Modeling Visit Potential of Geographic Locations Based on Mobility Data

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt
von
Christine Körner
aus
Leipzig

Bonn, 2011

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Stefan Wrobel
2. Gutachter: Prof. Dr. Armin B. Cremers
Tag der Promotion: 23.12.2011
Erscheinungsjahr: 2012

Christine Körner

Universität Bonn
Institut für Informatik III

und

Fraunhofer-Institut für Intelligente Analyse- und
Informationssysteme IAIS

Schloss Birlinghoven
53757 Sankt Augustin

christine.koerner@iais.fraunhofer.de

Declaration

I, Christine Körner, confirm that this work submitted for assessment is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (e.g. ideas, equations, figures, text, tables, programs) are properly acknowledged at the point of their use. A full list of the references employed has been included.

Abstract

Every day people interact with the environment by passing or visiting geographic locations. Information about such entity-location interactions can be used in a number of applications and its value has been recognized by companies and public institutions. However, although the necessary tracking technologies such as GPS, GSM or RFID have long found their way into everyday life, the practical usage of visit information is still limited. Besides economic and ethical reasons for the restricted usage of entity-location interactions there are also two very basic problems. First, no formal definition of entity-location interaction quantities exists. Second, at the current state of technology, no tracking technology guarantees complete observations, and the treatment of missing data in mobility applications has been neglected in trajectory data mining so far. This thesis therefore focuses on the definition and estimation of quantities about the visiting behavior between mobile entities and geographic locations from incomplete mobility data.

In a first step we provide an application-independent language to evaluate entity-location interactions. Based on a uniform notation, we define a family of quantities called *visit potential*, which contains the most basic interaction quantities and can be extended on need. By identifying the common background of all quantities we are able to analyze relationships between different quantities and to infer consistency requirements between related parameterizations of the quantities. We demonstrate the general applicability of visit potential using two real-world applications for which we give a precise definition of the employed entity-location interaction quantities in terms of visit potential.

Second, this thesis provides the first systematic analysis of methods for the handling of missing data in mobility mining. We select a set of promising methods that take different approaches to handling missing data and test their robustness with respect to different scenarios. Our analyses consider different mechanisms and intensities of missing data under artificial censoring as well as varying visit intensities. We hereby analyze not only the applicability of the selected methods but also provide a systematic approach for parameterization and testing that can also be applied to the analysis of other mobility data sets. Our experiments show that only two of the tested methods supply unbiased estimates of visit potential quantities and are applicable to the domain. In addition, both methods supply unbiased estimates only of a single quantity. Therefore, it will be a future challenge to design methods for the entire collection of visit potential quantities.

The topic of this thesis is motivated by applied research at the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS for business applications in outdoor advertisement. We will use the outdoor advertisement scenario throughout this thesis for demonstration and experimentation.

Zusammenfassung

Täglich interagieren Menschen mit ihrer Umgebung, indem sie sich im geografischen Raum bewegen oder gezielt geografische Orte aufsuchen. Informationen über derartige Besuche sind sehr wertvoll und können in einer Reihe von Anwendungen eingesetzt werden. Üblicherweise werden dazu die Bewegungen von Personen mit Hilfe von GPS, GSM oder RFID Technologien verfolgt. Durch eine räumliche Verschneidung der Trajektorien mit der Positionsangabe eines bestimmten Ortes können dann die Besuche extrahiert werden. Allerdings ist derzeitig die Verwendung von Besuchsinformationen in der Praxis begrenzt. Dies hat, neben ökonomischen und ethischen Gründen, vor allem zwei grundlegende Ursachen. Erstens existiert keine formelle Definition von Größen, um Besuchsinformationen einheitlich auszuwerten. Zweitens können aktuelle Technologien keine vollständige Erfassung von Bewegungsinformationen garantieren. Das bedeutet, dass die Basisdaten zur Auswertung von Besuchsinformationen grundsätzlich Lücken enthalten. Für eine fehlerfreie Auswertung der Daten müssen diese Lücken adäquat behandelt werden. Allerdings wurde dieses Thema in der bisherigen Data Mining Literatur zur Auswertung von Bewegungsdaten vernachlässigt. Daher widmet sich diese Dissertation der Definition von Größen zur Auswertung von Besuchsinformationen sowie dem Schätzen dieser Größen aus unvollständigen Bewegungsdaten.

Im ersten Teil der Dissertation wird eine anwendungsunabhängige Beschreibungssprache formuliert, um Besuchsinformationen auszuwerten. Auf Basis einer einheitlichen Notation wird eine Familie von Größen namens *visit potential* definiert, die grundlegende Besuchsgrößen enthält und offen für Erweiterungen ist. Die gemeinsame Basis aller Besuchsgrößen erlaubt weiterhin, Beziehungen zwischen verschiedenen Größen zu analysieren sowie Konsistenzanforderungen zwischen ähnlichen Parametrisierungen der Größen abzuleiten. Abschließend zeigt die Arbeit die generelle Anwendbarkeit der definierten Besuchsgrößen in zwei realen Anwendungen, für die eine präzise Definition der eingesetzten Statistiken mit Hilfe der Besuchsgrößen gegeben wird.

Der zweite Teil der Dissertation enthält die erste systematische Methodenanalyse für die Handhabung von unvollständigen Bewegungsdaten. Hierfür werden vier vielversprechende Methoden aus unterschiedlichen Bereichen zur Behandlung von fehlenden Daten ausgewählt und auf ihre Robustheit unter verschiedenen Annahmen getestet. Mit Hilfe einer künstlichen Zensur werden verschiedene Mechanismen und Grade von fehlenden Daten untersucht. Außerdem wird die Robustheit der Methoden für verschieden hohe Besuchsniveaus betrachtet. Die durchgeführten Experimente geben dabei nicht nur Auskunft über die Anwendbarkeit der getesteten Methoden, sondern stellen auch ein systematisches Vorgehen für das Testen und Parametrisieren weiterer Methoden zur Verfügung. Die Ergebnisse der Experimente belegen, dass nur zwei der vier ausgewählten Methoden für die Schätzung von Besuchsgrößen geeignet sind. Beide Methoden liefern jedoch nur für jeweils eine Besuchsgröße erwartungstreue Schätzwerte. Daher besteht eine zukünftige Herausforderung darin, Schätzmethoden für die Gesamtheit an Besuchsgrößen zu entwickeln.

Diese Arbeit ist durch anwendungsorientierte Forschung am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS im Bereich der Außenwerbung motiviert. Das Außenwerbeszenario sowie die darüber zur Verfügung gestellten Anwendungsdaten werden durchgängig zur Demonstration und für die Experimente in der Arbeit eingesetzt.

Acknowledgement

I sincerely thank Prof. Dr. Stefan Wrobel for his commitment during the supervision of my thesis. I am grateful for the many challenging questions and helpful suggestions that led me through the past years.

I am grateful to Dr. Michael May for many fruitful discussions and for the opportunity to study the scientific as well as the practical aspects of missing data in mobility mining. I am also thankful for the opportunity to participate in the FP6-IST/FET project GeoPKDD which allowed me to get into close contact with the European research community of mobility mining.

I especially thank our project partners Swiss Poster Research Plus (SPR+) in Switzerland and the Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma) in Germany for their approval to employ the project data within this thesis and for providing comprehensive material about the application domain. In addition, I am indebted to the British Trust for Ornithology (BTO), the Royal Society for the Protection of Birds (RSPB) and BirdWatch Ireland for their permission to employ the BirdTrack scenario and figures of BirdTrack in this thesis.

I sincerely thank Dr. Angelika Voß, Helga Körner, Dr. Kristian Kersting, Stephan Polster, Dr. Thomas Gärtner and Thomas Liebig for proof-reading this thesis in excerpts or in full length. Your comments have been invaluable!

I am grateful to my colleagues at Fraunhofer IAIS for always being open to discuss questions and for the pleasant research atmosphere. In particular I thank my colleagues of the Mobility Mining working group and Dirk Hecker for providing a “geographic newbie” with all information necessary to survive in GIS and for the constructive joint work during the past years. A special thank goes to Daniela Börner for many diverting pauses to breath and for becoming a true friend.

Finally, I am deeply grateful to my family and closely related friends, especially to my sister Henrike Körner, Thea Müller and my fiancé Thomas Kopp, for their constant support and care.

Chacun de nous a connu les joies les plus chaudes
là où rien ne les promettait.

All of us have had the experience of a sudden joy
that came when nothing in the world had forewarned us of its coming.

(Antoine de Saint-Exupéry, Terre des Hommes / Wind, Sand and Stars)

Contents

Abbreviations

aace	average absolute compound error
BC	Before Christ
BIPM	International Bureau of Weights and Measures
BTO	British Trust for Ornithology
CATI	computer assisted telephone interview
CDMAR	covariate-dependent missing at random
e.g.	exempli gratia (Latin, for example)
EM	Expectation Maximization
GeoPKDD	Geographic Privacy-aware Knowledge Discovery and Delivery
GIS	geographic information systems
GLM	general location model
GPS	Global Positioning System
GRP	gross rating points
GSM	Global System for Mobile Communications
GWR	geographically weighted regression
HMS	horizontal mean substitution
i.e.	id est (Latin, that is)
IERS	International Earth Rotation Service
iid	independent identically distributed
ILP	inductive logic programming
ISO	International Organization for Standardization
KM	Kaplan-Meier estimation
LBS	location based services
MAR	missing at random
MBR	minimum bounding rectangle
MCAR	missing completely at random
me	mean error
MI-GLM	multiple imputation via general location model
MI-Poisson	multiple imputation from a conditional Poisson distribution
MNAR	missing not at random
ML	maximum likelihood
OGC	Open Geospatial Consortium
OTS	opportunities to see
POI	points of interest
REMO	relative motion
RFID	radio frequency identification
rme	relative mean error
rmse	root mean squared error
RSPB	Royal Society for the Protection of Birds
SI	International System of Units
SI-SVR	single imputation via support vector regression
SVM	support vector machine
SVR	support vector regression

TAS	temporally annotated sequence
UTC	Coordinated Universal Time
VMS	vertical mean substitution
vs.	versus
WGS84	World Geodetic System of 1984

1. Introduction

The important thing is not to stop questioning.

(Albert Einstein)

1.1. Motivation

Personal mobility is one of the greatest achievements of the past century. Mass production in the automobile industry along with the provision of public transportation and infrastructure have led to a multitude of daily travel activities in commercial and private life. In Germany about 49.3 million licensed motor vehicles were registered in 2008 (Destatis, 2009), contributing to an average daily travel distance of 41 kilometers per person (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010). Thus, personal mobility has become one key component of daily life. In consequence, governmental as well as private organizations take an interest in mobile behavior in order to regulate and improve mobility as well as activities that are closely connected with it. A number of applications require, for example, information about the number of people that pass certain locations or the distribution of recurrent visitors. Such quantities, that base on the visits of persons - or more generally mobile entities - to locations in geographic space, require the evaluation of movement histories. Modern tracking technologies such as the Global Positioning System (GPS), Global System for Mobile Communications (GSM) or Radio Frequency Identification (RFID) seem able to provide all necessary information for such quantities. However, on closer consideration these technologies have one common drawback: they cannot guarantee complete observation. For example, consider a GPS survey, which is at present the predominant method to collect personal mobility data. For the survey, a group of test persons is recruited to carry GPS devices over a specified period of time. During this time persons easily forget to carry the device, to charge its battery or may even switch off the device on purpose, which leads to gaps in the data. Furthermore, technical defects of devices or weariness of test persons result in early dropouts of the study. Similar situations arise when GSM or RFID technology is applied. Thus, the underlying data from which visit quantities could be derived, contain inherently missing data. However, there exists no publication to-date which systematically analyzes techniques for the handling of missing movement data. This thesis therefore focuses on the estimation of quantities about the visiting behavior between mobile entities and geographic locations from incomplete mobility data.

1.2. Scientific Question, Research Goals and Challenges

The central question of this thesis is: How can quantities about entity-location visits be estimated from incomplete mobility data for applications under real-world conditions?

This question assumes that a general definition for the *measuring* of entity-location interactions exists. However, this is not the case. Although a number of companies and research institutions apply entity-location interaction quantities in their day-to-day business, these quantities are not generally defined. Typically, quantities are tailored to specific applications, use context-dependent terminology and are often only informally written down as, for example, in the environment of outdoor advertising. As a result, a number of quantities have evolved which are not suitable for methodological research and interdisciplinary exchange as their common background is hard to identify. The first goal of this thesis is therefore to provide a formalization and common notation of entity-location interaction quantities.

The second goal is to analyze and evaluate methods and algorithms for the estimation of the defined quantities given incomplete mobility data. Existing research on trajectory data mining neglects the treatment of incomplete data so far. It addresses predominately the analysis of mobility patterns as, for example, the clustering of (parts of) trajectories (Rinzivillo et al., 2008a; Pelekis et al., 2007; Nanni and Pedreschi, 2006), the detection of relative motion patterns (Gudmundsson et al., 2007; Hwang et al., 2005; Laube and Imfeld, 2002) or the sequential analysis of movement (Zheng et al., 2009; Giannotti et al., 2007; Yang and Hu, 2006). For further discussion of related work see Section 4.1.2.

Both research goals are challenging. Concerning the first goal, we seek a precise definition of visit quantities in terms of geographic and mathematic concepts. One goal hereby is to derive all quantities from a uniform foundation, which will explain relationships between the defined quantities. Further, the definitions must be able to express a variety of parameterizations, however, they shall not be unnecessary complex. Finally, the quantities shall be extensible for future application demands. Concerning the second goal, incomplete mobility data are challenging for the estimation of entity-location interaction quantities, because the quantities relate to the interactions over a specified period of time. If missing data are ignored, i.e. missing measurement periods are treated as immobility, the quantities will be underestimated. Other general approaches to handle missing data as the exclusion of incomplete data records or the estimation of missing values are also not truly applicable to mobility data. First, mobility data are expensive and the portion of incomplete records is typically significant. A removal of test persons with incomplete data records would lead to a significant reduction of data and may render further inference impossible. Second, the estimation of missing values from the distribution of available measurements is not obvious for mobility data as it implies the reconstruction of individual trajectories for the missing measurement periods. The handling of missing data may additionally be complicated by the fact that the absence of data relates to the mobility behavior of an entity. In statistics, such a dependence is known as *missing at random (MAR)* or *missing not at random (MNAR)*, depending on further criteria. Both cases require an explicit treatment in order to obtain unbiased results (Schafer and Graham, 2002). Finally, geographic and mobility data differ from other domains as the data is subject to spatial autocorrelation and dependencies, which may cause further difficulties.

1.3. Contribution

This thesis contributes to the field of computer-supported mobility analysis by a) providing a formal definition of entity-location interaction quantities and by b) analyzing methods and algorithms for the estimation of such quantities under incomplete mobility data.

First, we provide an application-independent language to evaluate entity-location interactions. Based on a uniform notation, we define a family of quantities called *visit potential*, which contains the most basic interaction quantities and can be extended on need. By identifying the common background of all quantities we are able to analyze relationships between different quantities and to infer consistency requirements when quantities are used with related parameterizations. We demonstrate the general applicability of visit potential using two real-world applications for which we give a precise definition of the employed entity-location interaction quantities in terms of visit potential. We hope that our formalization contributes to the interdisciplinary exchange of challenges and ideas in the area of entity-location interactions and stimulates further research.

Second, this thesis provides the first systematic analysis of methods for the handling of missing data in mobility mining. We select a set of promising methods that take different approaches to handle missing data and test their robustness with respect to different scenarios. Our analyses consider different mechanisms and intensities of missing data under artificial censoring as well as varying visit intensities. We hereby analyze not only the applicability of the selected methods but provide as well a systematic approach for parameterization and testing that can be applied also for the analysis of other mobility data sets. We conduct our experiments on a data set and application scenario from outdoor advertising.

Our experiments show that only two of the tested methods supply unbiased estimates for visit potential quantities and are applicable to the domain. These two methods, namely multiple imputation from a conditional Poisson distribution (MI-Poisson) and Kaplan-Meier (KM), are comparably simple, however, they have the advantage that they are designed for event data. Thus, they can be applied to the evaluation of visit potential without any additional data transformation. Further, both methods performed robustly under the induction of missing data under MAR. In the case of MI-Poisson we obtained unbiased results even without conditioning. However, both methods supply unbiased estimates only for a single quantity. Therefore, it will be a future challenge to design methods for the entire collection of visit potential quantities.

1.4. Application Scenario

The following scenario illustrates the importance of entity-location interactions in a prominent business application: outdoor advertisement. It generates yearly net sales of about 760 million Euro in Germany (Fachverband Außenwerbung e.V., 2011) and is therefore of high economic impact. Outdoor advertising, also called out-of-home advertising, is essentially any type of advertising that reaches a person outside of his or her home (Wikipedia, 2010a). In outdoor advertisement the pricing of poster sites is a business-critical task and must therefore be justified by objective performance indicators. This means - in simple words - that the more people pass a poster site, the higher is the price that a vendor can ask for rent of the site. Performance indicators in outdoor advertising are one example of entity-location interaction quantities: people represent mobile entities and poster sites represent geographic locations. An interaction takes place when a person passes a poster site. However, our results are not limited to outdoor advertising. For example, in Section 4.4 we show the applicability of visit potential quantities for the evaluation of bird recordings.

In this section we will introduce the outdoor advertising application scenario and illustrate with it the motivating tasks of this thesis. The presented application is closely connected to this thesis as it provides the data basis for our experiments, poses research questions, defines preconditions and supports modeling decisions. However, our results are not limited to outdoor advertising. They belong to the broader area of mobility data analysis and may be applied in similar contexts with potentially missing mobility information, such as the evaluation of travel diaries or GSM data.

Although the outdoor advertising industry has made a striking advancement during the past five years with respect to measurement techniques and the analysis of mobility data for poster evaluation, the business sector is also a typical example for the lack of precise definitions and the use of context-dependent terminology. First, existing definitions for performance indicators are typically designed to apply to various media types as print, radio, television or poster. Therefore, formulations are very general and do not reflect characteristics that apply in particular to poster advertisement. Second, the definitions use marketing-specific terms and are therefore hard to understand in other disciplines. The following examples shall illustrate this situation. They are definitions of the three leading performance indicators *gross rating points (GRP)*, *reach* and *opportunities to see (OTS)* as given by established Internet sources:

"[GRP is the] total of all rating points for an advertising schedule stated usually on a weekly basis." (WebFinance Inc., 2010)

"Reach refers to the total number of different people or households exposed, at least once, to a medium during a given period of time." (Wikipedia, 2010b)

"'Average OTS' is a measure of the number of chances an average member of the target audience will have of being exposed to an advertisement in an advertising campaign." (Westburn Publishers Ltd., 2010)

None of the definitions makes a reference to geographic or mobility data in order to clarify what "being exposed" to a campaign means. The involved population and poster sites are only informally mentioned - if they are mentioned at all - using application-specific terms such as "target audience" or "advertising schedule". Finally, other context-specific terms as "rating points" are used within the definitions, which are not generally known outside the advertising business sector. It is therefore difficult to understand poster performance indicators and to investigate them from a scientific point of view. In addition, the exchange with similar scenarios in other fields of application is restrained. In this thesis we provide a systematic formalization of entity-location interaction quantities and show how performance indicators in outdoor advertising can be precisely defined using these quantities.

A second complication in the estimation of poster performance indicators are incomplete measurement data. The outdoor advertising companies in Switzerland and Germany are pioneers in the usage of mobility data. Both of them conducted large GPS surveys with more than 10.000 participants to collect a representative sample of mobile behavior (Pasquier et al., 2008; Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma), 2009). However, both studies face the problem of missing measurements.

Missing data occur for several reasons and last different periods of time. Short interruptions are typically caused by loss of signal. Parts of the day are missing if the device is left at home for a single trip, for example, when going to the bakery. Finally, complete measurement days are missing if people forget to carry the device for a whole day or do not charge it previously. In addition, people may tire of the study and drop out early or the GPS device may be defective, which both results in missing measurements for all subsequent days. In this thesis we consider only missing data of complete measurement days. Short interruptions are typically treated during data preprocessing, and single missing trips can practically not be detected from the data. The only possibility to distinguish immobility from missing mobility information are follow-up interviews directly after the surveying period, which typically give evidence on a daily basis.

Missing measurement days are a serious problem in the application. They occur frequently in the data and cannot be ignored as performance indicators are defined over a continuous period of time. Figure 1.1 shows the distribution of valid measurement days for the German GPS mobility survey, which was conducted over seven days per person. Clearly, the removal of test persons with less than seven measurement days is not an option as it would reduce the data sample to one third of its original size. Also, the estimation of missing values from the distribution of available measurements is not feasible for mobility data, as it implies the reconstruction of individual trajectories for the missing measurement days. The last option

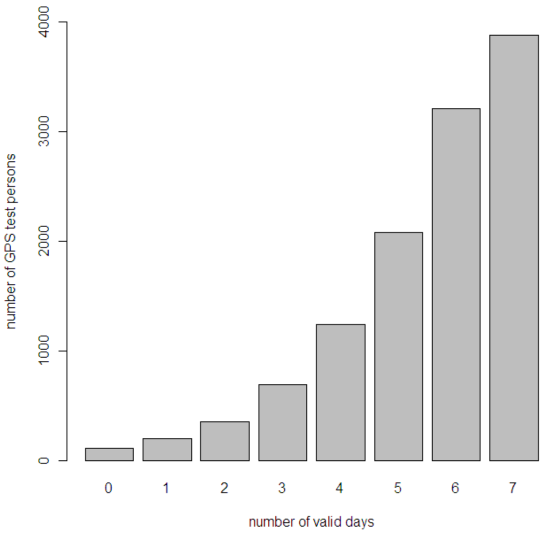


Figure 1.1.: Distribution of valid measurement days in the German GPS mobility survey. The survey was conducted over a period of seven days per participant.

is to treat missing data explicitly in the modeling step, which is the chosen approach in this thesis.

However, modeling may be complicated by a possible relationship between the absence of data and the mobile behavior of a person. It is well known that different groups of the population possess different mobile behaviors. For example, people between 30 and 39 years show with an average of 53 kilometers per day the highest mobility while teenagers travel around 30 kilometers per day and people above 74 years travel only 16 kilometers on average (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010). If certain characteristics of such groups relate to the intensity of missing data, for example, elder persons may be more reliable to carry the devices than teenagers, the pattern of missing data is not any more at random. Such a relationship violates the precondition for a large number of modeling techniques and may lead to biased results. In this thesis we therefore also assess the robustness of estimation methods with respect to various patterns and types of missing data.

1.5. Outline

The chapters of this thesis are arranged in consecutive order according to our main goals. We first lay the foundation of geographic data analysis and introduce the application context. Next, we formalize quantities for entity-location interaction. Subsequently, we analyze methods for the estimation of entity-location interaction quantities under missing data. The thesis concludes with a summary and an outlook on future work. More precisely, the chapters cover the following topics.

Chapter 2 gives an introduction to geographic data, presenting typical characteristics and methods of data handling. The chapter begins with data models, data structures, feature extraction methods and specialized analysis methods for spatial data, and then proceeds to mobility data.

Chapter 3 describes the application context that motivates the topic of this thesis. It defines performance indicators in outdoor advertising and introduces the audience measurement studies in Switzerland and Germany. The chapter concludes with a list of research challenges that arise within the audience measurement studies.

Chapter 4 formalizes visit potential quantities for entity-location interactions. It provides a systematic definition of entity-location interaction quantities and analyzes the relationships of the quantities to each other. In addition, the chapter demonstrates, based on two examples, how application-dependent entity-location interactions can be precisely defined using visit potential.

Chapter 5 addresses the estimation of visit potential quantities under missing data. It evaluates the usability and practicability of general estimation methods for missing data in the mobility domain and tests the robustness of a selection of estimation methods for particular visit potential quantities.

Chapter 6 gives a summary and outlook on future work and concludes the thesis.

1.6. Publications

The main contributions of this thesis have already been published in the following conference and workshop publications. All papers contain a significant contribution from the author of this thesis.

- C. Körner, D. Hecker, M. May, and S. Wrobel. Visit potential: A common vocabulary for the analysis of entity-location interactions in mobility applications. In M. Painho, M. Y. Santos, and H. Pundt, editors, *Geospatial Thinking*, Lecture Notes in Geoinformation and Cartography, pages 79-95. Springer, 2010.
- M. May, C. Körner, D. Hecker, M. Pasquier, U. Hofmann, and F. Mende. Handling missing values in GPS surveys using survival analysis: a GPS case study of outdoor advertising. In *ADKDD '09: Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 78-84. ACM, 2009.
- D. Hecker, H. Stange, C. Körner, and M. May. Sample bias due to missing data in mobility surveys. In *Proc. of the 2010 IEEE International Conference on Data Mining Workshops (ICDMW'10)*, pages 241-248. IEEE Computer Society, 2010.
- D. Hecker, C. Körner, and M. May. Räumlich differenzierte Reichweiten für die Außenwerbung. In *J. Strobl, T. Blaschke, and G. Griesebner, editors, Angewandte Geoinformatik 2010, Beiträge zum 22. Symposium für Angewandte Geoinformatik (AGIT) Salzburg*, pages 194-203. Wichmann, 2010.
- B. Guc, M. May, Y. Saygin, and C. Körner. Semantic annotation of GPS trajectories. In *Proc. of the 11th AGILE International Conference on Geographic Information Science (AGILE'08)*, 2008.

The paper of **Körner et al. (2010b)** was selected **second-best paper at Agile 2010**.

Chapters 2 and 3 are based on the author's contribution to the following book chapters.

- C. Körner, D. Hecker, M. Krause-Traudes, M. May, S. Scheider, D. Schulz, H. Stange, and S. Wrobel. Spatial data mining in practice: Principles and case studies. In C. Soares and R. Ghani, editors, *Data Mining for Business Applications.* IOS Press, 2010.

- S. Rinzivillo, F. Turini, V. Bogorny, C. Körner, B. Kuijpers, and M. May. Knowledge discovery from geographical data. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 9. Springer, Berlin Heidelberg, 2008.

- M. Nanni, B. Kuijpers, C. Körner, M. May, and D. Pedreschi. Spatiotemporal data mining. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 10. Springer, Berlin Heidelberg, 2008.

Additional material appeared in the following conference and workshop papers.

- D. Hecker, C Körner, and M. May. Robustness analyses for repeated mobility surveys in outdoor advertising. In *Proc. of the IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM'11)*, pages 148-153, 2011.

- D. Hecker, C Körner, and M. May. Challenges and advantages of using GPS data in outdoor advertisement. In *Proc. of the 3th Conference on Geoinformatik - Geochange*, pages 257-260. Akademische Verlagsgesellschaft, 2011.

- D. Hecker, C. Körner, H. Stange, D. Schulz, and M. May. Modeling micro-movement variability in mobility studies. In S. Geertman, W. Reinhardt, and F. Toppen, editors, *Advancing Geoinformation Science for a Changing World*, Lecture Notes in Geoinformation and Cartography, pages 121-140. Springer, 2011.

- D. Hecker, C. Körner, H. Streich, and U. Hofmann. A sensitivity analysis for the selection of business critical geodata in Swiss outdoor advertisement. In *GIScience 2010, Extended Abstracts Volume*, pages 194-203, 2010.

- T. Liebig, H. Stange, D. Hecker, M. May, C. Körner, and U. Hofmann. A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In *Proc. of the Third Workshop on Pervasive Advertising and Shopping*, 2010.

- T. Liebig, C. Körner, and M. May. Fast visual trajectory analysis using spatial Bayesian networks. In *ICDM Workshops*, pages 668-673. IEEE Computer Society, 2009.

- M. May, C. Körner, D. Hecker, M. Pasquier, Urs Hofmann, and Felix Mende. Modelling missing values for audience measurement in outdoor advertising using GPS data. In *GI Jahrestagung*, volume 154 of LNI, pages 3993-4006. GI, 2009.

- T. Liebig, C. Körner, and M. May. Scalable sparse Bayesian network learning for spatial applications. In *ICDM Workshops*, pages 420-425. IEEE Computer Society, 2008.

- M. May, D. Hecker, C. Körner, S. Scheider, and D. Schulz. A vector-geometry based spatial kNN-algorithm for traffic frequency predictions. In *Proc. of the 2008 IEEE International Conference on Data Mining Workshops (ICDMW '08)*, pages 442-447. IEEE Computer Society, 2008.

- M. Pasquier, U. Hofmann, F. H. Mende, M. May, D. Hecker, and C. Körner. Modelling and prospects of the audience measurement for outdoor advertising based on data collection using GPS devices (electronic passive measurement system). In *Proceedings of the 8th International Conference on Survey Methods in Transport*, 2008.

- D. Wegener, D. Hecker, C. Körner, M. May, and M. Mock. Parallelization of R-programs with GridR in a GPS-trajectory mining application. In *Proc. of the 1st Ubiquitous Knowledge Discovery Workshop (UKD'08)*, 2008.

- A. Zanda, C. Körner, F. Giannotti, D. Schulz, and M. May. Clustering of German municipalities based on mobility characteristics: An overview of results. In *Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS'08)*, pages 1-4. ACM, 2008.

- C. Körner and S. Wrobel. Bias-free hypothesis evaluation in multirelational domains. In *Proc. of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, (PAKDD'06)*, pages 668-672. Springer, 2006.

- C. Körner and S. Wrobel. Multi-class ensemble-based active learning. In *Proc. of the 17th European Conference on Machine Learning (ECML'06)*, pages 687-694. Springer, 2006.

The paper Zanda et al. (2008) received a best poster award at ACM GIS 2008.

2. Introduction to Geographic Data and Data Analysis

From henceforth, space by itself, and time by itself, have vanished into the merest shadows and only a kind of blend of the two exists in its own right.

(Herman Minkowski)

In this thesis we formalize quantities to measure the degree of spatiotemporal interaction of mobile entities and geographic locations and analyze methods for the estimation of these quantities. Hence, geographic and mobility data are the basic data types that are inseparably connected with all our activities. This chapter therefore provides an introduction to the nature, handling and analysis of geographic and mobility data. It is primarily intended for readers that are not familiar with geographic information science and mobility data. Sections 2.1 and 2.2 are based on each other in sequential order, however, they assume no prior knowledge of the named areas. Advanced readers may read these sections in excerpts or refer to them for definitions used in subsequent chapters.

In more detail the chapter is organized as follows. Section 2.1 contains a basic introduction to spatial data. It begins with positioning of objects in physical space and then proceeds to geographic data models and data structures, characteristics of geographic data and spatial feature extraction. It concludes with analysis tasks and methods for spatial data. Section 2.2 focuses on mobility data, i.e. a type of spatiotemporal data that provides information about the movement of objects and individuals. The section starts with a definition of movement, then presents trajectory data models and data structures and describes characteristics of mobility data. It further presents feature extraction, preprocessing and annotation methods as well as analysis tasks and methods for trajectory data. We conclude the chapter with a summary highlighting the most important concepts for this thesis.

Parts of this chapter are based on the author's contribution to (Guc et al., 2008) as well as to (Rinzivillo et al., 2008b) and (Nanni et al., 2008). The latter two have been published in the context of the EU-project *Geographic Privacy-aware Knowledge Discovery and Delivery*[1] (GeoPKDD).

2.1. Spatial Data

2.1.1. Positioning in Space

A georeference states the position of an object in physical space based on some reference system. In order to create a georeference we must therefore first agree on the structure of physical space and on a reference system. In the second step we can build the actual reference using direct or indirect positioning.

[1] http://www.geopkdd.eu

Physical space. The general theory on relativity presented by Albert Einstein in 1915 shapes our knowledge about physical space to date. The theory describes the influence of gravity upon space and time. In other words, it allows for a contortion of space by large masses. Einstein's theory generalizes classical mechanics, which perceives space as homogeneous and isotropic extent in three dimensions as proposed by Isaac Newton (Fritzsch, 2000). However, classical mechanics are sufficiently accurate when considering slow-moving objects in weak gravitational fields (Strobel, 2007). In this thesis we will therefore regard space in the classical meaning as three-dimensional Euclidean space.

Spatial reference systems. Several types of reference systems are used in geographic sciences to identify uniquely the position of a point in space. As we are interested in the position of objects relative to the surface of the Earth, we will focus on so-called terrestrial reference systems. Terrestrial reference systems anchor coordinate systems in such a way that they follow the daily rotation of the Earth and its annual circling around the sun. As a result, the position of points attached to the solid surface of the Earth varies only due to geophysical effects (e.g. motion of tectonic plates) and is small over time (McCarthy and Petit, 2004). The two predominately used coordinate systems in geographic sciences are Cartesian coordinates and geographic coordinates.

