweighted least squares algorithm.

This approach allows a great flexibility how η_i is linked to the conditional expectation, because as long as the technical requirements, the allowance for differentiability and bijectivity still hold, in principle any function may be used as response function. Common choices of response functions are due to mapping the space of the linear predictor into some subspace in which the response can take its values. Additionally, any specific distribution of the exponential family is parametrisable with respect to its natural parameter, which in turn leads to a specific link function, the so-called *canonical link function*. Choosing the canonical link function generally simplifies the involved calculations.

One should note that by the specification of the conditional distribution, the form of its variance is determined. Moreover, the covariance structure among the observations is fixed which in the setting presented till now, is zero, because of the assumed independence of the observations. However, such an assumption is unsuitable in situations with repetition measurements, e.g. longitudinal data with multiple observations of the same specifics of the same unit. To account for this, one could use models based on so-called *generalised estimation equations* (GEE), which still assume that the mean structure is correctly specified as in generalised linear regression models. But in contrast to the former, models based on GEE do not inherit the specification of the variance from the conditional distribution. Instead, the covariance structure may be freely specified in any reasonable and suitable way, which may be entirely data or application driven. In case of the longitudinal data, GEEs allow to specify a correlation structure between observations of the same observation. There exists also approaches modelling the conditional mean and the conditional variance separately by a regression model each, e.g. Rutemiller and Bowers (1968) for classical normal regression, in the context of generalised linear regression, e.g. Smyth (1989) and as special case for beta regression Smithson and Verkuilen (2006). In the latter, the expectation and variance parameters are independent due to the re-parametrisation of the beta distribution. In Hoff and Niu (2012) a connection to random effect models is created by utilising a special form of the regression model for the variance.

Stemming from basic ideas of generalised linear regression, several other models branched off. In nonparametric regression the linear predictor is now assumed to be a sum of unknown functions (to be estimated) of the covariates. The functions are estimated by a (penalised) spline-based approach, which transforms it back to the estimation of a generalised linear regression. Generalised additive regression (e.g. Hastie and Tibshirani (1999), Wood (2017, Chapter 4–6)) combines the aspects where some covariates are modelled in the predictor

[2] The inverse function h^{-1} is called *link function*.

by functions and others by linear combinations. These approaches allow for more flexible modelling of the linear predictor, but as in the contributions only standard generalised linear regression models are discussed, those are beyond the scope of this thesis.

2.2 Partial Identification

Partial identification maybe regarded as a cautious inference approach for a (parametric) statistical model. The historical development of partial identification is sketched, e.g. in Tamer (2010), which also covers the exemplary approach for dealing with missing outcomes, clearly advocating the use of partial identification as an inference approach.

In 'traditional' statistics, it is desired to select a single model from the assumed model class by selecting a unique parameter from the parameter space of the model class, and hence when such an unique parameter is achievable the model is said to be identifiable. If no unique solution exists, then the model is said to be *unidentifiable* and typically further assumptions are established in order to achieve identifiability. On a more formal note, a parameter is said to be *point-identified*[3] if the mapping from the elements of the parameter space to the model induced by those parameters is bijective[4]. This implies that if there are infinitely many observations available, one is able to obtain by inference, considering the inverse mapping, the underlying parameter of the model, as exactly one parameter will lead to the observed model. This concept does not state anything on the ability to estimate the parameter(s) of interest by a finite number of observations.

When there are latent variables in the model specification, one might not be able to distinguish the outcome on the observational level for different parameters of the latent variable. A typical and practically relevant situation is when the observation does not correspond to a single element in the sample space but to a set of elements, one of which is the true one. For instance, this may be due to pre-processing with anonymisation techniques (as in Chapter 6) or coarsening[5].
Considering a simple linear regression of n observations with interval-valued response $[\underline{Y}_i, \overline{Y}_i]$, where it is assumed that the unobservable Y_i lies within the interval, and one precise covariate X_i $(i = 1, \ldots, n)$, on the one hand a unique coefficient $\hat{\boldsymbol{\beta}}$ may not be obtainable, while on the other hand one could establish (questionable) assumptions on the representative ability of the intervals in order to enforce a unique estimate of the coefficients: Building a model only for the interval midpoint as a suitable representative, or building models on multiple characteristics of the interval, e.g. with the goal of prediction in Ferraro and Giordani (2012) and Giordani (2015).

However, such restricting assumption could be (severely) questionable in application. Partial identification offers an elegant escape of this dilemma: It breaks the binary concept of classical identifiability into a (semi-)continuous concept, by allowing for partial identification in form of so-called *identification region* (e.g. Tamer 2010) of the parameter of interest, i.e. a subspace of the parameter space of the models containing all model parameters that are compatible with the data at hand. Point-identification is included in this concept as the smallest of those subspaces, containing just one single parameter value, while the entire parameter space constitutes the other extreme case.
A natural identification region of the above example is the so-called *collection region*[6], which is the collection of all such regression coefficients, each of which is obtained by selecting a

[3] In 'traditional' statistics, it is simply called *identified*.

[4] In Lehmann and Casella (1998) this concept is indirectly defined by defining under what circumstances a parameter is said to be unidentified (cf. Definition 5.2 in Lehmann and Casella (1998), Chapter 1).

[5] The *missing* of an observation can be regarded as the most extreme form of coarsening.

[6] The name stems from Beresteanu and Molinari (2008) in the context of linear regression.

precise value $\tilde{Y}_i$ from $[\underline{Y}_i, \overline{Y}_i]$. This strategy breaks the dogma of hunting for only uniquely identified models.
Nonetheless, which set of parameters constitutes the identification region crucially depends on the parameter of interest and the understanding of the model. In the above context of linear regression with interval-valued response and precise covariate, Schollmeyer and Augustin (2015) contrasted three different identification regions, which depend on the understanding of the linear regression model itself. They prove that their marrow region is always within the collection region, which can be interpreted that the assumptions for obtaining the marrow region are stricter.

This shows a second advantage of the approach of partial identification, as it also allows to evaluate the influence of the assumptions established: The stronger the assumptions, the smaller the set of estimates and vice versa. Hence, it enables to distinguish between different strengths of partially identified parameters. The underlying idea is that assumptions, formulated purely for statistical and mathematical convenience, might produce results indicating a certain effect or trend, but they are just artefacts due to the – unfounded (in the context of application) – technical assumptions. In the more positive formulation this means that if there is a certain effect or trend in 'reality', then it will also be detectable by partially identified models.
In the setting of coarse data, assumptions on the coarsening mechanism may lead to either partial or point-identified parameters. For example, the strong assumption of coarsening at random (cf. Heitjan and Rubin 1991) allows for point identification, whereas weaker assumptions on the coarsening ratio (e.g. Plass et al. 2017, Section 4.3) do not longer guarantee point-identification.

A link between this and the following section was provided by Beresteanu et al. (2012), who explicitly formulated typical identification tasks in terms of random sets.

2.3 Random Set Theory

As presented in the previous section, partial identification works by collecting models (or rather their parameters), when there is not enough information allowing for point identification. Typically, the models involved are based on precise probabilities[7], which are turned into imprecise probabilities – as sets of precise probabilities – due to the collection. As such they rely on the concept of *random variables*, which are employed to make the involved features mathematically tractable. In order to familiarise the reader with the terminology used throughout this thesis, a verbal definition of a random variable is given: A random variable[8] is a measurable mapping from a measurable (source) space $(\Omega, \mathcal{A})$, in which the observations live, i.e. the sample space, to a measurable (target) space $(\Omega', \mathcal{A}')$ that houses the mathematical representation of the interested feature, i.e. the sample space.

In general, a random set can be seen as a generalisation of the concept of a random variable. The basic idea of random sets was already briefly sketched by Kolmogorov (1933), but its rigorous mathematical foundation was laid by Matheron (1975), in which he – at that time unaware of the existence – improved the independently developed work of Kendall (1974). For a random set the target space is not Ω', but the set of all subsets of it, the power set

[7] The use of precise probability models in partial identification approaches is by no means a requirement; models based on imprecise probability concepts could be used as well.

[8] Sometimes, e.g. in Nguyen (2006), the term *random variable* refers to a mapping into to measurable space consisting of the real numbers equipped with the Borel-σ-algebra. For a mapping into a generic measurable space the term *random element* is used instead.

$\mathcal{P}(\Omega')$. However, finding a suitable and natural σ-algebra on $\mathcal{P}(\Omega')$ for generic Ω' proves to be difficult, especially as it can become too large to be traceable.
If Ω' is a locally compact second countable Hausdorff space[9], i.e. every singleton in Ω' has a compact neighbourhood and its topology has a countable base and two distinct singletons of Ω' have disjoint neighbourhoods, an easier to trace variant, *random closed sets*, are definable. However, $\mathcal{P}(\Omega')$ itself is still too rich, so it is approximated by the class of all closed subset of Ω', $\mathcal{F}$. In order to specify a straightforward σ-algebra $\mathcal{B}(\mathcal{F})$, namely the Borel σ-algebra generated by the class of all closed subsets on Ω', the approximation is equipped with a suitable topology. It has been shown that the hit-miss-topology due to Matheron (1975) allows for this. Taking analogy from the definition of random variables, one can then define a *random closed set* as a measurable map Γ from $(\Omega, \mathcal{A})$ to $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$, where measurable means the compliance to the same criterion as for random variables (e.g. Nguyen 2006, Definiton 5.1, p.111). This can be equivalently formulated (e.g. Molchanov 2005) by requiring that for every compact set C' of Ω' holds

$$\Gamma^*(C') := \{\omega \in \Omega : \Gamma(\omega) \cap C' \neq \emptyset\} \in \mathcal{A}\ , \tag{2.1}$$

with Γ^* being the so-called *upper inverse* of Γ. If also the pre-image of every compact set C' of Ω' obtained by means of the so-called *lower inverse* Γ_* is within $\mathcal{A}$, i.e.

$$\Gamma_*(C') := \{\omega \in \Omega : \emptyset \neq \Gamma(\omega) \subseteq C'\} \in \mathcal{A}\ , \tag{2.2}$$

Γ is then said to be *strongly measurable* (with adjustment for the general case: cf. Miranda et al. 2005, Definition 2.1).
This definition relies on the hit-miss-topology of Matheron (1975), inducing implicitly a specific type of σ-algebra to ensure measurability for Γ. In analogy of the definition of a random variable, it is also directly expressible in terms of a requirement of a measurable mapping, but now with a pre-specified suitable σ-algebra.
Depending on the assumption posed on the target space, different specialised measurability criteria apply, e.g. if Ω' is a Polish space, equipped with a suitable σ-algebra, the multi-valued mapping Γ would be called a random closed set, if and only if it is Effros-measurable (cf. Molchanov 2005, p.26).

Fortunately, the complexity of the required mathematical formalism as previously described reduces somewhat if one assumes Ω' to be finite, which in turn implies that the power set $\mathcal{P}(\Omega')$ is also finite. A reasonable σ-algebra is then the power set of the power set of Ω', i.e. $\mathcal{P}(\mathcal{P}(\Omega'))$, giving rise to the definition of a *finite random set* [10].

As both contributions involving random sets (see Chapter 4 and Chapter 5) rely only on finite random sets, in the following the description is limited to this particular case. The image of the probability measure induced by the finite random set is completely characterisable by the probability mass function f on each element of $\mathcal{P}(\Omega')$. While in general it is not required that the finite random set assigns probability mass 0 to the empty set, there are practical considerations on why to do so, e.g. from a technical angle it allows to define a distribution function of the random set by utilising the structure induced by set inclusion, or from more practical consideration, when the finite random set arises from coarsening, i.e. where it is assumed that the outcome of the original but unobservable random variable is blurred, in the sense that it is an element of the random set outcome. Therefore, the set containing the empty set is usually removed from the target space in advance, leading to $\mathcal{P}(\Omega') \setminus \{\emptyset\}$ as target space of the finite random set.

[9] An example is the Euclidean space $\mathbb{R}^n$ with the topology, induced by the open balls of rational radii and rational centres, and the Euclidean metric. For the definitions, cf. Waldmann (2014), Chapters 2 and 5.