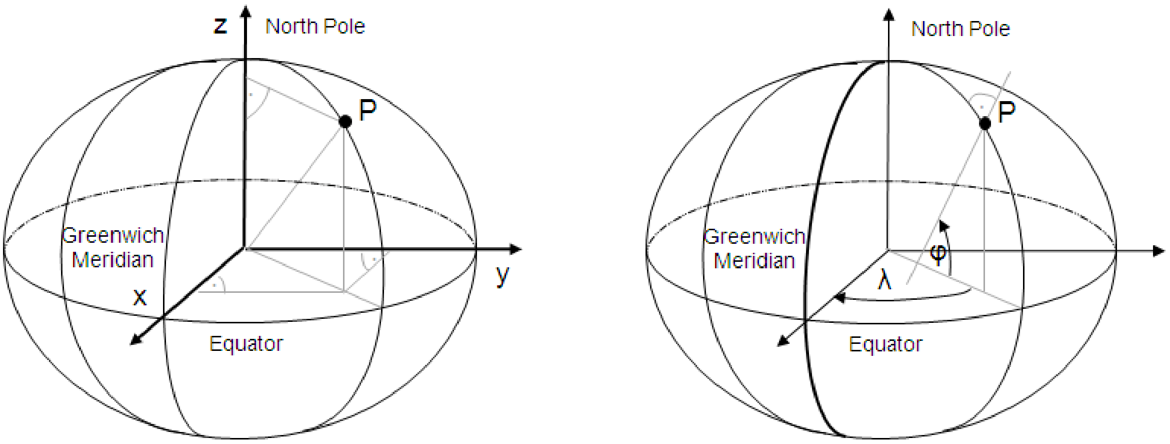
Definition 2.1.1 (Cartesian coordinate system) *A Cartesian coordinate system in three-dimensional space consists of three to each other orthogonal axes, usually named x, y, and z axis, which possess the same unit of length. The intersection of x, y, and z axis is called origin (Bronstein et al., 2001).*

In order to form a terrestrial reference systems, Cartesian coordinates are anchored such that their origin coincides with the Earth's center of mass. The z axis commonly equals the Earth's axis of rotation pointing to the North Pole, the x axis lies in the plane of the Greenwich Meridian perpendicular to the z axis, and the y axis lies in the equatorial plane forming a right-handed coordinate system (see Figure 2.1 left). Since 1988, the International Earth Rotation Service (IERS) is responsible for the definition, implementation and maintenance of such a reference system, called International Terrestrial Reference System (ITRS), for use in geographic sciences (Seeger, 1999; McCarthy and Petit, 2004).

While Cartesian coordinate systems are able to specify the position of any point of the Earth, Geographic coordinate systems define only points on the surface of the Earth. Hereby, the surface of Earth is assumed to approximate a flattened sphere. Although this is a simplification, it is a reasonable one. On globals scale, the surface of Earth is much smoother than we perceive locally. If we scaled Earth to about 25 cm in diameter, its highest peak, Mount Everest, would result in a rise of 0.176 mm only (Robinson et al., 1995a). In addition, the specification is convenient and sufficient for many applications. The flattening results from stronger centrifugal forces at the equator than toward the poles due to the rotation of Earth (Robinson et al., 1995a).

Definition 2.1.2 (Geographic coordinate system) *A geographic coordinate system specifies the position of points on the surface of the Earth based on a spherical polar coordinate system using the two angles longitude (λ) and latitude (φ) (Bollmann and Koch, 2001).*

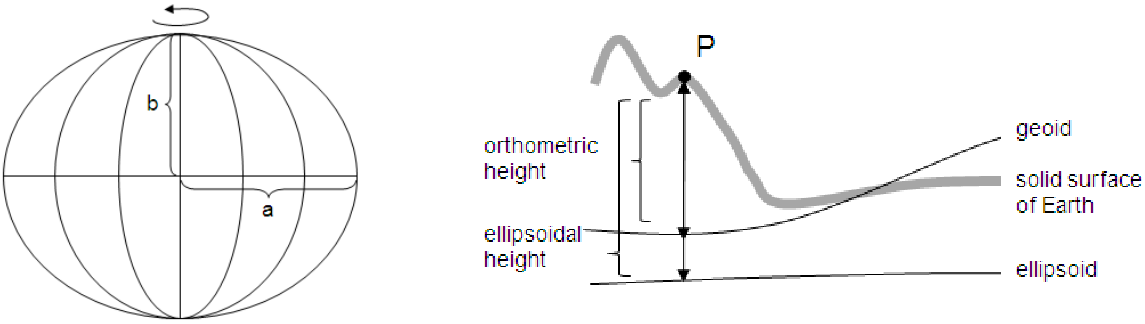
Note that geographic coordinate systems describe a two-dimensional non-Euclidean space, namely the surface of an ellipsoid of rotation. An ellipsoid of rotation is an ellipse that rotates - in this case - around its minor axis. The semi-minor axis b of the ellipse corresponds to the polar radius and the semi-major axis a corresponds to the equatorial radius of Earth,



(a) Cartesian coordinate system (b) Geographic coordinate system

Figure 2.1.: Spatial reference systems

both approximating mean sea level and intersecting at Earth's center of mass (see Figure 2.2 left). Today, a widely accepted ellipsoid is the World Geodetic System of 1984 (WGS84), however, a number of ellipsoids have been defined and are used for georeferencing in different parts of the world. For a reference list of ellipsoids see Longley et al. (2001b) or Robinson et al. (1995a). Given an ellipsoid of rotation, we can specify the east-west and north-south position of any point P on the surface of Earth using longitude and latitude, respectively. The longitude λ of some point P on Earth is the angle between the Greenwich Meridian plane and the meridian plane of P, where the meridian plane of P is the plane perpendicular to the equator which contains P. The latitude φ of P is the angle between the equator and the normal at P with respect to the surface of the ellipsoid (see Figure 2.1 right). Longitude and latitude are measured in degrees and range between 180°E (+180°) and 180°W (−180°) for longitude and between 90°N (+90°) and 90°S (−90°) for latitude (Robinson et al., 1995a). Note that geographic coordinates require the application of spherical geometry, for example, when calculating distances (see also Section 2.1.4, distance calculations).



(a) (b)

Figure 2.2.: (a) Ellipsoid of rotation and (b) ellipsoidal and orthometric height

If the specification of horizontal position (longitude, latitude) is not sufficiently accurate, geographic coordinates can be complemented with height. The height h of some point P is given with respect to a defined reference surface, which is typically either the ellipsoid (as defined previously) or the geoid. The geoid represents the equipotential surface of the

Earth's gravity field at mean sea level. This means, gravitational forces are equal for any point on the geoid and act perpendicular to its surface (see Figure 2.2 right). The geoid differs from an ellipsoidal surface because variations in rock density and topographic relief influence gravitational forces (Robinson et al., 1995a). Given a reference surface, we can define height h as the distance between some point P and its orthogonal projection on the reference surface equipped with a positive (negative) sign if P lies outside (inside) of the reference surface. If an ellipsoidal surface is used the height is termed ellipsoidal height, if a geoid reference surface is used it is called orthometric height (Seeger, 1999).

In summary, terrestrial reference systems allow to specify the position of any point (on the surface) of Earth using a given coordinate system. The relationship between the real world and the coordinate system is fixed by a set of parameters, e.g. the coordinate systems's origin, scale and orientation, which are also called a datum (Lott, 2004). In order to distinguish real world and coordinate space and to disambiguate them from other mathematical spaces, we will use the terms *geographic space* and *geographic coordinate space*, respectively. More formally, we define these terms as follows.

Definition 2.1.3 (Geographic space) *Geographic space $\mathcal{S}$ is the three-dimensional Euclidean space co-rotating with Earth and centered at its center of mass, i.e. the physical space that we observe in daily life.*

Definition 2.1.4 (Geographic coordinate space) *Geographic coordinate space $\mathcal{S_C}$ is the one-, two- or three-dimensional coordinate space of a terrestrial reference system used to specify a position in geographic space. Geographic coordinate space has the form*

- $\mathcal{S_C} = \mathbb{R}^3$ *in case of three-dimensional Cartesian coordinates,*
- $\mathcal{S_C} = [-180°, 180°] \times [-90°, 90°]$ *in case of two-dimensional geographic coordinates,*
- $\mathcal{S_C} = [-180°, 180°] \times [-90°, 90°] \times \mathbb{R}$ *in case of two-dimensional geographic coordinates supplemented with height.*

Obviously, geographic references of different referencing systems can be converted if the complete specifications of the reference systems are available. Mathematically we can interpret geographic coordinate space as a set of expressions, each of which can be evaluated to a point in geographic space by some function r_S, i.e. $r_S : \mathcal{S_C} \to \mathcal{S}$.

Direct and indirect positioning. Geographers distinguish two types of references for the specification of the position of some object: direct and indirect. Direct positioning refers to a coordinate specification of some object using a given terrestrial reference system. Indirect positioning identifies the position of some object by relating it to the (commonly-known) position of some other object. Typical examples for indirect positioning are postal addresses, place names or cadasters (Longley et al., 2001b). Note that objects may also possess an internal spatial structure. Andrienko et al. (2008) therefore differentiate indirect positioning further into division-based and linear referencing. The former refers to a possibly hierarchical division of space based on geometric or semantic criteria (e.g. the hierarchy: country, state, municipality). The latter specifies positions in relation to linear objects (e.g. a street name and a house number to indicate the position along the street).

In this thesis we will always assume that direct references are used to specify the position of some object. More specifically, we will use the term *geographic location* or short *location* to refer to the position of an object within this thesis.

Definition 2.1.5 *(Geographic location or location) Given a geographic coordinate space $\mathcal{S_C}$, a geographic location, or short location, is a non-empty subset $l \subseteq \mathcal{S_C}$.*

Note that we do not impose requirements on the shape of a geographic location (e.g. connectedness) as the definition is mathematically sufficient. However, the complexity of geographic locations is certainly limited in practice.

A closely related topic to positioning is the projection of geographic coordinates into two-dimensional Euclidean space. The most well-known application of such a projection is a paper map. However, map projections and related topics are only loosely connected to the central theme of this thesis and the interested readers are therefore referred to Robinson et al. (1995b) or Longley et al. (2001b) for further details.

Being now able to specify the position of some point in space, we will introduce spatial data models and data structures in the next section.

2.1.2. Geographic Data Models and Data Structures

Conceptual models of geographic space. Geographic data are data about objects or phenomena that are associated with a location relative to the surface of the Earth. On the conceptual level, two paradigms exist for the perception of geographic space: fields and objects (Burrough and McDonnell, 2000; Haining, 2003; Longley et al., 2001a). The former regards the spatial domain as a continuous surface with smooth variation of some attribute values. It represents a function of location in two- or three-dimensional space. Typical examples of such attributes are temperature, mineral or pollutant concentrations. The latter conceptual model identifies discrete objects in space, which possess a well-defined location, shape and other attributes. Examples of entities are trees, streets or cities.

Basic data structures. The two basic data structures to store geographic data are raster and vector (Burrough and McDonnell, 2000; Rigaux et al., 2001). A raster is a regular grid that consists, for example, of square or hexagonal cells (see Figure 2.3). Each cell contains a single value of a given attribute. Thus, all variation within a cell is lost and the size of the cells defines the level of resolution. Given recurring values in neighboring cells, a number of methods can be applied to achieve a compact representation of raster data. Among them are chain codes, run-length codes, block codes and quadtrees (Burrough and McDonnell, 2000). In general, raster structures possess the disadvantage that a high resolution results in large data volumes and requires in consequence also long computation times for spatial operations. The advantage of raster structures lies in their regularity and simplicity, which allows easy manipulation, filtering, modeling and analysis.

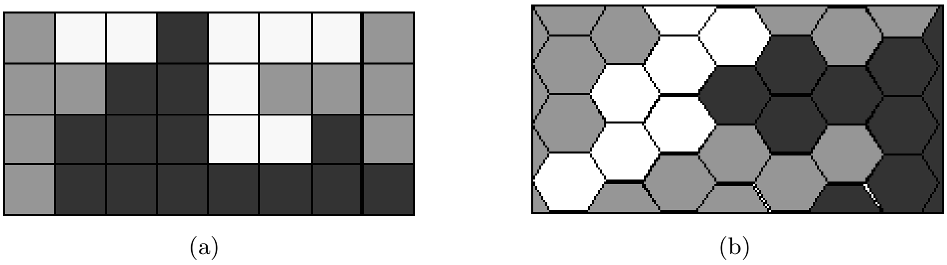

(a) (b)

Figure 2.3.: Raster of (a) square and (b) hexagonal cells

Vectors represent geographic objects according to their topological dimension as points, lines or areas. More precisely, a vector is a tuple of the form (*geometry*, *attributes*) with the geometry specifying the objects location and shape. A point is specified by its rational coordinates. Lines and areas are usually denoted by a sequence of points that are connected by straight lines, which has lead to the alternative terms polyline and polygon (Longley et al., 2001a). Lines and areas can reach different degrees of complexity. For example, a line can intersect with itself, and an area may contain holes or consist of several disconnected parts (see Figure 2.4). A standardization of the specification of geometries is given by the Open Geospatial Consortium (OGC) in the OpenGIS®Implementation Specification for Geographic Information - Simple Feature Access (Herring, 2006). In contrast to the field model, the world of the object model is empty except for places that are occupied by objects. Several possibilities exist to model a collection of vector objects. In the simplest case, the collection is a loose accumulation of objects, without explicit specification of topological relationships. This model is also called the spaghetti model. Relationships between points and lines can be expressed using network models. Finally, relationships between polygons can be specified explicitly using topological models (Rigaux et al., 2001). The advantage of the vector model is the concise representation of objects. However, it involves complex data structures, and the computation of spatial operations, such as intersection and overlay, may take considerable time and resources (Rigaux et al., 2001; Burrough and McDonnell, 2000).

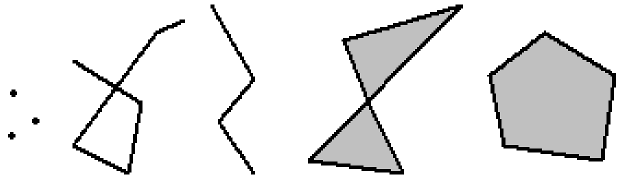


Figure 2.4.: Vector geometries: points, polylines and polygons

Representation of fields and objects using raster and vector. Although clearly related, both of the data structures raster and vector can be used to represent the concept of fields and objects.

Field data has to be discretized before it can be represented in raster or vector form. A partition of two-dimensional space into non-overlapping cells is called a tessellation. Tessellations may be regularly or irregularly spaced. If a regular grid is used (e.g. triangular, square or hexagonal grid), the obtained cells correspond directly to raster cells. In case of irregularly spaced cells (e.g. Voronoi polygons or triangulation networks), the vector model is used to store the geometries. Irregular tessellations possess the advantage that the density of cells can be adapted to the variation within the data (Burrough and McDonnell, 2000).

Objects are naturally represented by using the vector model. Vectors may be used to present single objects or irregular tessellations as, for example, the municipalities within a country. However, the geometry of an object can also be approximated by a raster data structure. For example, a street can be represented by a set of raster cells. Similarly, an area may be approximated by the smallest subset of cells that contains the area (see Figure 2.5). When more than one object intersects with a given cell, the cell is usually assigned to the object with the largest share of the cell's area. Alternatively, the object that covers the central point of the cell can be chosen (Longley et al., 2001a).

2.1.3. Characteristics of Geographic Data

Two predominant characteristics of geographic data are autocorrelation and variation. Both violate assumptions that are essential to traditional data mining techniques and must be taken into account during modeling and reasoning.

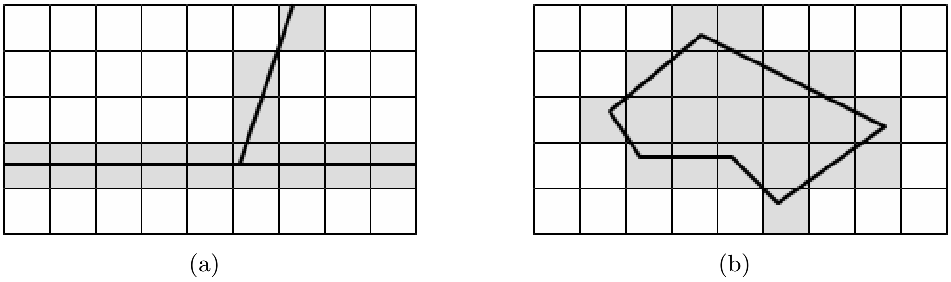


Figure 2.5.: Representation of (a) lines and (b) areas in raster format

Spatial Autocorrelation. Spatial autocorrelation is also known as *Tobler's Law* (Tobler, 1970), which states that "[...] everything is related to everything else, but near things are more related than distant things." It means that attribute values of spatial objects are the stronger correlated the closer two objects are in location. Usually, geographic objects exhibit positive autocorrelation and show similar values within their local neighborhood. This behavior directly contrasts the often made assumption of independent, identical distributions in classical data mining and causes poor performance of algorithms that ignore autocorrelation (Chawla et al., 2001).

Tests for spatial autocorrelation differ significantly from those for time series data. In time series dependence exists only in one direction, namely the past. Spatial autocorrelation, however, extends in all directions. The two most well-known statistics for spatial autocorrelation are *Moran's I* (Moran, 1950) and *Geary's c* (Geary, 1954). They can be applied to binary, ordinary and interval scaled data and are described below. Further details on spatial autocorrelation measures, including join count statistics for nominal scaled data, can be found in (Cliff and Ord, 1973; Upton and Fingleton, 1994).

Given a partition of space into n areas and a variable Z for observation. Let $\{z_i\}\,,\ i = 1..n$ denote the value of the observed variable in area i. Given further a weighting matrix $W \equiv \{w_{ij}\}\,,\ i, j = 1..n$ that states the distance between any two areas with $w_{ij} \geq 0$ and $w_{ii} = 0$. The generalized form of Moran's I and Geary's c statistics (Cliff and Ord, 1973) are then

$$I = \frac{n \displaystyle\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - \overline{z})(z_j - \overline{z})}{\mathcal{W} \displaystyle\sum_{i=1}^{n} (z_i - \overline{z})^2}, \tag{2.1}$$

$$c = \frac{(n-1) \displaystyle\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} w_{ij}(z_i - z_j)^2}{2\,\mathcal{W} \displaystyle\sum_{i=1}^{n} (z_i - \overline{z})^2} \tag{2.2}$$

where $\mathcal{W} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and $\overline{z}$ denotes the mean of the observed values. Both statistics are closely related as they rely on the cross-product of spatial proximity and the difference in the observed variable. Both denominators capture the variance of Z while the nominators measure covariance. However, Moran's I measures deviations from the mean $\overline{x}$ while Geary's c evaluates the squared differences between the x_i. In general, both statistics can be used interchangeably. However, Cliff and Ord (1975) show that Moran's I is consistently more powerful than Geary's c, i.e. the probability to reject a false null hypothesis is higher using Moran's I (smaller type II error). The expected values of the test statistics in case of no autocorrelation are $-1/(n-1)$ for Moran's I and 1 for Geary's c. Moran's I assumes a value of -1 for perfect negative autocorrelation and a value of 1 for perfect positive autocorrelation. In case of Geary's c, positive autocorrelation assumes a positive value of less than 1, approaching 0 in case of strong positive correlation. Negative autocorrelation is indicated by values larger than 1.

In the original definitions of Moran and Geary, binary weight matrices are applied. However, they have the disadvantage of being invariant under certain topological transformations and do not allow for higher order neighbors. Using a binary weight matrix, the weight of two areas is 1 if a given relation between the areas is true, else it is 0. Typically, one of the relations known as the rook's, bishop's and queen's case are used. Similarly to the figure movement in a chess play a neighborhood relationship exists in the rook's case if two cells share a common edge. In the bishop's case two cells are neighbors if they possess exactly one common vertex. The queen's case subsumes both approaches. Again, for all three approaches $w_{ii} = 0$. If only the type of connection between two areas is specified, neither the size of the areas nor the length of the common boundaries can be considered. Figure 2.6 shows several examples of different topological relations that all result in the same weight matrix assuming the rook's case. The second disadvantage of binary weights is the lack of defining different grades of relationship. For example, second and third degree neighbors may also want to be considered in the statistic but with a smaller weight than contiguous areas. Or a user could prefer the number of connecting streets between two areas. Applying the generalization of Cliff and Ord (1973), a very flexible weight matrix can be specified which is able, for example, to account for natural barriers or traffic infrastructure. Usually, the weight matrix W is not symmetric. The matrix also does not require to be standardized. However, Cliff and Ord (1973) suggest a standardization of W so that $\sum_{j=1}^{n} w_{ij} = 1$ for all $i = 1..n$. It follows that $\mathcal{W} = n$ and the weights allow an interpolation of values z_i using $z_i' = \sum_{j=1}^{n} w_{ij} z_j$.

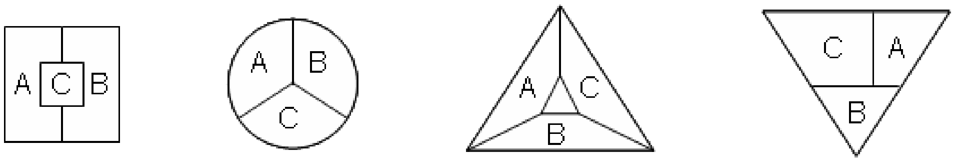


Figure 2.6.: Four topological compositions producing the same binary weight matrix

Spatial variation. A second characteristic of geographic data is its variation. In general, spatial variation is assumed to occur not completely at random but to contain also a structured component that reflects dependencies with the local environment. For spatial phenomena that spread in space a thorough theory, known as geostatistics or theory of regionalized variables, has been developed to describe and analyze the structure of spatial variation and to estimate unknown values (Matheron, 1971; Huijbregts, 1975). In the following, the basic concepts of modeling spatial variation are given.

In general, spatial variation can be of very complex form and is often too erratic to be expressed by a simple, smooth mathematical function (Burrough and McDonnell, 2000). Therefore, the underlying model to express spatial variation commonly assumes a stochastic form, as given in (Cressie, 1993).

Definition 2.1.6 (Stochastic process, random function) *Let $x \in \mathbb{R}^d$ denote a point in d-dimensional geographic coordinate space and $Z(x)$ a random variable of interest at location x. Given an index set $D \subset \mathbb{R}^d$, a stochastic process (or random function) is the family $\{Z(x) : x \in D\}$ of random variables that is generated when varying x over index set D.*

A sample of geo-referenced measurements forms an individual realization of a random process, denoted by $\{z(x) : x \in D\}$. In order to impose a structure on the variation of measurements within the set of locations, $Z(x)$ is generally decomposed into three terms: a structural component representing a mean or constant trend, a random but spatially correlated component, and random noise expressing measurement errors or variation inherent to the variable of interest (Burrough and McDonnell, 2000). More formally, the value of Z at location x is modeled as

$$Z(x) = m(x) + \epsilon'(x) + \epsilon''. \tag{2.3}$$

Note that the second, spatially correlated term corresponds to autocorrelation and models the dependency between individual random variables $Z(x)$. Figure 2.7 depicts the nature of the three terms by a stepwise accumulation of effects.

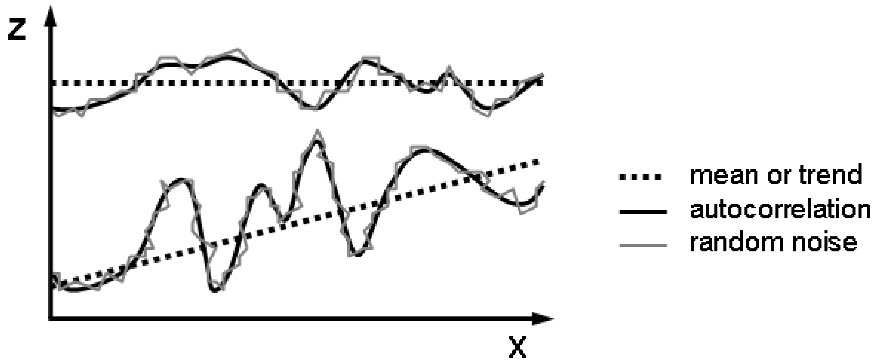


Figure 2.7.: Decomposition of spatial variation into a mean (upper curve) or trend (lower curve), a correlated term and random noise

Given a data sample, it is impossible to derive distribution functions for all combinations of variables without further assumptions. Therefore, some concept of stationarity must be introduced (Chilès and Delfiner, 1999; Paaß and Kindermann, 2000). Stationarity denotes some kind of invariance of a variable with respect to location. Typically, stationarity is defined with respect to the mean and variance of the data. Three types of stationarity are commonly distinguished: strict stationarity, weak stationarity and intrinsic stationarity. Strict stationarity demands that the distribution functions of the random variables $Z(x)$ are invariant under translation and rotation, i.e. for any vector $h \in \mathbb{R}^d$ and any finite set of variables $\{x_1, \ldots, x_k\}$ it is valid that

$$P\{Z(x_1) < z_1, \ldots, Z(x_k) < z_k\} = P\{Z(x_1 + h) < z_1, \ldots, Z(x_k + h) < z_k\}. \tag{2.4}$$

Strict stationarity means that a phenomenon spreads homogeneously in space. A less strict form of stationarity is weak stationarity (also called second-order stationarity). It requires stationarity for the first two moments of the distributions only, i.e.

$$\begin{aligned} E[Z(x)] &= m && (2.5)\\ E[(Z(x)-m)(Z(x+h)-m)] &= C(h) && (2.6)\end{aligned}$$

with m denoting a constant mean and $C(h)$ a covariance function which only depends on the difference in location h. Finally, intrinsic stationary relaxes the constraints further and requires weak stationarity only for the increment $Z(x+h) - Z()x$ for any vector h, i.e.

$$\begin{aligned} E[Z(x+h)-Z(x)] &= \langle a, h\rangle && (2.7)\\ Var[Z(x+h)-Z(x)] &= 2\gamma(h). && (2.8)\end{aligned}$$

The scalar product $\langle a, h\rangle$ of h with some constant a allows for a linear drift and $\gamma(h)$ is also known as the variogram function.

All three kinds of stationarity involve some kind of invariance under translation. However, the distributions are not necessarily invariant under rotation. Stochastic processes that are invariant under rotation are called isotropic and depend only on the length $|h|$ of the vector h, i.e. the phenomenon evolves uniformly in all directions.

2.1.4. Spatial Feature Extraction

Probably the most important question in geographic data analysis is how to extract and utilize spatial information. Geographic characteristics may relate to a single object, to the relation between two or more objects or to the neighborhood of the object in question. Common relational features are distance, topological and directional relationships. The extraction of neighborhood information is usually realized by aggregation using buffers, driving zones or Voronoi cells.

Unary features. Unary features are derived from a single spatial object. They include, for example, geometric characteristics as the length, perimeter or area.

Distance. The distance between two objects is fundamental to geographic modeling tasks. As the surface of the Earth is not flat its curvature has to be considered for (large) distance calculations. The most accurate calculation requires the usage of an ellipsoid as surface model. However, in order to simplify calculations spherical models are also applied. If we approximate the surface of the Earth by a sphere we can apply great circle distance calculation to determine the distance between any two points on the sphere's surface (Wikipedia, 2011).

Definition 2.1.7 (Great-circle distance on Earth) *Let $(\lambda_a \varphi_a)$ and (λ_b, φ_b) denote the geographic coordinates of two points a and b and let r denote the radius of Earth. The distance (in radian) between a and b is defined as*

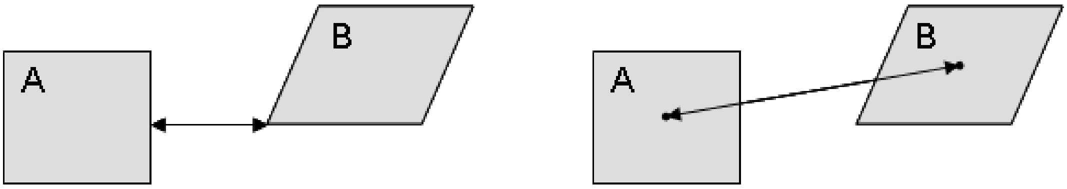
$$d(a,b) = r \cdot \arccos(\,\sin\varphi_a \sin\varphi_b + \cos\varphi_a \cos\varphi_b \cos(\lambda_b - \lambda_a)\,).$$

For short distances Euclidean distance can be applied.

Definition 2.1.8 (Euclidean distance) *The Euclidean distance between two points $a, b \in \mathbb{R}^n$ is defined as*

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}.$$

In the case of point objects, the calculation of distance is straightforward. In the case of lines and areas, the shortest distance between any two points of the objects or the distance between their centroids can be used, see Figure 2.8.



(a) $D(A, B) = min(d(a, b)|a \in A, b \in B)$ (b) $D(A, B) = d(centroid(A), centroid(B))$

Figure 2.8.: (a) Minimal and (b) centroid distance between two areas

However, both defined distances assume an unhindered development of spatial effects and are therefore not always appropriate for modeling purposes. Imagine to calculate the distance between two locations on opposite sides of a river. The beeline is very short. However, the travel distance depends on the location of the next bridge. This is an extreme example of restrictions induced by natural barriers or infrastructure. In practice distance is therefore often calculated based on the street network following two alternatives: travel distance and driving time. Travel distance is the length of the shortest path in the street network between two locations. Driving time is the shortest estimated time to drive between two locations. Both distances can be calculated on a graph representation of the street network using Dijkstra's algorithm. In the first case the edges (street sections) are weighted according to their length. In the second case the edges are additionally weighted by the reciprocal of average or maximum driving speed. Note that distances based on the street network are not necessarily symmetric (e.g. due to one-way streets) or conform to the triangle inequality.

Topological relations. All topological relations are invariant under topological transformations such as translation, rotation and scaling. The predominant formal model to describe topological relations between two objects has been developed by Egenhofer (Egenhofer, 1991) and is called the 9-intersection model. It relies on point-set topology (Willard (2004), for details see Appendix A.1) and utilizes the primitives *interior*, *boundary* and *exterior* to define all possible relations. Each object is represented by a point set. The interior A° of a set A is the largest open set contained in A. The exterior A^- is defined as the complement of A, i.e. all points that do not belong to A. The boundary δA, finally, is the intersection of the closure of A and the closure of A^-. The topological relationship between two objects A and B is then described by the 9-intersection matrix I which results from intersecting each set in $\{A^\circ, \delta A, A^-\}$ with each set in $\{B^\circ, \delta B, B^-\}$:

$$I(A, B) = \begin{pmatrix} \delta A \cap \delta B & \delta A \cap B^\circ & \delta A \cap B^- \\ A^\circ \cap \delta B & A^\circ \cap B^\circ & A^\circ \cap B^- \\ A^- \cap \delta B & A^- \cap B^\circ & A^- \cap B^- \end{pmatrix}.$$

In 2-dimensional space, eight different relations between two areas with connected boundaries can be distinguished. Figure 2.9 depicts the relations together with their belonging intersection matrices. A categorization of binary topological relations between point, line and area objects is given in (Egenhofer and Herring, 1990). Further extensions of the model exist to represent relations between complex areas containing holes (Egenhofer et al., 1994) and relations between areas and directed lines (Kurata and Egenhofer, 2007).

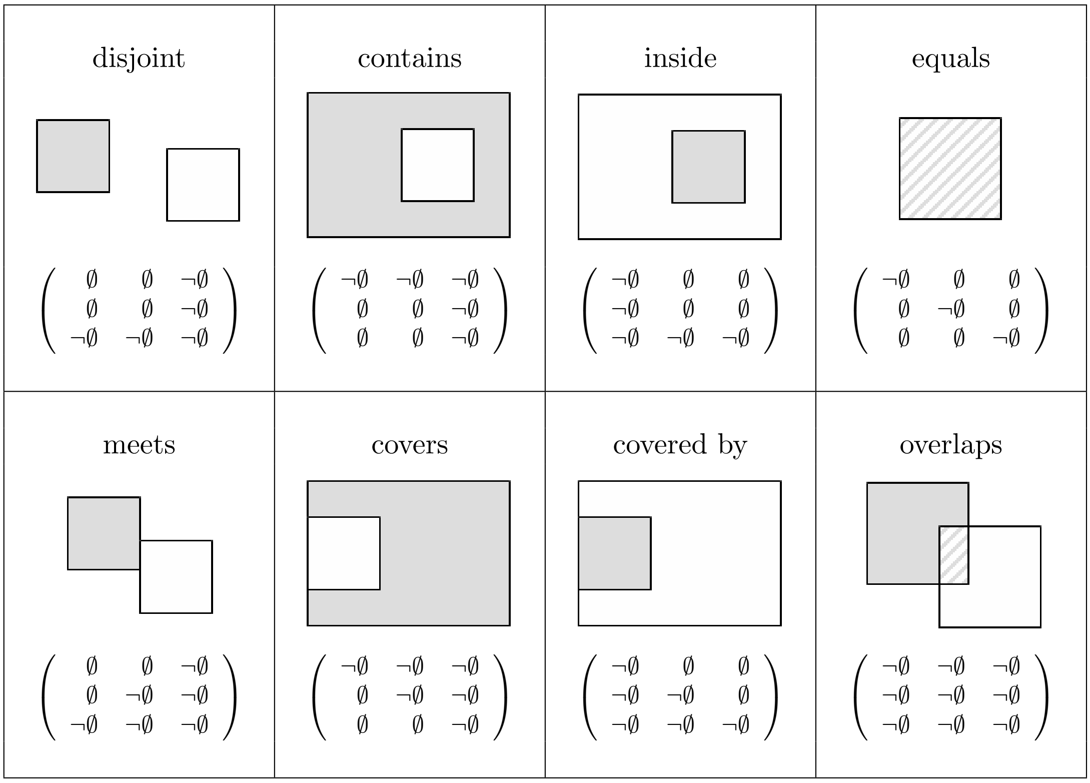

disjoint
$\begin{pmatrix} \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix}$
contains
$\begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix}$
inside
$\begin{pmatrix} \neg\emptyset & \emptyset & \emptyset \\ \neg\emptyset & \emptyset & \emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix}$
equals
$\begin{pmatrix} \neg\emptyset & \emptyset & \emptyset \\ \emptyset & \neg\emptyset & \emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix}$
meets
$\begin{pmatrix} \emptyset & \emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix}$
covers
$\begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \neg\emptyset & \neg\emptyset \\ \emptyset & \emptyset & \neg\emptyset \end{pmatrix}$
covered by
$\begin{pmatrix} \neg\emptyset & \emptyset & \emptyset \\ \neg\emptyset & \neg\emptyset & \emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix}$
overlaps
$\begin{pmatrix} \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \\ \neg\emptyset & \neg\emptyset & \neg\emptyset \end{pmatrix}$

Figure 2.9.: Topological relations of two areas with connected boundaries

Directional relations. Directional (or orientation) relations describe how objects are placed relatively to each other. In order to determine the orientation of a primary object with respect to a reference object, a frame of reference is also required (Hernández, 1994). The reference frame determines the perspective by which the relation is judged. Three reference frames are distinguished in the literature: intrinsic, extrinsic and deictic. An intrinsic reference frame imposes direction by some inherent property of the reference objects (e.g. the front of a car) while an extrinsic frame orientates the direction on external factors of the reference object (e.g. direction of a backwards moving car). Deictic reference frames assume orientation by the point of view from which the reference object is seen (e.g. from within the car).

In geographic context, an extrinsic reference frame with the cardinal directions north, south, east and west is usually used. Depending on the definition of sectors, the cone-based and the projection-based method can be distinguished (Frank, 1996). The cone-based definition assigns the name of the nearest direction to the primary object, resulting in cone-shaped sectors around the reference object (Figure 2.10 left). The projection-based definition relies on two orthogonal axes that each divide the space into two half-planes. The direction of a primary object then results from the combination of both directions (Figure 2.10 right). Both definitions were later generalized by the Star calculus, which allows for an arbitrary granularity of directions (Renz and Mitra, 2004).

The generalization of directional relations from points to areas is described by Papadias and Theodoridis (1997) for a projection-based cardinal reference system. They introduce the notion of strong and weak relations. For example, given a point set $P \subset \mathbb{R}^2$ of a primary object and a point set $Q \subset \mathbb{R}^2$ of a reference object and two basic relations $north : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \{0, 1\}$, $south : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \{0, 1\}$ that state for each pair of points (p_i, q_i) with $p_i \in P, q_i \in Q$ whether p_i is north respectively south of q_i. More precisely, given a functions $y : \mathbb{R}^2 \rightarrow \mathbb{R}$ that returns

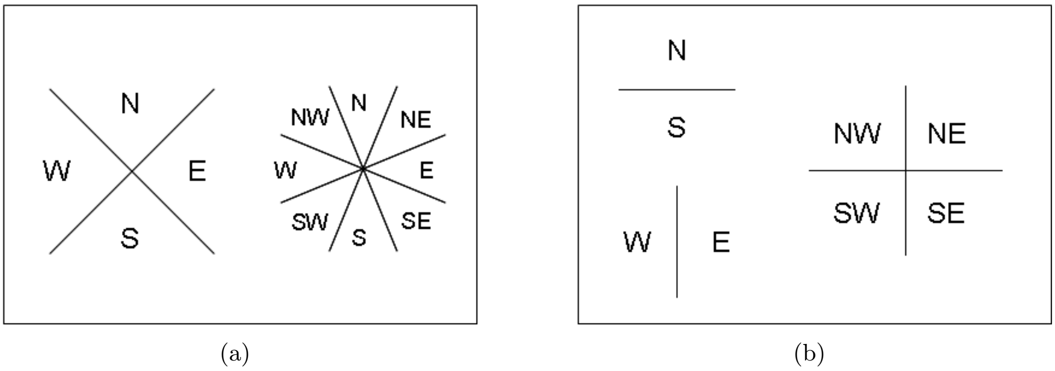


Figure 2.10.: (a) Cone-based and (b) projection-based cardinal sectors of different granularity

the y coordinate of some two-dimensional point, the relations *north* and south are defined as

$$north \equiv y(p_i) > y(q_j),$$
$$south \equiv y(p_i) < y(q_j).$$

The relations *strong_north* and *week_north* between the point sets P and Q are then defined as

$$\begin{aligned} strong\_north(P,Q) &\equiv \forall p_i\ \forall q_j\ \ north(p_i,q_j), \\ weak\_north(P,Q) &\equiv \exists p_i\ \forall q_j\ \ north(p_i,q_j)\ \wedge \\ & \ \forall p_i\ \exists q_j\ \ north(p_i,q_j)\ \wedge \\ & \ \exists p_i\ \exists q_j\ \ south(p_i,q_j). \end{aligned}$$

A strong relationship denotes that all pairs of points (p_i, q_j) abide by the relation *north* whereas a weak relation requires only that relation *north* holds at least once for each $p_i \in P$. An example of both relations is depicted in Figure 2.11.

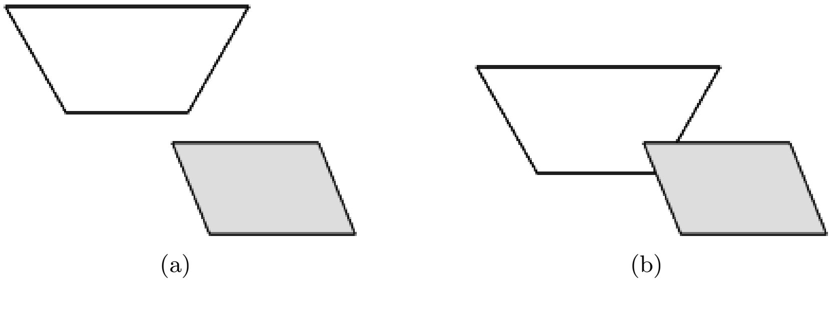


Figure 2.11.: Directional relations (a) *strong_north* and (b) *weak_north*

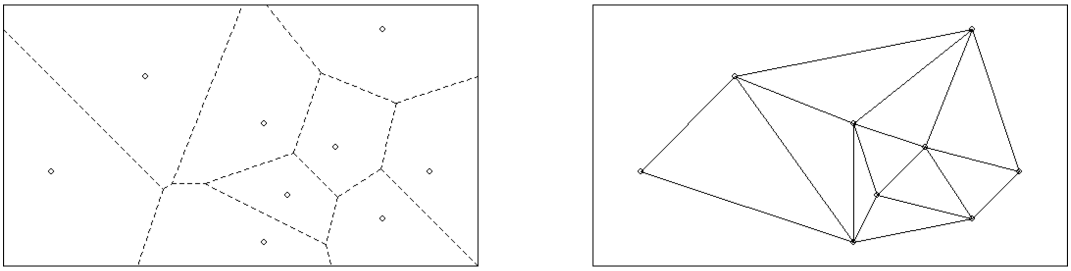
Spatial aggregation. Aggregation of data is commonly applied to summarize information and to derive attributes that cannot be measured at a single point. During feature extraction aggregation is used to attach information about the local environment to some entity. Aggregation units can be formed according to different criteria. Most commonly used are administrative borders, Voronoi polygons, buffers and drive time zones. Administrative borders and Voronoi polygons result in irregular tessellations of space while buffers and drive time zones are built for individual objects. Typical examples for administrative units are municipalities and post code areas.

Voronoi polygons, also called Thiessen or Dirichlet polygons, associate each point in space with the closest member of some given point set (Okabe et al., 1992). Thus, the partition solely depends on the number and distribution of points in the point set.

Definition 2.1.9 (Voronoi polygon) *Given a finite set of points* $P = \{p_1, p_2, \dots, p_n \mid p_i \in \mathbb{R}^2,\ i = 1..n,\ n \geq 2\}$ *and a function* $d(\cdot,\cdot)$ *denoting Euclidean distance, a Voronoi polygon is the area defined by*

$$V(p_i) = \left\{x \in \mathbb{R}^2 \mid d(x,p_i) \leq d(x,p_j) \ for\ i \neq j\ and\ i,j = 1..n\right\}.$$

The set of all Voronoi polygons $\mathcal{V} = \{V(p_1), \dots, V(p_n)\}$ is called a Voronoi diagram. Note that the Voronoi diagram is the dual tessellation of the Delauny triangulation. An example Voronoi diagram and its respective Delauny triangulation are depicted in Figure 2.12. Voronoi polygons are useful to model the influence of competitors in retail or the coverage of radio antennas (Körner et al., 2010a). However, they do neither consider spatial barriers (e.g. rivers, highways) nor the road network structure.



(a)

(b)

Figure 2.12.: (a) Voronoi diagram and (b) its associated Delauny triangulation

Buffers and drive time zones are distance and time based aggregation units, respectively. A buffer is a zone of specified width around a point, line or area.

Definition 2.1.10 (Buffer) *Given a point, line or area object* $A \subset \mathbb{R}^2$ *and a function* $d(\cdot,\cdot)$ *denoting Euclidean distance, the buffer of width* r *around* A *is the set of points*

$$\mathit{buffer}(A,r) = \left\{p \in \mathbb{R}^2 \mid \exists a \in A\ s.t.\ d(a,p) \leq r\right\}.$$

Buffers are often used during feature extraction to describe the neighborhood of an object. For example, May et al. (2008a,b) developed a model for the extrapolation of traffic frequencies. The model utilizes, amongst other attributes, the number of restaurants and public buildings within a buffer of specified length around each street segment. One disadvantage of buffers is, again, that they do neither consider spatial barriers nor the road network structure.

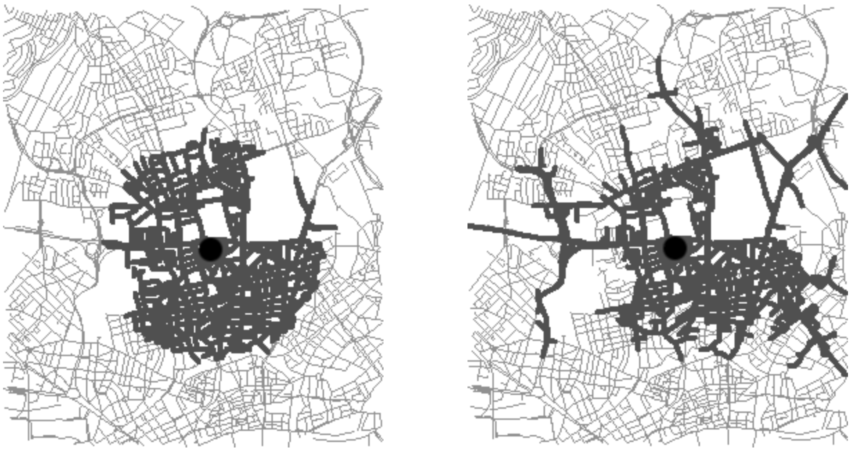
In contrast, drive time zones are defined on the street network. They account for barriers as well as speed limits on the street network.

Definition 2.1.11 (Street network) *A street network is a (possibly directed) graph* $N = (V,S)$ *with* $V \subset \mathbb{R}^2$ *and* $S \subseteq V \times V$.

Definition 2.1.12 (Street segment) *An edge* $s \in S$ *of a street network* N *is called street segment.*

Definition 2.1.13 (Drive time zone) *Given a street network N and a location s' on a street segment, a drive time zone is the set of street segments that be reached within a previously specified driving time from s'.*

Drive time zones are adequate aggregation units whenever an application involves personal mobility. In (Körner et al., 2010a; Krause-Traudes et al., 2008) drive time zones of varying sizes were used to model shopping linkage between retail locations. Figure 2.13 contrasts a 1500 m buffer and a 4 minute drive time zone in the middle of Frankfurt, Germany.



(a) 1500 m buffer

(b) 4 minute drive time zone

Figure 2.13.: Aggregations within Frankfurt

2.1.5. Analysis Tasks and Methods

In general, there are two alternatives how algorithms treat geographic data. The first approach uses traditional algorithms and includes spatial attributes either as ordinary variables or requires feature extraction during preprocessing. The second approach relies on specialized algorithms that incorporate feature extraction or are able to handle geographic dependencies directly. In the remaining section several algorithms for geographic data are presented and their strategy for feature extraction and ability to handle autocorrelation is emphasized.

Clustering. Clustering divides a given set of objects into non-overlapping groups, so that similar objects are within the same group and objects of different groups are most heterogeneous. As clustering relies on the distance between objects, it is inherently spatial. Yet, the assumption of convex clusters (e.g. k-means) is inappropriate for many geographical data sets (see Figure 2.14). Ester et al. (Ester et al., 1996) developed a density based algorithm for point data that finds clusters of arbitrary shape. The idea of this approach is that a cluster can be recognized by a high density of points within, while only few points are found in the surrounding environment. It requires the definition of a *neighborhood*, which is used to iteratively join points, and a *density* which is used to delineate the borders of a cluster. In (Sander et al., 1998) this approach is extended to cluster vector data (e.g. polygons).

Classification and regression. In classification and regression the unknown target value of some object is predicted given a set of training instances. If the target variable is discrete, the learning task is called classification. If it is continuous, it is referred to as regression. We start

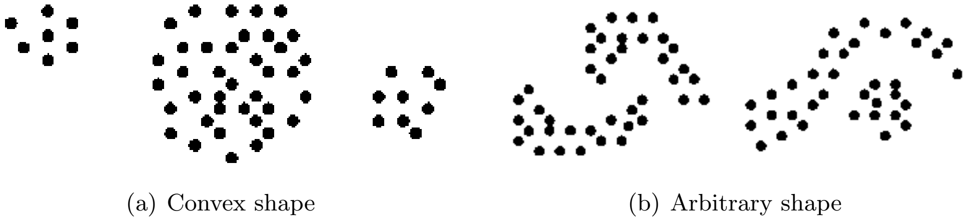

(a) Convex shape (b) Arbitrary shape

Figure 2.14.: Spatial clusters

with the well-known k-nearest neighbor method, which can be applied to both, classification and regression tasks. The second part presents spatial model trees, geographically weighted regression, and we conclude this section with Kriging. Kriging is a popular regression method in geostatistics and takes explicitly advantage of autocorrelation.

The k-nearest neighbor algorithm (kNN) is an instance based learning method that classifies unknown instances according to the target value of the k most similar training examples. It assumes that objects with similar characteristics also possess similar class values. In case of classification, the most frequent target value among the neighbors will be assigned to the instance. In case of regression, a (weighted) mean is calculated. In order to determine the similarity between two objects, kNN requires a distance function for each attribute. As geographic coordinates can be used to determine the distance between two locations, they can be directly included in the algorithm. Thus, kNN relies on objects that are within the geographic neighborhood and exploits positive autocorrelation of the target variable.

May et al. (2008a) developed a true spatial version of the kNN algorithm, called s-kNN, which can be generally applied to spatial objects with vector geometries. The algorithm performs on-the-fly distance calculations and improves performance efficiency using a partial evaluation scheme. While differences in numerical attributes can be determined very fast, the geographic distance between geometric objects is computationally expensive. The authors therefore perform a step-wise distance calculation when searching for the k nearest neighbor objects. First, only non-spatial attributes of a neighbor candidate are evaluated. If the summarized distance of all non-spatial attributes already exceeds the maximal total distance of the current top-k neighbors, the candidate neighbor can be safely discarded and no spatial calculation is necessary. Else, the distance between the minimum bounding rectangles (MBRs) of the geometries is calculated. The MBR distance is a lower bound for the actual distance between two geometries is and less expensive to calculate. Again, if the distance of the non-spatial attributes plus the distance between the MBRs is greater or equal to the threshold, the instance can be discarded. Only if both tests are passed, the actual spatial distance is determined. May et al. (2008a) applied the approach for the prediction of traffic frequencies in Germany. For the city of Frankfurt, for example, the specialized algorithm obtained a significant reduction in computation time from nearly one day to about two hours.

Model trees Wang and Witten (1997) operate similar to decision trees, but possess leaves that are associated with (linear) functions instead of fixed values. While internal nodes of the tree partition the sample space, leave nodes construct local models for each part of the sample space. Malerba, Ceci and Appice Malerba et al. (2005) developed a spatial model tree which is able to model local as well as global effects. Their induction method, Mrs-SMOTI (Multi-relational Spatial Stepwise Model Tree Induction), places regression nodes also within inner nodes of the tree and passes these regression parameters to all child nodes. Mrs-SMOTI exploits spatial relationships over several layers and possesses a tight database integration to extract spatial relations during the induction phase.

Geographically weighted regression (GWR) (Fotheringham et al., 2002) extends the traditional regression framework such that all parameters are estimated within a local context. The model for some variable z at location i then takes the following form:

$$z_i = \beta_0(x_{ix}, x_{iy}) + \sum_k \beta_k(x_{ix}, x_{iy})x_{ik} + \varepsilon_i.$$

In the equation above, (x_{ix}, x_{iy}) denotes the pair of coordinates at location i, $\beta_k(x_{ix}, x_{iy})$ is the localized parameter for attribute k, x_{ik} is the value of attribute k at location i and ε_i denotes random noise. The GWR model assumes that all parameters are spatially consistent. Therefore, parameters at location i are estimated from measurements close to i. This is realized by the introduction of a diagonal weight matrix $\boldsymbol{W_i}$ which states the influence of each measurement for the estimation of regression parameters at i:

$$\widehat{\boldsymbol{\beta}}(x_{ix}, x_{iy}) = (\boldsymbol{X}^T\boldsymbol{W_i}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W_i}\ \boldsymbol{z}.$$

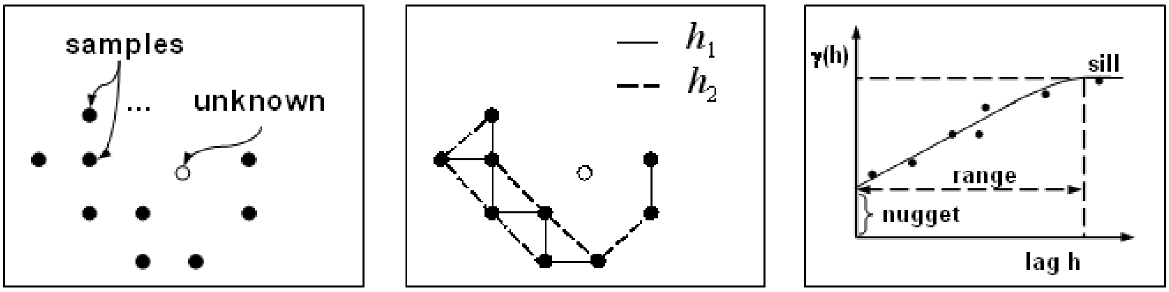
The weight matrix can be built according to several weighting schemes, such as a Gaussian or bi-square function. GWR is a local regression method which takes advantage of positive autocorrelation between neighboring points in space.

Kriging (Wackernagel, 1998; Chilès and Delfiner, 1999; Cressie, 1993) is an optimal linear interpolation method to estimate unknown values in geographic field data. Let x denote a location in an index set $D \subset \mathbb{R}^n$ in n-dimensional geographic coordinate space and $Z(x)$ a random variable of interest at location x. Generally, each variable $Z(x)$ can be decomposed into three terms: a structural component representing a mean or constant trend, a random but spatially correlated component that denotes autocorrelation, and random noise expressing measurement errors or variation inherent to the variable of interest (Burrough and McDonnell, 2000), see also Figure 2.7.

A technique most widely used is Ordinary Kriging, which assumes intrinsic stationarity with an unknown but constant mean of the random target variable $Z(x)$. Given a set of measurements, Kriging estimates unknown values as weighted sum of neighboring sample data (Figure 2.15(a)) and uses the variogram to determine optimal weights (Figure 2.15(c)). Variograms model spatial dependency between locations and are a function of distance for any pair of sites:

$$\gamma(h) = \frac{1}{2}Var[Z(x+h) - Z(x)].$$

A variogram of the data can be obtained in two steps. First, the experimental variogram is calculated from the sample by calculating the variance between samples for all increments h. Figure 2.15(b) shows all pairs of sample points with a lag h_1 (solid lines) and a second lag h_2 (dashed lines). In a second step the experimental variogram serves to fit a theoretical variogram which is used in Ordinary Kriging. Depending on the data, different model types may be appropriate for the theoretical variogram. Often a spherical model is used and its parameters are adapted to reflect the experimental variogram. Each variogram is characterized by three parameters: nugget, range and sill as depicted in Figure 2.15(c). The nugget effect represents random noise, as by definition $\gamma(0) = 0$. Within the range the variance of increments increases gradually with distance in this example. It directly shows the spatial dependency. The closer two points are the more likely is it that their values are similar. Finally, the curve levels off at the sill. The variance has reached its maximum value and is independent of distance.



(a) Data sample (b) Lag between sample points (c) Variogram

Figure 2.15.: Variance of sample increments

Ordinary Kriging estimates the unknown value at a location x_0 as weighted sum of neighboring sample points x_i $(i = 1..n)$:

$$Z^*(x_0) = \sum_{i=1}^{n} w_i \, Z(x_i).$$

The weights w_i are determined in conformance with two restrictions. First, $Z^*(x_0)$ must be an unbiased estimate of the true value $Z(x_0)$, which means that on average the prediction error for location x_0 is zero. Because the model assumes a constant mean $m = E\left[Z(x_i)\right]$ $(i = 0..n)$, this claim bounds the sum of weights to one.

$$\begin{aligned} 0 &= E\left[Z^*(x_0) - Z(x_0)\right] = E\left[\sum_{i=1}^{n} w_i \, Z(x_i) - Z(x_0)\right] \\ &= m\left(\sum_{i=1}^{n} w_i - 1\right) \qquad \Rightarrow \sum_{i=1}^{n} w_i = 1 \end{aligned}$$

Second, we require an optimal estimate which minimizes the error variance σ_E^2 of the estimate. The second equation expresses the variance in terms of the variogram.

$$\begin{aligned} \sigma_E^2 &= Var\left(Z^*(x_0) - Z(x_0)\right) = E\left[\left(Z^*(x_0) - Z(x_0)\right)^2\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(x_i - x_j) - 2\sum_{i=1}^{n} w_i \gamma(x_i - x_0) + \gamma(x_0 - x_0) \end{aligned}$$

The derivatives of the error variance with respect to $w_i (i = 1..n)$ yield a linear system of n equations. In combination with the restriction on the weights, a Lagrange parameter ϕ is introduced and a total of $n+1$ equations is obtained. For each location x_0, the optimal weights w_i are estimated using the following system of equations, given in matrix form:

$$\begin{pmatrix} \gamma(x_1 - x_1) & \dots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & 1 \\ \gamma(x_n - x_1) & \dots & \gamma(x_n - x_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \phi \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}$$

Note that Ordinary Kriging is an exact interpolator. If the value of a location in the data sample is estimated, it will be identical with the measured value. Several variants of Kriging have been developed which extend interpolation to data that contains a trend (Universal Kriging (Chilès and Delfiner, 1999; Cressie, 1993)), involves uncertainty (Bayesian Kriging (Chilès and Delfiner, 1999)) or contains temporal relations (Spatiotemporal Kriging).

Association rules. Spatial association rules extend the common definition of association rules as given by Agrawal et al. (1993) insofar as they require at least one spatial predicate in the antecedent or consequent of the association rule (Koperski and Han, 1995). Hereby, a spatial predicate expresses, for example, a topological relation. More formally, given a conjunction of predicates $P = P_1, \ldots, P_m$ and $Q = Q_1, \ldots, Q_n$ that contain at least one spatial predicate, a spatial association rule is a rule of the form

$$P_1 \wedge \ldots \wedge P_m \rightarrow Q_1 \wedge \ldots \wedge Q_n \quad (s\%, c\%)$$

with s and c denoting the support and confidence of the rule, respectively. The support of a rule in a set A of objects is the percentage of objects in A that satisfy the antecedent, i.e.

$$s(P \rightarrow Q, S) \;=\; \frac{\mid P(A) \mid}{\mid S \mid}$$

where $P(A)$ is the set of objects for which P is true and $| \cdot |$ denotes the number of elements in a set. The confidence of a rule in a set A states the conditional probability that objects satisfying the antecedent also satisfy the consequent, i.e.

$$c(P \rightarrow Q, S) \;=\; \frac{\mid P(A) \cap Q(A) \mid}{\mid P(A) \mid}.$$

Several algorithms to mine spatial association rules have been proposed, among them (Koperski and Han, 1995; Appice et al., 2003). The approach of Appice et al. (2003) relies on inductive logic programming (ILP) to represent spatial relations and background knowledge. It is a relational approach and has been extended to multiple relations and multi-level association rules in (Appice et al., 2005). All of the above approaches posses a loose coupling to spatial objects and perform feature extraction prior to data mining. This has an advantage for repeated (similar) calculations, requires, however, high computational effort prior to data mining and redundant data storage. One disadvantage of spatial association rule mining is the discovery of trivial spatial relations, such as cities with ports are located at a water resource (Bogorny et al., 2006). Bogorny et al. (2006) use prior knowledge to eliminate such dependencies during frequent set generation, which leads to fewer frequent sets as well as improved computation time.

A second type of association rules for geographic data are co-location rules (Shekhar and Chawla, 2003). Co-location rules (Huang et al., 2004) find types of spatial objects that are frequently co-located in geographic space. They differ from association rules as they are not transaction-based but rely on neighborhood relations and operate directly on the (point) coordinates of geographic objects. Co-location rules can be used, for example, in ecology to identify symbiotic relationships between plants and animals.

Subgroup discovery. Subgroup discovery analyzes dependencies between a target variable and several explanatory variables. It detects groups of objects that show a significant deviation in their target value with respect to the whole data set. For example, given a discrete target attribute, a subgroup displays an over-proportionally high or low share of a specific target value. More precisely, the quality q of a subgroup h accounts for the difference of target share between the subgroup p and the whole data set p_0, as well as the size n of the subgroup (Klösgen, 2002):

$$q(h) = \frac{|p - p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

Subgroups are usually defined by simple conjugation of attribute values, which are then applied to the data set in question. Spatial subgroups are formed if the subgroup definition involves operations on spatial components of the objects. For example, a spatial subgroup could consist of all city districts that are intersected by a river (Klösgen and May, 2002). However, spatial operations are expensive and, due to early pruning, it may not be necessary to compute all relations in advance. Klösgen and May (2002) developed a spatial subgroup mining system which integrates spatial feature extraction into the mining process. Spatial joins are performed separately on each search level. Thus, the number of spatial operations can be reduced and redundant storage of features is avoided.

2.2. Mobility Data

2.2.1. Movement in Space and Time

Movement is the change of location over time. We therefore need a geographic space as well as a temporal space to trace the movement of some object. In the previous section we have already defined the term geographic space (see Definitions 2.1.3 and 2.1.4). However, temporal space has not been defined yet, and we will do so in the first part of this section. Afterwards we proceed to the movement of objects.

Relativistic and Newtonian time. Albert Einstein's theories on relativity influence not only our view on physical space but also our view on time. In his special theory on relativity in 1905 Einstein postulates that time differs for two observers which move relative to each other (Wikipedia, 2010d). In addition, Einstein describes the influence of gravitation upon the flow of time in his general theory on relativity in 1915 (Fritzsch, 2000). However, both effects are visible only in extreme situations, for example, when considering velocities close to the speed of light or strong differences in gravitational fields. For observations of slow-moving objects in weak gravitational fields as in every-day life, classical mechanics are sufficiently accurate (Strobel, 2007). We will therefore consider time as absolute, uniform and from external effects independent extent as postulated by Newton (Wikipedia, 2010d).

Definition 2.2.1 (Time) *Time $\mathcal{T}$ is an absolute, uniform and from external effects independent extent in one dimension, i.e. the temporal space that we observe in daily life.*

Temporal reference systems. Temporal reference systems have been developed in many cultures over time. Already 3000 to 4000 years BC solar or lunar calendar systems existed, for example, in Egypt, China or India (Richmond, 1956). However, the rotation of Earth is subject to fluctuations due to physical effects and therefore not suitable for uniform time measurement. In 1967 the SI second was defined as part of the International System of Units (SI) based on a fixed number of radiation periods of a caesium 133 atom and has become the international time unit standard (International Bureau of Weights and Measures, 2006). Today the standard temporal reference system is Coordinated Universal Time (UTC), maintained by the International Bureau of Weights and Measures (BIPM). UTC uses the SI second as basic unit, however, it is kept in accordance with the Earth's rotation. By insertion of leap seconds the BIPM ensures that on average the sun crosses the Greenwich meridian at noon UTC with a deviation below 0.9 seconds (International Bureau of Weights and Measures, 2010).

UTC uses the Gregorian calender to reference days and divides a day further into hours, minutes and seconds (Wikipedia, 2010c). A temporal reference is typically given as a string that consists of the individual date and time components. Commonly these components are formatted according to ISO 8601, an international standard for the specification of dates and times provided by the International Organization for Standardization (ISO). One possible format for a combined date and time representation according to ISO 8601 is the following (International Organization for Standardization (ISO), 2004):

$\pm\,\underline{\text{Y}}\text{YYYY-MM-DD T hh:mm:ss.}$

Hereby, each letter represents one digit of a date or time component. In case of an underlined letter, zero or more digits can be specified. If a component is shorter than the defined format length, the space is filled by leading zeros. In the above string the term "$\pm\,\underline{\text{Y}}\text{YYYY-MM-DD}$" specifies the date using a minimum of four digits for the *year*, two digits for the *month* and two digits for the *day*. "T" is a separator of the date and time component. The string "hh:mm:ss" specifies the *hour*, *minute* and *second* using two digits each. The range of the date and time components can be specified further as follows: $year \in \mathbb{Z}$, $month \in [1, 12]$, $day \in [1, 31]$ yet dependent upon month and type of year (common/leap year), $hour \in [0, 23]$, $minute \in [0, 59]$ and $second \in [0, 59]$.

Next to references to a fixed time scale as stated above, temporal references can also be specified relative to an arbitrary point in time. Such references are therefore sometimes called *relative* time references. A relative reference system requires a given starting point in time and some convenient unit of time which is used to measure the difference between the considered time point and the starting time point. Typically, the time units seconds, minutes, days or years are used. Note that the obtained time differences may be positive as well as negative. However, in many statistical applications the starting point is selected such that temporal references are non-negative natural numbers. Similar to geographic references, a conversion between temporal references of different reference systems is possible if all referencing information is available.

To summarize, temporal reference systems allow to specify points in time based on a fixed or relative time scale. The relationship between the specification and the time dimension is established by the origin of the chosen scale and the unit of measurement. In the following, we will call the space of all possible references for a given temporal reference system *temporal coordinate space.*

Definition 2.2.2 (Temporal coordinate space) *Temporal coordinate space $\mathcal{T}_\mathcal{C}$ is the one-dimensional coordinate space of a temporal reference system used to specify a unique point in time. Temporal coordinate space has the form*

- $\mathcal{T}_\mathcal{C} = \{\pm\,\underline{Y}YYYY\text{–}MM\text{–}DD\ T\,hh : mm : ss\}$ *in case of the UTC temporal reference system formatted according to ISO 8601,*
- $\mathcal{T}_\mathcal{C} = \mathbb{Z}$ *or* $\mathcal{T}_\mathcal{C} = \mathbb{N}_0$ *in case of a relative temporal reference system.*

Similar to geographic references, we can define a function r_T, which assigns each temporal reference to some point in time, i.e. $r_T : \mathcal{T}_\mathcal{C} \to \mathcal{T}$.

Given a precise definition of geographic space and time as well as their respective coordinate spaces, we can now define movement in space and time. In this thesis we will use the term *trajectory* or *trajectory function* to refer to the path that a moving object traverses over time, and specify it using the respective reference systems. Note that the terms *space-time path* (Hägerstrand, 1970) and *geospatial lifeline* (Hornsby and Egenhofer, 2002) are also used in time geography to refer to the movement of an object.

Definition 2.2.3 (Trajectory or trajectory function) *Given a temporal coordinate space* $\mathcal{T}_\mathcal{C}$ *and a geographic coordinate space* $\mathcal{S}_\mathcal{C}$*, a trajectory is a function* $tr : \mathcal{T}_\mathcal{C} \to \{\,\{s\} \mid s \in \mathcal{S}_\mathcal{C}\,\} \cup \emptyset$ *which is continuous on all intervals* $[t_1, t_2]$ *with* $t_1, t_2 \in \mathcal{T}_\mathcal{C}, t_1 < t_2$ *for which the following holds* $tr(t) \neq \emptyset \; \forall t \in [t_1, t_2]$.