[10] cf. Definition 3.1 Nguyen (2006), p.35

For the correct interpretation of (finite) random sets, as stressed in Couso et al. (2014), Couso and Dubois (2014) and also already mentioned in Zadeh (1978), the actual nature of the data itself, as outcomes of the random set, needs to be considered. If the vague observation consists of mutually exclusive values, which are options for an unknown, only ill-perceived precise quantity, one takes the so-called *epistemic view* on the observation. This assumes that the coarseness is induced by a lack of information, which would have vanished, i.e. resulting in a precise observation, if sufficient information had been collected. If so the according random set would be called a *disjunctive random set.* Quite contrary is the so-called *ontic view*: In this perspective, the vague observation represents an indivisible piece of information, which is a conjunction of precise quantities which all are correct for the observation in question. Hence in this view the vague data may be regarded as a precise characterisation of the information at hand. A random set associated with the ontic view is termed *conjunctive random set.* In essence, the distinction can be reduced to the following: 'Typically, while the output of an ontic model is precise (but possibly wrong), an epistemic model delivers an imprecise output (hopefully consistent with the reality it accounts for).' (Couso and Dubois 2014, p.1503) Depending on the view taken, the same probability assessment may lead to different values. In this way, this distinction provides a pitfall for those unaware of the necessity of it.

For the ontic view, an ordering with respect to set inclusion does not make sense, as it would contradict the assumption that the elements are own entities. However, if taking the epistemic view, the elements of $\mathcal{P}(\Omega')$ are comparable with respect to set-inclusion, therefore it is allowed by this interpretation to define a distribution function for some $A' \subseteq \Omega'$ as the probability that the random set Γ is included within A', $P(\Gamma \subseteq A')$. This distribution function then has the property that it is monotone of infinite order (e.g. Choquet 1954, Chapter III), and due to its construction by set inclusion assigns zero to the empty set and one to Ω'. It is linked to the probability mass function f by the following:

$$P(\Gamma \subseteq A') = F(A') = \sum_{B \subseteq A'} f(B) \quad \text{for } A' \subseteq \Omega' \ .$$

Alternatively it can be expressed by means of the image measure of the lower inverse (2.2):

$$P_*(A') = P(\Gamma_*(A')) \quad \text{for } A' \subseteq \Omega' \ .$$

By this construction it becomes evident that either the distribution function or the probability mass function is sufficient to describe the probability law induced by the random set. A further characterisation can be achieved by means of the upper inverse (2.1) by

$$P^*(A') = P(\Gamma^*(A')) \quad \text{for } A' \subseteq \Omega' \ ,$$

which is a capacity function and dual to the distribution function. It has the same properties as the distribution function, but the monotonicity property is replaced by the property of alternating of infinite order.

There exists a very close connection to the Dempster-Shafer theory of belief functions (e.g. Dempster 1967; Shafer 1976): The probability mass function of the finite random set is the basic probability assignment[11] m. Although the formalism in Dempster-Shafer theory and random set theory in the unconditional setting are mathematically equivalent, as shown in Nguyen (1978), they come with different pre-dominant interpretations and contexts in which they are applied. As pointed out by Moral (2014), the distinction between the different views can also be made in terms of belief functions, where the original interpretation of Dempster (1967) is interpretable as an epistemic position, whereas the interpretation of belief functions due to Shafer (1976) corresponds to the ontic view, as he stresses that '[the basic probability assignment] $m(A)$ measures the total portion of belief [. . .] that is confined to A yet none of which is confined to any proper subsets of A.' (Shafer 1976, p.40)

[11] Sometimes it is also called *mass function.*

The distinction between the interpretations is also vital for conditioning of random sets. Furthermore, different notions of conditioning are present, the choice made is dependent on the context of application. In the setting of a single probability measure those concepts of conditioning lead to the same result, but they differ for sets of probability measures. This is especially relevant for the epistemic view, as stated in Couso and Dubois (2014), where such a set is obtained by the set of the image probability measures of the almost surely-selectors, i.e. the underlying precise random variables which are almost surely contained within the random set. A popular rule in the context of information fusion is Dempster's conditioning rule, which is a special case of Demptser's rule of Combination (Dempster 1967), which essentially re-evaluates the basic probability assignment in the light of the condition. Another conditioning scheme, more familiar to statisticians, involves the application of the traditional conditioning rule for any precise probability measure which lies within the set of precise probability measures induced by the random set on the (transformed) target space (e.g. Couso et al. 2014, Equations (11) and (12) in Section 4.1).

Further implications and specialities of random sets, which are unique for either the ontic or epistemic view are discussed in Chapter 4 and Chapter 5, respectively.

As random sets are a generalisation of random variables, they are applicable in any context, where there is some set-valued outcome of a random quantity. Beyond the application in context of voting behaviour as in Chapter 4, or in the analysis of set-valued imputed data, finite random sets are a popular tool in the (honest) analysis of sensor data, as sensors are typically not able to pin down their measurement exactly to a single value in an error-free way. Therefore, some basic concepts were developed in Mahler (1994), and later summarised in Mahler (2000), which are based on the work of Goodman and Nguyen (1985). The general model is the so-called *FISST* (finite set statistics), in which additionally to the occurrence of some values within the outcome of the random set, the cardinality is also explicitly modelled by a probability distribution (cf. Mahler 2013).
A selection of further applications of closed random set is given in Stoyan (1998), in which a slight focus on the use of random sets in the context of stochastic geometry models is apparent.

3 Entropy based classification trees

The first contribution is about classification trees, where the focus is mainly on prediction. As such it is one extreme of a statistical learning process, where the data generating process itself is of minor interest, and only the predictive ability matters. Focussing on prediction usually simplifies statistical learning, as one is no longer restricted to data generating processes which are easily tractable, both with regard to estimation and interpretation. It corresponds to the view, which Breiman (2001b) favoured in his widely disputed article about the two cultures of statistical modelling.

3.1 Specific Theoretical Background

The contribution Fink and Crossman (2013) develops on the one hand an algorithm for finding a probability distribution with minimal entropy within a credal set[1], generated by the categorical Nonparametric Predictive Inference model (cf. Augustin and Coolen 2004; Coolen and Augustin 2005; Coolen and Augustin 2009), and on the other hand it also introduces a splitting criterion for imprecise classification trees, based on entropy ranges utilising the aforementioned algorithm.

The ultimate goal of classification is to predict the outcome of a nominal class(ification) variable on the basis of some independent observed attribute variables[2]. The framework of classification trees provides a solution of this task, by recursively dividing the sample space into disjoint subspaces in a way that the subspaces are becoming more homogeneous with each splitting, with respect to the class variable. If considering only a single classification tree it also provides a structural interpretation, i.e. the obtained splitting structure, which is lacking when applying ensembles of trees, like bagging (e.g. Breiman 1996) or random forests (e.g. Breiman 2001a), which are more focussed on obtaining a better accuracy of the prediction. In this sense, ensembles are often used in a black-box perspective for which the underlying data generating process does not matter (at all): One is satisfied as long as the prediction is accurate.

There are multiple algorithms to obtain reasonable classification trees, which all involve estimation of precise probability distributions when deciding on splits, but differ on how they decide on splitting and in the way the splits are performed: ID3 and its successors C4.5 and C5.0 (cf. Quinlan 1986; Quinlan 1993; RuleQuest Research 2017), Breiman's CART and its successors (cf. Breiman et al. 1984; Breiman 1996; Breiman 2001a), and also, exemplary for more statistically driven algorithms, CHAID (Kass 1980) and the framework of unbiased recursive partitioning (Hothorn et al. 2006). The setting of (imprecise) classification trees is presented in Section 1 of the contribution.

Abellán and Moral (2003) developed imprecise classification trees which are based on the ID3 algorithm (cf. Quinlan 1986). In the ID3 setting the splitting criterion can be formulated in terms of the information gain of CART, which is based on the estimated precise probability distribution of the class label within the node of consideration. Abellán and Moral (2003)

[1] a (convex) set of probability measures

[2] In the literature they are also called *feature variables*.

modified the splitting strategy firstly by replacing the estimated precise probability distribution with a credal set, estimated by an imprecise Dirichlet model (IDM) (Walley 1996), and secondly by adapting the splitting criterion to measure the uncertainty within the credal set. Initially their criterion consisted of two parts, one for the uncertainty of the credal set by calculating the maximum entropy, i.e. selecting the precise probability distribution within the credal set that gives rises to the maximal entropy, while the other was used for the non-specificity within the credal set. However, they settled to use only the first, as the two measures are highly related (cf. Abellán and Moral 2005). In recent developments the splitting algorithm was updated to C4.5 and its behaviour in context of noisy data is analysed: In Mantas and Abellán (2014) the concept is introduced, while in Mantas et al. (2016) the impact of the hyper-parameter s of the IDM is evaluated.

The approach of Nonparametric Predictive Inference (NPI), initially developed for numerical data (cf. Coolen 1998), relies heavily on the exchangeability of the (future) observations as random quantities and is based on Hill's $A_{(n)}$ assumption (Hill 1968). Exemplary for a real-valued single variable, this means that the already observed quantities induce a ranking and hence a partition of the data space, i.e. intervals. Under the assumption of exchangeability the probability that a further observation takes any rank is the same, i.e. expressed in terms of the data space means that the probability that it belongs to a certain element of the partition is the same for all the elements of the partition. Roughly speaking the partitions are all equally likely. As only the ranking matters, this concept is also directly suitable for only ordered data that are not measured on a numerical scale. Typically this assumption is not strong enough to derive precise probability statements for an event of interest. Augustin and Coolen (2004) embedded the inferences obtained by assuming $A_{(n)}$ into the concepts of imprecise probability, in the spirit of a subjective/behavioural interpretation as in the approach of Walley (1991), and also into *interval-probability*, a generalisation of classical Kolmogorovian probability due to Weichselberger (cf. Weichselberger 2000; Weichselberger 2001). This concept has then been generalised by Coolen and Augustin (2005) to be suitable for multinomial variables by transforming the $A_{(n)}$ assumption to so-called *circular-*$A_{(n)}$, denoted by $\textcircled{A}_{(n)}$, representing the multinomial data on a probability wheel.

The family of NPI models has been applied in various contexts: In the analysis of economic time series (e.g. Baker et al. 2017), in the context of analysis of receiver operating characteristic (ROC) curves or surfaces (e.g. Coolen-Maturi et al. 2014; Coolen-Maturi 2017), and especially in the context of risk and reliability analyses (e.g. Yin et al. 2017; Coolen and Coolen-Maturi 2015). Because of its predictive nature, NPI has also been used to estimate credal sets in the process of building classification trees as local models within the nodes as it was first introduced in Baker (2010) and later further investigated in Abellán et al. (2014), relying on algorithms to calculate the maximal entropy of a credal set as developed in Abellán et al. (2011): One is an approximate algorithm, which does not necessarily comply with the underlying assumption/representation of a probability wheel, while the other is an exact one. Subsequently, the NPI was involved in building imprecise classification trees with entropy ranges (cf. Crossman et al. 2011): In their contribution the authors considered the ordinal variant of the NPI model and developed an algorithm to obtain minimal entropy of the credal set generated by this model. Furthermore, they adapted the split criterion within each node, basing it on entropy intervals. They compared the different entropy intervals by interval dominance[3] and built independent sub-trees for any of the variables belonging to the non-dominated ones. This led to an explosion in complexity as by this strategy the algorithm did not return a single tree, but rather a bunch of trees, having a similar/same structure in the nodes close to the root.

[3] cf. Definition 8.5, Huntley et al. (2014), p.194

Fink and Crossman (2013) continues in this spirit, but there the algorithm calculating the minimal entropy for multinomial NPI generated credal sets is developed and its appropriateness proven (cf. Fink and Crossman 2013, Section 2). It is shown that the obtained probability distribution with minimal entropy is indeed conform with the theory as it is representable on the probability wheel. Furthermore, in this contribution in Section 3 the ideas of using the full information contained within the entropy interval as well as having a single split variable are combined. Criteria based on both the upper and lower bound of the entropy interval are introduced and then added in a convex combination, whose weights may be tuned by parameters. This allows the user to specify the degree of their optimism: By putting full weight on the upper bound of the results in the original pessimistic splitting strategy of Abellán and Moral (2003), while increasing the weight on the lower bound one becomes more optimistic with respect to the information gain. In order to keep the desirable behaviour of entropy reduction going from root notes to leaves, this convex combination of criteria is post-processed, which in the same step allows to also exclude splitting variables whose entropy interval is interval dominated. Furthermore, for the post-processed criterion a threshold level is introduced which a variable needs to attain in order to qualify as a splitting candidate. When following the more optimistic strategy, the trees will get larger, as the observations within each leaf are typically more homogeneous, which may lead to a higher overfitting. In this sense, being too optimistic could even be more harmful with respect to the generalisation error than directly using a precise classification tree grown to full size. The proposed split criterion was then evaluated on real world data sets with regard to the sensitivity when varying the introduced parameters, i.e. weighting parameters and threshold level (cf. Fink and Crossman 2013, Section 5). The results of the evaluation showed that, under certain data set specific combinations of the tuning parameters, the accuracy increased, yet there was no pattern which would allow to guess them in advance easily.