We define a trajectory as a function which maps each point in temporal coordinate space $\mathcal{T}_\mathcal{C}$ to a point in geographic coordinate space $\mathcal{S}_\mathcal{C}$ or to the empty set. We thus follow the common simplification to represent the movement of an object by the movement of its center of mass. Note that we represent a point in geographic coordinate space as a singleton in order to be compliant with the definition of topological relations as defined by Egenhofer (1991), see also Section 2.1.4. Further, we include the empty set in the co-domain in order to handle time periods before and after an object exists. This simplifies the definition of topological relations for trajectory data. Finally, our definition requires that a trajectory is *continuous* on all intervals $[t_1, t_2]$ with $t_1, t_2 \in \mathcal{T}_\mathcal{C}$, $t_1 < t_2$ during which the function value of the trajectory is not the empty set. Continuity is an important restriction which characterizes movement in geographic space. Mathematically the continuity of a trajectory corresponds to the continuity of a function between two metric spaces because both coordinate spaces $\mathcal{T}_\mathcal{C}$ and $\mathcal{S}_\mathcal{C}$ are metric spaces when provided with an appropriate distance function. A precise definition of continuity on a given time interval for a function between two metric spaces is provided in Appendix A.1.2, Definition A.1.9.

Having defined a trajectory, we will now give a precise definition of a *mobile entity* or short *entity*, which is the formal term that we use in this thesis to refer to a moving object. Mobile entities will play a central role in all following chapters.

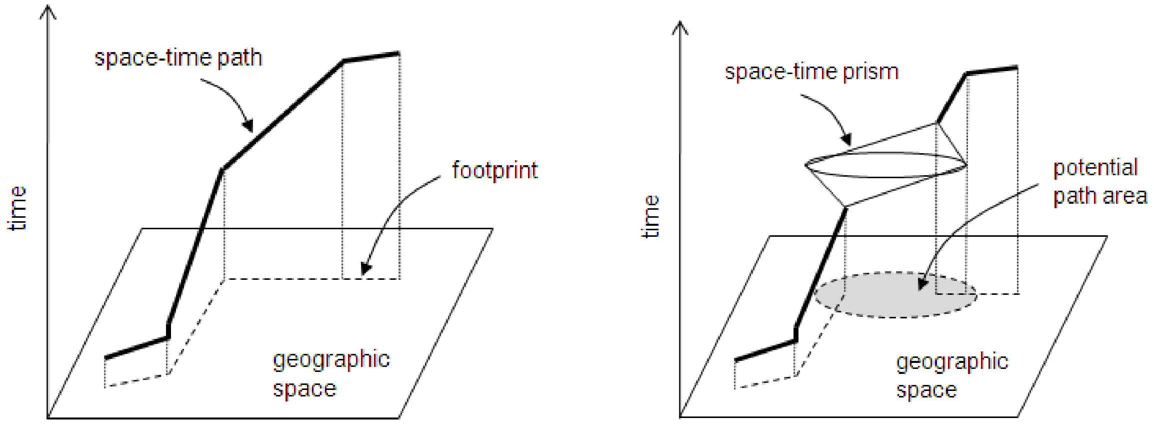
Definition 2.2.4 (Mobile entity or entity) *Given a temporal coordinate space* $\mathcal{T}_\mathcal{C}$ *and a geographic coordinate space* $\mathcal{S}_\mathcal{C}$*, a mobile entity, or short entity, is an object* e *with lifetime interval* $[t_1, t_2]$ *with* $t_1, t_2 \in \mathcal{T}_\mathcal{C}$*,* $t_1 < t_2$ *which possesses a trajectory function* $tr(t) : \mathcal{T}_\mathcal{C} \to \{\,\{s\} \mid s \in \mathcal{S}_\mathcal{C}\,\} \cup \emptyset$ *for which the following holds*

$$\begin{aligned} &tr(t) \neq \emptyset \quad \forall t \in [t_1, t_2]\,, \\ &tr(t) = \emptyset \quad \forall t \notin [t_1, t_2]\,. \end{aligned}$$

2.2.2. Trajectory Data Models and Data Structures

The first space-time concept to study daily human activities was introduced by Hägerstrand in 1970 (Hägerstrand, 1970). It postulates that time should be considered along with space when studying human behavior, which is today a prominent feature in the research area known as time geography.

Hägerstrand proposes a model that uses temporal constraints to limit the area of possible movements for a person. He defines a space-time path for each person, which portrays the person's movements from birth to death. The path may be broken into smaller units reflecting, for example, daily or weekly movement. The paths are visualized in a so-called space-time cube. A space-time cube depicts the course of movement using both horizontal axes to represent space (hereby collapsing 3-dimensional geographic space into a 2-dimensional plain) and the vertical axis to represent time. Figure 2.16 left illustrates this concept. It shows the movement of a person, for example, starting at his or her home, walking a short distance to the bus stop, taking the bus and continuing by foot to his / her place of work. Vertical lines represent a stay at a certain location while inclining lines represent movement. The steeper the gradient of a line is, the slower is the movement. For example, compare the sections of movement by foot and by bus. A projection of the space-time path to the horizontal plain is then called footprint of the path. Note that the diagram abstracts from true movement behavior as straight lines represent constant velocity. In our example this could be the average velocity for each part of the space-time path.



(a) Space-time path

(b) Space-time prism

Figure 2.16.: Concepts of time geography

A second important concept introduced by Hägerstrand is the space-time prism. It is the volume in space-time in which a person can possibly move when starting from a given position and going to another position within a limited period of time (see Figure 2.16 right). The space-time prism results from assuming a maximum speed that a person can reach while traveling. When starting from or going to a specified position, the speed naturally limits the places within reach or from which the given position can be reached, respectively. The projection of the space-time prism to the plain is called potential path area. Note that also space-time prisms are an abstraction of reality. As outlined in Section 2.1.4 (see spatial aggregation), individual movement does not spread evenly into all directions. It is usually bound to the road network and can be hindered by natural barriers.

The concepts of space-time paths and space-time prisms were formalized by Miller (2005), specifying a measurement theory for time-geographic concepts.

Note that several synonyms are used in the literature for space-time paths as introduced by Hägerstrand. Among them are the term *lifeline* (Hornsby and Egenhofer, 2002) and *trajectory*, the latter of which we will use in this thesis (see Definition 2.2.3). Note further that in order to focus on movement behavior and to keep a simple notation we assume that each entity is represented by its center of mass.

Several data models exist in order to store trajectory data and to analyze it in a trajectory database or a trajectory data warehouse. All models have to decide between a discrete and a continuous temporal representation of movement. As positioning techniques are bound to record position data at discrete points in time, it is natural to choose a discrete representation of trajectories, too. Typically such a trajectory in 2-dimensional geographic coordinate space is represented simply as a sequence of tuples

$$(\,(x_1, y_1, t_1),\ \ldots,\ (x_n, y_n, t_n)\,)$$

with $(x_i, y_i) \in \mathcal{S_C}$, $t_i \in \mathcal{T_C}$ and $t_1 < \ldots < t_n \; \forall i \in 1..n$. A more advanced model is the moving objects data model (Güting et al., 2000; Forlizzi et al., 2000). The model was designed to express continuous changes of objects over time, both in their position and extent (Erwig et al., 1999). It is thus not only able to model moving point objects but also to model moving regions. The model approximates movement in continuous space by piecewise linear movements. In the case of moving points the data structure consists of a sequence of tuples (*time interval*, *moving point*), each tuple determining the linear movement of a point within the specified time interval. A moving point is further specified as follows

$$\textit{moving point} = \{\, (x_0, x_1, y_0, y_1) \mid x_0, x_1, y_0, y_1 \in \mathcal{S_C} \,\}\,.$$

Hereby, the position of the point at time t within the specified interval of time is obtained through the function

$$f(t) = (x_0 + x_1 \cdot t,\; y_0 + y_1 \cdot t).$$

A comprehensive overview of the moving objects data model is given by Güting and Schneider (2005). A comprehensive comparison of other spatiotemporal data models is given by Pelekis et al. (2004) and Macedo et al. (2008).

Specifically for the support of trajectory data analysis Pelekis et al. (2011b) implemented a trajectory database engine called HERMES. It relies on a moving point data structure and provides state-of-the-art indexing and querying functionalities. More details about the implementation of trajectory databases and trajectory data warehouses can be found in (Fretzos et al., 2008) and (Pelekis et al., 2008).

Note that the preprocessing of trajectories may change the format of their location reference. Many traffic-related applications require, for example, to match trajectories to the underlying street network. A trajectory then consists of a sequence of traversed street segments, each provided with the time of entry and exit. In addition, trajectories may be enriched with semantic information (Guc et al., 2008), for example, stating the aim of the trip or the means of transportation. Different preprocessing techniques for trajectory data will be discussed in detail in Section 2.2.5.

2.2.3. Characteristics of Human Movement Behavior

Human movement does not resemble a random walk. Instead, it is characterized by regular visits to places as our home, work, facilities for shopping or leisure activities. At least two possibilities exist to study movement behavior. The first method is to analyze constraints of human movement and to draw conclusions about the resulting behavior. The second method is, of course, observation. Hägerstrand (1970) has been a pioneer of the constraint-based approach. In his work he emphasized, on the one hand, the functional characteristics of motion, i.e. each individual is at some location during each point in his or her lifetime, and the locations of consecutive points in time are spatially related. On the other hand he defined three general groups of constraints that dominate individual movement behavior: capability, coupling and authority constraints. Capability constraints subsume any restrictions due to the physiology of the human body (e.g. maximum walking speed, need for sleep) or means of transportation a person can use. Coupling constraints refer to space-time restrictions due to necessary interaction with other individuals or tools in order to complete a task (e.g. being at a meeting or operating a machine at work). Finally, movement limitations may be imposed by authorities by restricting the access to certain places or domains. One of the main characteristics that Hägerstrand deduced from the constraints is regularity of movement. This regularity originates on the one hand from the habit of a home base and the human need for sleep (capability constraint). On the other hand regularity is imposed by fixed time tables that repeat over the week (coupling constraint).

The regularity of human movement behavior was verified by studies which observed the mobility of persons over a long period of time. Schlich and Axhausen (2003) analyzed a six-week travel diary of 361 persons, González et al. (2008) and Song et al. (2010) analyzed location records from mobile phone usage over several months. Although the studies used different types of mobility data and were conducted over different periods of time, their results with respect to repetitive human movement behavior were similar. Both studies found out that although individuals visit up to 60 (Schlich and Axhausen, 2003) respectively 50 (González et al., 2008) different places, they spent most of their time in only a few locations. More precisely, Schlich and Axhausen (2003) noted that 70% of all visits are made to two to four locations and 90% of all visits of a person are made to eight different locations. Song et al. (2010) determined that people spend about 60% of their time in their top two visited locations (typically home and work) and about 75% in their top five visited locations. Thus, humans shows a high regularity in their daily travel patterns. In addition, Song et al. (2010) affirmed that most people move within a neighborhood of 1 to 10 km and only few people travel large distances of over 100 km per day. This means that human movement behavior is not only repetitive but most of the movement also takes place locally.

Travel behavior further depends on sociodemographic characteristics. In Germany the Federal Ministry of Transport, Building and Urban Affairs commissions a nationwide survey to evaluate the day-to-day travel behavior of German people in regular intervals of several years (Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2010). The study analyzes, amongst others, the travel distance, means of transportation and purpose of traveling with respect to sociodemographic characteristics. The results clearly show differences in travel behavior of different sociodemographic groups. For example, starting at teenager age the mobility of a person increases, while a first decrease is visible in age group 50-59 years. Also, the mobility in terms of average daily travel distance of male and female persons as well as persons living in cities or in rural areas differs.

Clearly, characteristics of human movement behavior will be directly visible in entity-location interaction quantities. For example, a poster campaign in Hamburg will be frequently seen by inhabitants of Hamburg. However, only few people from Berlin will see the campaign. Knowledge about movement behavior therefore helps to interpret visit potential quantities, and the study of visit potential may help to characterize movement behavior. Examples about the relationship of movement behavior and visit potential are given in Section 4.5 when applying visit potential to the application data set.

2.2.4. Trajectory Feature Extraction

For trajectory data analysis characteristics of a single trajectory as well as the relationship between two or more trajectories are important. In this section we will introduce unary trajectory features as well as the relational features distance and topological relationship.

Unary features. Andrienko et al. (2008) group unary features of a trajectory into two basic categories: *moment-related* and *overall* characteristics. Moment-related characteristics can be extracted for each point in time whereas overall characteristics rely on a trajectory interval. The following examples are taken from (Andrienko et al., 2008).

Examples of moment-related characteristics of a trajectory are a moment's

- spatial reference (position),
- temporal reference (time),
- direction and speed of movement,
- change of direction (turn),
- change of speed (acceleration) as well as
- accumulated travel time and travel distance.

Examples of overall characteristics of a trajectory are the trajectory's

- geometry,
- length and duration,
- direction from initial to final position,
- minimum, average and maximum speed as well as
- dynamics of speed and direction.

In case of speed, dynamics refer to periods with constant speed, acceleration or deceleration. In case of direction, dynamics refer to periods of straight, curvilinear or circular movement.

Distance. As trajectories are spatiotemporal objects, distance functions may be defined both over space and time or over one of the dimensions only. Pelekis et al. (2007) make a basic distinction between distance functions relying on spatiotemporal characteristics or on spatial characteristics only. In the case of spatiotemporal characteristics a small distance is given if mobile entities follow similar routes concurrently whereas in the case of spatial characteristics only the similarity of routes is decisive. In addition, derived characteristics of a trajectory such as speed and direction can be considered in the distance function (Pelekis et al., 2007).

Spatial distance functions are, for example, Locality In-between Polylines (LIB) which is a weighted distance of areas in-between trajectories (Pelekis et al., 2011a, 2007), route similarity which iteratively searches for the closest pair of points between two trajectories and augments the average distance of the pairs by a penalty term for unmatched points (Andrienko et al., 2007) and common origin and/or destination (Rinzivillo et al., 2008a). Distance functions considering spatiotemporal characteristics are, for example, average Euclidean distance between the positions of mobile entities over time (Nanni and Pedreschi, 2006), Spatiotemporal Locality In-between Polylines (STLIB) which is the temporal extension of LIB (Pelekis et al., 2011a, 2007) and route similarity + dynamics which is the temporal extension of route similarity (Andrienko et al., 2007). A number of distance functions from the area of time series or sequence analysis have also been applied to trajectory data. For a comprehensive overview see Nanni et al. (2008) and Pelekis et al. (2011a).

Topological relations. In order to define topological relations between trajectories we have to consider their relation in space as well as in time. We have already introduced topological relations between spatial objects in Section 2.1.4. For time intervals Allen (1984) has defined the seven temporal relations, which are depicted in Figure 2.17. Note that including inverse relationships 13 relations exist in total.

Claramunt and Jiang (2001, 2000) combine spatial and temporal topological relations to express spatiotemporal topological relations between regions along with possible transitions over time. Figure 2.18 shows an excerpt of all possible topological relations based on the spatial topological relations equals and meets and the temporal topological relations equals,

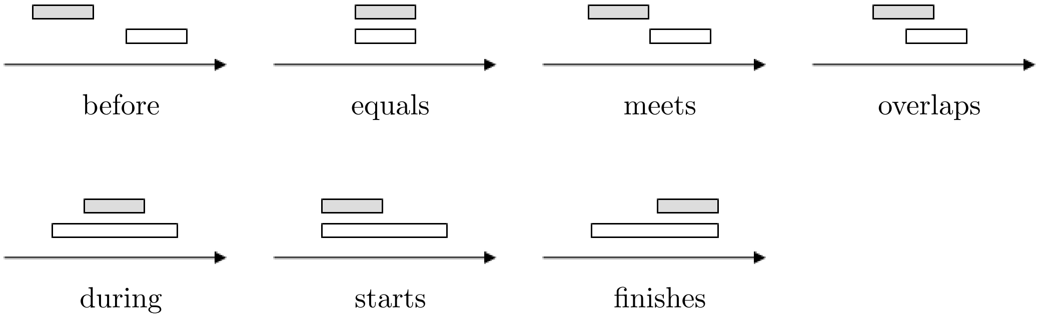


Figure 2.17.: Topological relations of two time intervals

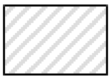
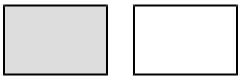
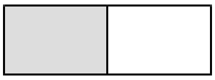
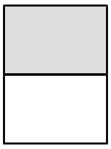
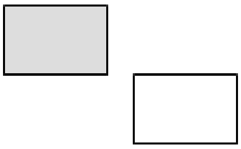
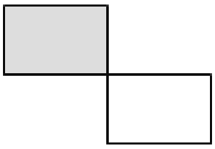
	temporal rel.		
spatial rel.	equals	before	meets
equals	 equals	 disjoint	 meets
meets	 meets	 disjoint	 meets

Figure 2.18.: Examples of spatiotemporal topological relations of two regions

before and meets. Note that for spatiotemporal objects a total of eight topological relations exists (equals, meets, inside, contains, covers, covered by, overlaps, disjoint).

For trajectory data the complexity of spatiotemporal topological relations and possible transitions can be reduced. For a given moment in time two trajectories are represented by point objects. Thus, the possible spatial topological relations reduce to equals and disjoint (Hallot and Billen, 2008). Hallot and Billen (2008) use this fact to represent all possible topological relations between two trajectories in a two-dimensional space, one dimension representing degenerated geographic space (i.e. equals and disjoint) and the other dimension representing time. In this space Hallot and Billen (2008) analyze all possible relations between two lines as defined by Egenhofer and Herring (1990) and remove impossible relations. Such relations arise, for example, because an object cannot move backwards in time. As a result, Hallot and Billen (2008) obtain 25 relations that describe all possible topological relations between two trajectories. Figure 2.19 depicts four example relations along with their natural language interpretations. The x axis represents time whereas the y axis represents the spatial topological relation at each time instance.

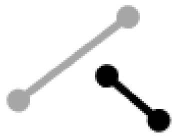
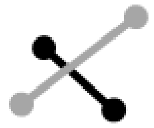
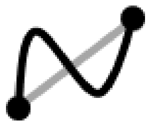
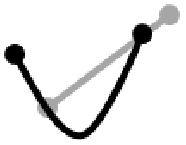
			
A and B never meet.	A and B meet during their coexistence.	A and B come into and pass out of existence together. They meet during their coexistence.	A meets B when B comes into existence, B meets A when A passes out of existence.

Figure 2.19.: Examples of spatiotemporal topological relations of two trajectories

2.2.5. Trajectory Preprocessing and Annotation

Several technologies such as the Global Positioning System (GPS), Global System for Mobile Communications (GSM) or Radio Frequency Identification (RFID) can be used to record the position of a mobile entity over time. These positions are sample points of the true trajectory function of the entity and may differ in their temporal and spatial resolution. In addition, the recorded trajectories are subject to measurement errors. In this section we will focus on the preprocessing and annotation of trajectory data collected via GPS as this is the most commonly provided form of mobility data. Note that in the literature the term *trajectory* is overloaded with several meanings. On a conceptual level it can refer to the trajectory function of an object, on an data focused level it can refer to the sequence of timestamped position records or a further processed form of the data.

In the database literature trajectory preprocessing is also referred to as *trajectory construction*. It comprises the steps data cleaning, data compression and data segmentation (Yan et al., 2011a). During data cleaning measurement noise and outliers are removed. Data compression reduces the amount of trajectory sample points because frequent position records easily lead to large volumes of data. Meratnia and de By (2004) group compression techniques into four categories following existing work on compression of time series data (Keogh et al., 2001): top-down, bottom-up, sliding window and open window. Top-down and bottom-up approaches partition a trajectory respectively merge data points of the trajectory until some stop criteria are met. Sliding window approaches compress data according to a fixed window size while open window approaches allow for a variable size of the window. A recent evaluation of compression techniques is given in (Muckell et al., 2010). Data segmentation is the division of a trajectory into meaningful sub-trajectories. The main two reasons to perform trajectory segmentation in the literature are the extraction of movement sequences and the annotation of trajectories. Movement sequences are commonly used in trajectory data mining where it is more meaningful to compare individual trips of persons than the movement of a whole day as, for example, in cluster analysis. The annotation of trajectories requires to identify homogeneous sub-trajectories with respect to certain characteristics. For example, a stop indicates the performance of some activity while a movement sequence may be further divided according to the mode of transportation. Different strategies for trajectory segmentation exist, including the detection of stops (Schuessler and Axhausen, 2009; Stopher, 2009; Marketos et al., 2008), the detection of sequences with homogeneous movement characteristics (Buchin et al., 2010) or the detection of representative sub-trajectories in a trajectory database (Panagiotakis et al., 2011).

Trajectory annotation means to lift a trajectory from its representation in physical space to a semantic space. A semantic trajectory has the advantage that it contains information about *why* and *how* people move. Trajectories can be annotated with different types of semantic information, among them are stop locations and activities (Alvares et al., 2007; Liao et al., 2007; Zhou et al., 2007), means of transportation (Schuessler and Axhausen, 2009) or the traversed segments of the street network (Newson and Krumm, 2009; Quddus et al., 2007; Brakatsoulas et al., 2005), see Figure 2.20. The latter task, assigning sequences of timestamped positions to the street network, is also known as map matching. Computer supported or even automated annotation of trajectory data is very helpful for mobility studies because it reduces the amount of interview time and the burden on the survey participant (Wolf et al., 2001). Guc et al. (2008) developed a tool to support the manual trajectory annotation. The annotation process is facilitated by features such as a zoom-enabled timeline, trajectory animation and storage of placemarks. An automated framework for trajectory annotation has recently been proposed by Yan et al. (2011a), and a first step toward online segmentation and annotation of trajectories is presented by Yan et al. (2011b).

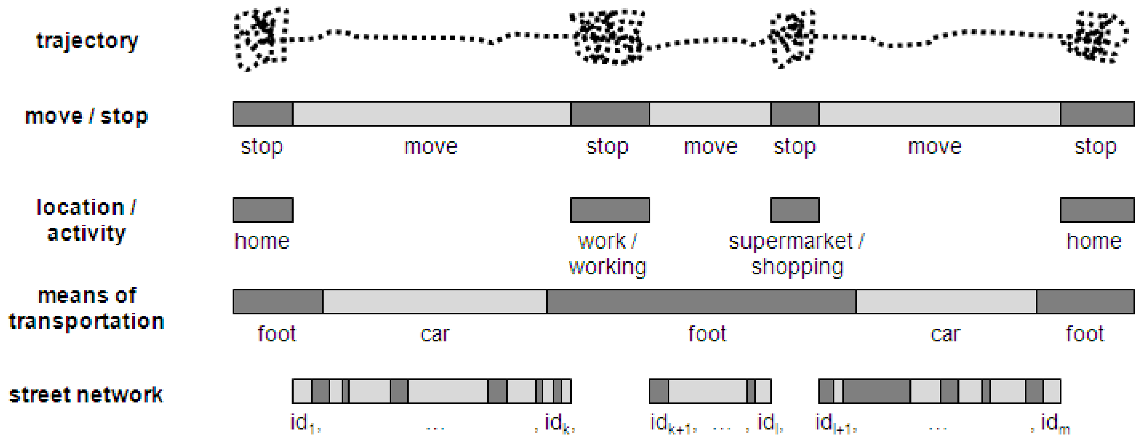


Figure 2.20.: Different types of trajectory annotation

2.2.6. Analysis Tasks and Methods

Mobility mining analyzes the movement of mobile entities and their interaction with the environment. It is thus not a generic extension of spatial data mining to the temporal dimension, but instead a subsection of the broader field of spatiotemporal data mining. In general, mobility mining assumes that mobile entities (in the database literature also called moving objects (Güting and Schneider, 2005)) are represented by moving point objects, reducing the objects to their center of mass. In this section we give an introduction to the most prominent analysis tasks and data mining algorithms for mobile entities. For a comprehensive overview of the topic see (Nanni et al., 2008).

Clustering. The clustering of trajectories, i.e. the segmentation of trajectories into groups with similar movement characteristics and determination of group representatives (Giannotti and Pedreschi, 2008), generally takes place on a set of trajectory sections rather than on the lifelong trajectories of entities. These sections, by definition trajectories themselves (see Definition 2.2.3), are typically selected to represent semantically meaningful movements as, for example, the trip from home to work (Spaccapietra et al., 2008; Guc et al., 2008). Techniques to identify such sections rely on the analysis of stops and moves (Spaccapietra et al., 2008; Alvares et al., 2007) as described in Section 2.2.5.

State-of-the-art clustering techniques for trajectories rely on traditional clustering algorithms and put their main effort into the definition of meaningful similarity functions. Nanni and Pedreschi (2006) recommend the usage of density-based algorithms for trajectory clustering because they are able to form clusters of arbitrary shape, are robust to noise and do not require the number of resultant clusters as input parameter. All three characteristics are important for trajectory data. Nanni and Pedreschi (2006) as well as Rinzivillo et al. (2008a) use the OPTICS algorithm by Ankerst et al. (1999) for density-based trajectory clustering.

Depending on the analysis goal, different similarity functions for trajectories can be applied. As stated in Section 2.2.4 distance functions relying on spatial or on spatiotemporal characteristics can be distinguished. In (Pelekis et al., 2011a, 2007) and (Andrienko et al., 2007) the authors first define similarity functions based on spatial distance and then extend the functions to the spatiotemporal domain.

A common approach for the clustering of trajectory data is the stepwise, visually aided application of clustering algorithms. The gradual refinement of clusters has the advantage that it breaks down complexity with respect to comprehensibility as well as to computational resources (Rinzivillo et al., 2008a; Andrienko et al., 2009).

One further research direction of trajectory clustering, which the interested reader may like to follow, is the clustering of trajectories under uncertainty. Location uncertainty is an inherent characteristic of trajectory data due to measurement errors. Pelekis et al. (2011c) introduce a representation for the uncertainty in trajectory data and provide a clustering algorithm based on fuzzy logic.

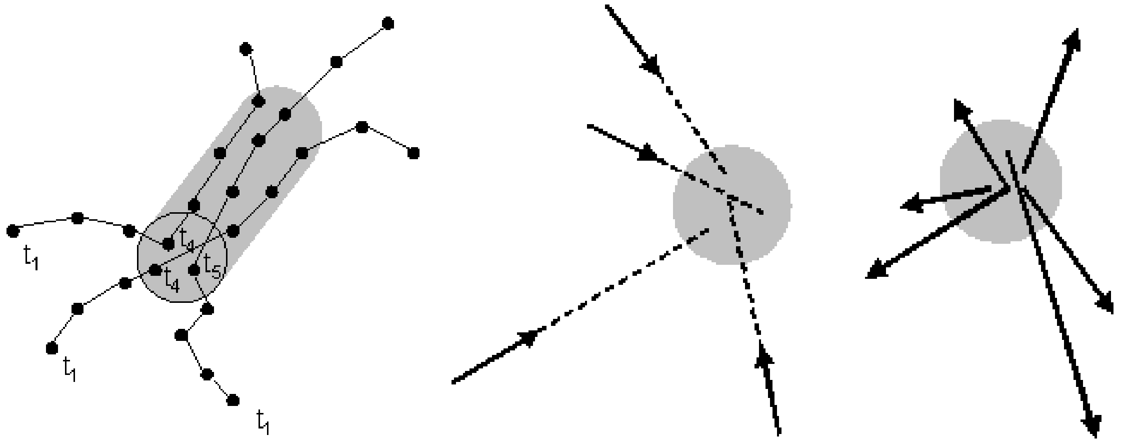
Pattern Analysis. Trajectory patterns describe interesting behaviors of groups of moving objects. Hereby, two tasks are considered in the literature: the detection of frequent movement patterns and the detection of pattern occurrences. In the first case the goal is to identify the pattern itself, for example, a frequent movement from location A to location B to location C. In the second case the goal is to identify when and where a specific pattern occurs and which entities participate in it, for example, the convergence of a group of entities to some location. In the following we will discuss both data mining tasks in more detail.

Mining frequent trajectory patterns is the task to extract (parts of) routes that are frequently followed by the objects of interest. Similar to the task of trajectory clustering, the mining of frequent trajectory patterns can rely on the spatial characteristics of the trajectories only or on their spatiotemporal characteristics. In the first case only the sequence of the visited locations are considered as implemented by Cao et al. (2005) and Yang and Hu (2006). In the second case also the transition times between the locations are important as implemented by Giannotti et al. (2006), Giannotti et al. (2007) and Kang and Yong (2010). All three authors follow the concept of temporally annotated sequences (TAS) first introduced by Giannotti et al. (2006). A TAS is a sequence of items along with a sequence of transition times (i.e. the temporal annotations) between the items. The items hereby represent geographic locations. Giannotti et al. (2007) generalize the concept of TAS to trajectory patterns (also called T-patterns) by substituting items with pairs of coordinates in two-dimensional geographic coordinate space and by using a neighborhood function in order to specify the containment of a T-pattern in a trajectory. One challenge of trajectory pattern mining is the handling of continuous geographic coordinate space. Clearly, two persons that travel along a street will not yield trajectories with the same coordinates. Therefore trajectories are either generalized or discretized in the literature. Cao et al. (2005) and Kang and Yong (2010) generalize trajectories by applying line simplification techniques whereas Giannotti et al. (2007) discretize trajectories based on regions of interest. These regions are either obtained from external points of interest (POI) databases or by mining often visited locations from the trajectory data set. Finally, Lee et al. (2009) uses a regular grid in order to discretize trajectories. The authors provide a graph-based mining algorithm which, however, is restricted to patterns of adjacent cells. On large trajectory data sets the extraction of frequent patterns can be very time consuming. Leonardi et al. (2009) therefore developed a method to aggregate and store frequent trajectory patterns in a trajectory data warehouse.

The detection of pattern occurrences naturally requires a specification of the pattern to be detected. In the literature the most commonly described patterns for this task are group patterns. Group patterns refer to objects that conform to a specified collective behavior and may involve derived information concerning the whole group of objects (e.g. average speed). Intuitively, a group is formed by a number of objects that stay close in space for a meaningful period of time. In Wang et al. (2003) physical proximity is delimited by a maximum distance threshold between each pair of objects. If k objects stay close for a given minimal threshold of time, they form a so-called k-group pattern. The algorithm of Wang et al. (2003) discovers mobile group patterns on trajectory data where the location is recorded at fixed, regularly spaced points in time. A generalization to irregularly spaced trajectories, assuming linear movement, is provided in Hwang et al. (2005). In addition to the general definition of spatiotemporal closeness, a group can be specified by some characteristic internal structure. For example, a group could be headed by some individual which anticipates the group motion. This pattern is called leadership (Figure 2.21 left) and was introduced by Laube and Imfeld (2002) under the general concept of *relative motion* (REMO). Other basic spatiotemporal group patterns of REMO are flock, convergence and divergence. A flock is a group of objects which move in the same direction, while convergence and divergence describe the simultaneous motion of objects to or from some point in space (see Figure 2.21 middle and right). Algorithms for the efficient computation of REMO patterns are provided in Laube et al. (2004) and Gudmundsson et al. (2007). One disadvantage of REMO is that the patterns are detected only for single snapshots in time. Extensions were therefore proposed by Benkert et al. (2008) for flock patterns and by Andersson et al. (2008) for leadership. Both extensions follow the same principle and require that each pattern lasts for a given interval of time. In addition, Gudmundsson and van Kreveld (2006) provide algorithms for the computation of flocks of maximal duration. One further extension of the flock pattern is provided by Wachowicz et al. (2011), who define a *moving flock* for the analysis of pedestrian movement.

So far, the described patterns rely on a stable group of objects. Yet, a pattern may continue over time although its group members change. For example, a traffic jam can prevail for several hours while new cars continuously arrive at one end and escape at the other. This phenomenon is called a *moving cluster* and refers to a cluster that retains its density (or other similar properties, like cluster size or diameter) although different objects participate in the cluster during its lifetime (Kalnis et al., 2005). Another example of patterns with changing objects are *mixed-drove patterns* (Celik et al., 2008, 2006). These patterns describe relationships between types of objects and therefore allow the exchange of individuals of the same type.

Location Prediction. During the past years the reliable prediction of future locations of moving objects has been of interest in mainly two research areas, namely moving object database systems and wireless communication networks. Moving object databases employ future locations of objects for example in range or nearest neighbor searches of *forecasting queries*. These queries require sophisticated structures for indexing future positions of moving objects. In wireless networks, the anticipation of future movement is important to enable an efficient



(a) Leadership

(b) Convergence

(c) Divergence

Figure 2.21.: Relative motion patterns

allocation of network resources.

In the database literature, forecasting queries rely on indexing structures for current positions and motion vectors. Given the current location l_c and the velocity vector v_c of an object, the future position after time Δt can be computed as $l_f = l_c + v_c \Delta t$. The TPR-tree (Saltenis et al., 2000) and its optimized version TPR* (Tao et al., 2003a) have been developed to handle predictive range queries (Saltenis et al., 2000), time-parameterized nearest neighbor queries (Tao and Papadias, 2002) or reverse nearest neighbor queries (Benetis et al., 2006) over the future positions of moving objects. The underlying assumption of all techniques is that the involved objects continue their motion with the given velocity vector until the ending time of the query interval. This assumption applies for linear movement in unobstructed spaces, as for example for ships, planes or weather phenomena. However, it is not reasonable for street networks where objects change their direction and speed within short time intervals (Tao et al., 2003b).