3.2 Remarks and Perspectives

In comparison to the unpruned classification trees obtained by the classical algorithms, e.g. CART or ID3, the unpruned imprecise classification trees are more robust to the threat of overfitting. They achieve the robustness by basing their local inferences on multiple models and select the worst-case scenario. As stated in Abellán and Moral (2005), the performance is comparable with fine-tuned pruned classification trees, yet with a lesser complexity. As one is always free to not classify observations at all, those imprecise classification trees may also be used as a tool to detect hard to classify observations, i.e. difficult cases. However, this pessimism of the method, selecting always the worst-case scenario within the local models, might also be hold against it: If one trusts in the model, there is little need to be pessimistic about it. In this sense the contribution allows the user more freedom, as it is his or her choice if the focus should be set to pessimism or optimism, depending on how trustworthy and representable the data at hand are seen. However, it should be noted that by allowing a split criterion also relying on minimum entropy, the obtained probability distribution is not unique. In the contribution this issue is circumvented by selecting randomly one split variable if multiple were tied with respect to its value. In hindsight, this strategy is still improvable, as it goes to some extent against the idea of using imprecise methodology which implicitly allows to give set-valued results in case no reasonable decision between precise options can be made.

In a more abstract view on the developed criterion, one can see it as an approach to deal with the comparison of entropy intervals: It forces a precise decision – if needed also by means of randomisation – to avoid the complexity explosion inherent in the approach in Crossman et al. (2011). Additionally, even if the complexity issue can be dealt with, the question remains on how to combine the results, as the obtained trees according to that approach are highly correlated. An approach accounting for the combination was firstly developed in Abellán and Masegosa (2010) and then enhanced in Abellán and Masegosa (2012). They concluded that for data sets with high classification noise their proposed ensembles perform better, however, as the finding in Fink (2012) are hinting at, in cases with little classification noise there is not much to gain from using an ensemble of imprecise classification tress instead of a single imprecise classification tree. In any case, the actual implementation into software of the proposed method in Crossman et al. (2011), and subsequently its application in a simulation study, would be the next steps to evaluate its usefulness, beyond the theoretical justification.

A general issue that still remains, although some progress has been made, is how to compare the results of an imprecise classifier to a precise one. Clearly, the accuracy is not a helpful tool as with the vacuous classification for every observation one can achieve total accuracy, while the classifier itself is less than entirely useless. Zaffalon et al. (2012) developed an utility based approach to overcome the shortcoming of the discounted-accuracy, which does not distinguish well enough between a vacuous (imprecise) classifier and a random precise classifier. They proposed reasonable utility functions, however those are just few out of a huge class of other possible ones. Choosing a suitable utility function as soon as at least one imprecise classifier enters the comparison, may be objected, as it can be seen to defeat the purpose of achieving an objective and fair comparison. Therefore, it is a difficult task to convince people to apply imprecise classification to achieve accurate prediction. However, if the purpose is to identify hard to classify observations, it can be argued that imprecise classifiers are highly suited, as one can refrain from prediction if there is no clear answer[4].

[4] This is indirectly mentioned by Abellán and Moral (2005) when they argue for a fair comparison with precise classifiers.

4 Statistical modelling in surveys without neglecting 'The Undecided'

This contribution links with the previous one in the way that it is an application example for the presented imprecise classification trees. But in it also a rather general methodological aspect is discussed, namely that the presence of vague data does not necessarily imply a coarseness inducing deficiency process, but those data could also arise if the data generating process directly allows to generate such data. Hence, the labelling of such vague data as *deficient* is highly misleading. In this situation, the deficiency transformation would be the identity transformation, which means that there is no deficiency present at all, a fact that is typically ignored in application as described later on. The data are formalised by finite random sets in the ontic view (cf. Section 2.3).

4.1 Specific Theoretical Background

The motivational application data of Plass et al. (2015a) is the German Longitudinal Election Study (GLES) of the year 2013 (Rattinger et al. 2014). It is an umbrella for some very detailed electoral surveys, including, amongst others, specific pre- and post-election surveys, focussing on the federal elections in Germany in 2009, 2013 and 2017. The GLES 2013 pre-election cross sectional study, on which the contribution focused, was a 3-step random sample design, collecting the voting intentions, which needed to be provided as the name of a single party, for the federal election along with the certainty of the given intention, as well as an assessment of their proximity to each political party in Germany. Typically the certainty was used to check if the respondent would be included within the analyses and in this way excluding all with little to no certainty. By this procedure some rather important information is certainly lost, as respondents are dropped who were not certain for a specific party, but from their given proximity to the political parties it was clearly visible that they were favouring a likely coalition, and thus only indecisive which of the coalition partners to vote for[1]. Prior to an election those respondents are of particular interest for parties in order to attract them to go to vote (for them).

In the political literature there are multiple models on forecasting the results of the German federal elections (e.g. for the election in 2013: Ganser and Riordan 2015; Kayser and Leininger 2016; Graefe 2015; Selb and Munzert 2016; Küntzler 2017). Their predictions are based on precisely given voting intentions or expectations, either on a party or coalition level and either as individual observations or aggregate numbers, and other socio-demographic factors. However, the preprocessing of the raw data, which is required to obtain those precise numbers, is seldom published (cf. Schnell and Noack 2014, p.9), and hence it is not clear neither how many observations are discarded because of not being able to provide a precise intention, nor how trustworthy those precise answers are. Furthermore, the focus on coalitions is natural, as in Germany it is highly unlikely that a single party will be in power. The forecasts for coalitions can be constructed aggregating the precise individual party values, or directly assessing the values for specific coalitions[2]. In Debus (2013) the voting intentions are

[1] In such a way party proximity was operationalised in Kayser and Leininger (2016), for instance.

[2] Typically, only few coalitions are assessed which appear to be reasonable. All other coalitions are excluded by design.

predicted by a multinomial logistic regression model in which the independent variables were the preferences for some coalitions, while in Thurner (2000) a multinomial logistic regression model is argued for and applied in which the independent variables are distances to bipolar political statements. In Elff and Roßteutscher (2011) the links between class, religion and the voting behaviour are studied for the federal elections in Germany from 1994 to 2009, relying on the precisely given names of the parties the respondents voted for.

All those studies have in common that they force respondents to answer precisely. They do – by design of the questionnaire – not allow for indecisiveness, which in the context of pre-election polls/surveys is to be expected.

Therefore, in the contribution the vague data need to be (artificially) constructed by taking the certainty and the party proximity into account: If respondents were certain, their voting intention was solely used, but for lesser degrees of certainty, additionally their proximity was considered, by adding to their given voting intention the parties to which they reported a close proximity. It is a more honest operationalisation as the individual uncertainty is reflected in the multiplicity of obtained intentions per person.

As already stated in Section 2.3, the framework of (finite) random (closed) sets provides a reasonable toolbox, when analysing vague data from any background. As also mentioned there, the nature of the vague data should be decided prior to analysing them.

In the context of voting preferences as constructed for the contribution, if the question is 'How many percent of the votes is party X to attain at the next election?', one will take an epistemic view point (and hope that if multiple party preferences were given, the final vote is contained within). If the pre-election setting is taken seriously, the voting preferences comprise the actual and complete information about the voting behaviour of the respondents and hence an ontic view should be taken. With this view, the previous question is actually not suitable to ask, as it neglects the conjunctive nature of the observations. A legitimate question in the ontic setting would be, for instance, if there are some external effects present, which have an influence on the voting preferences, while this time treating each multi-valued voting preference as an entity of its own.

In the contribution the ontic view was taken, therefore in the following the formalisation in the ontic setting is presented[3], using the notation as in Section 2.3.

Within this finite setting the finite random set can be interpreted as an ordinary random variable, where each (possibly) multi-valued outcome lives now as a point in the space $\mathcal{P}(\Omega')$, which allows to switch to the usual measurability condition of (ordinary) random variables[4]. This view that the outcomes of the random set are considered as points in the power set of the original target space rather than as sets of elements from the original target space, relates strongly to the ontic view: By treating the outcome as a point, one directly states that one sees this outcome as something indivisible and therefore as an entity of its own. In this sense, the multi-valued mapping, i.e. random set, is transferable to an ordinary random variable by transforming the sample space of the target space prior to any statistical analysis. If done so the source space is linked to the target space via an ordinary random variable. The contribution presents a relevant use-case of the ontic view, and also exemplary demonstrates its usefulness as the situation with vague data reduces to a situation with precise data, which is more appealing to classical statisticians.

[3] The epistemic view playing a crucial role in the estimation of the quantities involved in Chapter 5 will be discussed there.

[4] In the contribution the term *state space* is used for Ω'. For reason of internal consistency the term *sample space* is used herein instead.

In the contribution in Section 5 the vague data were analysed by means of a multinomial logistic regression and imprecise classification trees[5]. It was shown that for the regression model there were differences in the effect of the considered covariates among the undecided respondents itself and also the decided ones, as is clearly visible in Table 6 in Appendix A.1 of the contribution. Despite the construction this strategy reveals information, which in turn are directly usable by the parties involved to target specific groups, which in the pre-election setting is essential in order to attract votes. The application of imprecise classification trees achieved a reasonable predictive power, however, the main advantage was the ability to detect persons, who were hard to classify with respect to the variables used. This information is also usable in the field, as it can be seen as a possible characterisation of persons, who are not attracted by the main parties or coalitions.

4.2 Remarks and Perspectives

As demonstrated, e.g. in Couso et al. (2014), the finite random set in the ontic interpretation also behaves as an ordinary random variable with respect to conditioning, as well as with respect to (in)dependence statements. In the contribution only an exemplary analysis by application of an (adapted) multinomial logistic regression model and imprecise classification trees were presented, however, one could in principle use any other statistical model for precise data, which would have been suitable in this context. This is an extremely strong result of the contribution, as it is in general not necessary to use complex (imprecise) statistical models which account for the uncertainty, seemingly induced by having vague instead of precise observations, because there is no uncertainty at all with respect to the data imprecision.

By the transformation of the target space from Ω' to $\mathcal{P}(\Omega')$ the sample space for the analysis is exponentially increased. Without any further assumptions this can be an issue for statistical models which estimate parameter(s) for each component of the sample space, e.g. the multinomial logistic regression model with at least one category-specific component, i.e. category-specific intercept or slope. If $|\Omega'|$ is small then the possible combinations of categories to be considered are still reasonably small with respect to usual sample sizes in surveys. In the current German Bundestag, elected in 2017, six parliamentary groups are present, which could (hypothetically) lead to possibly $2^6 - 1$ different constellations of groups getting into executive power. However, if one had taken the sample space consisting of all parties which could be voted for with the second vote in the election in 2017[6], $|\Omega'| = 34$, then the vast number of 17 179 869 183 combinations would have been to be considered, surpassing the number of people in Germany eligible to vote by a multiple. The strategy for dealing with this overly large number could be twofold: One way would be to reduce the transformed sample space directly by limiting it to 'reasonable' combinations, while the other would be to adjust the probability distribution on the transformed sample space, namely by assigning probability zero to the highly unlikely categories. From a practitioners point of view these approaches may seem equivalent, which by and large may be true, however, they come with subtle differences: When excluding categories, one could be surprised if in a follow-up study such an excluded category is reported, yet a benefit of this approach is that statistical quantities taking the number of categories into account, e.g. correction terms, are directly using the reasonable subset. Contrarily, the other strategy would require to limit the categories on those with probability larger than zero in order to be meaningful, while surprise observations of categories with probability zero are easily to deal with.
In either case, the implications of such a change on the equivalence of the random set perspective and the perspective of an ordinary random variable on a transformed sample space,

[5] Their respective theoretical backgrounds are sketched in Section 2 of the contribution.

[6] cf. Der Bundeswahlleiter 2017.

still need to be evaluated from a theoretical point of view.

Another point, which again was only hinted at in the contribution, is the extension to incorporate ordinal information. It seems like a immediate generalisation of the nominal case, however, one should notice that it also bears some pitfalls.
For the finite case, one could question if the power set of the sample space would still be a reasonable space. Consider the following example as a compelling argument against it: Let the original sample space be $\Omega'_{10} = \{1, 2, 3, \ldots, 10\}$. This set is totally ordered by the usual ordering of the natural numbers. Furthermore, it holds that $\{2, 3, 5\} \in \mathcal{P}(\Omega'_{10})$, but taking this as an observation seems obscure: It means that there is indecision between categories 2 and 3, so far all is fine, but also between category 5. Taking the implied ordering seriously, the set should include also 4, as only by then the entire indecision between the ordered categories 2 and 5 is captured. With this argumentation, the power set should be reduced to only contain sets, whose elements are all directly neighboured in the original space. After this reduction step the proposed methodology of contribution would be suitable again.
Independent from the above discussion, one should also notice that by transforming the sample space the strict ordering is lost, but only a partial order is preserved. This has to be accounted for in the statistical analyses. One way to deal with it is a partial identification approach, which can be summarised as follows: One takes the set of strict orderings, which are compatible with the original strict ordering of the untransformed sample space, and then for each ordering within this set statistical inference for ordinal data is performed in the classical way. If the model applied is parametric, the obtained estimates could be united in a collection region. Alternatively, one could evaluate the likelihood function and report only those estimates for which the likelihood is maximal (or above a certain threshold).

5 Imprecise imputation

In this contribution the statistical matching is treated as a missing data task and therefore three related imprecise imputation strategies are proposed to generate a partially synthetic micro data set. It is finally demonstrated how this approach can be embedded into the methodology of finite random sets in order to obtain probability assessments. This is a link to the previously described contribution. However, in this one the interpretation of the outcome of the finite random sets (cf. Section 2.3) is fundamentally different, as now an epistemic view is taken.

5.1 Specific Theoretical Background

As reflected in the introduction of the contribution, statistical matching[1] is the task of combining two or more data sets[2] of different sources, which share only a subset of their variables. Whereas in the methodology of record linkage[3] it is assumed that many of the observations in the data sets to link are identical, in statistical matching it is assumed that they are distinct[4]. Nonetheless, they are still assumed to come from the same population, such that the joint probability distribution is the same irrespective of the data set. The variables within the intersection of the variables of the individual data sets are termed *common variables*, and those that are not in the intersections are called *specific variables*. A schematic representation of the task of statistical matching for two data sets A and B is depicted in Figure 5.1, where it can be seen that there does not exist a single observation, which houses the joint information of X, Y and Z. In this scenario, Y and Z are the specific and X the common variables.

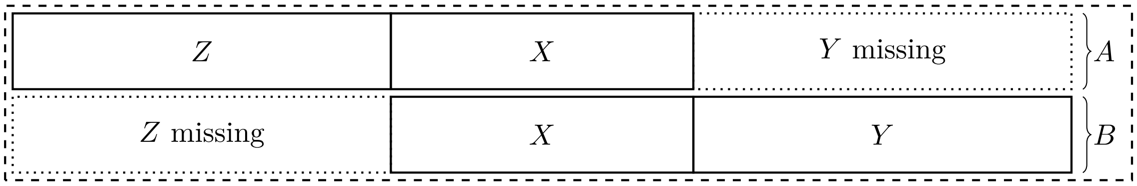

Figure 5.1: Schematic representation of statistical matching of two data sets A and B in the mirco approach (adaptation of e.g. Rässler 2002, p.3; D'Orazio et al. 2006b, p.5)

The application contexts[5], in which statistical matching has been already applied, are limited. However, there is interest especially in official statistics, e.g. the application of it at Statistics Canada, as described in Singh et al. (1993), and also at Eurostat, e.g. Eurostat (2013), in which two experimental studies were conducted. The interest of official statistics roots in combining the data the agencies have already collected in order to obtain and then to communicate further insights. But it may also be of interest for private market research companies as they would be enabled by statistical matching to save resources as they could

[1] It is also known under the term *data fusion* or *data integration*.

[2] In the literature the term *data file* is commonly used instead.

[3] cf. Fellegi and Sunter (1969) for the elaborated foundations of probabilistic record linkage

[4] cf. Oropallo and Inglese (2004) in which record linkage is used to identify identical units and later on statistical matching is used for the not matched ones.

[5] For an exemplary overview confer to Rässler (2002), Chapter 3.

use already gathered data, instead of conducting a new survey in the field, provided that such a procedure is admissible by the privacy regulations.

In general, there exist two distinct, but related aims of statistical matching. The so-called *macro approach* aims at inferring on the joint probability of all the variables, while the so-called *micro approach* has the goal to obtain a complete, yet partially synthetic data set. Naturally the aims are interrelated, as with knowing the joint probability distribution one can sample a complete data set, and also the other way round, when the joint probability distribution is inferred from the synthetic data set of the micro approach. Even though they are related, for each of the approaches there exist methods which are specifically tailored for it.

Nonetheless, every statistical matching approach needs to tackle the identification problem of the joint probability law, stemming from the fact that there is no observation with information on all variables simultaneously. D'Orazio et al. (2006b, p.9) stated different approaches to possibly overcome it:

i. Assuming conditional independence between the specific variables conditional on the common variables, the so-called *conditional independence assumption*,
ii. using auxiliary or proxy information or
iii. accounting for the uncertainty by providing regions for quantities involving the specific variables.

Irrespective of the chosen approach, in the actual computation the fact is utilised that the specific form of missingness can be regarded as MCAR (missing completely at random)[6]. This allows by basing the inference on the actually observed data to consistently estimate those quantities of interest that involve only jointly observed variables.

The first two approaches have issues, as firstly the assumption of conditional independence is not testable and in many practical situations not correctly assumable, and secondly there may be no auxiliary data at hand, which could be utilised. In the greater framework of this thesis the third approach is the best as it is the most cautious one. However, especially the first approach is attractive as it involves 'only' estimation of precise quantities, and also allows for a straightforward imputation procedure. Exemplary for a macro approach in Endres and Augustin (2016), the dependency structure amongst the variables is estimated by a Bayesian network. Yet, as in D'Orazio et al. (2006a) listed, there have been several applications in the past which led to questionable results, as conditional independence is assumed without further reflection upon it.

The third approach which can be regarded as a partial identification approach, has gained increasing attention in the recent years, however, mostly in terms of a macro approach, e.g. D'Orazio et al. (2006a) and Di Zio and Vantaggi (2017) for (misclassified) categorical variables. But e.g. in Rässler (2002) a parametric micro approach for data following a multivariate normal distribution to obtained bounds by means of multiple imputation is applied.

The contribution considers only categorical data and proposes three related micro approaches (Section 3 of the contribution) introducing the concept of imprecise imputation, i.e. instead of imputing a single value[7], it is proposed to impute a set of plausible values. In this sense it is a multi-valued single imputation. It extends the concept of hot-deck imputation techniques

[6] For the classification of missingness, cf. Rubin (1976).

[7] If imputation of a single value is done only once it would be called *single imputation*, and the repeatedly imputation of single values (which may take different values at each imputation) is termed *multiple imputation*.

(cf. Little and Rubin 2002, Chapter 4), based on so-called *donation classes*. The donation classes partition the observations into groups such that the observations within each group are similar with respect to some pre-specified criterion. In the context of statistical matching the similarity is evaluated on some or all common variables[8].

By switching to a single observation the difference between the three imputation approaches becomes evident. Considering the situation as in Figure 5.1, without loss of generality, let the observations be in donation class g and the values of the variables in $Y = (Y_1, Y_2)$ be missing and their values be imputed simultaneously, i.e. by $(\tilde{Y}_1, \tilde{Y}_2)$. As Y_1 and Y_2 are categorical, with sample spaces (domains) $\mathcal{Y}_1$ and $\mathcal{Y}_2$, respectively, a single observation value is a point in the product of their sample spaces $\mathcal{Y}_1 \times \mathcal{Y}_2$. Hence, in case of single (value) imputation, one would pick a point within $\mathcal{Y}_1 \times \mathcal{Y}_2$ to act as the imputed value. For imprecise imputation in the spirit of partial identification the most cautious approach (domain imputation) is to impute the collection of all points, i.e. the entire product space $\mathcal{Y}_1 \times \mathcal{Y}_2$. This procedure guarantees error-freeness, as the unobserved value has to lie within it, and is actually independent from the specific donation class. A less cautious approach (variable-wise imputation) imputes only a subspace of $\mathcal{Y}_1 \times \mathcal{Y}_2$, namely the product space of the per variable already observed values[9] of Y_1 and Y_2 within the donation class g. The least cautious of the three imprecise imputation approaches (observation-wise imputation), imputes the set of already observed live values within the donation class g, which is a subspace of the two other ones.

The domain imputation may be regarded as micro approach equivalent to the approach of consistent completions of Ramoni and Sebastiani (2001), with the difference that they perform imputation independently for each (single) value of the domain, whereas in the domain imputation the entire domain is used as a substitute. Furthermore, in the setting of only discrete variables, domain imputation may be considered as the hull of multiple imputation in the sense that it contains all data completions. Hence all inference results, which would be obtained for multiple imputation, are contained within the set of inferences based on the domain approach. Variable- and observation-wise imputation follow similar strategies as for the two-pattern case described in Andridge and Little (2010), with the notable differences that in the contribution the donation classes are not specifically tailored for the imputation[10] and there is no longitudinal structure present advocating sequential imputation.

This makes the link to finite random sets evident, as now each variable can take values within the power set of its sample space U excluding the empty set[11]. However, in distinction to Chapter 4, the finite random set comes in this contribution with the interpretation of the epistemic view. It means that the (unobserved) value of the unobserved random variable is contained within the value of the random set with probability one, i.e. there exists at least one unobserved random variable which is an almost surely selector of the random set[12]. Stated differently, a random set in this epistemic interpretation induces a set of precise probability measures on the target space by a virtual set of compatible random variables, the almost surely selectors. A welcome nicety of finite random sets is the ability to characterise the finite random set entirely by its probability mass function, which in case of a categorical target space, i.e. no order on the target space is assumed (and imposed), is a usual probability mass function f on the transformed target space $(\mathcal{P}(U), \mathcal{P}(\mathcal{P}(U)))$. This is the same starting

[8] In the contribution all common variables are involved.

[9] Those are called *live values* in the contribution.

[10] In a practical application, it may be favourable to tailor the donation classes for the imputation of a (set of) certain specific variable(s).

[11] In the contribution a slightly different notation is used: The sample space is denoted by $\mathcal{W}$ therein.

[12] Especially this view makes it necessary from the interpretation point of view to exclude the empty set as outcome of the random set.

point as with random sets in the ontic interpretation. But now an additional structure is imposed, the ordering with respect to set inclusion, which allows for the definition of a probability distribution which utilises this additional structure, as described in Section 2.3. Alternatively, following the constructive definitions of Dempster (1967), upper (and lower) probability measures on $(U, \mathcal{P}(U))$ can be derived (event wise), which fully characterise the set of precise probability measures on $(U, \mathcal{P}(U))$, which are compatible with the random set. The functional of the lower (upper) probability measure corresponds to the distribution (capacity) function. For the relevant theorems and proofs, also showing the equivalence and the link to belief functions, see e.g. Dempster (1967), Nguyen (1978), Couso and Dubois (2014), Nguyen (2006, Chapter 2) and Miranda et al. (2010). The rigorous embedding of imprecise imputation into the theory of finite random set is presented in Section 4 of the contribution.

Therefore, in the contribution in Section 4.2 it is demonstrated which conditioning rule would be suitable, noting that mathematically there are different options on how to condition: Adopting the statistical perspective, conditioning is performed for each probability measure in the compatible set and then the hull thereof is taken. This is the recommended conditioning rule by Couso et al. (2014) to be applied for disjunctive random sets.

By drawing analogy from the probability assessment in the precise case, it is demonstrated in Section 4.3 of the contribution how probability assessments for an event are obtainable in case of random sets. This strategy is cautious as it avoids statements on the distribution of the random sets, which may be not identical for the individual observations. However, by taking the assumption seriously that all observations come from the same precise joint distribution[13], it may also be argued that the outcomes of the random sets are from the same joint distribution with respect to its almost-surely selectors, i.e. underlying random variables. Adhering to this perspective, the distribution of the random set is consistently estimable by the set of empirical probability distributions, by taking the variables in the matched data set as outcomes of random sets (cf. Nguyen and Wu 2006, p.80). As in any context this requires that the sampling designs for obtaining the data sets to be matched are chosen in such a way that the data sets can be regarded as representative for the underlying population (e.g. Rässler 2002, Chapter 2.3), which in statistical matching has even more importance as typically data sets with only similar sampling designs are to be matched.

In a simulation study the appropriateness of the proposed imprecise imputation approaches are evaluated with respect to different dependency structures and strengths of the individual dependencies, as detailed in Section 6 of the contribution. Special care was taken to design the simulation study in such a way to eliminate any sampling error, as detailed in the appendices of the contributions. The simulation study showed the already expected results that the domain approach was the most cautious in terms of wide intervals, but therefore had a guaranteed coverage of the estimator that would have been obtained, if the data had been jointly observed in a precise way[14]. The others approaches yielded shorter intervals, but therefore did not cover very few components of the joint probability distribution[15].

[13] In fact, this assumption is crucial for the justification of statistical matching

[14] Due to the specific simulation design in the contribution, it was also guaranteed to hold for the underlying true value of the population.

[15] Hence the estimated set of probabilities did not contain the one underlying. However, one should notice that such a situation arises when at least one component of the joint probability distribution (overall 46656) is not covered. Given the amount of data used in the simulation this behaviour was not surprising.