Such unstable conditions are met in wireless communication networks where mobility management serves mainly two tasks. First, appropriate resources must be allocated to guarantee a smooth transfer of service if a user changes from one cell to the other. Second, when an incoming call arrives, the network should page as few cells as possible within a given location area. Both tasks require to anticipate the motion of users for the near future. Several algorithms have been investigated to accomplish this task. Biesterfeld et al. (1997) and Liou and Huang (2005) train neuronal networks based on the location area or x,y-coordinates respectively. Liang and Haas (2003) apply Gauss-Markov models based on the location and velocity of objects. A common approach for location prediction is to analyze historic trajectories, derive predominant pattens and apply the most similar pattern to the trajectory in question. Such an approach is followed by Katsaros et al. (2003) and Yavas et al. (2005), who apply clustering and sequential pattern mining respectively to extract patterns. A comprehensive study and comparison of methods for location prediction in wireless networks can be found in (Cheng et al., 2003) and (Song and He, 2006). Outside the area of wireless communication networks Monreale et al. (2009) presented an approach to predict the next location of a user based on trajectory patterns of frequently visited locations. In a first step Monreale et al. (2009) derive trajectory patterns from historic data in a spatial and temporal region relevant to the trajectory in question. Next, they build a decision tree from the patterns and, finally, predict a location based on the best mapping path in the tree with respect to the given trajectory.

In addition to location prediction in the near future, an important research task is to anticipate the most likely route and destination of a moving object. For example, location based services can offer more sophisticated services when knowing which locations a user will pass and whether the user is on the way to work or to the supermarket. The general assumption behind the prediction of routes and destinations is that people follow daily or weekly routines. Usually, people visit only a few places frequently, as for example their home, workplace or favorite restaurant. In addition, people are creatures of habit and select their present route from a small set of candidate routes. Karimi and Liu (2003) adapt a transition matrix to personal preferences and are thus able to predict the most likely route and destination of a single person within a given time frame. While Karimi and Liu (2003) base their predictions solely on routing information, Laasonen (2005) incorporates residence times into his model. The author first detects places where a user spends a comparatively large amount of time. These places form the set of all possible destinations and delimit individual routes. Similar to (Katsaros et al., 2003), Laasonen (2005) clusters historic routes and compares the obtained types with the present trajectory. The predicted destination belongs to the most similar trajectory type and can optionally be conditioned on the time of day and day of week.

2.3. Summary

In this chapter we provide an overview to spatial and mobility data analysis. We introduce basic geographic concepts and provide a precise definition of all concepts that are relevant for the later chapters of this thesis. Most important are the definitions of mobile entities (Definition 2.2.4), trajectories (Definition 2.2.3), geographic locations (Definition 2.1.5), geographic coordinate space (Definition 2.1.4) and temporal coordinate space (Definition 2.2.2). Furthermore we describe important characteristics of spatial and mobility data which are important to understand spatial phenomena and movement behavior and which have an impact on data analysis and data mining algorithms. The most important characteristic of spatial data is spatial autocorrelation, which defines the high correlation of objects that are close in geographic space. The two most important characteristics of human movement behavior are repetitiveness and locality. Finally, we describe feature extraction and preprocessing methods as well as the most important data mining tasks and methods for spatial and mobility data.

Equipped with this background knowledge the reader will be well prepared for the formalization and estimation of entity-location interaction quantities as described in the following chapters.

3. Application Context

A theory is the more impressive the greater is the simplicity of its premises, the more different are the kinds of things it relates and the more extended the range of its applicability.

(Albert Einstein)

This section presents the application context of this thesis. As already mentioned in the introduction, this thesis is motivated by two commercial projects at Fraunhofer IAIS which are concerned with audience measurement in outdoor advertising. Both projects provide the data basis for our experiments and will be used throughout the thesis for demonstration purposes. However, our results are not limited to outdoor advertising. They belong to the broader area of mobility data analysis and may be applied in similar contexts. However, the application context will help the reader to gain an understanding of the applicability of the theoretical and practical results of this thesis. In the following we therefore provide an overview on performance measurement in outdoor advertising and the audience measurement studies in Switzerland and Germany.

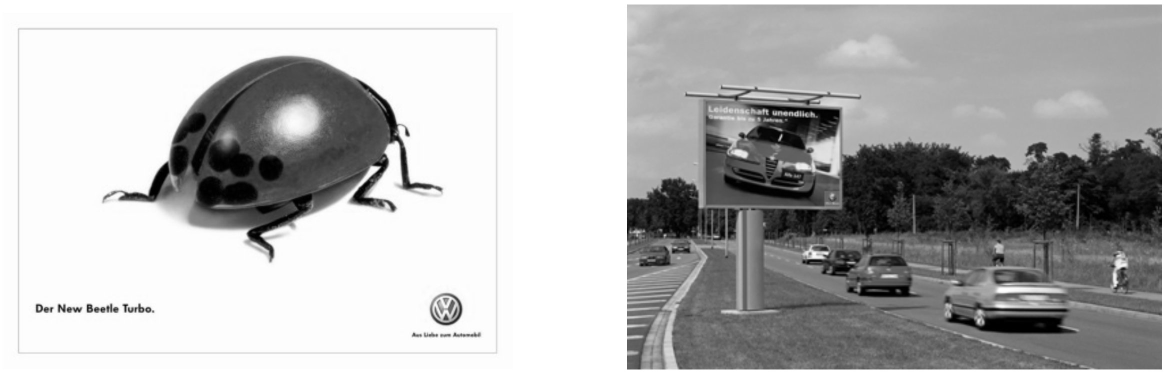
Section 3.1 introduces poster performance indicators. Section 3.2 presents the two audience measurement studies in more detail and Section 3.3 presents a collection of research challenges which are related to the applications and may interest the scientific community. We conclude the chapter with a short summary.

3.1. Audience Measurement in Outdoor Advertising

3.1.1. Motivation

Outdoor advertising is one of the oldest forms of promotion for goods, services or events. Figure 3.1 shows a typical example of todays advertising campaigns and their locations. Until today outdoor advertisement plays a major role in the advertising landscape. In Switzerland and Germany, the home countries of the two commercial projects at Fraunhofer IAIS, the outdoor advertising industries generated net sales of 608 million Swiss Franc (about 497 million Euro) (Stiftung Werbestatistik Schweiz, 2011) and 766 million Euro (Fachverband Außenwerbung e.V., 2011) in 2010, respectively. These numbers correspond to about 17% and 4% of the total national advertisement net sales, respectively.

Consequently, the pricing of poster sites is a critical business task and must be justified by performance indicators. A poster site is the more valuable the more people look at the advertisement and, hence, the more people that pass the location. The Swiss and German outdoor advertising industries commissioned large mobility surveys to capture the movement behavior of the population, resulting in unique mobility data sets. Given the mobility data and the location of poster sites, individual passages with advertising campaigns can be calculated and performance indicators can be derived. In the next section we will introduce the most common poster performance indicators in outdoor advertising.



(a) (b)

Figure 3.1.: (a) PlakaDiva award winner 2005, category best poster and (b) poster site along a street; source of left figure Out-of-Home Research & Services GmbH, Fachverband Aussenwerbung e.V. (2009)

3.1.2. Performance Indicators

Which quantities measure the performance of an advertising campaign? The number of people that see the advertisement is obviously relevant. Next, one may ask which percentage of the population sees an advertisement and how often a person sees it on average. In addition, advertisers are interested to refine such indicators with respect to sociodemographic groups in order to perform targeted advertising.

Principally, the above quantities can be divided into indicators measuring the dispersion of the displayed information and the frequency with which it is received (Sissors and Baron, 2002). In outdoor advertisement the indicators reach, coverage, effective reach, opportunities to see, gross rating points and gross impressions have become standard criteria to evaluate the performance of poster campaigns. In general, all indicators relate to a given target audience (e.g. all residents of a specified city, all female residents, etc.), a given poster campaign and a given period of time. We will explain the different indicators and their relationships below. We have assembled the descriptions using the following sources: Sissors and Baron (2002), Koschnick (2011), Swiss Poster Research Plus (SPR+) (2011a) and Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma) (2011). Table 3.1 contains a short summary of the performance indicators. A precise definition of the indicators will be given in Chapter 4 after providing the formal concept of entity-location interaction quantities.

Coverage. Coverage is an indicator of dispersion. It states how many different people pass at least one of the posters of a campaign within a specified period of time. It is usually defined as percentage of a given population, but it may also report the absolute number of affected persons. Coverage is a preliminary state of reach (as defined below) because it requires only that people pass through the visibility area of a poster. It does not request that people actually look at the advertisement. Coverage is calculated using poster passages, i.e. the geographic intersection of trajectories and poster locations.

Reach. Reach is similar to coverage an indicator of dispersion. However, it states how many different people actually see an advertisement within a specified period of time. It is one of the predominant indicators used in outdoor advertising. As trajectory and poster location data only reveal geographic passages, weights are introduced to model the attention of passers-by. The weights account for speed and angle of passage, the size of a poster, illumination criteria, etc. and range between zero and one. Weighted passages are also called poster contacts, or simply contacts.

There are two possibilities to interpret poster contacts: as dose or as probability. In the dose model, each (even unconscious) contact is said to contribute to the perception of an advertisement. Therefore the contacts are accumulated over time. A value larger than one indicates that a person is reached by a campaign. This model is applied in the Swiss audience measurement study. The second possibility interprets the weights as contact probabilities given that a poster location has been passed. This approach requires a simulation model similar to the German audience measurement study to evaluate poster contacts. It results either in a full contact or in no contact per passage.

Effective reach. How often does a person need to see an advertisement until it is affected by the message? This question is not resolved until today. Some studies show that at least two or three repetitions are necessary to pass a message, some practitioners object that already a single contact causes noticeable response. Some studies show that response declines when the contact frequency becomes too high (Sissors and Baron, 2002). In general, the optimal number of contacts depends on several factors such as the type of product, the brand awareness within the population, the creative message, the advertisement medium, the personal attitude of a target consumer, etc. Therefore marketing experts use individual criteria when planning a campaign.

Effective reach is an answer to this dilemma. It simply states the reach of a campaign at a specified level of repetition. If a planner requires three contacts, the effective reach states the percentage of population which have seen the advertisement at least three times. Naturally, with increasing contact numbers, the effective reach declines. In order to use a clear terminology, we will state reach according to different contact classes, with contact class one corresponding to the general meaning of reach.

Opportunities to see (OTS). OTS forms the counterpart of reach on the frequency side of evaluation. It states the average number of poster contacts of all persons that see a poster of the campaign. Similarly to reach, OTS can be calculated with respect to different contact classes. The OTS of contact class k specifies the average number of visits for persons with at least k contacts. When considered over contact classes, OTS increases monotonically with increasing contact class.

Gross rating points (GRP). Gross rating points refer to a contact volume. They state the average number of contacts that 100 persons of the whole target audience produce. Given the reach (in percent) and OTS for contact class one of a campaign, GRP is calculated as:

$$GRP = reach \cdot OTS. \tag{3.1}$$

Gross impressions. Gross impressions state the total number of poster contacts that a target audience achieves within a given period of time. Thus, gross impressions are similarly to GRP a contact volume, however, they are not normalized to a fixed number of persons. Given the GRP and the population of the target audience, gross impressions are calculated as:

$$gross\ impressions = \frac{GRP \cdot \mid target\ audience \mid}{100}. \tag{3.2}$$

Table 3.1.: Summary of performance indicators in outdoor advertisement (for a given campaign, target audience and time span)

coverage	Percentage or size of target audience which passes at least one poster of a given campaign.
reach	Percentage or size of target audience which sees at least one poster of a given campaign.
effective reach	Reach according to a specified contact class.
opportunities to see	Average number of contacts a person of a specified contact class produces.
gross rating points	Number of contacts that 100 persons of the total target audience produce on average.
gross impressions	Total number of poster contacts produced by the target audience.

In order to illustrate the defined indicators, consider the fictitious contact distribution in Table 3.2. The table shows for each contact class the number of reached persons in absolute numbers and in percent as well as the total number of generated contacts. The target population consists of 10,000 people, that generate 28,100 poster contacts (gross impressions) in the given time period, for example, one week. During this time 20% of the target audience does not see any poster of the campaign, resulting in a reach of 80%. On average each person with more than one contact sees 3.5 posters of the campaign (OTS). Considering also persons without poster contacts, 100 persons of the audience see on average 281 posters (GRP).

If a single poster contact is not sufficient to carry an advertising message, the performance indicators can be evaluated for a higher contact level. For example, assuming a minimum of three poster contacts, the effective reach drops to 46.6% and OTS increases to five contacts per person.

Table 3.2.: Example contact distribution

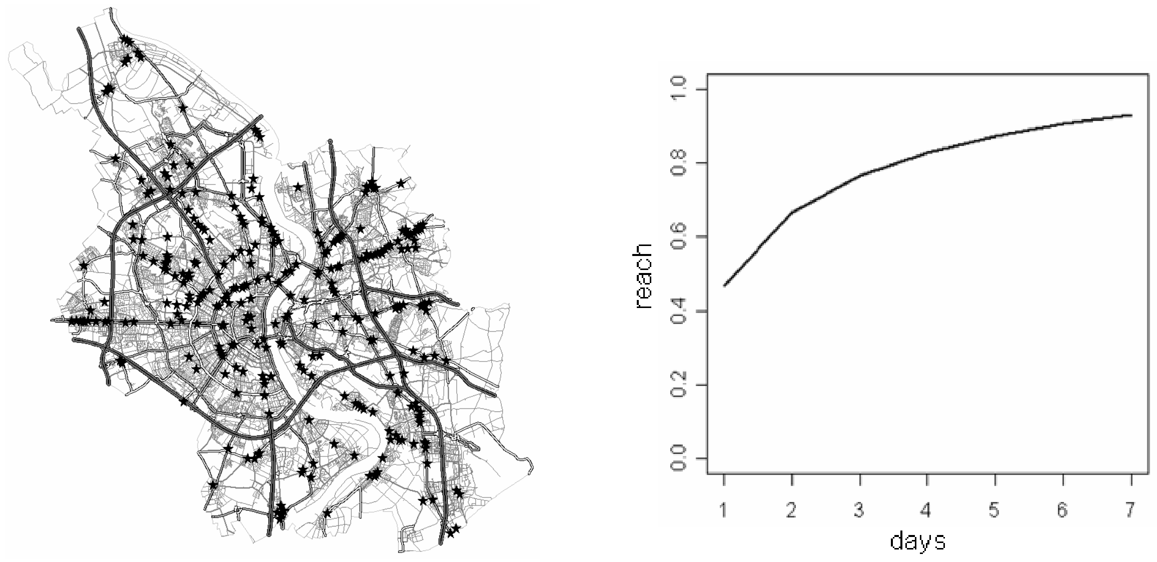
frequency of exposure	number of exposed persons	percent of exposed persons	generated contacts
0	2,000	20.0 %	0
1	1,800	18.0 %	1,800
2	1,540	15.4 %	3,080
3	1,160	11.6 %	3,480
4	980	9.8 %	3,920
5	740	7.4 %	3,700
6	750	7.5 %	4,500
7	620	6.2 %	4,340
8	510	4.1 %	3,280
total	10,000	100.0 %	28,100

3.1.3. Characteristics of Poster Performance by Example

Having introduced poster performance indicators in the previous section, we will now demonstrate how the spatial distribution and location selection of poster sites influences the performance of a campaign. This is also the reason why the individual evaluation of poster campaigns is so important in outdoor advertising.

The performance of a poster campaign clearly depends on the size of the campaign and the traffic frequency of the chosen locations (May et al., 2008a,b). However, it also depends on the distribution of poster locations. An example calculation of the German audience measurement study which was published in a press conference in 2008 (Fachverband Außenwerbung e.V., January 16th, 2008) shall demonstrate this characteristic.

Figures 3.2 and 3.3 show on their left side the poster locations of two campaigns in Cologne, Germany. Both campaigns consist of 321 poster sites. However, the first campaign is spread over the whole city while the second campaign concentrates in the northeast of Cologne. Table 3.3 shows the corresponding performance indicators for the population of Cologne and a duration of one week. The performance in terms of gross rating points is similar for both campaigns. However, the reach of the campaigns differs considerably. While the dispersed campaign reaches nearly all inhabitants of Cologne, the clustered campaign reaches little more than half of the population.



(a)

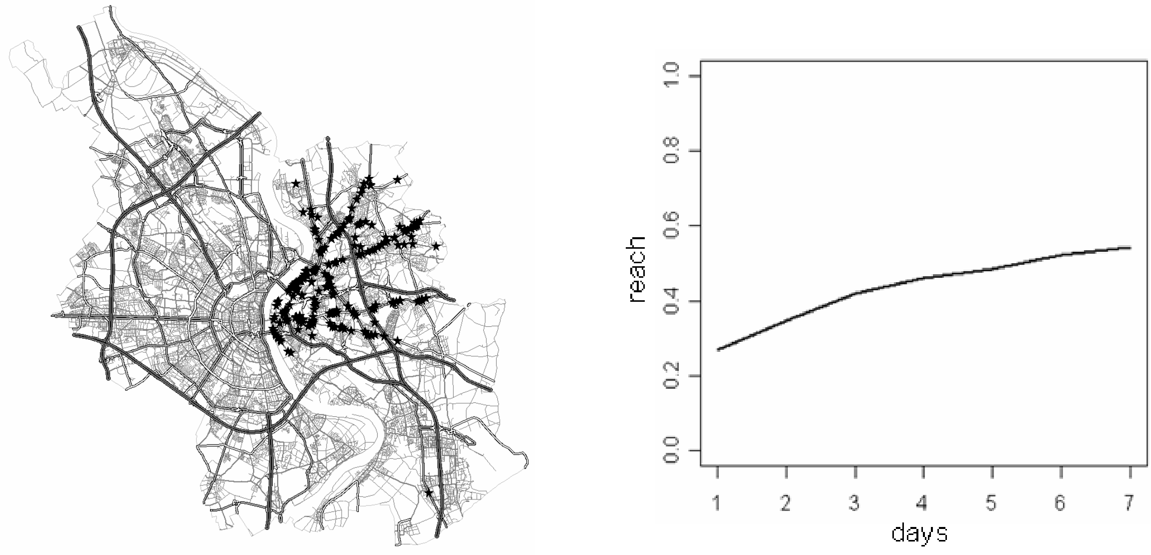
(b)

Figure 3.2.: (a) Dispersed poster campaign in Cologne and (b) reach of the campaign over one week; source of figures: Fachverband Außenwerbung e.V. (January 16th, 2008)

Table 3.3.: Performance indicators of dispersed and clustered campaigns

campaign	**# posters**	**reach**	**OTS**	**GRP**
dispersed	321	92%	10.2	953
clustered	321	54%	20.6	1115

This result is not surprising because people move mostly in a restricted geographic space (see also Section 2.2.3), and people of southern Cologne are unlikely to visit the northeastern part very often. This is especially true as Cologne is divided from north to south by the river Rhine. The major part of the city center resides on the western side of the river, and therefore the usual traffic between all circumjacent periphery and the center of the city contributes only little to the performance of the clustered campaign. However, if a person passes a poster of the clustered campaign, he / she is likely to pass it several times. This is due to the spatial proximity of the posters, the spatial limitation and repetitive structure of daily routes, and it is reflected by high opportunities to see. Further examples demonstrating the influence of



(a) (b)

Figure 3.3.: (a) Clustered poster campaign in Cologne and (b) reach of the campaign over one week; source of figures: Fachverband Außenwerbung e.V. (January 16th, 2008)

the spatial distribution of poster campaigns using the Swiss mobility study can be found in (Hecker et al., 2010a) and will also be given in Section 4.5.

Figures 3.2 and 3.3 on the right side show the development of reach over time. For contact class one it is characteristic that the increments of growth are steep in the beginning and level off with increasing time. Again, this results from the spatial limitation and repetitive structure of individual movement. Some people visit a certain geographic area regularly and will be reached within the first days. Some people visit the given area seldom and it requires more time until they are reached. Finally some people never visit the area and the reach levels off.

3.2. Two Case Studies

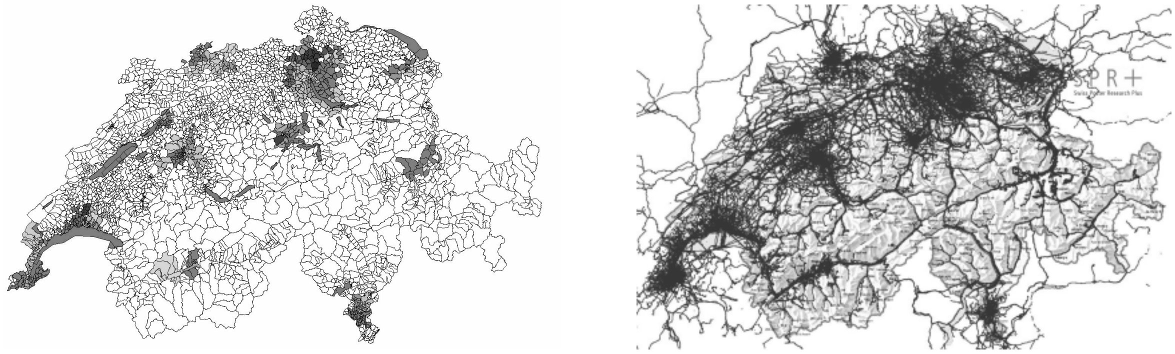
In this section we will introduce the two commercial projects at Fraunhofer IAIS with the Swiss and German outdoor advertising industries. In particular we will give an introduction to the data sets of the application, which are used throughout this thesis for demonstration and experimentation.

3.2.1. Swiss Audience Measurement Study

Swiss Poster Research Plus[1] (SPR+) is a neutral research organization of the Swiss outdoor advertising industry. The first pilot mobility study to measure the performance of poster campaigns was conducted in the conurbation of Winterthur in 2003. A representative sample of persons was selected and equipped with a GPS device for a period between 7-10 days. Since then further GPS studies have been conducted including the largest metropolitan areas in Switzerland as well as a number of smaller conurbations. In total the survey includes more than 10,000 participants which form a representative sample for about two thirds of the Swiss

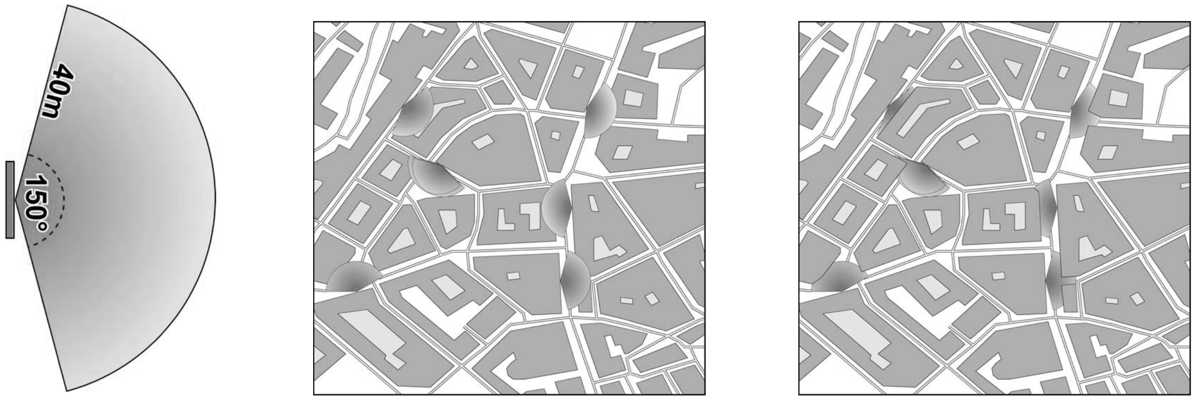
[1] http://www.spr-plus.ch

population. Figure 3.4 displays the 12 Swiss conurbations with GPS measurements and the resulting GPS traces.



(a) (b)

Figure 3.4.: (a) Swiss conurbations included in the GPS survey and (b) resulting total mobility measurements; source of figure on the right: Swiss Poster Research Plus (SPR+) (2011a)



(a) (b) (c)

Figure 3.5.: (a) Standardized visibility area of a panel, (b) overlay of visibility areas and building layer and (c) visibility areas after intersection with building layer; source of figures: Swiss Poster Research Plus (SPR+) (2011a)

A second part of the empirical data contains information about poster sites. In total, the study includes about 50,000 sites. Besides geographic coordinates, a visibility area for each panel is defined from within which the poster is likely to be seen (see Figure 3.5 left). In order to adapt the visibility area to individual location criteria of a poster site, the visibility areas are intersected with a building layer (see Figure 3.5 middle and right). By the intersection viewing obstacles and the resulting dead angles are cut out.

Given the trajectories of an individual and the visibility area of a poster panel, all resulting passages can be calculated by geographic intersection. However, passing the visibility area of a panel does not imply that a person actually looks at the poster. Depending on passage angle, speed, time of day (only some posters are illuminated at night) and the number of panels at the location (many panels increase the distraction), each passage is weighted. For example, Figure 3.6 shows three passages with differing orientation through the visibility area. Depending on the angle with respect to the normal of the poster panel, different weights are assigned. A thus qualified passage constitutes a poster contact, which serves as basis to evaluate reach and gross impressions of poster campaigns.

Further details on the Swiss audience measurement study can be found in (Swiss Poster Research Plus (SPR+), 2011b) and (Pasquier et al., 2008) as well as on the SPR+ website of the mobility study (Swiss Poster Research Plus (SPR+), 2011a). In addition, the following research results have been published in connection with the Swiss mobility study (Hecker et al., 2011a, 2010a,b,c; Liebig et al., 2010; May et al., 2009a,b).

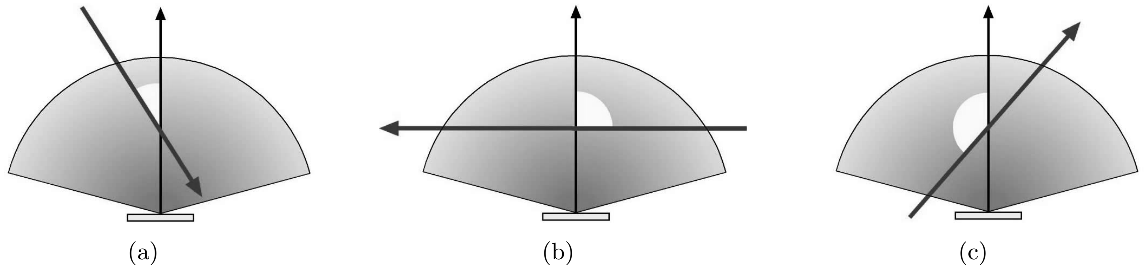


Figure 3.6.: (a) Frontal poster contact: angle < 45°, (b) parallel poster contact: angle 45° - 110° and (c) no poster contact: angle > 110°; source of figures: Swiss Poster Research Plus (SPR+) (2011a)

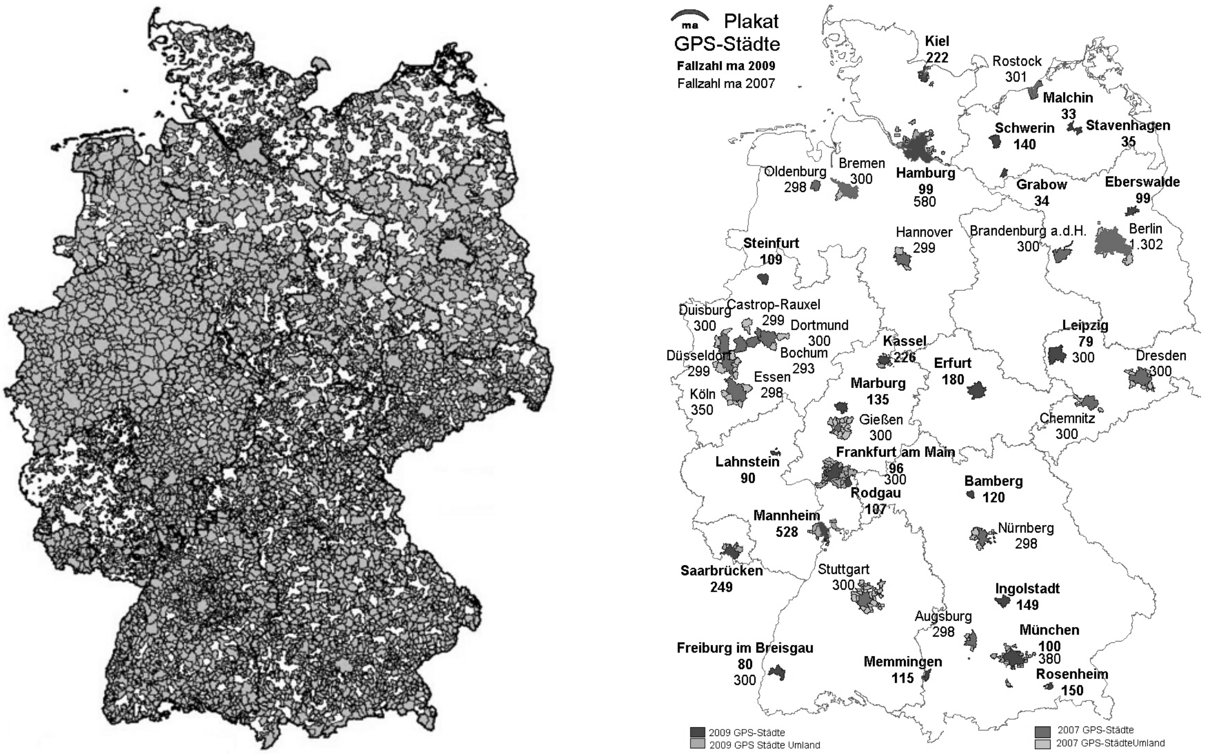
3.2.2. German Audience Measurement Study

The Arbeitsgemeinschaft Media-Analyse e.V.[2] (ag.ma) is a joint industry committee of German advertising vendors and customers. Starting in 2006 it commissioned a yearly nationwide mobility survey as basis for an objective performance evaluation of outdoor advertisements in Germany. The surveys were conducted by using two different observation policies. On the one hand, persons were queried about their movements on the previous day in a Computer Assisted Telephone Interview (CATI). These interviews were performed nationwide to yield a representative sample of the German population. On the other hand, persons from 42 primarily large cities were provided with GPS devices for a period of 7 days to obtain movement information for a longer period of time. The survey has been designed as rolling system and will be continuously extended over the next years. Until mid 2011 41,106 surveys were conducted via CATI and 11,770 surveys via GPS (Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma), 2011). Figure 3.7 left shows the German municipalities with CATI test persons and Figure 3.7 right shows the German municipalities with GPS test persons along with the number of participants.

For the evaluation of poster passages both the mobility data and the poster locations are mapped to the street network. During a CATI interview the mobility information is recorded with the assistance of a routing system and is directly available as sequence of street segments. The GPS data pass through a preprocessing step which closes small gaps by routing and afterwards matches trajectories to the street network. In addition, gaps that last for a whole measurement day are classified within a follow-up survey, where test persons can specify whether they truly stayed at home or maybe forgot (to switch on) the GPS device on the respective day.

In total about 268,000 poster locations are surveyed in the study. According to the type of poster different visibility areas are defined. In order to represent poster locations on the street network the visibility areas are intersected with the network, and each poster location

[2] http://www.agma-mmc.de



(a) (b)

Figure 3.7.: (a) Municipalities with CATI test persons and (b) municipalities with GPS test persons; source of figure on the right: Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma) (2011)

is assigned a set of street segments. Similar to the Swiss measurement study, poster passages are weighted according to visibility criteria. The weights are interpreted as contact probability and are evaluated using repeated simulations during the modeling step.

Further details on the German audience measurement study can be found in (Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma), 2011). In addition, the following research results have been published in connection with the German mobility study (Hecker et al., 2011c; May et al., 2008a,b).

3.3. Application Challenges

In this section we present a number of application challenges (and first solutions) that arise when modeling poster performance (Hecker et al., 2011b). Although addressed for the industrial use, we would like to persuade the scientific reader that these challenges present highly interesting research questions which appear in a number of different problem settings in mobility data analysis.

Missing measurements. Given trajectory data of a representative set of test persons and location information of poster sites, the passages of the test persons with a given poster campaign can easily be extracted by spatial intersection, and the performance of the campaign can be calculated. However, the challenge in calculating poster performance lies in the incompleteness of mobility information. Poster performance indicators are defined over a continuous period of time. Therefore, problems arise if measurements are missing within this time span. GPS mobility studies inherently contain different types of missing measurement data. First, short interruptions occur due to tunnels, street canyons or the warm-up phase of a GPS device. Second, single trips within a day may not be recorded. For example, people may easily forget to carry the device during a short trip to the bakery. Third, complete measurement days may be missing due to several reasons. People may forget to carry or to charge the device. Devices may be defective or people may simply tire of the study and drop out early.

Depending on the kind of missing data, different courses of action must be taken. Short interruptions can be detected during data preprocessing and can be closed using routing algorithms. The second and third type of missing data pose serious problems because they cannot be identified from the data itself. Given a longer period of time without GPS measurements or measurements that stem from a single location, it is impossible to determine whether a person stayed at home or left without the device. A differentiation can only be given by the test persons themselves, for example, in a follow-up survey.

The estimation of poster performance indicators from incomplete mobility data has motivated the research in this thesis. We contribute a formal framework for the definition of entity-location interaction quantities and perform a systematic analysis of missing data methods for the estimation of entity-location interaction quantities. We have published parts of our results in (May et al., 2009a) and (May et al., 2009b).

Extrapolation over time and space. The conduction of GPS surveys is very expensive. Therefore mobility studies are restricted in the measurement period and typically concentrate on the most important locations (e.g. large cities). Consequently, challenges of extrapolation of performance indicators over time and space arise. How can performance indicators be determined for time spans that are longer than the surveying period? How can geographically sparse data be used for performance evaluation? Can performance indicators be predicted for locations without mobility measurements, i.e. is it possible to infer the performance of a campaign from the mobility of another (similar) city? One approach to handle sparse movement data has been developed by Hecker et al. (2011c). The approach achieves a better representation of population movement by increasing the micro-movement variability of a mobility sample.

Pedestrian Movement Model for Indoor Poster Campaigns. GPS technology has the drawback that it cannot be applied indoors due to loss of signal. In Germany and Switzerland many highly frequented posters are situated in public buildings such as train stations or shopping malls and their evaluation is of high interest. Liebig et al. (2010) introduced a method that allows performance measurements for indoor poster sites. The approach is based on obtaining a number of comparably inexpensive frequency counts manually, and on subsequently generating a model for indoor pedestrian movements based on the counts and a network of the possible pathways through the objects.

Optimization of poster campaigns and poster locations. Once poster performance indicators are determined, further questions arise regarding the optimization of poster campaigns. How many posters and at which locations shall be selected to optimize one or several performance indicators? What are good places for new poster sites? Such queries require efficient search and pruning strategies as the set of possible locations is large and location combinations are exponential in number. Can an exact solution be found? Can the exploitation of spatial correlation between trajectories speed up the search?

Usage of secondary data sources. As mentioned above, GPS surveys are very expensive. Therefore, the wish to utilize secondary data sources (e.g. GSM data) has developed over the past years. The usage of such data, however, poses two major challenges. First, movement information is very sensitive and any use of such data has to ensure the privacy of an individual. Therefore, privacy-preserving data mining methods have to be developed in order to exploit secondary data sources. Second, secondary data sources are not necessarily representative for a given population of interest and the contained mobility may be skewed as a result. Therefore the question arises: How can such a bias be detected and, even more important, be compensated?

3.4. Summary

In this section we introduced the application context of this thesis. We provided an introduction to poster performance indicators and demonstrated the dependence of performance indicators on the spatial distribution of a campaign. Further, we introduced the Swiss and German audience measurement studies. The data of both studies will be used in this thesis for demonstration and experimentation. Finally, we presented a number of challenges connected to performance measurement in outdoor advertising which pose demanding research questions to the field of mobility mining.

4. Formalization of Visit Potential

We need to make new symbols
Make new signs
Make a new language
With these we'll redefine the world

(Tracy Chapman, New Beginning)

In the previous chapter we presented the application domain which motivates this thesis and which provides challenging research questions concerning the estimation of entity-location interaction quantities from mobility data. However, this domain is only one among others that employ entity-location interactions. Usually the measured quantities are tailored to specific applications, use context-dependent terminology and are often only informally defined. As a result, a number of quantities have evolved which are not suitable for methodological research and interdisciplinary exchange as their common background is hard to identify. In this chapter we therefore present a systematic definition of entity-location interaction quantities and provide a common vocabulary under the name *visit potential*. We further analyze the relationship of the provided quantities to each other and study their behavior under partitioning of entity and location sets.

This chapter expects that the reader possesses a basic knowledge in statistics and set theory as may be taught in the course of studies in computer science. It further assumes familiarity with spatial and mobility data as presented in Chapter 2. However, we provide references for definitions given earlier so that the reader may recall them individually. The formalization in this chapter is one contribution to this thesis and fundamental for the understanding of the next chapter. Readers that are mostly interested in the estimation of visit potential quantities from incomplete data might want to read Sections 4.1 and 4.2 and possibly Section 4.5 before continuing.

The chapter is organized as follows. Section 4.1 introduces the concept of entity-location interactions and reviews related work. Section 4.2 defines the various visit potential quantities and analyzes their relationships to each other. Section 4.3 studies the defined quantities under partitioning of entity and location sets. In Section 4.4 we show how the general framework of visit potential can be applied to precisely define application-dependent entity-location interaction quantities. As example we use two real-world application domains, namely outdoor advertisement as presented in Chapter 3 and bird tracking. In Section 4.5 we provide example calculations of visit potential. The examples shall help to clarify the provided framework and to develop a better understanding for the demeanor of visit potential when applied to human mobility data. We conclude the chapter with a short summary.

Excerpts of this chapter have been published in (Körner et al., 2010b) and (Hecker et al., 2010a).

4.1. Entity-Location Interaction

This section gives a first introduction to entity-location interactions and visit potential quantities. It provides examples for the usage of visit potential, delineates related work and defines the basic interaction between a mobile entity and a location.

4.1.1. Introduction

Every day people interact with the environment by passing or visiting geographic locations. Such interactions can be recorded by various technologies. On the one hand, technologies can be installed locally as, for example, induction loops, surveillance cameras, Bluetooth sensors or RFID. On the other hand, positioning technologies, such as GPS or GSM, can be used to trace individual movements. Given the geographic position of locations, spatiotemporal interactions can be reconstructed from the trajectories afterwards. By spatiotemporal interactions we simply mean the passage of an area or the visit of a location. We will give a precise definition of such a visit later on. Spatiotemporal interactions between a mobile entity and a given location form only a small part of the daily mobility, yet the knowledge about such interactions is extremely useful. In the following we describe a number of short scenarios in order to demonstrate the capabilities of entity-location interaction quantities.

Typically, applications do not consider the interactions between a single entity and a single location, but are interested in the quantity and structure of the interactions between a set of entities and a set of locations. Such sets can be selected by various characteristics. For example, we may analyze visits with respect to the origin or sociodemographic characteristics of the entities or with respect to the type of location. A traffic application could thus determine which part of traffic in a city is caused by locals and which part by commuters. An example for the structuring of interactions is the identification of regular visitors and their number of repeated visits. Such information is interesting for the owners of restaurants or of retail chains, which can analyze the portion of customers that return on a regular basis. If movement histories are available, such an analysis is not restricted to a single location but can be applied to a set of subsidiaries as well. Further, given movement histories the analysis of interactions allows to model the dependency within a group of locations. Imagine a drug store chain which plans to open a new subsidiary. Of course, the chain prefers highly frequented locations. However, it is not interested to provide alternative shopping facilities for existent customers. Instead, it aims at reaching people which rarely pass any of their present subsidiaries. Such a location can be identified by applying an interaction quantity that states the percentage of people that pass one or more locations of a given set.

These few examples already show the usefulness of mobility-based interaction quantities. However, they also show the variety of domains in which they can be applied and let assume the patchwork of quantities which has emerged. In this chapter we therefore provide a systematic definition and common vocabulary of quantities that express entity-location interactions, which we name *visit potential*.

4.1.2. Related Work

Mobility data in form of trajectories, which can be collected via positioning technologies such as GPS, RFID or GSM, have drawn the attention of the data mining community recently. However, current developments in trajectory data mining concentrate on the analysis of mobility patterns and not on quantities to measure interactions between mobile entities and locations. Algorithms are predominately presented for clustering of (parts of) trajectories (Rinzivillo et al., 2008a; Pelekis et al., 2007; Nanni and Pedreschi, 2006), detection of relative motion patterns (Gudmundsson et al., 2007; Hwang et al., 2005; Laube and Imfeld, 2002) or sequential analysis of movement (Zheng et al., 2009; Giannotti et al., 2007; Yang and Hu, 2006). The latter is in part related to visit potential. Frequent spatiotemporal sequential patterns are sequences of geographic locations that occur at least a given number of times in the trajectories of some mobile entities. Spatiotemporal sequential pattern mining requires in the beginning a set of disjoint geographic locations which represent relevant places for some application. These locations are used to transform the trajectories into sequences of visited locations. Afterwards frequent transitions between the locations are detected. The locations may be provided externally (e.g. a collection of points of interest (POI)), by an analysis of trajectories to identify regularly visited regions or by a combination of both approaches (Giannotti et al., 2007). Giannotti et al. (2007) and Palma et al. (2008) provide algorithms to extract such a set of locations directly from a set of trajectories. Alvares et al. (2007) provide an algorithm to extract all stops from the trajectories of a mobile entity for a given set of locations. Algorithms for the sequential analysis of movements are related to visit potential insofar, as they also consider movement information only with respect to a set of relevant locations. Our definition of visits, as will be given below (see Definition 4.1.5), is similar to the definition of stops by Alvares et al. (2007), however, it is made on a conceptual level and is not restricted to data in the form of space-time points. The work of Zheng et al. (2009) differs from frequent sequential pattern mining as they consider not the frequency of patterns but the interest in some location or movement sequence. The authors adapt the concept of hubs and authorities by Kleinberg (1999) for the rating of geographic locations. Their quantity of interest may be interpreted as a semantically enriched entity-location interaction as the quantity estimates and includes the local expertise of each person. However, the quantity has been designed for recommender systems and is thus suitable only to measure entity-location interactions for established locations. In addition, when comparing entity-location interactions of arbitrary locations in a large area (e.g. a country), the concept of local experience weakens and the interpretation of the quantity is not clear.

4.1.3. Visits: Defining Entity-Location Interaction

A *visit* is a spatiotemporal interaction between a geographic location and a mobile entity. We have already defined both objects in Chapter 2 (Definitions 2.1.5 and 2.2.4). However, before we proceed to their interaction, we will introduce the notation of the universal sets and selected subsets.

Definition 4.1.1 *(Universal location set) For a given application the finite, non-empty set of relevant geographic locations is called universal location set $\mathcal{L}$. A geographic location is hereby defined according to Definition 2.1.5.*

Definition 4.1.2 *(Universal entity set) For a given application the finite, non-empty set of relevant mobile entities is called universal entity set $\mathcal{E}$. A mobile entity is hereby defined according to Definition 2.2.4.*

Definition 4.1.3 *(Location set) A location set is a non-empty subset $L \subseteq \mathcal{L}$.*

Definition 4.1.4 *(Entity set) An entity set is a non-empty subset $E \subseteq \mathcal{E}$.*

We will denote the cardinality of $\mathcal{L}$, $\mathcal{E}$, L and E with $|\mathcal{L}|$, $|\mathcal{E}|$, $|L|$ and $|E|$, respectively. The location set L and entity set E are important for all following definitions because they contain the objects of interest whose interactions we will analyze using visit potential quantities.

Naturally, the interaction between a location and an entity has a spatial as well as a temporal dimension. We define the interaction based on the geographic extent of the location as well as the entity's trajectory function. Please recollect therefore the definition of a trajectory function (Definition 2.2.3) as given in Chapter 2. Note that a visit refers to the interaction between a location and entity in the real world. However, as both objects are specified using geographic and temporal coordinate space, a visit is also specified in coordinate space.

Definition 4.1.5 (Visit) *Given a geographic coordinate space $\mathcal{S_C}$, a temporal coordinate space $\mathcal{T_C}$, a location $l \subseteq \mathcal{S_C}$, $l \neq \emptyset$, a mobile entity $e \in \mathcal{E}$ along with the entity's trajectory function $tr : \mathcal{T_C} \rightarrow \{\{s\} \mid s \in \mathcal{S_C}\} \cup \emptyset$ and a time interval $\varepsilon > 0$, a visit is the tuple (l, e, t_1, t_2) with $t_1, t_2 \in \mathcal{T_C}$, $t_1 < t_2$ for which the following holds*

1. *the intersection of l and $tr(t)$ is non-empty for all $t \in [t_1, t_2]$, i.e.* $l \cap tr(t) \neq \emptyset \quad \forall t \in [t_1, t_2]$,
2. *the time span $[t_1, t_2]$ is maximal, i.e. there exists no time interval $[t_1^*, t_2^*] \supseteq [t_1, t_2]$ so that* $l \cap tr(t) \neq \emptyset \quad \forall t \in [t_1^*, t_2^*]$,
3. *the time interval of intersection is greater or equal to ε, i.e. $t_2 - t_1 \geq \varepsilon$.*

In the above definition t_1 and t_2 define the lower and upper temporal bound of the visit. The maximality criterion ensures that an uninterrupted stay at some location cannot be split into an infinite number of visits of shorter duration. Finally, we require that a visit lasts a given minimum period of time. Parameter ε is application dependent and may have to be determined in a separate study. The motivation behind the introduction of a temporal threshold is that visits are most often associated with a specific activity. For example, we may want to monitor shopping behavior or leisure activities. In such a case we have to distinguish between the passage of a location and a stay in order to perform a given activity. For example, assume that we want to determine the number of regular theater visitors. The introduction of a minimum visit duration allows to distinguish the actual visiting of a performance from a simple passage or the picking up of tickets. Figure 4.1 illustrates the definition of a visit.

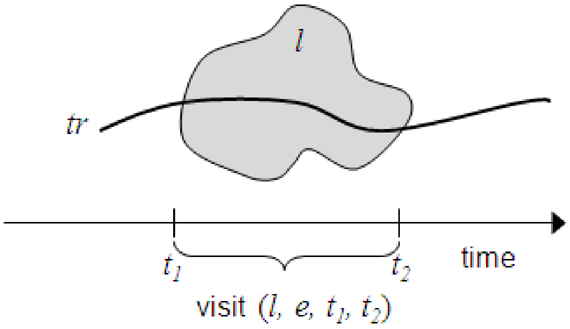


Figure 4.1.: A visit (l, e, t_1, t_2) of an entity e to a location l in time interval $[t_1, t_2]$

We have restricted the definition of visits to very basic characteristics of the entity-location interaction. Note that these characteristics may not be sufficient for all applications, and a further specialization of a visit with respect to other properties of the entity, location or passage may be required. For example, in outdoor advertising the passage of a poster visibility area does not automatically generate a poster contact. Rather, passage characteristics as angle and speed or poster characteristics as size and illumination have to be considered as well. For the purpose of this thesis, we will not go into details about visit specialization because it is strongly application dependent and does not affect the general problem setting.

Note further that our visit definition relies on a functional description of movement. It is independent of the trajectory data format and the quality of the measurements. Depending on the data characteristics appropriate preprocessing and querying methods have to be applied. Such methods are provided, for example, by trajectory databases (see also Sections 2.2.2 and 2.2.5).

4.2. Visit Potential

This section contains the formal definition of visit potential quantities. It introduces the underlying concepts, defines the quantities and analyzes their relationships to each other.

4.2.1. Basic Concepts of Visits

In this section we define basic concepts of visits that underlie the definition of visit potential quantities. We will introduce the concept of visit counts, k-visiting entities and k-visited locations as well as the frequency and probability distribution of the latter two concepts.

The visit of an entity to a location is the fundamental event of interest when analyzing visit potential. Considered over time, the number of visits can be modeled as a counting process.

Definition 4.2.1 (Counting process) *For any $t \in \mathbb{N}_0$ let $N(t)$ be a random variable which denotes the number of occurrences of a specified event within time span $(0, t]$, the set $\{N(t)\}$ of random variables forms a counting process. Without loss of generality we define $N(0) = 0$.*

In our case the variable of interest is the number of visits that an entity $e \in \mathcal{E}$ realizes with a single location $l \in \mathcal{L}$ within time span $(0, t]$. In order to emphasize that *visits* form the event of interest, we will add the letter V to the variable name. Additionally, we will include l and e in the arguments to differentiate counting processes of different location-entity pairs. Note that definition 4.2.1 implicitly assumes that time is specified using a relative temporal reference system with $\mathcal{T}_C = \mathbb{N}_0$ as specified in Section 2.2.1.

Given the definition of a visit and the mathematical concept of a counting process, we can now formally define a random variable for the number of visits of an entity e to a location l.

Definition 4.2.2 (Elementary visit count) *Given an arbitrary time moment $t \in \mathbb{N}_0$, a location $l \in \mathcal{L}$, a mobile entity $e \in \mathcal{E}$ and the resulting visits $\{\, (l, e, t_1, t_2) \,\}$ of e to l according to some minimum visiting time span $\varepsilon > 0$, the elementary visit count of e and l in time span $(0, t]$ is defined as*

$$NV(t, l, e) = \Big| \{\, (l, e, t_1, t_2) \mid 0 < t_1 \leq t \,\} \Big|.$$

We will refer to the elementary visit count also as the visit count of an entity-location pair. Note that the number of visits refers to a given time moment $t \in \mathbb{N}_0$ while visits themselves take place over a time span $[t_1, t_2]$. Definition 4.2.2 therefore states that all visits that start before or at t will be considered for the elementary visit count at t.

Clearly, $NV(t, l, e)$ increases monotonically over time, i.e. if $t_1 < t_2$ then $NV(t_1, l, e) \leq NV(t_2, l, e)$. Figure 4.2 exemplarily shows the counting processes behind three elementary visit counts and the evaluation of the elementary visit count for time moment $t = 9$.

Given the elementary visit count of all entity-location pairs, we can define three further variables by aggregating visits for sets of entities and locations. We begin by aggregating the number of visited locations for a single entity. Next, we aggregate the visits of a single location and finally we aggregate the number of visits over sets of entities and locations. Figure 4.3 depicts all four cases using a space-time cube.

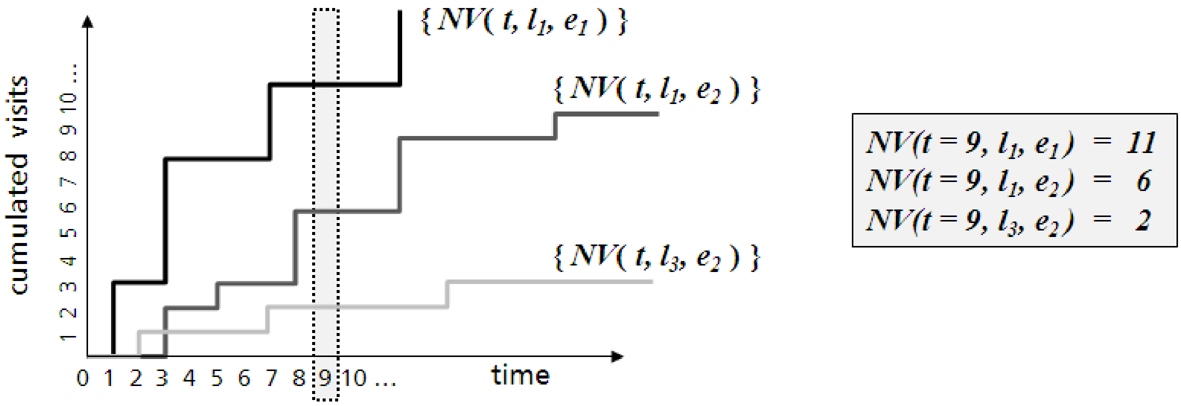


Figure 4.2.: Counting process behind elementary visit count and evaluation for $t = 9$

Definition 4.2.3 (Visit count of an entity) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity $e \in \mathcal{E}$ and the elementary visit count $NV(t,l,e)$ $\forall l \in L$, the visit count of entity e is defined as*

$$NV(t,L,e) = \sum_{l \in L} NV(t,l,e).$$

Definition 4.2.4 (Visit count of a location) *Given a time moment $t \in \mathbb{N}_0$, a location $l \in \mathcal{L}$, an entity set E and the elementary visit count $NV(t,l,e)$ $\forall e \in E$, the visit count of location l is defined as*

$$NV(t,l,E) = \sum_{e \in E} NV(t,l,e).$$

Definition 4.2.5 (Visit count of an entity set and a location set) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E and the elementary visit count $NV(t,l,e)$ $\forall l \in L$ and $\forall e \in E$, the visit count of entity set E and location set L is defined as*

$$NV(t,L,E) = \sum_{l \in L} \sum_{e \in E} NV(t,l,e).$$

The three sets of random variables $\{\, NV(t,L,e) \,|\, t \in \mathbb{N}_0 \,\}$, $\{\, NV(t,l,E) \,|\, t \in \mathbb{N}_0 \,\}$ and $\{\, NV(t,L,E) \,|\, t \in \mathbb{N}_0 \,\}$ are counting processes and thus we know that $NV(t_1,L,e) \leq NV(t_2,L,e)$, $NV(t_1,l,E) \leq NV(t_2,l,E)$ and $NV(t_1,L,E) \leq NV(t_2,L,E)$ for all $t_1 < t_2$.

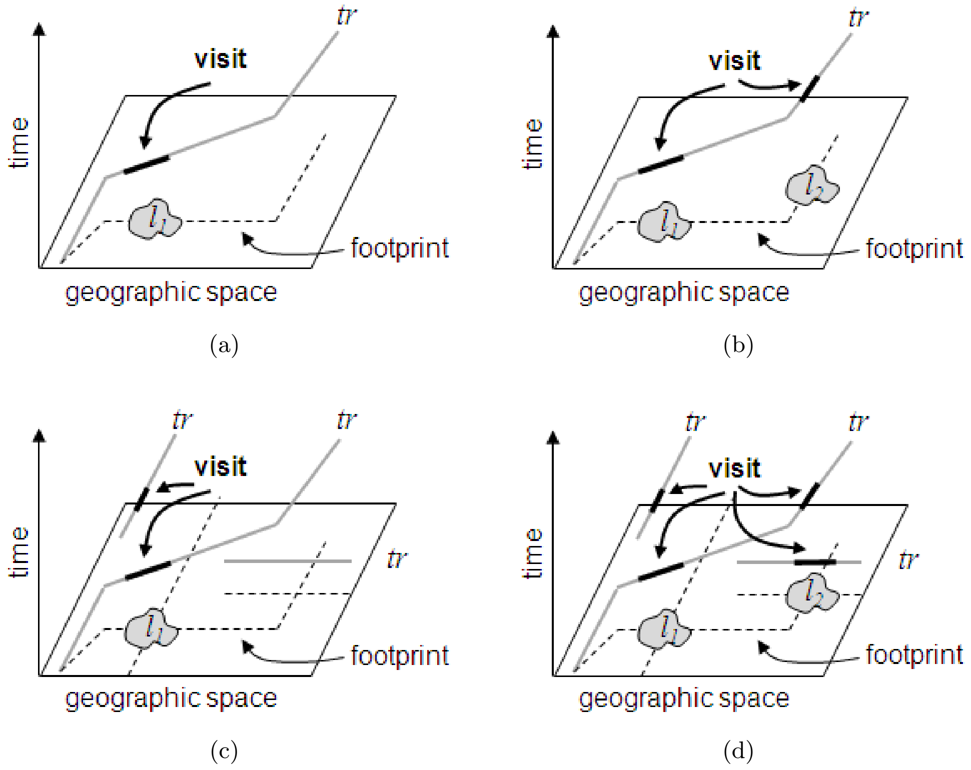
The second important concept when analyzing interactions between an entity set and a location set marks entities and locations with a given visit count. This allows to determine their frequency and frequency distribution.

Definition 4.2.6 (k-visiting entity) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E, the entities' visit count $NV(t,L,e)$ $\forall e \in E$ and a non-negative integer $k \in \mathbb{N}_0$, a k-visiting entity is an entity $e \in E$ such that*

$$NV(t,L,e) = k.$$

Definition 4.2.7 (k-visited location) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E, the locations' visit count $NV(t,l,E)$ $\forall l \in L$ and a non-negative integer $k \in \mathbb{N}_0$, a k-visited location $l \in L$ is a location such that*

$$NV(t,l,E) = k.$$



(a)

(b)

(c)

(d)

Figure 4.3.: Example visits for (a) elementary visit count (b) visit count of an entity (c) visit count of a location (d) visit count of an entity set and a location set

If we count for a given $k \in \mathbb{N}_0$ the number of k-visiting entities or locations, we obtain the k-visiting entity frequency respectively k-visited location frequency.

Definition 4.2.8 (k-visiting entity frequency) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E, the entities' visit count $NV(t, L, e)$ $\forall e \in E$ and a non-negative integer $k \in \mathbb{N}_0$, the k-visiting entity frequency $f_E^{=k}(t, L, E)$ is defined as the number of entities with a visit count of exactly k, i.e.*

$$f_E^{=k}(t, L, E) = \Big| \{e \in E \mid NV(t, L, e) = k\} \Big|.$$

Definition 4.2.9 (k-visited location frequency) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E, the locations' visit count $NV(t, l, E)$ $\forall l \in L$ and a non-negative integer $k \in \mathbb{N}_0$, the k-visited location frequency $f_L^{=k}(t, L, E)$ is defined as the number of locations with a visit count of exactly k, i.e.*

$$f_L^{=k}(t, L, E) = \Big| \{l \in L \mid NV(t, l, E) = k\} \Big|.$$

If we are given the k-visiting entity frequency (k-visited location frequency) for all $k \in \mathbb{N}_0$, we obtain the frequency distribution of k-visiting entities (k-visited locations).

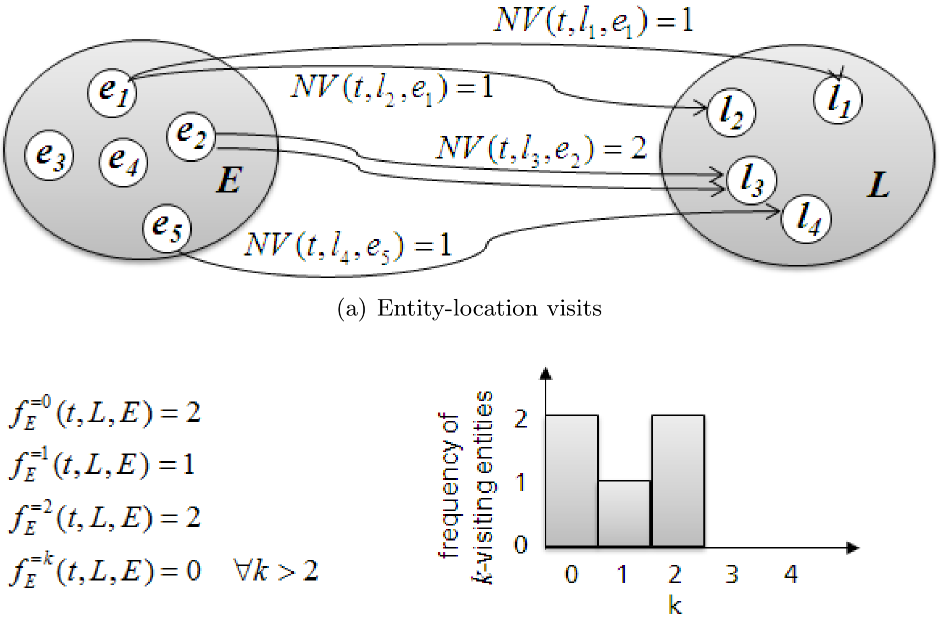
Definition 4.2.10 (Frequency distribution of k-visiting entities) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E and the k-visiting entity frequency $f_E^{=k}(t, L, E)$ $\forall k \in \mathbb{N}_0$, the frequency distribution of k-visiting entities is the ordered set*

$$D_{kE}(t, L, E) = (\, f_E^{=k}(t, L, E) \mid k \in \mathbb{N}_0 \,).$$

Definition 4.2.11 (Frequency distribution of k-visited locations) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E and the k-visited location frequency $f_L^{=k}(t, L, E)$ $\forall k \in \mathbb{N}_0$, the frequency distribution of k-visited locations is the ordered set*

$$D_{kL}(t, L, E) = (\, f_L^{=k}(t, L, E) \mid k \in \mathbb{N}_0 \,).$$

Figure 4.4 illustrates the given definitions using the entity point of view. Starting with the elementary visit count for all entity-location pairs of the given entity set and location set (see Figure 4.4(a)), it shows the k-visiting entity frequencies (see Figure 4.4(b)) and a histogram of the frequency distribution of k-visiting entities (see Figure 4.4(c)).



(a) Entity-location visits

(b) Frequencies of k-visiting entities

(c) Frequency distribution of k-visiting entities

Figure 4.4.: From entity-location visits to frequency distribution of k-visiting entities

From the frequency distribution of k-visiting entities and k-visited locations, which contain absolute frequencies, it is only a small step to the respective probability distributions. The probability distributions play only a minor role in this thesis, however, we will provide their definition for completeness.

Definition 4.2.12 (Probability distribution of k-visiting entities) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E and the frequency distribution of k-visiting entities $D_{kE}(t, L, E)$, the probability distribution of k-visiting entities is a probability distribution with probability function*

$$g_{kE}(k \,|\, t, L, E) = P(NV(t, L, e) = k) = \begin{cases} \frac{f_E^{=k}(t,L,E)}{|E|} & k \geq 0, \\ 0 & k < 0; \end{cases} \qquad k \in \mathbb{Z}$$

and the distribution function

$$G_{kE}(k \,|\, t, L, E) = P(NV(t, L, e) \leq k) = \sum_{k_i \leq k} g_{kE}(k_i \,|\, t, L, E) \qquad k, k_i \in \mathbb{Z}.$$

Definition 4.2.13 (Probability distribution of k-visited locations) *Given a time moment $t \in \mathbb{N}_0$, a location set L, an entity set E and the frequency distribution of k-visited locations $D_{kL}(t, L, E)$, the probability distribution of k-visited locations is a probability distribution with probability function*

$$g_{kL}(k \,|\, t, L, E) = P(NV(t, l, E) = k) = \begin{cases} \frac{f_L^{=k}(t,L,E)}{|L|} & k \geq 0, \\ 0 & k < 0; \end{cases} \qquad k \in \mathbb{Z}$$

and the distribution function

$$G_{kL}(k \,|\, t, L, E) = P(NV(t, l, E) \leq k) = \sum_{k_i \leq k} g_{kL}(k_i \,|\, t, L, E) \qquad k, k_i \in \mathbb{Z}.$$

Note that if we consider the frequency and probability distribution of k-visiting entities (k-visited locations) at two points in time t_1, t_2 with $t_1 < t_2$, the distributions typically show a right-shift, i.e. the cumulated frequencies respectively probabilities of k-visiting entities (k-visited locations) until a given number of visit counts of an entity (location) $v \in \mathbb{N}_0$ decrease monotonically over time. More formally,

$$\sum_{k=0}^{v} f_E^{=k}(t_1, L, E) \;\geq\; \sum_{k=0}^{v} f_E^{=k}(t_2, L, E), \tag{4.1}$$

$$\sum_{k=0}^{v} f_L^{=k}(t_1, L, E) \;\geq\; \sum_{k=0}^{v} f_L^{=k}(t_2, L, E) \tag{4.2}$$

and

$$G_{kE}(v \,|\, t_1, L, E) \;\geq\; G_{kE}(v \,|\, t_2, L, E), \tag{4.3}$$

$$G_{kL}(v \,|\, t_1, L, E) \;\geq\; G_{kL}(v \,|\, t_2, L, E). \tag{4.4}$$

The proof follows directly from the counting process-based definition of visit counts of an entity respectively location, for which we know that $NV(t_1, L, e) \;\leq\; NV(t_2, L, e) \;\; \forall e \in E$ and $NV(t_1, l, E) \;\leq\; NV(t_2, l, E) \;\; \forall l \in L$ given $t_1 < t_2$.

4.2.2. Visit Potential Quantities

In this section we define central visit potential quantities based on the general concepts of visits introduced in the previous section. These quantities are gross visits, average visits per entity, average visits per location, entity coverage and location coverage. As the structure of the names already suggests, the quantities belong to three categories of visit potential quantities. However, as we can implement the quantities within each category from an entity and a location point of view, we obtain five quantities in total.

The first and most basic quantity is *gross visits*, which denotes the total number of visits for a given location and entity set within a given time span. It corresponds to the already defined auxiliary variable visit count of an entity set and a location set (Definition 4.2.5), however, now obtains its status as a quantity. Note that in the case of gross visits the entity and location perspective result in the same quantity. However, when we extend the quantity to visit classes as described in the next section, we will obtain two separate quantities.

Definition 4.2.14 (Gross visits) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E and the elementary visit count $NV(t,l,e)$ $\forall l \in L$ and $\forall e \in E$, the number of total visits until time t is called gross visits:*

$$GV(t,L,E) = \sum_{l \in L} \sum_{e \in E} NV(t,l,e).$$

Corollary 4.2.15 *Gross visits can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t,L,E)$ or the frequency distribution of k-visited locations $D_{kL}(t,L,E)$:*

$$\begin{aligned} GV(t,L,E) &= \sum_{k \geq 0} k \cdot f_E^{=k}(t,L,E) \\ &= \sum_{k \geq 0} k \cdot f_L^{=k}(t,L,E). \end{aligned}$$

Gross visits reflect a contact volume and strongly depend on the number of locations and entities in the given sets. In order to obtain the average contribution of an entity or location in the compared sets, we can normalize gross visits by the size of the entity or location set, respectively. This leads us to the visit potential quantities of the second category, namely average visits per entity and average visits per location.

Definition 4.2.16 (Average visits per entity) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E and the elementary visit count $NV(t,l,e)$ $\forall l \in L$ and $\forall e \in E$, the number of average visits per entity until time t is defined as:*

$$AV_E\,(t,L,E) = \frac{\sum_{l \in L} \sum_{e \in E} NV(t,l,e)}{|E|}.$$

Corollary 4.2.17 *Average visits per entity can be expressed using gross visits:*

$$AV_E\,(t,L,E) = \frac{GV(t,L,E)}{|E|}.$$

Corollary 4.2.18 *Average visits per entity can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t,L,E)$:*

$$AV_E\,(t,L,E) = \frac{\sum_{k \geq 0} k \cdot f_E^{=k}(t,L,E)}{|E|}.$$

Similarly, the average number of visits per location can be calculated.