5.2 Remarks and Perspectives

The contribution explores new territory as it is the first developing cautious micro approaches which are nonparametric for categorical data. However, the quality and appropriateness of the resulting synthetic matched data set mainly depends on the following factors: The first is the sampling error that is induced by looking only at a fraction of the whole population. As stated previously, such an error is hoped to be controlled for and thus in the literature typically neglected, yet, if present, it still may crucially influence the outcome. The second factor is the so-called *uncertainty* (e.g. D'Orazio et al. 2006b; Conti et al. 2017) which arises from the identification problem that there are no observations containing joint information on the specific variables. There are different ways to quantify this uncertainty e.g. Conti et al. (2012), yet like intuition tells, the uncertainty can be reduced by adding further information, e.g. auxiliary information or constraints (e.g. Conti et al. 2017).

In the contribution one possible way to analyse the resulting (imprecise) synthetic data sets by means of finite random sets is presented, but the analysis is in general not confined to random sets. By using random sets with the epistemic interpretation one automatically assumes the error-freeness, which is required for the construction of the set of image measures on the original target space. For the proposed domain imputation approach this assumption is guaranteed to hold, yet for the other two it is an assumption, which in spirit of partial identification should be critically evaluated. However, this is usually infeasible in practice as one does not know the joint underlying data set or joint data generating process, and consequently also applies for the so-called *matching noise*. Nonetheless, with the aid of simulations, situations can be evaluated on what is affecting the uncertainty, regarding the simulation presented in the contribution as a starting point. Moreover, as the assumption of multivariate normal data, which is frequently used in the literature on statistical matching, is rarely questioned, despite being questionable for real applications.

Due to the explorative nature of the contribution further pitfalls occurring in practical applications were not considered. A strong assumption posed in the contribution is that the observations are coming from independent and identically distributed random variables/sets, which means that all observations within a data set share the same weights. However, in practice when matching surveys, the observations are typically weighted to account for e.g. population strata or non-response. This can be addressed with techniques already developed for imputation, e.g. Conti et al. (2016) (for at least ordinal data), Renssen (1998) (for mainly categorical data) and Rubin (1986) (idea generalisable to any scale of measurement), in order to re-calculate the weights. Due to the imprecise nature of the imputed data, the mere adjustment of individual weights may not be sufficient as values within the imputed imprecise value may come from observations with different weight. This can be addressed by taking the sets of imputed values to be fuzzy sets and thus adjusting for the weighting via the membership function. Moreover, by 'tuning' the membership function one is also able to account for some probabilistic constraints[16], the latter being briefly sketched in Section 5 of the contribution. For sake of simplicity all common variables were used to generate donation classes, although in practice it has to be critically evaluated which ones should actually be used (e.g. D'Orazio et al. 2017). Additionally, one could use different variables in the generation of donations classes for different sets of specific variables like descried in Andridge and Little (2010) for the general situation of missing data .

Regarding the analyses of the synthetic micro data set, one may draw randomly from the set-valued imputed values to obtain a precise value and perform the statistical analysis based on this representation. In order to create statements on the variability, one should conduct the

[16] Structural zeros, i.e. impossible events can be accounted for directly in the imputation step.

statistical analyses multiple times based on different random draws. Yet, such an approach would reduce the imprecise imputation to absurdity as the now sketched strategy would be a multiple imputation strategy in the first place. Taking the uncertainty more seriously, a partial identification approach to construct collection regions for the quantities of interest (cf. Section 2.2) is more reasonable. Another fruitful alternative, especially in the context of fuzzy imputed sets, is the approach of Hüllermeier (2014), which is different to a partial identification approach as in the former the model is fixed first and then evaluated in the light of the most compatible (partially synthetic) observation configuration.

6 (Generalized) Linear regression on microaggregated data

The fourth contribution contains the other 'culture' of statistical modelling according to Breiman (2001b), in which one is interested in the internal mechanics of the data generating process. Thus it provides a link to the first contribution, filling the circle. In order to be meaningful the model class is restricted to such models that allow for an interpretability of their components (typically model specific parameters, which need to be estimated from data, or transformations thereof). But this usually comes along with loosing some of the predictive power. In this view, a single classification tree is inferior to an ensemble of multiple trees and the imprecise classification trees of Chapter 3 are superior to ordinary classification trees. Due to their imprecise nature they suffer less in predictive power. However, in this contribution generalised linear regression was used as model class. The goal is to construct reliable generalised linear regression models on microaggregated data. This requires in some sense the opposite interpretation of data as in Chapter 4: For the analysis the imprecise data in there are to be interpreted as actually precise data after the transformation of the sample space, while in here the seemingly precise data are actually imprecise, as detailed later.

6.1 Specific Theoretical Background

In order to make use of statistical matching of (micro) data[1] as in the last chapter, the data itself need to be available to the institution, e.g. (research) agency or company, which want to analyse them. While each institution is rather free to do anything with self-collected data, as long as they comply with law and ethics standards, the sharing of micro data between different institutions is more regulated for the sake of privacy protection[2]. Therefore, prior to sharing or dissemination of the micro data, the individual records need to be anonymised appropriately[3]. As privacy protection may clashes with data utility, i.e. the ability to analyse the data in a statistical manner and derive meaningful results, a balance between both needs is required. Yet not necessarily all variables within the micro data are sensible and hence required to be protected. One can distinguish four types of variables:

i. Variables acting as *direct identifiers*, e.g. the name and surname combination or any identification number,
ii. variables which are not unique but in combination allow for disclosure (so-called *quasi-identifiers*), e.g. sex or other socio-demographic characteristics,
iii. variables that contain sensitive information on the subjects which should not allow for attribution to specific subjects after anonymisation, e.g. information on disease(s) or income of a person, and
iv. variables that are not considered sensitive and are not required to be anonymised, e.g. intentions or opinions on specific questions.

[1] Each record in micro data is typically associated with a natural or legal person, providing also sensitive information about the person.

[2] In the European Union privacy of personal data is regulated by Regulation (EU) 2016/679 of the European Parliament and of the Council (2016), which will get into full effect on 25 May 2018.

[3] e.g. Chapter IV and V in Regulation (EC) No 223/2009 of the European Parliament and of the Council (2009) as amended by Regulation (EU) 2015/759 of the European Parliament and of the Council (2015) for institutions of the European Statistical Systems

Direct identifiers are never to be shared and usually instantly removed, but even though each quasi-identifier for itself does not allow for an identification of a subject, by means of combining those quasi-identifiers, possibly also with external information, subjects maybe identifiable again. Without the presence of any (quasi-)identifiers in the data set, the sensitive information usually does not allow for the identification of subjects. Hence, it is essential to protect the sensitive information either directly, or indirectly by anonymisation of the quasi-identifiers, or even both.

For indirect protection the concept of k-anonymity by Sweeney (2002) is well established, which requires that in the anonymised data set each combination of quasi-identifiers should occur at least k times. One can also apply this concept to the direct protection. Microaggregation is a class of anonymisation techniques, which achieves k-anonymity by perturbing the original data by first grouping the records into clusters of a least size k and then replacing each individual record with an appropriate representative of the group. In the literature on statistical disclosure control microaggregation is therefore classified as perturbative method (e.g. Willenborg and Waal 2001; Templ 2017). Initially, it was developed for variables on numerical scale, but there are now modifications which allow for any scale of measurement (e.g. Torra 2004; Domingo-Ferrer and Torra 2005).

The multiplicity of different microaggregation techniques stems from the various way groups can be formed. The replacement step is mostly the same, only depending on the scale of measurement of the variables involved. The grouping is performed in such a way that similar records are within the same group, yet due to the minimal group size restriction that is not guaranteed, leading to artefacts (e.g. Domingo-Ferrer and Mateo-Sanz 2002, Figure 1), even for data-driven group sizes.

The simplest form of microaggregation techniques relies on the concept of a sorting variable for a fixed group size k[4]: *Single-axis sorting* simultaneously orders all variables to microaggregate according to a sorting variable, which may be of the variables in question, an external variable or a score. *Individual ranking* is a variation thereof with the difference that each variable to be microaggregated is ordered according to itself and the clusters are defined on a per-variable basis. Other techniques do not rely on a sorting variable, as they perform the clustering in a multivariate manner, e.g. *MDAV* (Maximum Distance to Average Vector) by Domingo-Ferrer and Mateo-Sanz (2002) (and its subsequent evolutions of Domingo-Ferrer and Torra (2005)). Furthermore, the constraint on the group size can be allowed to depend on the data, relaxing 'fixed k' to 'at least k' records, e.g. *k-Ward-algorithm* (Mateo-Sanz and Domingo-Ferrer 1998; Mateo-Sanz and Domingo-Ferrer 1999), *Hansen-Mukherjee-algorithm* (Hansen and Mukherjee 2003) and multivariate Hansen-Mukherjee-based algorithms (Domingo-Ferrer et al. 2006). Additionally, algorithms applying fuzzy clustering techniques have been proposed as basis for microaggregation (e.g. Torra and Miyamoto 2004; Torra 2017). A more detailed listing of microaggregation techniques including their according references can be found in Hundepool et al. (2012, Chapter 3.7.3). As in recent years there has been an explosion in data at hand, the polynomial computational complexity of the distance based approaches does matter and more efficient techniques have been developed to reduce it (e.g. Mortazavi and Jalili 2014).

The effect of specific microaggregation techniques on the estimation of coefficient of linear regression models has already been studied, though rarely in the context of generalised linear regression. All focus on the performance of the naive ordinary least squares estimator, i.e. the ordinary least squares estimator that is obtained when treating the microaggregated data as

[4] If the number of records is not a multiple of k, then one group, typically one in 'the middle', is allowed to contain more than k records

independent observations, neglecting the inter-observational dependency introduced by the microaggregation. In Ronning et al. (2005, Chapter 23) the bias and the consistency of the estimator are studied for various microaggregation scenarios. The authors discovered that only in few situations there is no bias present and the estimation is consistent. A closer investigation was conducted in Schmid and Schneeweiss (2005) by means of a simulation study with similar results[5]. More specialised investigations have been conducted for single-axis sorting (e.g. Schmid et al. 2007; Schmid 2007, Chapter 4), individual ranking (Schmid and Schneeweiss 2008), k-ward-microaggregation (Fink 2009) and microaggregation techniques which aim at preserving the variance (Höhne 2010). Knowing the bias, a correction of the naive estimator leading to unbiased estimation is achievable (e.g. Schmid 2007). All investigations rely on the existence of the closed form solution for the ordinary least squares estimator. In Höhne (2010) it is conjectured and demonstrated by means of examples that for non-linear regression models a bias is naturally introduced. However, therein the term 'non-linear regression model' means that a linear regression model is estimated by ordinary least squares for non-linearly transformed response and covariates.

As stated in the introduction of the contribution, the basic idea is to regard microaggregation as a mapping m from the original data x to the anonymised data $\tilde{x} = m(x)$. In order to fulfil its purpose of privacy protection, it is clear that this mapping is not injective. In a naive way one can estimate a generalised linear regression model on the anonymised data, but by this approach one grasps only the structure present in the microaggregated data, and not necessarily that of the inaccessible underlying data. Therefore, in the contribution the view is shifted to the set of all compatible underlying data situations

$$\mathbb{X}(\tilde{x}) = \{x : m(x) = \tilde{x}\} ,$$

which is the pre-image of $\tilde{x}$. As each individual value within a group is replaced by the representative, each element of $\mathbb{X}(\tilde{x})$ must fulfil this aggregation equality constraint on a per group basis and depending on the used microaggregation technique also on a per variable basis (cf. p.160f in the contribution). As the contribution dealt with numeric variables on a metrical scale, the mean of the individual values within each group was chosen as representative. Depending on the actually used microaggregation technique, another element (in form of inequality constraints) aids in the description of $\mathbb{X}(\tilde{x})$: Some microaggregation techniques, e.g. individual ranking or MDAV, allow for an educated guess of the subregion of the data space to which the underlying values belong.

In the contribution the theoretical foundations for the coefficient estimation of generalised linear regression models on microaggregated data are laid out. It is assumed that only the covariates are microaggregated, but the actually applied microaggregation technique is left unspecified; just for the purpose of the simulation the actual technique is specified. In order to compare with the finding for the ordinary least squares estimator a classical linear regression model in the formulation of generalised linear regression is considered. The formulas for the score function components, as derived in Section 3 of the contribution, are specific for classical linear regression, but the strategy is applicable to the general case.

In order to account for the previously described multiplicity of data situations two different strategies for dealing with it are proposed[6]: The first takes the unknown underlying values directly as nuisance parameters into the model and utilises a maximax[7] optimisation strategy

[5] The situations analysed in the simulation study only overlap with the ones considered in Ronning et al. (2005).

[6] In fact, the same strategies have been proposed by Manski and Tamer (2002), in a (generic) setting of interval-valued covariates.

[7] Iteratively maximising the likelihood with respect to the regression coefficients and the nuisance parameters in an alternating fashion.

to obtain concise estimates for the regression coefficients (Section 3 of the contribution). The second strategy is a pure partial identification approach by constructing the collection region for the regression coefficients[8] (Section 4 of the contribution). It is more cautious than the first approach as it does not assess the plausibility of the underlying non-aggregated values in the light of the likelihood.

In the simulation study both approaches are evaluated and it is demonstrated that they achieve the expected results which are also consistent with the findings for the ordinary least squares estimator, as presented in Section 5 of the contribution.

6.2 Remarks and Perspectives

One structural aspect of microaggregation that was utilised in the contribution is the ability to make an educated guess on the microaggregation technique, its parameter k and for some techniques even regions of the original values from looking at the microaggregated data only. However, having obtained such information puts the privacy at risks and allows for so-called transparency attacks which may lead to exact or reasonably close de-anonymisation. Furthermore, as stated in the discussion on consistency in the contribution (Section 3.2), the number of available observations (n) has an influence on the level of protection. For example looking at data which are microaggregated by means of MDAV or individual ranking: For a fixed k the region within the neighbourhood which consists of all compatible values for a microaggregated one is expected to get smaller with increasing n; in the limit case the microaggregation is practically without effect, even though k-anonymity is still satisfied. This clearly shows the shortcomings of the concept of k-anonymity. There are different general statistical disclosure control approaches which try to overcome them: e.g. p-sensitive k-anonymity (e.g. Truta and Vinay 2006) requires that there are at least p different values of the sensitive variables within each group, l-diversity (e.g. Machanavajjhala et al. 2007; Jian-Min et al. 2008, in context of MDAV microaggreation) enforces that the sensitive values are different to a certain extend with respect to a discrepancy measure, ε-differential privacy (e.g. Dwork 2006) limiting the influence of an individual on the outcome.

Shortcomings of different microaggregation approaches have already been analysed by e.g. Domingo-Ferrer and Torra (2001), Nin et al. (2008), Nin and Torra (2009) and Schmid (2007), and has also been acknowledged by official statistics (D'Acquisto et al. 2015).
In the context of microaggregation, especially when the sensitive variables are on a metrical scale, neither p-sensitive k-anonymity nor l-diversity are able to appropriately protect against disclosure for sufficiently high n. However, a promising protection strategy based on ε-differential privacy is proposed in Sánchez et al. (2016), which adds a random noise to the microaggregated values (by individual ranking), where the variation is dependent on the dissimilarity of the values within each group. A similar strategy was previously proposed in Soria-Comas et al. (2014) requiring (multivariate) so-called *insensitive* microaggregation techniques for which the authors also sketch a generic strategy to obtain such techniques besides their actual adaptation of MDAV to its insensitive variant.

Nonetheless, from the statistical viewpoint data utility should also be considered. As in the contribution the approaches are developed in full generality, they are applicable to any of the previously mentioned. Nonetheless, for the ϵ-differential privacy approaches, one could additionally incorporate the noise addition.

An immediate further research step is the evaluation of the proposed approaches in the contribution beyond classical linear regression. There are already first ideas for dealing

[8] In the contribution actually a hyper cube as an outer approximation to it was estimated.

with binary response in logistic regression. Though, due to the non-linearity of the score function in the regression parameters, the optimisation is less straightforward. Furthermore, for logistic regression without any region constraints, the mixed groups, i.e. those groups which have observations of 0s and 1s, are troublesome, as in their ‘optimum’ they have a likelihood contribution of one by numerically shifting the 0s close to $-\infty$ and the 1s close to ∞, such that the aggregation equality constraint still holds. This in turn means for the maximax approach that they do not contribute to the estimation and need to be removed, reducing the number of effective observations even more, and increasing the risk of complete separation. The findings of Manski and Tamer (2002) in context of bounded covariates may be too pessimistic, as in the case of microaggregation the aggregation equality constraint plays a crucial role, which does in general not allow to freely shift the values to the boundaries. It is conjectured here that for the partial identification approach those mixed groups are equally fatal in the sense that the collection region is the entire parameter space.

7 Final Remarks

As the chapter of each contribution already contains specific remarks and research perspectives, in the following some global remarks are given.

This thesis encompassed different methods on how to deal with data in a cautious way. It is demonstrated that the meaningful interpretation of the data itself is crucial in order to obtain credible analysis results, which is especially relevant for vague data. It is shown in case of anonymisation that seemingly precise data are actually to be treated as imprecise (Chapter 6), while also the converse may hold (Chapter 4). Furthermore, practical methods for analysis of both types of data are proposed, also in cases when they are inherently imprecise (Chapter 5). But cautious modelling is not limited to vague data, even precise data may benefit from inferences which do not neglect different kinds of uncertainty (Chapter 3).

In terms of the statistical learning scheme of Chapter 1, one should refrain as much as possible from imposing assumptions which are questionable with respect to the background of application and are only justified by technical convenience. As a consequence of this, some reflection on the methods applied is necessary in order to decide on assumptions to actually maintain, as seen in the application on microaggregation and imprecise imputation. Building statistical models which allow for vague data provides a twofold benefit: Firstly, it allows to drop assumptions which are only necessary in order to squeeze the vague data into a precise model. Secondly, the model may be used for similar kinds of data, invalidating the common claim that only precise data, which actually reflect only half the truth as they neglect their inherent uncertainty, are produced in surveys because of a lack of models being able to deal with vague ones. Figuratively speaking, the chicken (the model) is provided to lay eggs (the vague data).

As seen in the elicitation of (subjective) probability statements (Mosteller and Youtz 1990; Budescu et al. 2009) the imprecise representation of the underlying probability value is more honest. Furthermore, as convincingly demonstrated in Smithson and Segale (2009) that representation allows to overcome some issues the precise probability judgements suffer from. Taking it a step further into surveys, the commonly used discrete scales, e.g. for measuring (dis)agreement to statements, could be broken up in order to allow assessments via the staircase method (Tversky and Koehler 1994), similar to the assessment of probability judgements.

In either generalisation, the produced data would be vague, and depending on the context, one could analyse them either from the ontic or epistemic view, as elaborated on in Section 2.3. By this approach inherent indecisiveness of the respondents can be captured and appropriately accounted for. In this spirit the present thesis contains recipes, which were exemplary executed in the presented contexts, but are general enough to be utilised in other situations.

Despite the power of cautious methods, it should be emphasised that the still quality, meaningfulness and interpretation of the data at hand are essential for the trustworthiness of statistical analyses and their results.

Further References

J. Abellán, R. M. Baker and F. P. A. Coolen (2011). 'Maximising entropy on the nonparametric predictive inference model for multinomial data'. In: *European Journal of Operational Research* 212(1), pages 112–122. DOI: 10.1016/j.ejor.2011.01.020.

J. Abellán, R. M. Baker, F. P. Coolen, R. J. Crossman and A. R. Masegosa (2014). 'Classification with decision trees from a nonparametric predictive inference perspective'. In: *Computational Statistics & Data Analysis* 71, pages 789–802. DOI: 10.1016/j.csda.2013.02.009.

J. Abellán and A. R. Masegosa (2010). 'An ensemble method using credal decision trees'. In: *European Journal of Operational Research* 205(1), pages 218–226. DOI: 10.1016/j.ejor.2009.12.003.

J. Abellán and A. R. Masegosa (2012). 'Bagging schemes on the presence of class noise in classification'. In: *Expert Systems with Applications* 39(8), pages 6827–6837. DOI: 10.1016/j.eswa.2012.01.013.

J. Abellán and S. Moral (2003). 'Building classification trees using the total uncertainty criterion'. In: *International Journal of Intelligent Systems* 18(12), pages 1215–1225. DOI: 10.1002/int.10143.

J. Abellán and S. Moral (2005). 'Upper entropy of credal sets. Applications to credal classification'. In: *International Journal of Approximate Reasoning* 39(2-3), pages 235–255. DOI: 10.1016/j.ijar.2004.10.001.

R. R. Andridge and R. J. A. Little (2010). 'A review of hot deck imputation for survey non-response'. In: *International Statistical Review* 78(1), pages 40–64. DOI: 10.1111/j.1751-5823.2010.00103.x.

T. Augustin and F. P. A. Coolen (2004). 'Nonparametric predictive inference and interval probability'. In: *Journal of Statistical Planning and Inference* 124(2), pages 251–272. DOI: 10.1016/j.jspi.2003.07.003.

T. Augustin, F. P. A. Coolen, G. de Cooman and M. C. M. Troffaes, editors (2014). *Introduction to Imprecise Probabilities*. Chichester: Wiley.

R. M. Baker (2010). 'Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory Data'. PhD thesis. Durham University, GB. Available via http://www.npi-statistics.com/pdfs/theses/RB10.pdf.

R. M. Baker, T. Coolen-Maturi and F. P. Coolen (2017). 'Nonparametric Predictive Inference for Stock Returns'. In: *Journal of Applied Statistics* 44(8), pages 1333–1349. DOI: 10.1080/02664763.2016.1204429.

A. Beresteanu, I. Molchanov and F. Molinari (2012). 'Partial identification using random set theory'. In: *Journal of Econometrics* 166(1), pages 17–32. DOI: 10.1016/j.jeconom.2011.06.003.

A. Beresteanu and F. Molinari (2008). 'Asymptotic properties for a class of partially identified models'. In: *Econometrica* 76(4), pages 763–814. DOI: 10.1111/j.1468-0262.2008.00859.x.

L. Breiman (1996). 'Bagging predictors'. In: *Machine Learning* 24(2), pages 123–140. DOI: 10.1007/BF00058655.

L. Breiman (2001a). 'Random forests'. In: *Machine Learning* 45(1), pages 5–32. DOI: 10.1023/A:1010933404324.

L. Breiman (2001b). 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)'. In: *Statistical Science* 16(3), pages 199–231. DOI: 10.1214/ss/1009213726.

L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone (1984). *Classification and Regression Trees*. Boca Raton: Chapman & Hall/CRC.

D. V. Budescu, S. Broomell and H.-H. Por (2009). 'Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change'. In: *Psychological Science* 20(3), pages 299–308. DOI: 10.1111/j.1467-9280.2009.02284.x.

G. Choquet (1954). 'Theory of capacities'. In: *Annales de l'Institut Fourier* 5, pages 131–295. DOI: 10.5802/aif.53.

P. L. Conti, D. Marella and M. Scanu (2012). 'Uncertainty analysis in statistical matching'. In: *Journal of Official Statistics* 28(1), pages 69–88.

P. L. Conti, D. Marella and M. Scanu (2016). 'Statistical matching analysis for complex survey data with applications'. In: *Journal of the American Statistical Association* 111(516), pages 1715–1725. DOI: 10.1080/01621459.2015.1112803.

P. L. Conti, D. Marella and M. Scanu (2017). 'How far from identifiability? A systematic overview of the statistical matching problem in a non parametric framework'. In: *Communications in Statistics - Theory and Methods* 46(2), pages 967–994. DOI: 10.1080/03610926.2015.1010005.

T. Coolen-Maturi (2017). 'Predictive inference for best linear combination of biomarkers subject to limits of detection'. In: *Statistics in Medicine* 36(18), pages 2844–2874. DOI: 10.1002/sim.7317.

T. Coolen-Maturi, F. F. Elkhafifi and F. P. A. Coolen (2014). 'Three-group ROC analysis: A nonparametric predictive approach'. In: *Computational Statistics & Data Analysis* 78, pages 69–81. DOI: 10.1016/j.csda.2014.04.005.

F. P. A. Coolen (1998). 'Low structure imprecise predictive inference for Bayes' problem'. In: *Statistics & Probability Letters* 36(4), pages 349–357. DOI: 10.1016/S0167-7152(97)00081-3.

F. P. A. Coolen and T. Augustin (2005). 'Learning from multinomial data: a nonparametric predictive alternative to the imprecise Dirichlet model'. In: *ISIPTA '05: Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications.* Edited by F. Cozman, R. Nau and T. Seidenfeld. Pittsburgh, Manno: Carnegie Mellon University, SIPTA, pages 125–135. URL: http://www.sipta.org/isipta05/proceedings/037.html.

F. P. A. Coolen and T. Augustin (2009). 'A nonparametric predictive alternative to the imprecise Dirichlet model: The case of a known number of categories'. In: *International Journal of Approximate Reasoning* 50(2), pages 217–230. DOI: 10.1016/j.ijar.2008.03.011.