Definition 4.2.19 (Average visits per location) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E and the elementary visit count $NV(t,l,e)$ $\forall l \in L$ and $\forall e \in E$, the number of average visits per location until time t is defined as:*

$$AV_L\,(t,L,E) = \frac{\sum_{l \in L} \sum_{e \in E} NV(t,l,e)}{|L|}.$$

Corollary 4.2.20 *Average visits per location can be expressed using gross visits:*

$$AV_L\,(t,L,E) = \frac{GV(t,L,E)}{|L|}.$$

Corollary 4.2.21 *Average visits per location can be expressed using the frequency distribution of k-visited locations $D_{kL}(t, L, E)$:*

$$AV_L\,(t, L, E) = \frac{\sum_{k\geq 0} k \cdot f_L^{=k}(t, L, E)}{|L|}.$$

The last category of visit potential quantities is coverage, which again can be defined from an entity or location point of view. It is a quantity measuring dispersion and states the percentage of mobile entities that generate visits with a location set or, respectively, the percentage of locations that are visited by entities of a given entity set.

Definition 4.2.22 (Entity coverage) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E and the entities' visit count $NV(t, L, e)\ \forall e \in E$, entity coverage is defined as the proportion of entities which visit at least one location of the location set until time t:*

$$C_E\,(t, L, E) = \frac{|\,\{e \in E \mid NV(t, L, e) \geq 1\}\,|}{|E|}.$$

Corollary 4.2.23 *Entity coverage can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t, L, E)$:*

$$C_E\,(t, L, E) = \frac{\sum_{k\geq 1} f_E^{=k}(t, L, E)}{|E|}.$$

Corollary 4.2.24 *Entity coverage can be expressed using the distribution function of k-visiting entities $G_{kE}(k\,|\,t, L, E)$:*

$$C_E\,(t, L, E) = 1 - G_{kE}(k = 0\,|\,t, L, E).$$

Definition 4.2.25 (Location coverage) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E and the locations' visit count $NV(t, l, E)\ \forall l \in L$, location coverage is defined as the proportion of locations which are visited by at least one entity of the entity set until time t:*

$$C_L\,(t, L, E) = \frac{|\,\{l \in L \mid NV(t, l, E) \geq 1\}\,|}{|L|}.$$

Corollary 4.2.26 *Location coverage can be expressed using the frequency distribution of k-visited locations $D_{kL}(t, L, E)$:*

$$C_L\,(t, L, E) = \frac{\sum_{k\geq 1} f_L^{=k}(t, L, E)}{|L|}.$$

Corollary 4.2.27 *Location coverage can be expressed using the distribution function of k-visited locations $G_{kL}(k\,|\,t, L, E)$:*

$$C_L\,(t, L, E) = 1 - G_{kL}(k = 0\,|\,t, L, E).$$

Note that entity coverage and location coverage can be expressed in terms of the distribution function of k-visiting entities (Corollary 4.2.24) or, respectively, k-visited locations (Corollary 4.2.27). We simply have to exclude the proportion of entities (locations) without any visits from the total probability.

Due to the characteristics of the frequency distribution of k-visiting entities (k-visited locations), which forms the basis of the defined visit potential quantities, the quantities themselves also increase monotonically over time.

4.2.3. Visit Potential Quantities for Visit Classes

The definitions of visit potential quantities so far allow only for a very general description of entity-location interactions and the resulting frequency distributions of k-visiting entities and k-visited locations. Easily two different frequency distributions may result in the same values of the quantities. In order to disclose more characteristics of the underlying frequency distributions, we therefore extend the above defined visit potential quantities with respect to different *visit classes.*

Visit classes restrict the entity (location) set by introducing a lower bound for the number of visits an entity (location) must show in order to be included in some quantity. For example, the number of average visits per entity for visit class $vc = 2$ states the average number of visited locations for entities with at least two visits. The quantity results by averaging the visits of all entities with a visit count $NV(t, L, e) \geq 2$. In the following we refine the definitions for gross visits, average visits per entity, average visits per location, entity and location coverage with respect to visit classes.

In the case of gross visits the extension to visit classes has two interpretations. On the one hand, we can consider the visit volume of all entities with at least vc visits. On the other hand, we can examine the visit volume of all locations with at least vc visits.

Definition 4.2.28 (Gross visits of entities for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the entities' visit count $NV(t, L, e)$ $\forall e \in E$ and a visit class $vc \in \mathbb{N}_0$, the number of gross visits of entities for visit class vc until time t is defined as:*

$$GV_E\,(t, L, E, vc) = \sum_{e \in E \,|\, NV(t,L,e) \geq vc} NV(t, L, e).$$

Corollary 4.2.29 *Gross visits of entities for visit class vc can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t, L, E)$:*

$$GV_E\,(t, L, E, vc) = \sum_{k \geq vc} k \cdot f_E^{=k}(t, L, E).$$

Definition 4.2.30 (Gross visits of locations for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the locations' visit count $NV(t, l, E)$ $\forall l \in L$ and a visit class $vc \in \mathbb{N}_0$, the number of gross visits of locations for visit class vc until time t is defined as:*

$$GV_L\,(t, L, E, vc) = \sum_{l \in L \,|\, NV(t,l,E) \geq vc} NV(t, l, E).$$

Corollary 4.2.31 *Gross visits of locations for visit class vc can be expressed using the frequency distribution of k-visited locations $D_{kL}(t, L, E)$:*

$$GV_L\,(t, L, E, vc) = \sum_{k \geq vc} k \cdot f_L^{=k}(t, L, E).$$

Similarly, we extend the definitions of average visits per entity and average visits per location to visit classes.

Definition 4.2.32 (Average visits per entity for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the entities' visit count $NV(t, L, e)$ $\forall e \in E$ and* *a visit class $vc \in \mathbb{N}_0$, the number of average visits per entity for visit class vc until time t is defined as:*

$$AV_E\,(t, L, E, vc) = \frac{\sum_{e \in E\,|\,NV(t,L,e) \geq vc} NV(t, L, e)}{|\,\{e \in E \mid NV(t, L, e) \geq vc\}\,|}.$$

If no entity exists which reaches visit class vc, i.e. $|\,\{e \in E \mid NV(t, L, e) \geq vc\}\,| = 0$, then the average visits per entity for visit class vc are undefined.

Corollary 4.2.33 *Average visits per entity for visit class vc can be expressed using the gross visits of entities for visit class vc:*

$$AV_E\,(t, L, E, vc) = \frac{GV_E\,(t, L, E, vc)}{|\,\{e \in E \mid NV(t, L, e) \geq vc\}\,|}.$$

Corollary 4.2.34 *Average visits per entity for visit class vc can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t, L, E)$:*

$$AV_E\,(t, L, E, vc) = \frac{\sum_{k \geq vc} k \cdot f_E^{=k}(t, L, E)}{\sum_{k \geq vc} f_E^{=k}(t, L, E)}.$$

Definition 4.2.35 (Average visits per location for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the locations' visitt count $NV(t, l, E)$ $\forall l \in L$ and a visit class $vc \in \mathbb{N}_0$, the number of average visits per location for visit class vc until time t is defined as:*

$$AV_L\,(t, L, E, vc) = \frac{\sum_{l \in L\,|\,NV(t,l,E) \geq vc} NV(t, l, E)}{|\,\{l \in L \mid NV(t, l, E) \geq vc\}\,|}.$$

If no location exists which reaches visit class vc, i.e. $|\,\{l \in L \mid NV(t, l, E) \geq vc\}\,| = 0$, then the average visits per location for visit class vc are undefined.

Corollary 4.2.36 *Average visits per location for visit class vc can be expressed using the gross visits of locations for visit class vc:*

$$AV_L\,(t, L, E, vc) = \frac{GV_L\,(t, L, E, vc)}{|\,\{l \in L \mid NV(t, l, E) \geq vc\}\,|}.$$

Corollary 4.2.37 *Average visits per location for visit class vc can be expressed using the frequency distribution of k-visited locations $D_{kL}(t, L, E)$:*

$$AV_L\,(t, L, E, vc) = \frac{\sum_{k \geq vc} k \cdot f_L^{=k}(t, L, E)}{\sum_{k \geq vc} f_L^{=k}(t, L, E)}.$$

Note that the maximum reached visit class for a given entity set and location set is different when evaluated from entity or location perspective. For example, consider a location set of size five and an entity set with a single entity which visits each location once. The maximum number of visits of the entity is five while the maximum number of visits of each location is one.

For coverage the extended definitions are very similar to Definitions 4.2.22 and 4.2.25, because these definitions already required a visit count of at least one of the considered entities (locations).

Definition 4.2.38 (Entity coverage for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the entities' visit count $NV(t, L, e)$ $\forall e \in E$ and a visit class $vc \in \mathbb{N}_0$, entity coverage for visit class vc is defined as the proportion of entities which have a visit count of at least vc until time t:*

$$C_E\left(t, L, E, vc\right) = \frac{|\left\{e \in E \mid NV(t, L, e) \geq vc\right\}|}{|E|}.$$

Corollary 4.2.39 *Entity coverage for visit class vc can be expressed using the frequency distribution of k-visiting entities $D_{kE}(t, L, E)$:*

$$C_E\left(t, L, E, vc\right) = \frac{\sum_{k \geq vc} f_E^{=k}(t, L, E)}{|E|}.$$

Corollary 4.2.40 *Entity coverage for visit class vc can be expressed using the distribution function of k-visiting entities $G_{kE}(k \,|\, t, L, E)$:*

$$C_E\left(t, L, E, vc\right) = 1 - G_{kE}(vc - 1 \,|\, t, L, E).$$

Definition 4.2.41 (Location coverage for visit class vc) *Given a time moment $t \in \mathbb{N}_0$, a location set L, a set of entities E, the locations' visit count $NV(t, l, E)$ $\forall l \in L$ and a visit class $vc \in \mathbb{N}_0$, location coverage for visit class vc is defined as the proportion of locations which have a visit count of at least vc until time t:*

$$C_L\left(t, L, E, vc\right) = \frac{|\left\{l \in L \mid NV(t, l, E) \geq vc\right\}|}{|L|}.$$

Corollary 4.2.42 *Location coverage for visit class vc can be expressed using the frequency distribution of k-visited locations $D_{kL}(t, L, E)$:*

$$C_L\left(t, L, E, vc\right) = \frac{\sum_{k \geq vc} f_L^{=k}(t, L, E)}{|L|}.$$

Corollary 4.2.43 *Location coverage for visit class vc can be expressed using the distribution function of k-visited locations $G_{kL}(k \,|\, t, L, E)$:*

$$C_L\left(t, L, E, vc\right) = 1 - G_{kL}(vc - 1 \,|\, t, L, E).$$

Note that coverage for visit class $vc = 0$ always amounts to one, i.e.

$$C_E\left(t, L, E, vc = 0\right) = C_L\left(t, L, E, vc = 0\right) = 1. \tag{4.5}$$

If we use visit potential quantities in the remaining thesis without explicitly specifying a visit class, we will assume that gross visits of entities (locations) and average visits per entity (location) are given according to $vc = 0$ and entity (location) coverage according to $vc = 1$ as defined in Section 4.2.2.

4.2.4. Interrelations of Visit Potential Quantities

All visit potential quantities are derived from the frequency distribution of k-visiting entities or k-visited locations, which again rely on the elementary visit count of each entity-location pair $(l \in L, e \in E)$. Due to this common background a number of relationships exist between the quantities. In addition, the frequency distribution of k-visiting entities and k-visited locations can be completely derived from the visit potential quantities when considered over all visit classes.

The first relationship concerns only the quantity gross visits and exists due to the integer character of visits. For visit classes $vc = 0$ and $vc = 1$ the following property holds:

$$\begin{aligned} GV(t, L, E) &= GV_E\,(t, L, E, vc = 0) = GV_L\,(t, L, E, vc = 0) \\ &= GV_E\,(t, L, E, vc = 1) = GV_L\,(t, L, E, vc = 1). \end{aligned} \tag{4.6}$$

The second relationship provides a linkage between average visits per entity (location) of contact classes $vc = 0$ and $vc = 1$ using the quantity entity (location) coverage.

$$AV_E\,(t, L, E, vc = 0) = AV_E\,(t, L, E, vc = 1)\,\cdot\,C_E\,(t, L, E, vc = 1), \tag{4.7}$$

$$AV_L\,(t, L, E, vc = 0) = AV_L\,(t, L, E, vc = 1)\,\cdot\,C_L\,(t, L, E, vc = 1). \tag{4.8}$$

The relationship relies on the equivalence of gross visits for visit classes $vc = 0$ and $vc = 1$ as stated in Equation 4.6, and uses coverage to adapt the number of considered entities respectively locations. This becomes clear, when substituting the quantities with their definitions, as exemplary shown for Equation 4.7:

$$\frac{GV_E\,(t, L, E, vc = 0)}{|E|} = \frac{GV_E\,(t, L, E, vc = 1)}{|\{e \in E \mid NV(t, L, e) \geq 1\}|} \cdot \frac{|\{e \in E \mid NV(t, L, e) \geq 1\}|}{|E|}. \tag{4.9}$$

The third relationship is already hidden in Corollaries 4.2.17 and 4.2.20. It simply states that the number of average visits per entity (location) for visit class $vc = 0$ can be deduced from gross visits by the size of the entity (location) set:

$$GV(t, L, E) = AV_E\,(t, L, E) \cdot |E|, \tag{4.10}$$

$$GV(t, L, E) = AV_L\,(t, L, E) \cdot |L|. \tag{4.11}$$

The above equations can also be generalized for all visit classes:

$$GV_E\,(t, L, E, vc) = AV_E\,(t, L, E, vc) \cdot |E| \cdot C_E\,(t, L, E, vc), \tag{4.12}$$

$$GV_L\,(t, L, E, vc) = AV_L\,(t, L, E, vc) \cdot |L| \cdot C_L\,(t, L, E, vc). \tag{4.13}$$

Hereby, the product of the size of the entity (location) set and entity (location) coverage results in the number of entities (locations) that show at least vc visits. A further multiplication with the average number of visits per entity (location) in this visit class then results in the gross visits of entities (locations) of the visit class.

Finally, given the sequence of gross visits of entities (locations) for all visit classes $vc \geq 0$, the frequency distribution of k-visiting entities (k-visited locations) can be derived. In a first step we have to subtract from gross visits of visit class $vc = k$ the gross visits of visit class $vc = k - 1$, which results in the visit volume for entities (locations) with exactly k visits. Dividing this volume by the k of the according visit class leads to the underlying number of entities (locations). More precisely, from the entity point of view the frequency of k-visiting entities can be obtained from gross visits of entities as follows.

Frequency of k-visiting entities for $vc > 0$:

$$\begin{aligned} f_E^{=vc}(t, L, E) &= \frac{GV_E\,(t, L, E, vc) - GV_E\,(t, L, E, vc + 1)}{vc} \\ &= \frac{\sum_{k \geq vc} k \cdot f_E^{=k}(t, L, E) - \sum_{k \geq vc+1} k \cdot f_E^{=k}(t, L, E)}{vc} \\ &= \frac{vc \cdot f_E^{=vc}(t, L, E)}{vc} = f_E^{=vc}(t, L, E). \end{aligned} \tag{4.14}$$

Frequency of k-visiting entities for $vc = 0$:

$$f_E^{=vc}(t, L, E) = |E| - \sum_{k \geq 1} f_E^{=k}(t, L, E). \tag{4.15}$$

The frequency of k-visited locations can be obtained from gross visits of locations analogously.

4.3. Visit Potential under Partitioning of Location and Entity Set

4.3.1. Overview of Partitioning

In Section 4.2.4 we considered the relationship of visit potential quantities to each other given a constant location and entity set. In this section we describe the behavior of visit potential quantities under a changing location and entity sets. More precisely, we create a partition of the location and entity set and analyze the relationship of visit potential quantities when applied to the whole location and entity set versus quantities that result from subsets of the location and entity set. Partitioning is a very useful analysis technique in practice because it allows to trace mobile behavior with respect to different geographical areas and at several levels of resolution.

The knowledge about relationships between visit potential quantities on the complete and partitioned sets is helpful in practice to reduce either computational complexity or to perform tests. On the one hand, given large data sets partitioning may allow to split one large problem into a number of subproblems. On the other hand, by partitioning we can verify that estimates of visit potential quantities are consistent over different levels of granularity with respect to the location and entity set. As we will see in Chapter 5 visit potential quantities are often estimated due to incompleteness of the data.

Definition 4.3.1 (Partition) *Given a set A, a partition of $Part(A)$ is a set of non-empty subsets $\{A_1, A_2, \ldots, A_n\}$, $n \in \mathbb{N}$, such that*

$$\begin{aligned} \bigcup_{i=1}^{n} A_i &= A \quad \textit{and} \\ A_i \cap A_j = \emptyset & \quad \forall i, j = 1..n,\ i \neq j. \end{aligned}$$

Given a partition $Part(L) = \{L_1, L_2, \ldots, L_u\}$ of a location set, respectively a partition $Part(E) = \{E_1, E_2, \ldots, E_w\}$ of an entity set, we are interested in the behavior of some visit potential quantity $h(t, L, E, vc)$ when applied to the entire location (entity) set versus its application to the subsets of the partition. More specifically, we are interested to know whether the following relation holds, with $\circ$ denoting some binary operator:

$$h(t, L, E, vc) = h(t, L_1, E, vc) \circ h(t, L_2, E, vc) \circ \ldots \circ h(t, L_u, E, vc), \tag{4.16}$$

$$h(t, L, E, vc) = h(t, L, E_1, vc) \circ h(t, L, E_2, vc) \circ \ldots \circ h(t, L, E_w, vc). \tag{4.17}$$

If such a relation does not exist, we look for relaxations of the equation and try to find an upper and / or lower bound for the left hand side in terms of the right hand side.

We begin with partitioning of the location set and subsequently consider visit potential quantities under partitioning of the entity set. However, due to the reverse relationship of visit potential quantities when considered from an entity or a location point of view, the results of partitioning of the entity set are analogous to the results of partitioning of the location set. In the second part we therefore relinquish the derivation of relationships and directly show results.

In order to make the relationships easy to comprehend, we will illustrate the effects of partitioning by the following example. We are given a location set with five locations and an entity set with four entities. Both sets produce elementary visits as shown in Figure 4.5 within time period t.

	e_1	e_2	e_3	e_4	$NV(t,l,E)$
l_1	1	1	0	0	2
l_2	0	2	0	0	2
l_3	0	0	0	0	0
l_4	0	0	0	1	1
l_5	1	0	0	1	2
$NV(t,L,e)$	2	3	0	2	

Figure 4.5.: Visit example showing elementary visit count

4.3.2. Visit Potential under Partitioning of Location Set

Given a partition $Part(L) = \{L_1, L_2, \dots, L_u\}$ of the location set we will first consider the impact of partitioning on the frequency of k-visited locations and k-visiting entities. Afterwards we will consider all introduced visit potential quantities for varying visit classes.

Frequency of k-visited locations and k-visiting entities The visit count $NV(t,l,E)$ of a single location remains unchanged if we assign locations to subsets of the partition. Therefore, the following equation holds:

$$f_L^{=k}(t,L,E) = \sum_{i=1}^{u} f_L^{=k}(t,L_i,E). \tag{4.18}$$

Proof

$$\begin{aligned}
f_L^{=k}(t,L,E) &= |\,\{l \in L \mid NV(t,l,E) = v\}\,| \\
&= |\,\{l \in L_1 \cup L_2 \cup \ldots \cup L_u \mid NV(t,l,E) = k\}\,| \\
&= \sum_{i=1}^{u} |\,\{l \in L_i \mid NV(t,l,E) = k\}\,| \\
&= \sum_{i=1}^{u} f_L^{=k}(t,L_i,E)
\end{aligned}$$

The last but one line follows from Definition 4.3.1, as each location belongs to exactly one location subset. □

In contrast to the unchanged visit count $NV(t,l,E)$ of each location, the visit count $NV(t,L,e)$ of each entity splits up between the location subsets, i.e.

$$NV(t,L,e) = \sum_{i=1}^{u} NV(t,L_i,e) \qquad \forall e \in E. \tag{4.19}$$

This results in a decrease of entities with high visit frequencies and an increase of entities with low visit frequencies in the location subsets and equals a left-shift of the frequency distribution of k-visiting entities within the subsets when compared to the total set, i.e. given a maximum number of considered visit counts $v \in \mathbb{N}_0$ the following holds

$$\sum_{k=0}^{v} f_E^{=k}(t, L, E) \leq \sum_{k=0}^{v} f_E^{=k}(t, L_i, E) \qquad \forall i = 1..u. \tag{4.20}$$

When evaluating the distribution of k-visiting entities, the identity of entities is naturally lost. In consequence, the total visit count of an entity cannot be deduced from the frequency distributions of k-visiting entities of the subsets, and the frequency distribution of k-visiting entities of the complete location set cannot be derived.

The effects on the frequency of k-visited locations and k-visiting entities can be pictured in the visit example. Figure 4.6 shows a location partition of our example, and tables 4.1 and 4.2 contain the original and resulting frequency distributions. The frequency distribution of k-visited locations of the complete location set results from a join of the frequency distributions of k-visited locations of the partition. In contrast, the combination of k-visiting entity frequencies is not obvious. Note that the frequency of k-visiting entities for a given k of a subset can lie above or below the frequency of the unpartitioned location set, for example, $f_E^{=2}(t, L, E) = 2 \geq f_E^{=2}(t, L_1, E) = 0$ while $f_E^{=1}(t, L, E) = 0 \leq f_E^{=1}(t, L_1, E) = 1$.

$$\begin{array}{cc|c|c|c|c|c}
 & & e_1 & e_2 & e_3 & e_4 & NV(t,l,E) \\
\cline{3-6}
 & l_1 & 1 & 1 & 0 & 0 & 2 \\
\cline{3-6}
L_1 & l_2 & 0 & 2 & 0 & 0 & 2 \\
\cline{3-6}
 & l_3 & 0 & 0 & 0 & 0 & 0 \\
\cline{3-6}
NV(t,L_1,e) & & 1 & 3 & 0 & 0 & \\
 & & & & & & \\
\cline{3-6}
L_2 & l_4 & 0 & 0 & 0 & 1 & 1 \\
\cline{3-6}
 & l_5 & 1 & 0 & 0 & 1 & 2 \\
\cline{3-6}
NV(t,L_2,e) & & 1 & 0 & 0 & 2 &
\end{array}$$

Figure 4.6.: Location partition of visit example in Figure 4.5

Table 4.1.: Frequency distribution of k-visited locations for complete and partitioned location set of visit example

	$k=0$	$k=1$	$k=2$	$k=3$	$k \geq 4$
$f_L^{=k}(t, L, E)$	1	1	3	0	0
$f_L^{=k}(t, L_1, E)$	1	0	2	0	0
$f_L^{=k}(t, L_2, E)$	0	1	1	0	0

Table 4.2.: Frequency distribution of k-visiting entities for complete and partitioned location set of visit example

	$k=0$	$k=1$	$k=2$	$k=3$	$k \geq 4$
$f_E^{=k}(t, L, E)$	1	0	2	1	0
$f_E^{=k}(t, L_1, E)$	2	1	0	1	0
$f_E^{=k}(t, L_2, E)$	2	1	1	0	0

Given the relationships of the frequency distribution of k-visited locations and k-visiting entities, we now proceed to visit potential quantities under location partitioning.

Gross visits of entities and locations for visit class vc Gross visits of entities (locations) measure a visit volume. Both quantities are defined based on the frequency distribution of k-visiting entities or k-visited locations, respectively. In case of contact classes $vc = 0$ and $vc = 1$ both quantities state the total number of visits, and we can formulate the following equation:

$$\begin{aligned} GV_E\,(t, L, E, vc \in \{0,1\}) &= GV_L\,(t, L, E, vc \in \{0,1\}) \\ &= \sum_{k \geq 0} k \cdot f_L^{=k}(t, L, E) \\ &= \sum_{k \geq 0} k \cdot \sum_{i=1}^{u} f_L^{=k}(t, L_i, E) \\ &= \sum_{i=1}^{u} \sum_{k \geq 0} k \cdot f_L^{=k}(t, L_i, E) \\ &= \sum_{i=1}^{u} GV_E\,(t, L_i, E, vc \in \{0,1\}). \end{aligned} \tag{4.21}$$

For higher visit classes gross visits of locations can be summarized similar to Equation 4.21 because the underlying frequencies of k-visited locations are maintained in the subsets of the partition. A restriction of locations due to the visit class thus affects the same locations in the complete and partitioned location set.

$$GV_L\,(t, L, E, vc > 1) = \sum_{i=1}^{u} GV_L\,(t, L_i, E, vc > 1). \tag{4.22}$$

In contrast, the number of gross visits of entities for the complete location set cannot be obtained from the gross visits of entities of the partition for visit classes $vc > 1$. As the visits are distributed over several location sets, most entities reach lower maximal visit classes in the location subsets than in the complete set. This leads to an early elimination of entities during the calculation of gross visits for a given visit class. Therefore, the sum of gross visits of entities of the partition starting at a given visit class is always smaller or equal to the gross visits of entities of the complete location set:

$$GV_E\,(t, L, E, vc > 1) \geq \sum_{i=1}^{u} GV_E\,(t, L_i, E, vc > 1). \tag{4.23}$$

Proof

$$NV(t,L,e) = \sum_{i=1}^{u} NV(t,L_i,e) \qquad \forall e \in E \qquad \text{(Equation 4.19)}$$

Let th define a threshold function:

$$th(a,b) = \begin{cases} a & \text{if } a \geq b, \\ 0 & \text{if } a < b. \end{cases}$$

$$\begin{aligned}
\Rightarrow th\big(NV(t,L,e),vc\big) &= th\big(\sum_{i=1}^{u} NV(t,L_i,e),vc\big) \qquad vc \in \mathbb{N}_0 \\
\Rightarrow th\big(NV(t,L,e),vc\big) &\geq \sum_{i=1}^{u} th\big(NV(t,L_i,e),vc\big) \\
\Rightarrow \sum_{e\in E} th\big(NV(t,L,e),vc\big) &\geq \sum_{e\in E}\sum_{i=1}^{u} th\big(NV(t,L_i,e),vc\big) \\
\Rightarrow \sum_{k\geq vc} k \cdot f_E^{=k}(t,L,E) &\geq \sum_{k\geq vc}\sum_{i=1}^{u} k \cdot f_E^{=k}(t,L_i,E) \\
\Rightarrow GV_E\,(t,L,E,vc) &\geq \sum_{i=1}^{u} GV_E\,(t,L_i,E,vc)
\end{aligned}$$

□

For example, consider the gross visits of entities for visit class $vc = 2$ in the visit example. In the complete location set one entity with visit count three and two entities with visit count two exist, resulting in $GV_E\,(t,L,E,vc=2) = 7$ gross visits. Under partitioning of the location set all visits of entity e_1 belong to the first location subset and all visits of e_4 belong to the second location set while the two visits of entity e_2 split up between the subsets (see Figure 4.6). The summarized gross visits of both location sets under visit class $vc = 2$ are thus reduced to $GV_E\,(t,L_1,E,vc=2) + GV_E\,(t,L_2,E,vc=2) = 3+2 = 5$.

Average visits per entity for visit class vc The average number of visits per entity are derived from gross visits of entities as stated in Corollaries 4.2.17 and 4.2.33. We therefore derive the relationship of this quantity between a complete and a partitioned location set following the considerations above.

The average number of visits that an entity produces with the complete location set for visit class $vc = 0$ equals the sum of average visits per entity under location partitioning:

$$\begin{aligned}
AV_E\,(t,L,E,vc=0) &= \frac{GV_E\,(t,L,E,vc=0)}{|E|} = \frac{\sum_{i=1}^{u} GV_E\,(t,L_i,E,vc=0)}{|E|} \\
&= \sum_{i=1}^{u} \frac{GV_E\,(t,L_i,E,vc=0)}{|E|} = \sum_{i=1}^{u} AV_E\,(t,L_i,E,vc=0). \qquad (4.24)
\end{aligned}$$

For visit class $vc = 1$ the number of gross visits in the calculation for the complete and partitioned location set remain the same as in Equation 4.24. However, the number of entities that visit at least once a given location subset may decrease. Hereby, the reduced entity set used for the complete location set forms a superset of the reduced entity sets used for the location subsets. Thus, the denominator of each subset is potentially smaller than for the complete set and the following relationship exists:

$$AV_E\left(t, L, E, vc = 1\right) \leq \sum_{i=1}^{u} AV_E\left(t, L_i, E, vc = 1\right) \tag{4.25}$$

$\forall i$ with $AV_E\left(t, L_i, E, vc = 1\right)$ defined.

Note that for visit classes $vc \geq 1$ average visits per entity do not need to be defined for all subsets, although a value for the given visit class may be defined for the complete location set.

For visit classes $vc > 1$ in addition to the denominator also the numerator may decrease. However, the gross visits of entities of the complete location set form an upper bound of the sum of gross visits of entities of the location subsets. Thus, the numerator of each subset is potentially smaller than the subset's share of gross visits of entities in the complete location set, and a conclusion about the relationship of average visits per entity can no longer be drawn.

Table 4.3.: Example average visits per entity for complete and partitioned location set

	$k=0$	$k=1$	$k=2$	$k=3$
$AV_E\left(t, L, E, vc\right)$	7/4	7/3	7/3	3/1
$AV_E\left(t, L_1, E, vc\right)$	4/4	4/2	3/1	3/1
$AV_E\left(t, L_2, E, vc\right)$	3/4	3/2	2/1	$--$

Let us relate the above statements to our visit example. Table 4.3 shows the average visits per entity for the complete and partitioned location set. For visit class $vc = 0$ all elementary visits $NV(t, l, e)$ are included in the calculations, and the denominator amounts counts four entities for L, L_1 and L_2:

$$AV_E\left(t, L, E, vc = 0\right) = \frac{2+3+0+2}{4} = \frac{7}{4},$$

$$\sum_{i=1}^{2} AV_E\left(t, L_i, E, vc = 0\right) = \frac{1+3+0+0}{4} + \frac{1+0+0+2}{4} = \frac{7}{4}.$$

For visit class $vc = 1$ the calculation on the complete location set excludes entity e_3 because it does not produce any visits. Entity e_3 is also excluded in the calculations for L_1 and L_2. However, the calculation for L_1 excludes additionally e_4 and the calculation for L_2 excludes e_2:

$$AV_E\left(t, L, E, vc = 1\right) = \frac{2+3+2}{3} = \frac{7}{3},$$

$$\sum_{i=1}^{2} AV_E\left(t, L_i, E, vc = 1\right) = \frac{1+3}{2} + \frac{1+2}{2} = \frac{7}{2}.$$

For visit class $vc = 2$ the calculation for L still includes entities e_1, e_2 and e_3 while for the average visits per entity of L_1 and L_2 entity e_1 drops out because the elementary visits of e_1 split up between the location subsets and do not suffice for visit class $vc = 2$ in any of the subsets. Finally, only one entity reaches visit class $vc = 3$. All three visits take place with location set L_1, therefore $AV_E\left(t, L, E, vc = 3\right) = 3$ and $AV_E\left(t, L_1, E, vc = 3\right) = 3$, however, $AV_E\left(t, L_2, E, vc = 3\right)$ is undefined as no entity reaches a visit count of three in L_2.

Average visits per location for visit class vc The average visits per location of some location set depend, on the one hand, on the number of gross visits of locations (see Corollaries 4.2.20 and 4.2.36) and, on the other hand, on the frequency distribution of k-visited locations (see Corollaries 4.2.21 and 4.2.37). For visit class $vc = 0$, we can derive the average visits per location of the complete set from a weighted average of average visits per location of the partitioned location set. We hereby assume that the sizes of the location subsets are known, which is a reasonable assumption. The weight $|L_i|/|L|$ hereby changes the denominator of the average visits per location of the subsets from the cardinality of the subsets to the cardinality of the total location set.

$$\begin{aligned} AV_L\,(t, L, E, vc = 0) &= \frac{GV_L\,(t, L, E, vc = 0)}{|L|} \\ &= \frac{\sum_{i=1}^{u} GV_L\,(t, L_i, E, vc = 0)}{|L|} \\ &= \sum_{i=1}^{u} \frac{|L_i|}{|L|} \cdot \frac{GV_L\,(t, L_i, E, vc = 0)}{|L_i|} \\ &= \sum_{i=1}^{u} \frac{|L_i|}{|L|} \cdot AV_L\,(L_i). \end{aligned} \tag{4.26}$$

For higher visit classes, the size of location subsets has to be additionally adapted to the number of locations that reach the given visit class. This can be achieved by multiplication with the location coverage of the given subset and visit class.

$$\begin{aligned} AV_L\,(t, L, E, vc > 0) &= \frac{GV_L\,(t, L, E, vc > 0)}{\sum_{k \geq vc} f_L^{=k}(t, L, E)} \\ &= \frac{\sum_{i=1}^{u} GV_L\,(t, L_i, E, vc > 0)}{\sum_{k \geq vc} f_L^{=k}(t, L, E)} \\ &= \sum_{i=1}^{u} \frac{\sum_{k \geq vc} f_L^{=k}(t, L_i, E)}{\sum_{k \geq vc} f_L^{=k}(t, L, E)} \cdot \frac{GV_L\,(t, L_i, E, vc > 0)}{\sum_{k \geq vc} f_L^{=k}(t, L_i, E)} \\ &= \sum_{i=1}^{u} \frac{\sum_{k \geq vc} f_L^{=k}(t, L_i, E)}{\sum_{k \geq vc} f_L^{=k}(t, L, E)} \cdot AV_L\,(t, L_i, E, vc > 0) \\ &= \sum_{i=1}^{u} \frac{C_L\,(t, L_i, E, vc > 0) \cdot |L_i|}{C_L\,(t, L, E, vc > 0) \cdot |L|} \cdot AV_L\,(t, L_i, E, vc > 0) \\ &\quad \forall i \text{ with } AV_L\,(t, L_i, E, vc > 0) \text{ defined.} \end{aligned} \tag{4.27}$$

Note that Equation 4.27 generalizes Equation 4.26 as location coverage for visit class $vc = 0$ is always one. Equation 4.27 assumes that in addition to the number of average visits per location the location coverage of each subset as well as the location coverage of the complete location set are also known. Most likely, the first assumption is met because we aim to derive visit potential quantities from the partitioned location set for the complete location set. It is therefore reasonable that further visit potential quantities are available for the location subsets. The second assumption is always fulfilled because location coverage of the complete location set can be derived from the location coverage of the partitioned set as will be shown in the next but one paragraph (see Equation 4.31).