F. P. A. Coolen and T. Coolen-Maturi (2015). 'Predictive inference for system reliability after common-cause component failures'. In: *Reliability Engineering & System Safety* 135, pages 27–33. DOI: 10.1016/j.ress.2014.11.005.

I. Couso and D. Dubois (2014). 'Statistical reasoning with set-valued information: Ontic vs. epistemic views'. In: *International Journal of Approximate Reasoning* 55(7), pages 1502–1518. DOI: 10.1016/j.ijar.2013.07.002.

I. Couso, D. Dubois and L. Sánchez (2014). *Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables.* Cham: Springer.

R. J. Crossman, J. Abellán, T. Augustin and F. P. A. Coolen (2011). 'Building imprecise classification trees with entropy ranges'. In: *ISIPTA '11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications.* Edited by F. P. A. Coolen, G. de Cooman, T. Fetz and M. Oberguggenberger. Manno: SIPTA, pages 129–138. URL: http://www.sipta.org/isipta11/index.php?id=paper&paper=028.html.

G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye and A. Bourka (2015). *Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics.* ENISA Report. European Union Agency for Network and Information Security (ENISA). DOI: 10.2824/641480.

M. D'Orazio, M. Di Zio and M. Scanu (2006a). 'Statistical matching for categorical data: Displaying uncertainty and using logical constraints'. In: *Journal of Official Statistics* 22(1), pages 137–157.

M. D'Orazio, M. Di Zio and M. Scanu (2006b). *Statistical Matching: Theory and Practice.* Chichester: Wiley.

M. D'Orazio, M. Di Zio and M. Scanu (2017). 'The use of uncertainty to choose matching variables in statistical matching'. In: *Soft Methods for Data Science.* Edited by M. B. Ferraro, P. Giordani, B. Vantaggi, M. Gagolewski, M. Ángeles Gil, P. Grzegorzewski and O. Hryniewicz. Cham: Springer, pages 149–156. DOI: 10.1007/978-3-319-42972-4_19.

M. Debus (2013). 'Koalitionspräferenzen als erklärende Komponente des Wahlverhaltens: Eine Untersuchung anhand der Bundestagswahl 2009'. In: *Koalitionen, Kandidaten, Kommunikation: Analysen zur Bundestagswahl 2009.* Edited by T. Faas, K. Arzheimer, S. Roßteutscher and B. Weßels. Wiesbaden: Springer Fachmedien, pages 57–76. DOI: `10.1007/978-3-531-94010-6_4`.

A. P. Dempster (1967). 'Upper and lower probabilities induced by a multivalued mapping'. In: *The Annals of Mathematical Statistics* 38(2), pages 325–339. DOI: `10.1214/aoms/1177698950`.

Der Bundeswahlleiter (8th Aug. 2017). *42 Parteien nehmen an der Bundestagswahl 2017 teil.* Pressemitteilung Nr. 10/17. URL: `https://www.bundeswahlleiter.de/info/presse/mitteilungen/bundestagswahl-2017/10_17_teilnahme.html`.

M. Di Zio and B. Vantaggi (2017). 'Partial identification in statistical matching with misclassification'. In: *International Journal of Approximate Reasoning* 82, pages 227–241. DOI: `10.1016/j.ijar.2016.12.015`.

J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz and F. Sebé (2006). 'Efficient multivariate data-oriented microaggregation'. In: *The VLDB Journal* 15(4), pages 355–369. DOI: `10.1007/s00778-006-0007-0`.

J. Domingo-Ferrer and J. M. Mateo-Sanz (2002). 'Practical data-oriented microaggregation for statistical disclosure control'. In: *IEEE Transactions on Knowledge and Data Engineering* 14(1), pages 189–201. DOI: `10.1109/69.979982`.

J. Domingo-Ferrer and V. Torra (2001). 'Disclosure control methods and information loss for microdata'. In: *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Edited by J. I. Lane, P. Doyle, L. Zayatz and J. Theeuwes. Amsterdam: North-Holland, pages 91–110.

J. Domingo-Ferrer and V. Torra (2005). 'Ordinal, continuous and heterogeneous k-anonymity through microaggregation'. In: *Data Mining and Knowledge Discovery* 11(2), pages 195–212. DOI: `10.1007/s10618-005-0007-5`.

C. Dwork (2006). 'Differential privacy'. In: *Automata, Languages and Programming.* Edited by M. Bugliesi, B. Preneel, V. Sassone and I. Wegener. Berlin, Heidelberg: Springer, pages 1–12. DOI: `10.1007/11787006_1`.

M. Elff and S. Roßteutscher (2011). 'Stability or decline? Class, religion and the vote in Germany'. In: *German Politics* 20(1), pages 107–127. DOI: `10.1080/09644008.2011.554109`.

E. Endres and T. Augustin (2016). 'Statistical matching of discrete data by Bayesian networks'. In: *Proceedings of the Eighth International Conference on Probabilistic Graphical Models.* Edited by A. Antonucci, G. Corani and C. de Campos. Volume 52. Proceedings of Machine Learning Research. PMLR, pages 159–170. URL: `http://proceedings.mlr.press/v52/endres16.html`.

European Parliament, Council of the European Union (31st Mar. 2009). 'Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities'. In: *Official Journal* L 87, pages 164–173. URL: `http://data.europa.eu/eli/reg/2009/223/oj`.

European Parliament, Council of the European Union (19th May 2015). 'Regulation (EU) 2015/759 of the European Parliament and of the Council of 29 April 2015 amending Regulation (EC) No 223/2009 on European statistics'. In: *Official Journal* L 123, pages 90–97. URL: `http://data.europa.eu/eli/reg/2015/759/oj`.

European Parliament, Council of the European Union (4th May 2016). 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)'. In: *Official Journal* L 119, pages 1–88. URL: `http://data.europa.eu/eli/reg/2016/679/oj`.

Eurostat (2013). 'Statistical matching: a model based approach for data integration'. In: *Methodologies & Working papers.* Luxembourg: Publications Office of the European Union. DOI: `10.2785/44822`.

L. Fahrmeir, T. Kneib, S. Lang and B. Marx (2013). *Regression: Models, Methods and Applications.* Berlin: Springer.

I. P. Fellegi and A. B. Sunter (1969). 'A theory for record linkage'. In: *Journal of the American Statistical Association* 64(328), pages 1183–1210. DOI: 10.1080/01621459.1969.10501049.

M. B. Ferraro and P. Giordani (2012). 'A multiple linear regression model for imprecise information'. In: *Metrika* 75(8), pages 1049–1068. DOI: 10.1007/s00184-011-0367-3.

P. Fink (2009). 'K-Ward-microaggregated data in linear models: An analytical approach to obtain unbiased estimators under anonymized data'. Bachelor's thesis. Ludwig-Maximilians-Universität München, DE. URL: https://epub.ub.uni-muenchen.de/11132/1/BA_Fink.pdf.

P. Fink (2012). 'Ensemble methods for classification trees under imprecise probabilities'. Master's thesis. Ludwig-Maximilians-Universität München, DE. URL: https://epub.ub.uni-muenchen.de/25521/1/MA_Fink_Paul.pdf.

C. Ganser and P. Riordan (2015). 'Vote expectations at the next level. Trying to predict vote shares in the 2013 German federal election by polling expectations'. In: *Electoral Studies* 40, pages 115–126. DOI: 10.1016/j.electstud.2015.08.001.

P. Giordani (2015). 'Lasso-constrained regression analysis for interval-valued data'. In: *Advances in Data Analysis and Classification* 9(1), pages 5–19. DOI: 10.1007/s11634-014-0164-8.

I. R. Goodman and H. T. Nguyen (1985). *Uncertainty Models for Knowledge-Based Systems.* Amsterdam: North Holland.

A. Graefe (2015). 'German election forecasting: Comparing and combining methods for 2013'. In: *German Politics* 24(2), pages 195–204. DOI: 10.1080/09644008.2015.1024240.

S. L. Hansen and S. Mukherjee (2003). 'A polynomial algorithm for optimal univariate microaggregation'. In: *IEEE Transactions on Knowledge and Data Engineering* 15(4), pages 1043–1044. DOI: 10.1109/TKDE.2003.1209020.

T. J. Hastie and R. J. Tibshirani (1999). *Generalized Additive Models.* Boca Raton: Chapman & Hall/CRC.

D. F. Heitjan and D. B. Rubin (1991). 'Ignorability and coarse data'. In: *The Annals of Statistics* 19(4), pages 2244–2253. DOI: 10.1214/aos/1176348396.

O. Hellevik (2009). 'Linear versus logistic regression when the dependent variable is a dichotomy'. In: *Quality & Quantity* 43(1), pages 59–74. DOI: 10.1007/s11135-007-9077-3.

B. M. Hill (1968). 'Posterior distribution of percentiles: Bayes' theorem for sampling from a population'. In: *Journal of the American Statistical Association* 63(322), pages 677–691. DOI: 10.2307/2284038.

P. D. Hoff and X. Niu (2012). 'A covariance regression model'. In: *Statistica Sinica* 22, pages 729–753. DOI: 10.5705/ss.2010.051.

J. Höhne (2010). *Verfahren zur Anonymisierung von Einzeldatenn.* Volume 16. Statistik und Wissenschaft. Wiesbaden: Statistisches Bundesamt.

T. Hothorn, K. Hornik and A. Zeileis (2006). 'Unbiased recursive partitioning: A conditional inference framework'. In: *Journal of Computational and Graphical Statistics* 15(3), pages 651–674. DOI: 10.1198/106186006X133933.

E. Hüllermeier (2014). 'Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization'. In: *International Journal of Approximate Reasoning* 55(7), pages 1519–1534. DOI: 10.1016/j.ijar.2013.09.003.

A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer and P.-P. de Wolf (2012). *Statistical Disclosure Control.* Chichester: Wiley.

N. Huntley, R. Hable and M. C. M. Troffaes (2014). 'Decision making'. In: *Introduction to Imprecise Probabilities.* Edited by T. Augustin, F. P. A. Coolen, G. de Cooman and M. C. M. Troffaes. Chichester: Wiley, pages 190–206.

H. Jian-Min, C. Ting-Ting and Y. Hui-Qun (2008). 'An improved V-MDAV algorithm for l-diversity'. In: *Proceedings of International Symposium on Information Processes (ISIP 2008)*. IEEE, pages 733–739. DOI: 10.1109/ISIP.2008.110.

G. V. Kass (1980). 'An exploratory technique for investigating large quantities of categorical data'. In: *Applied Statistics* 29(2), pages 119–127. DOI: 10.2307/2986296.

M. A. Kayser and A. Leininger (2016). 'A predictive test of voters' economic benchmarking: The 2013 German Bundestag election'. In: *German Politics* 25(1), pages 106–130. DOI: 10.1080/09644008.2015.1129531.

D. G. Kendall (1974). 'Foundations on a theory of random sets'. In: *Stochastic Geometry*. Edited by E. F. Harding and D. G. Kendall. New York, NY: Wiley, pages 322–376.

A. N. Kolmogorov (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer-Verlag.

T. Küntzler (2017). 'Using data combination of fundamental variable-based forecasts and poll-based forecasts to predict the 2013 German election'. In: *German Politics*, pages 1–19. DOI: 10.1080/09644008.2017.1280781.

E. L. Lehmann and G. Casella (1998). *Theory of Point Estimation*. 2nd. New York: Springer.

R. J. A. Little and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. 2nd. Hoboken: Wiley.

A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam (2007). 'l-diversity: Privacy beyond k-anonymity'. In: *ACM Transactions on Knowledge Discovery from Data* 1(1), page 3. DOI: 10.1145/1217299.1217302.

R. P. S. Mahler (1994). 'Random-set approach to data fusion'. In: *SPIE's International Symposium on Optical Engineering and Photonics in Aerospace Sensing*. DOI: 10.1117/12.181026.

R. P. S. Mahler (2000). *An introduction to multisource-multitarget statistics and its applications*. Lockheed Martin Technical Monograph.

R. P. S. Mahler (2013). '"Statistics 102" for multisource-multitarget detection and tracking'. In: *IEEE Journal of Selected Topics in Signal Processing* 7(3), pages 376–389. DOI: 10.1109/JSTSP.2013.2253084.

C. F. Manski (2003). *Partial Identification of Probability Distributions*. Berlin: Springer.

C. F. Manski (2007). *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

C. F. Manski and E. Tamer (2002). 'Inference on regressions with interval data on a regressor or outcome'. In: *Econometrica* 70(2), pages 519–546.

C. J. Mantas and J. Abellán (2014). 'Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data'. In: *Expert Systems with Applications* 41(10), pages 4625–4637. DOI: 10.1016/j.eswa.2014.01.017.

C. J. Mantas, J. Abellán and J. G. Castellano (2016). 'Analysis of Credal-C4.5 for classification in noisy domains'. In: *Expert Systems with Applications* 61, pages 314–326. DOI: 10.1016/j.eswa.2016.05.035.

J. M. Mateo-Sanz and J. Domingo-Ferrer (1998). 'A comparative study of microaggregation methods'. In: *Qüestiió* 22(3), pages 511–526.

J. M. Mateo-Sanz and J. Domingo-Ferrer (1999). 'A method for data-oriented multivariate microaggregation'. In: *Statistical Data Protection*. Edited by J. Domingo-Ferrer. Luxembourg: Office for Official Publications of the European Communities, pages 89–99.

G. Matheron (1975). *Random Sets and Integral Geometry*. New York: Wiley.

E. Miranda, I. Couso and P. Gil (2005). 'Random sets as imprecise random variables'. In: *Journal of Mathematical Analysis and Applications* 307(1), pages 32–47. DOI: 10.1016/j.jmaa.2004.10.022.

E. Miranda, I. Couso and P. Gil (2010). 'Approximations of upper and lower probabilities by measurable selections'. In: *Information Sciences* 180(8), pages 1407–1417. DOI: 10.1016/j.ins.2009.12.005.

I. Molchanov (2005). *Theory of Random Sets*. London: Springer.

S. Moral (2014). 'Comments on "Statistical reasoning with set-valued information: Ontic vs. epistemic view" by Inés Couso and Didier Dubois'. In: *International Journal of Approximate Reasoning* 55(7), pages 1578–1579. DOI: 10.1016/j.ijar.2014.04.004.

R. Mortazavi and S. Jalili (2014). 'Fast data-oriented microaggregation algorithm for large numerical datasets'. In: *Knowledge-Based Systems* 67, pages 195–205. DOI: 10.1016/j.knosys.2014.05.011.

F. Mosteller and C. Youtz (1990). 'Quantifying probabilistic expressions'. In: *Statistical Science* 5(1), pages 2–12.

J. A. Nelder and R. W. M. Wedderburn (1972). 'Generalized Linear Models'. In: *Journal of the Royal Statistical Society. Series A (General)* 135(3), pages 370–384. DOI: 10.2307/2344614.

H. T. Nguyen (1978). 'On random sets and belief functions'. In: *Journal of Mathematical Analysis and Applications* 65(3), pages 531–542. DOI: 10.1016/0022-247X(78)90161-0.

H. T. Nguyen (2006). *An Introduction to Random Sets.* Boca Raton: Chapman & Hall/CRC.

H. T. Nguyen and B. Wu (2006). 'Random and fuzzy sets in coarse data analysis'. In: *Computational Statistics & Data Analysis* 51(1), pages 70–85. DOI: 10.1016/j.csda.2006.04.016.

J. Nin, J. Herranz and V. Torra (2008). 'On the disclosure risk of multivariate microaggregation'. In: *Data & Knowledge Engineering* 67(3), pages 399–412. DOI: 10.1016/j.datak.2008.06.014.

J. Nin and V. Torra (2009). 'Analysis of the univariate microaggregation disclosure risk'. In: *New Generation Computing* 27(3), pages 197–214. DOI: 10.1007/s00354-007-0061-1.

F. Oropallo and F. Inglese (2004). 'The development of an integrated and systematized information system for economic and policy impact'. In: *Austrian Journal of Statistics* 33(1&2), pages 211–235. DOI: 10.17713/ajs.v33i1&2.439.

J. Plass, M. E. G. V. Cattaneo, G. Schollmeyer and T. Augustin (2017). 'On the testability of coarsening assumptions: A hypothesis test for subgroup independence'. In: *International Journal of Approximate Reasoning* 90, pages 292–306. DOI: 10.1016/j.ijar.2017.07.014.

J. R. Quinlan (1986). 'Induction of decision trees'. In: *Machine Learning* 1(1), pages 81–106. DOI: 10.1007/BF00116251.

J. R. Quinlan (1993). *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann Publishers Inc.

M. Ramoni and P. Sebastiani (2001). 'Robust learning with missing data'. In: *Machine Learning* 45(2), pages 147–170. DOI: 10.1023/A:1010968702992.

S. Rässler (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches.* New York: Springer.

H. Rattinger, S. Roßteutscher, R. Schmitt-Beck, B. Weßels and C. Wolf (2014). *Vorwahl-Querschnitt (GLES 2013).* GESIS Datenarchiv, Köln. ZA5700 Datenfile Version 2.0.0. DOI: 10.4232/1.12000.

R. H. Renssen (1998). 'Use of statistical matching techniques in calibration estimation'. In: *Survey Methodology* 24(2), pages 171–183.

G. Ronning, R. Sturm, J. Höhne, R. Lenz, M. Rosemann, M. Scheffler and D. Vorgrimler (2005). *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten.* Volume 4. Statistik und Wissenschaft. Wiesbaden: Statistisches Bundesamt.

D. B. Rubin (1976). 'Inference and missing data'. In: *Biometrika* 63(3), pages 581–592. DOI: 10.1093/biomet/63.3.581.

D. B. Rubin (1986). 'Statistical matching using file concatenation with adjusted weights and multiple imputations'. In: *Journal of Business & Economic Statistics* 4(1), pages 87–94. DOI: 10.2307/1391390.

RuleQuest Research (2017). *See5/C5.0.* Version 2.11. URL: https://www.rulequest.com/see5-info.html.

H. C. Rutemiller and D. A. Bowers (1968). 'Estimation in a Heteroscedastic Regression Model'. In: *Journal of the American Statistical Association* 63(322), pages 552–557. DOI: 10.2307/2284026.

D. Sánchez, J. Domingo-Ferrer, S. Martínez and J. Soria-Comas (2016). 'Utility-preserving differentially private data releases via individual ranking microaggregation'. In: *Information Fusion* 30, pages 1–14. DOI: 10.1016/j.inffus.2015.11.002.

M. Schmid (2007). *Estimation of a Linear Regression with Microaggreated Data.* Munich: Verlag Dr. Hut.

M. Schmid and H. Schneeweiss (2005). *The effect of microaggregation procedures on the estimation of linear models: A simulation study.* Technical report 443. München: Institut für Statistik, Sonderforschungsbereich 386. URL: https://epub.ub.uni-muenchen.de/1831/.

M. Schmid and H. Schneeweiss (2008). 'Estimation of a linear model in transformed variables under microaggregation by individual ranking'. In: *AStA Advances in Statistical Analysis* 92(4), pages 359–374. DOI: 10.1007/s10182-008-0087-9.

M. Schmid, H. Schneeweiss and H. Küchenhoff (2007). 'Estimation of a linear regression under microaggregation with the response variable as a sorting variable'. In: *Statistica Neerlandica* 61(4), pages 407–431. DOI: 10.1111/j.1467-9574.2007.00366.x.

R. Schnell and M. Noack (2014). 'The accuracy of pre-election polling of German general elections'. In: *methods, data, analyses* 8(1), pages 5–24. DOI: 10.12758/mda.2014.001.

G. Schollmeyer and T. Augustin (2015). 'Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data'. In: *International Journal of Approximate Reasoning* 56, pages 224–248. DOI: 10.1016/j.ijar.2014.07.003.

P. Selb and S. Munzert (2016). 'Forecasting the 2013 German Bundestag election using many polls and historical election results'. In: *German Politics* 25(1), pages 73–83. DOI: 10.1080/09644008.2015.1121454.

G. Shafer (1976). *A Mathematical Theory of Evidence.* Princeton: Princeton University Press.

A. C. Singh, H. J. Mantel, M. D. Kinack and G. Rowe (1993). 'Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption'. In: *Survey Methodology* 19(1), pages 59–79.

M. Smithson and E. C. Merkle (2013). *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables.* New York: Chapman & Hall/CRC.

M. Smithson and C. Segale (2009). 'Partition priming in judgments of imprecise probabilities'. In: *Journal of Statistical Theory and Practice* 3(1), pages 169–181. DOI: 10.1080/15598608.2009.10411918.

M. Smithson and J. Verkuilen (2006). 'A better lemon squeezer? Maximum-likelihood regression with Beta-distributed dependent variables'. In: *Psychological Methods* 11(1), pages 54–71. DOI: 10.1037/1082-989X.11.1.54.

G. K. Smyth (1989). 'Generalized Linear Models with Varying Dispersion'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 51(1), pages 47–60.

J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez (2014). 'Enhancing data utility in differential privacy via microaggregation-based k-anonymity'. In: *The VLDB Journal* 23(5), pages 771–794. DOI: 10.1007/s00778-014-0351-4.

D. Stoyan (1998). 'Random Sets: Models and Statistics'. In: *International Statistical Review* 66(1), pages 1–27. DOI: 10.1111/j.1751-5823.1998.tb00403.x.

L. Sweeney (2002). 'k-Anonymity: A model for protecting privacy'. In: *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5), pages 557–570. DOI: 10.1142/S0218488502001648.

E. Tamer (2010). 'Partial identification in econometrics'. In: *Annual Reviews in Economics* 2(1), pages 167–195. DOI: 10.1146/annurev.economics.050708.143401.

M. Templ (2017). *Statistical Disclosure Control for Microdata: Methods and Applications in R.* Cham: Springer.

P. W. Thurner (2000). 'The empirical application of the spatial theory of voting in multiparty systems with random utility models'. In: *Electoral Studies* 19(4), pages 493–517. DOI: 10.1016/S0261-3794(99)00025-6.

V. Torra (2004). 'Microaggregation for categorical variables: A median based approach'. In: *Privacy in Statistical Databases.* Edited by J. Domingo-Ferrer and V. Torra. Berlin, Heidelberg: Springer, pages 162–174. DOI: `10.1007/978-3-540-25955-8_13`.

V. Torra (2017). 'Fuzzy microaggregation for the transparency principle'. In: *Journal of Applied Logic* 23, pages 70–80. DOI: `10.1016/j.jal.2016.11.007`.

V. Torra and S. Miyamoto (2004). 'Evaluating fuzzy clustering algorithms for microdata protection'. In: *Privacy in Statistical Databases.* Edited by J. Domingo-Ferrer and V. Torra. Berlin, Heidelberg: Springer, pages 175–186. DOI: `10.1007/978-3-540-25955-8_14`.

T. M. Truta and B. Vinay (2006). 'Privacy protection: p-sensitive k-anonymity property'. In: *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW 2006).* IEEE, page 94. DOI: `10.1109/ICDEW.2006.116`.

A. Tversky and D. J. Koehler (1994). 'Support theory: A nonextensional representation of subjective probability'. In: *Psychological Review* 101(4), pages 547–567. DOI: `10.1037/0033-295X.101.4.547`.

S. Waldmann (2014). *Topology: An Introduction.* Cham: Springer.

P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities.* London: Chapman & Hall.

P. Walley (1996). 'Inferences from multinomial data: Learning about a bag of marbles'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1), pages 3–34. DOI: `10.2307/2346164`.

K. Weichselberger (2000). 'The theory of interval-probability as a unifying concept for uncertainty'. In: *International Journal of Approximate Reasoning* 24(2-3), pages 149–170. DOI: `10.1016/S0888-613X(00)00032-3`.

K. Weichselberger (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept.* Heidelberg: Physica.

L. Willenborg and T. de Waal (2001). *Elements of Statistical Disclosure Control.* New York: Springer.

S. N. Wood (2017). *Generalized Additive Models: An Introduction with R.* 2nd edition. Boca Raton: Chapman & Hall / CRC.

Y.-C. Yin, F. P. A. Coolen and T. Coolen-Maturi (2017). 'An imprecise statistical method for accelerated life testing using the power-Weibull model'. In: *Reliability Engineering & System Safety* 167, pages 158–167. DOI: `10.1016/j.ress.2017.05.045`.

L. A. Zadeh (1978). 'PRUF–a meaning representation language for natural languages'. In: *International Journal of Man-Machine Studies* 10(4), pages 395–460. DOI: `10.1016/S0020-7373(78)80003-0`.

M. Zaffalon, G. Corani and D. Mauá (2012). 'Evaluating credal classifiers by utility-discounted predictive accuracy'. In: *International Journal of Approximate Reasoning* 53(8), pages 1282–1301. DOI: `10.1016/j.ijar.2012.06.022`.

Affidavit

Hereby I, Paul Fink, declare in lieu of an oath that the present dissertation was composed autonomously without any illicit aid.

Munich, 1 March 2018

(Paul Fink)

Eidesstattliche Versicherung

Hiermit erkläre ich, Paul Fink, an Eides statt, dass die Dissertation von mir selbstständig und ohne unerlaubte Hilfe oder Hilfsmittel angefertigt worden ist.

München, 1. März 2018

(Paul Fink)