For illustration we calculate average visits per location for visit class $vc = 1$ in our example from the total and partitioned location set. In the last line the rebasing of the denominator becomes visible, which substitutes the number of locations with a visit count of at least one in a given subset with the total number of locations with a visit count of at least one.

$$AV_L\,(t, L, E, vc = 1) = \frac{GV_L\,(t, L, E, vc = 1)}{\sum_{k\geq 1} f_L^{=k}(t, L, E)} = \frac{2+2+1+2}{4} = \frac{7}{4},$$

$$\begin{aligned} AV_L\,(t, L, E, vc = 1) &= \sum_{i=1}^{2} \frac{C_L\,(t, L_i, E, vc = 1) \cdot |L_i|}{C_L\,(t, L, E, vc = 1) \cdot |L|} \cdot AV_L\,(t, L_i, E, vc = 1) \\ &= \frac{2/3 \cdot 3}{4/5 \cdot 5} \cdot \frac{2+2}{2} + \frac{2/2 \cdot 2}{4/5 \cdot 5} \cdot \frac{1+2}{2} \\ &= \frac{2}{4} \cdot \frac{2+2}{2} + \frac{2}{4} \cdot \frac{1+2}{2} = \frac{4}{4} + \frac{3}{4} = \frac{7}{4}. \end{aligned}$$

Entity coverage for visit class vc Entity coverage states the proportion of entities that visit a location set at least a minimum number of times. By Corollaries 4.2.23 and 4.2.39 entity coverage depends on the frequency distribution of k-visiting entities. As the frequency distributions of k-visiting entities of the location subsets do not allow to infer the frequency distribution of the complete location set, an equivalence relationship for entity coverage cannot be established between the partitioned and unpartitioned location set.

Basically two problems arise when trying to combine the entity coverage of the subsets. First, the overlap of entities that reach a given visit class in any pair of location subsets is unknown. Second, entities that do not reach the specified visit class in any location subset, may qualify for the visit class in the complete location set. Both problems can be shown in our example. The entity coverage for visit class $vc = 2$ is 3/4 in L, and 1/4 in L_1 and L_2. Given only the entity coverage of the location subsets, it is unknown whether both coverage values originate from a single entity or two different entities. In addition, entity e_1 only reaches visit class $vc = 2$ by summation of its visits in L_1 and L_2. It is thus not included in the entity coverage of the location subsets.

Due to the possible overlap of entities that reach a given visit class in the various subsets, however, we can establish an inequality relationship. The entity coverage of the complete location set is always greater than or equal to the maximum entity coverage of the location subsets:

$$C_E\,(t, L, E, vc > 0) \geq max \left\{ \; C_E\,(t, L_i, E, vc > 0) \mid i = 1..u \; \right\}. \tag{4.28}$$

For visit class $vc = 1$ we can also form an upper bound for entity coverage of the complete location set because any entity with at least one visit has to appear in at least one location subset. Given the entity coverage for all location subsets, the maximum coverage of the complete location set is produced if any entity visits at most one location subset, i.e. there is no overlap in entity coverage between the location subsets, and the entity coverage of the complete location set equals the sum of entity coverage of the subsets:

$$C_E\,(t, L, E, vc = 1) \leq min \left\{ \; 1, \; \sum_{i=1}^{u} C_E\,(l, L_i, E, vc = 1) \; \right\}. \tag{4.29}$$

For visit classes $vc > 1$ an upper bound for entity coverage cannot be derived from the location subsets due to the second problem stated above. Entities that do not reach a given visit class in the subsets can still be present in the entity coverage of the complete location set. The upper bound of entity coverage in the complete set thus defaults to 1.

Note that for the trivial case of visit class $vc = 0$ it still holds that

$$C_E\,(t, L, E, vc = 0) = C_E\,(t, L_i, E, vc = 0) = 1 \qquad \forall i = 1..u. \tag{4.30}$$

Location coverage for visit class vc Location coverage is calculated from the distribution of k-visited locations (see Corollaries 4.2.26 and 4.2.42). As we have proved the composition of the frequency of k-visited locations of the complete location set from the frequency of k-visited locations of the location subsets in Equation 4.18, location coverage for the complete location set can simply be derived from location coverage of the subsets by the following equation:

$$\begin{aligned} C_L\,(t, L, E, vc \geq 0) &= \frac{\sum_{k \geq vc} f_L^{=k}(t, L, E)}{|L|} \\ &= \frac{\sum_{k \geq vc} \sum_{i=1}^{u} f_L^{=k}(t, L_i, E)}{|L|} \\ &= \sum_{i=1}^{u} \frac{|L_i|}{|L|} \cdot \frac{\sum_{k \geq vc} f_L^{=k}(t, L_i, E)}{|L_i|} \\ &= \sum_{i=1}^{u} \frac{|L_i|}{|L|} \cdot C_L\,(t, L_i, E, vc \geq 0). \end{aligned} \tag{4.31}$$

We can again depict this relationship in our example. Consider the location coverage for visit class $vc = 2$. Locations l_1, l_2 and l_5 fulfill the given visit class, leading to a location coverage for the complete location set of $C_L\,(t, L, E, vc = 2) = 3/5$. Considering the location subsets, the location coverage is $C_L\,(t, L_1, E, vc = 2) = 2/3$ and $C_L\,(t, L_2, E, vc = 2) = 1/2$ for L_1 and L_2, respectively. Applying Equation 4.31, we obtain:

$$\begin{aligned} C_L\,(t, L, E, vc \geq 0) &= \sum_{i=1}^{u} \frac{|L_i|}{|L|} \cdot C_L\,(t, L_i, E, vc \geq 0) \\ &= \frac{3}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot \frac{1}{2} = \frac{3}{5}. \end{aligned}$$

Summary Location Partitioning Table 4.4 summarizes the behavior of visit potential quantities under location partitioning as discussed in this section. It contains the relationships of the quantities on the complete location set to the quantities on the location subsets. The relations have been divided according to visit classes $vc = 0$, $vc = 1$ and $vc \geq 1$. Note that for notational convenience, we have dropped all parameters in the visit potential quantities that remain constant. The symbol $--$ means that no direct relationship can be established between individual quantities of the location partition and the quantity of the complete location set.

Table 4.4.: Relationship of visit potential quantities under location partitioning

	$\boldsymbol{vc = 0}$	$\boldsymbol{vc = 1}$	$\boldsymbol{vc > 1}$
$\boldsymbol{GV_E\,(L)}$	$= \sum_{i=1}^{u} GV(L_i)$		$\geq \sum_{i=1}^{u} GV_E\,(L_i)$
$\boldsymbol{GV_L\,(L)}$			$= \sum_{i=1}^{u} GV_L\,(L_i)$
$\boldsymbol{AV_E\,(L)}$	$= \sum_{i=1}^{u} AV_E\,(L_i)$	$\leq \sum_{i=1}^{u} AV_E\,(L_i)$	$--$
$\boldsymbol{AV_L\,(L)}$	$= \sum_{i=1}^{u} \frac{\lvert L_i\rvert}{\lvert L\rvert} AV_L\,(L_i)$	$= \sum_{i=1}^{u} \frac{C_L\,(L_i) \cdot \lvert L_i\rvert}{C_L\,(L) \cdot \lvert L\rvert} \cdot AV_L\,(L_i)$	
$\boldsymbol{C_E\,(L)}$	$= C_E\,(L_i) = 1$	$\geq max\,\{C_E\,(L_i)\},$ $\leq min\left\{1, \sum_{i=1}^{u} C_E\,(L_i)\right\}$	$\geq max\,\{C_E\,(L_i)\}$ ≤ 1
$\boldsymbol{C_L\,(L)}$	$= \sum_{i=1}^{u} \frac{\lvert L_i\rvert}{\lvert L\rvert} C_L\,(L_i)$		

Table 4.5.: Relationship of visit potential quantities under entity partitioning

	$\boldsymbol{vc = 0}$	$\boldsymbol{vc = 1}$	$\boldsymbol{vc > 1}$
$\boldsymbol{GV_E\,(E)}$	$= \sum_{j=1}^{w} GV(E_j)$		$= \sum_{j=1}^{w} GV_E\,(E_j)$
$\boldsymbol{GV_L\,(E)}$			$\geq \sum_{j=1}^{w} GV_L\,(E_j)$
$\boldsymbol{AV_E\,(E)}$	$= \sum_{j=1}^{w} \frac{\lvert E_j\rvert}{\lvert E\rvert} AV_E\,(E_j)$	$= \sum_{j=1}^{w} \frac{C_E\,(E_j) \cdot \lvert E_j\rvert}{C_E\,(E) \cdot \lvert E\rvert} \cdot AV_E\,(E_j)$	
$\boldsymbol{AV_L\,(E)}$	$= \sum_{j=1}^{w} AV_L\,(E_j)$	$\leq \sum_{j=1}^{w} AV_L\,(E_j)$	$--$
$\boldsymbol{C_E\,(E)}$	$= \sum_{j=1}^{w} \frac{\lvert E_j\rvert}{\lvert E\rvert} C_E\,(E_j)$		
$\boldsymbol{C_L\,(E)}$	$= C_L\,(E_j) = 1$	$\geq max\,\{C_L\,(E_j)\},$ $\leq min\left\{1, \sum_{j=1}^{w} C_L\,(E_j)\right\}$	$\geq max\,\{C_L\,(E_j)\}$ ≤ 1

4.3.3. Visit Potential under Partitioning of Entity Set

Similar considerations can me made for a partition $Part(E) = \{E_1, E_2, \ldots, E_w\}$ of entity set E. In fact, due to the analogous definition of visit potential quantities when considered from an entity or a location point of view, the relationships between gross visits of entities and gross visits of locations, average visits per entity and average visits per location, and between entity coverage and location coverage are reversed under partitioning of the entity set. We will therefore skip the derivation of the relationships and depict only the summary of results in Table 4.5. The symbol $--$ means again that no direct relationship can be established between individual visit potential quantities of the entity subsets and the quantity of the complete entity set.

4.4. Application Scenarios for Visit Potential

Visit potential is a family of generic quantities that can be applied to a wide range of applications that deal with interactions between mobile objects and locations. In this section we show the generalization capability of visit potential by application in two real-world domains. The implementations shall also help to clarify the definitions given in the previous sections. The first domain is outdoor advertising where we apply visit potential to define performance indicators of poster sites precisely. The second domain analyzes migration patterns of birds. Here visit potential is used to evaluate the distribution and frequency of bird sighting reports. While the first domain takes an entity point of view, the second domain utilizes the location point of view.

4.4.1. Visit Potential for the Evaluation of Poster Performance

In Chapter 3 we introduced the motivating application of this thesis only informally. After specifying the formal framework of visit potential in the previous sections, we will now define the application more precisely using the above introduced terminology.

First, we need to specify the objects of interest and their type of interaction, i.e. locations, entities, trajectories and visits. The universal set $\mathcal{L}$ of discrete geographic locations contains all poster locations under consideration, for example, all poster locations of a country. A location set $L \subseteq \mathcal{L}$ is instantiated by the selection of a specific poster campaign. Furthermore, the population $\mathcal{E}$ of mobile entities represents the population of the country of interest, and an entity set $E \subseteq \mathcal{E}$ specifies the target group for which performance indicators shall be evaluated. For example, the target group could contain only the inhabitants of a given city or of a specific sociodemographic group. A trajectory displays the movements of a person and a visit denotes the (weighted) passage of a person past a poster location. Remember that a poster passage may be weighted by several factors (e.g. passage speed, angle of passage, poster size, illumination) in order to consider the attention of the passers-by. However, as such quality factors are not considered further in this thesis, we will simply use the terminology *passage* or *contact* to refer to the one or other interpretation (see also Section 3.1.2).

Note that although we instantiate the objects of interest with real-world objects, they are represented by expressions of one or more reference systems. Thus, a poster location is specified in a geographic coordinate space $\mathcal{S}_\mathcal{C}$ using an appropriate data structure. Entities are referred to by some identifier, and the trajectory of an entity is given as function mapping a point from a temporal coordinate space $\mathcal{T}_\mathcal{C}$ to a point in a geographic coordinate space $\mathcal{S}_\mathcal{C}$. In consequence, visits are calculated using the provided coordinates spaces. Due to the unique mapping from each reference system to the real world we can evaluate the meaning of the given expressions. For details on and definitions of physical and temporal space as well as geographic and temporal reference systems see Sections 2.1.1 and 2.2.1.

Having defined the basic units of interest, we now define the performance indicators coverage, reach, effective reach, opportunities to see, gross rating points and gross impressions in terms of visit potential. Usually performance indicators are evaluated for a period of time of $t = 7$ days. However, longer periods of 10, 14 or even 21 days are also possible.

Definition 4.4.1 (Coverage) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$ and the visit count of an entity $NV(t, L, e)$ $\forall e \in E$ obtained by the evaluation of poster passages, coverage corresponds to the visit potential quantity entity coverage:*

$$coverage := C_E\,(t, L, E) = \frac{|\{e \in E \mid NV(t, L, e) \geq 1\}|}{|E|}.$$

Definition 4.4.2 (Reach) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$ and the visit count of an entity $NV(t, L, e)$ $\forall e \in E$ obtained by the evaluation of poster contacts, reach corresponds to the visit potential quantity entity coverage:*

$$reach := C_E\,(t, L, E) = \frac{|\{e \in E \mid NV(t, L, e) \geq 1\}|}{|E|}.$$

Note that the definitions of coverage and reach differ only by the specification of a visit.

Definition 4.4.3 (Effective reach) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$, the visit count of an entity $NV(t, L, e)$ $\forall e \in E$ obtained by the evaluation of poster contacts and a contact class $vc \in \mathbb{N}_0$, effective reach corresponds to the visit potential quantity entity coverage for visit class vc:*

$$effective\ reach := C_E\,(t, L, E, vc) = \frac{|\{e \in E \mid NV(t, L, e) \geq vc\}|}{|E|}.$$

Definition 4.4.4 (Opportunities to see (OTS)) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$, the visit count of an entity $NV(t, L, e)$ $\forall e \in E$ obtained by the evaluation of poster contacts and a contact class $vc \in \mathbb{N}_0$, opportunities to see correspond to the visit potential quantity average visits per entity for visit class vc:*

$$OTS := AV_E\,(t, L, E, vc) = \frac{\sum_{e \in E \,|\, NV(t,L,e) \geq vc} NV(t, L, e)}{|\{e \in E \mid NV(t, L, e) \geq vc\}|}.$$

Definition 4.4.5 (Gross rating points (GRP)) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$ and the elementary visit count $NV(t, l, e)$ $\forall l \in L$ and $\forall e \in E$ obtained by the evaluation of poster contacts, gross rating points correspond to the hundredfold of the visit potential quantity average visits per entity:*

$$GRP := AV_E\,(t, L, E) \cdot 100 = \frac{\sum_{l \in L} \sum_{e \in E} NV(t, l, e)}{|E|} \cdot 100.$$

Definition 4.4.6 (Gross impressions) *Given a poster campaign L, a target audience E, a time span $t \in \mathbb{N}_0$ and the elementary visit count $NV(t, l, e)$ $\forall l \in L$ and $\forall e \in E$ obtained by the evaluation of poster contacts, gross impressions correspond to the visit potential quantity gross visits:*

$$gross\ impressions := GV(t, L, E) = \sum_{l \in L} \sum_{e \in E} NV(t, l, e).$$

In Section 3.1.2 we stated a fundamental relationship between gross impressions, reach and opportunities to see (Equation 3.1). Substituting the terms with the above definitions, the relationship as stated in Equation 4.7 appears. Note that Equation 3.1 assumes that reach is given as percentage, therefore an additional multiplication with factor 100 is introduced on the right hand side of the equation.

$$\begin{aligned} GRP &= reach \cdot OTS \\ AV_E\left(t, L, E, vc = 0\right) \cdot 100 &= C_E\left(t, L, E, vc = 1\right) \cdot 100 \cdot AV_E\left(t, L, E, vc = 1\right) \end{aligned} \tag{4.32}$$

Note that the mobility surveys conducted in Switzerland and Germany capture the mobility of only a sample of the population. Thus, visit potential quantities are actually calculated for the data sample and afterwards extrapolated to the complete population. Furthermore, the indicators are intended to state the average performance of a given time span. The surveys therefore take place over a longer period of time and provide different start days to the persons in the survey. For evaluation the data is synchronized by these dates as described in Section 4.2.1 and additionally permuted in order to avoid sampling-related patterns within the measurement days.

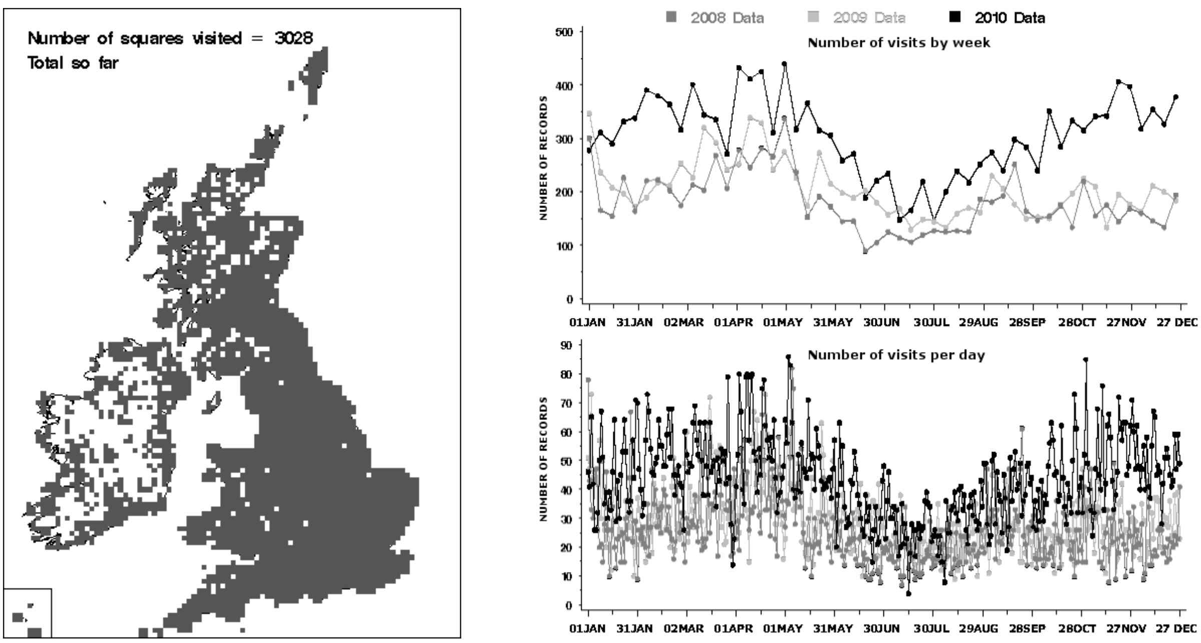
4.4.2. Visit Potential for the Evaluation of Bird Recordings

BirdTrack (BirdTrack, 2011) is a joint project of the British Trust for Ornithology (BTO), the Royal Society for the Protection of Birds (RSPB) and BirdWatch Ireland to record the migration movements and distribution of birds in Great Britain and Ireland. The project relies on volunteers from all over Great Britain and Ireland to watch and record the observation of bird species. The records can be entered over an online form and are evaluated on a daily basis. The results comprise, amongst others, maps and graphs showing the distribution of bird observations as well as time-based animations that allow to trace the arrival and departure of migrating species and their movements across the country.

BirdTrack aims at the nationwide observation of birds and therefore relies on volunteers to report sightings. This means that the origin and number of reports on bird sightings cannot be controlled and are certain to vary over time and space. BirdTrack therefore analyzes not only the reported species but also the distribution and frequency of submissions. It provides maps about the covered areas and time series diagrams about the number of records. We will examine these two statistics in more detail in this section. In order to gain a better understanding of the project itself, we begin with a description of the data collection method. Afterwards we show examples of the two mentioned statistics, and finally we define the statistics in terms of visit potential.

All persons who join the project have to register one or more sites for which they record bird sightings. Such sites may be of various form and range from gardens over local parks to (parts of) nature reserves. The sites can be specified either by map grid reference, postcode or interactive search using Google Maps. Later on, most evaluations are conducted on a 10 by 10 km grid unto which all sites are mapped. When a person enters his or her sightings, the visited site, date and time of the bird watching trip are queried and a list of birds is presented. The person can then mark all species that he or she encountered during the trip. In addition, information about the numbers, age, sex or breeding status of each sighted species can be entered.

BirdTrack evaluates the reports on a daily basis. It provides regional and national statistics about the reported species as well as about the submitted records. The latter are of interest in this thesis. Statistics about the submitted records consist of two types which are depicted in Figure 4.7. The left figure is a map which shows all sites that have been visited within the year 2010. This map is called a coverage map and available for each single month as well as for the complete year. Figure 4.7 right shows the number of submitted records by week and per day for Wales. These graphs are called coverage graphs and are available for different regions as well as for the whole of Great Britain and Ireland.



(a)

(b)

Figure 4.7.: (a) BirdTrack coverage map for Great Britain and Ireland Jan. 1st - Dec. 31th 2010 (b) BirdTrack coverage graphs for Wales by week and per day 2008, 2009 and 2010; source of figures: BirdTrack (2011)

Although provided as graphics, coverage map and coverage graph are closely related to visit potential. Both statistics can be defined more formally in terms of visit potential. The universal set $\mathcal{L}$ of discrete geographic locations consists of all squares in a 10 by 10 km grid of Great Britain and Ireland. The location set L is equal to the universal set $\mathcal{L}$ and may be partitioned into regional subsets $Part(L) = \{L_{nSc}, L_{sSc}, L_{neE}, L_{nwE}, L_{YH}, L_{eMl}, L_{wMl}, L_W,$ $L_{eE}, L_L, L_{seE}, L_{swE}, L_{NI}, L_I\}$. Hereby, the stated subsets refer to the regions listed in Table 4.6.

Table 4.6.: BirdTrack regional location subsets

L_{neE} - North East England	L_{nSc} - North Scotland	L_W - Wales
L_{nwE} - North West England	L_{sSc} - South Scotland	L_{eE} - East of England
L_{seE} - South East England	L_{eMl} - East Midlands	L_{NI} - Northern Ireland
L_{swE} - South West England	L_{wMl} - West Midlands	L_I - Republic of Ireland
L_{YH} - Yorkshire and the Humber		L_L - London

The population $\mathcal{E}$ consists of all people that live or travel through Great Britain and Ireland, and the entity set E consists of all registered users of the BirdTrack project. Note that users join the project on a voluntary basis and are therefore not necessarily representative for the population. However, representativity of volunteers is not the primary aim of BirdTrack. For the project it is more important that a sufficient number of reports are regularly available for all locations of the grid.

The trajectories are given by the sequence of submitted reports per volunteer. They take the form of snapshots in time and space and are collected over the whole year. The trajectories are already reduced to the visits, as only the locations of bird watching are of interest. Note that in case of BirdTrack a visit cannot be defined by passing a location for a minimum time span ε alone. A visit also requires the observation of birds. For example, a person may cross a park where he or she regularly watches birds on his / her way to work. During the journey, however, he or she will not have the time to observe birds. Note also that all visits to individual bird watching sites are mapped to a 10 by 10 km grid before evaluation.

Having defined the basic units of interest, we now translate the statistics displayed in the BirdTrack coverage maps and graphs. The coverage map statistic is available for $t = 1$ month or for $t = 1$ year.

Definition 4.4.7 (Coverage map statistic) *Given a 10 by 10 km grid over Great Britain and Ireland L, a set of bird watching volunteers E, a time span $t \in \mathbb{N}_0$ and the visit count of a location $NV(t, l, E)$ $\forall l \in L$ obtained by the evaluation of bird watching reports, the coverage map statistic corresponds to the visit potential quantity location coverage:*

$$\textit{coverage map statistic} := C_L\,(t, L, E) = \frac{|\,\{l \in L \mid NV(t, l, E) \geq 1\}\,|}{|L|}.$$

Definition 4.4.8 (Coverage graph statistic) *Given a 10 by 10 km grid over Great Britain and Ireland L, a set of bird watching volunteers E, a time span $t \in \mathbb{N}_0$ and the elementary visit count $NV(t, l, e)$ $\forall l \in L$ and $\forall e \in E$ obtained by the evaluation of bird watching reports, the coverage graph statistic corresponds to the visit potential quantity gross visits:*

$$\textit{coverage graph statistic} := GV(t, L, E) = \sum_{l \in L} \sum_{e \in E} NV(t, l, e).$$

Coverage graph statistics are displayed as time series and show aggregated visits for $t = 1$ day or $t = 1$ week. The graphs are available for the complete location set L or for regional subsets, i.e. the graphs can be obtained for any subset of the grid which is listed in Table 4.6.

4.5. Visit Potential by Example

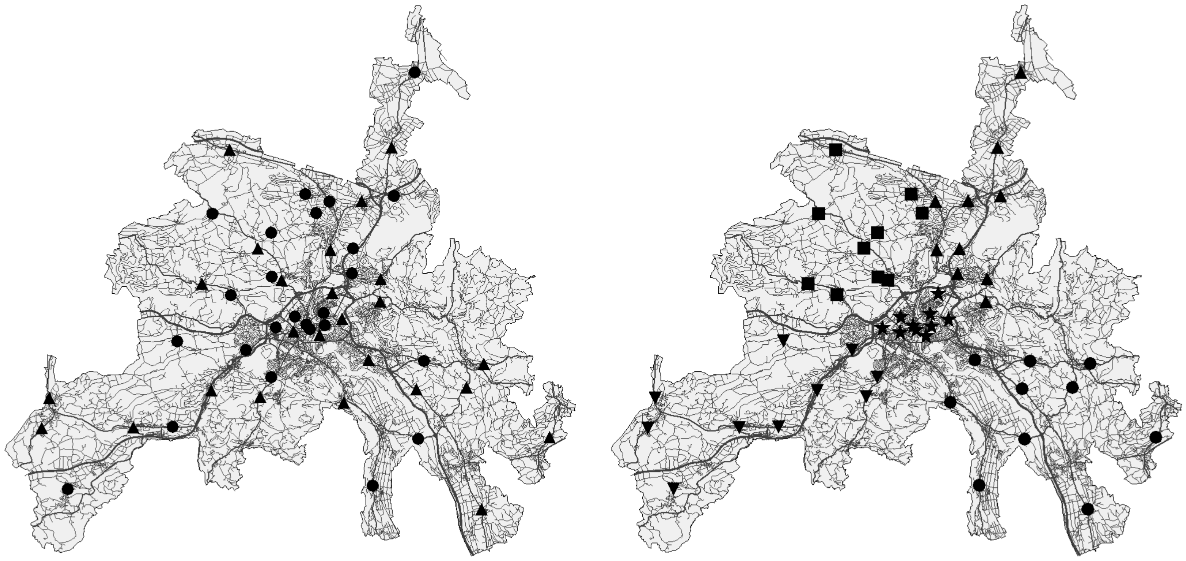
This section illustrates typical characteristics of visit potential quantities by example calculations on real-world application data. The examples underline, on the one hand, the formal relationships between visit potential quantities. On the other hand, they show the behavior of visit potential quantities as induced by mobility patterns that are typical for humans. Especially the locality and repetitive character of human movement behavior as described in Section 2.2.3 becomes visible. Finally, the examples show the usefulness of the defined quantities in a real-world application domain.

For the following examples we use data of the conurbation Bern from the Swiss audience measurement study. Note that although we use real-world application data, for the purpose of this thesis we do not incorporate the full complexity of the actual performance model. The results therefore do not conform to the true performance measurements and are meant for demonstration purpose only.

Our entity set E is a subset of the test persons in Bern and contains 635 persons with seven or more measurement days. We selected two different poster campaigns for the locations sets. The first (L_1) is spread over the whole conurbation and consists of 50 posters. The second (L_2)

is concentrated in the city center and consists of 10 posters. For simplicity, we only consider poster passages and do not account for visibility criteria in the examples. In addition, we calculate all visit potential quantities for the data sample and not for the true population.

We calculate the visit potential quantities gross visits, average visits per entity, average visits per location, entity and location coverage for visit classes $vc = 0$ and $vc = 1$ over a period of $t = 7$ days. The first experiment calculates the visit potential for location set L_1 and entity set E. In the second experiment we randomly partition L_1 into two subsets L_{R1}, L_{R2} of 25 posters each. The partition is depicted in Figure 4.8(a). Experiment 3 again partitions L_1, however, according to geographic characteristics. The posters are grouped according to the city center and the cardinal directions northwest, northeast, southwest and southeast (see Section 2.1.4). The location subsets are referred to by the variables L_C, L_{NW}, L_{NE}, L_{SW} and L_{SE} and are depicted in Figure 4.8(b). The fourth experiment uses location subset L_C where posters are dispersed over the city center and compares the results to the outcome of location set L_2 where all posters are clustered in the city center. Both location sets are depicted in magnification in Figure 4.9. Finally, Experiment 5 uses location subset L_{SE} and reports visit potential quantities under entity partitioning. In this experiment we grouped the entity set according to the place of living of the test persons. We used the city center and the cardinal directions west and east for partitioning. The obtained entity subsets E_C, E_W and E_E are depicted in Figure 4.10. The visit potential quantities for the various experiments are shown in Tables 4.7-4.11.


(a)

(b)

Figure 4.8.: Location partitioning in Bern (a) random partition (b) directional partition

We begin with a general description of results for Experiment 1, which are depicted in Table 4.7. In total, our 635 test persons produce 4730 visits with the selected 50 locations within one week. On average each person performs 7.4 visits and each location is visited 94.6 times. However, not all test persons and locations produce visits. Only 86.1 percent of our test persons (corresponding to 547 persons) interact with the location set. If we distribute the gross visits only among these persons, we obtain an average of 8.6 visits per person. Similarly, only 84 percent of the locations are visited (corresponding to 42 locations). If we restrict the evaluation to these locations, their average number of visits per location increases to 112.6.

Experiment 2 (see Table 4.8) calculates quantities for a random partition of the location set

Table 4.7.: Visit potential quantities for Experiment 1 with location set L_1, entity set E and $t = 7$ days

$\boldsymbol{GV}$ $vc = 0$	$\boldsymbol{AV_E}$ $vc = 0$	$\boldsymbol{AV_E}$ $vc = 1$	$\boldsymbol{C_E}$ $vc = 1$	$\boldsymbol{AV_L}$ $vc = 0$	$\boldsymbol{AV_L}$ $vc = 1$	$\boldsymbol{C_L}$ $vc = 1$	$\boldsymbol{\|L\|}$	$\boldsymbol{\|E\|}$
4,730	7.4	8.6	0.861	94.6	112.6	0.840	50	635

used in Experiment 1. Naturally, gross and average visits per entity of Experiment 2 therefore add up to the gross visits and average visits per entity of Experiment 1. Average visits per location and location coverage of Experiment 2 lead also to the numbers in Experiment 1 when applying the formulas of Table 4.4. When we compare quantities only across the location subsets of Experiment 2, it is noticeable that their values vary only slightly. This results from the random partitioning of locations. Locations with numerous visits are as likely to appear in one subset as locations with few visits. In addition, both subsets are again dispersed over the conurbation. This is reflected in the entity coverage. Although entity coverage declines in both location subsets, the obtained values are still high.

Table 4.8.: Visit potential quantities for Experiment 2 with random partitioning of location set L_1, entity set E and $t = 7$ days

	$\boldsymbol{GV}$ $vc = 0$	$\boldsymbol{AV_E}$ $vc = 0$	$\boldsymbol{AV_E}$ $vc = 1$	$\boldsymbol{C_E}$ $vc = 1$	$\boldsymbol{AV_L}$ $vc = 0$	$\boldsymbol{AV_L}$ $vc = 1$	$\boldsymbol{C_L}$ $vc = 1$	$\boldsymbol{\|L\|}$	$\boldsymbol{\|E\|}$
L_{R1}	2,417	3.8	5.3	0.712	96.7	105.1	0.920	25	635
L_{R2}	2,313	3.6	5.5	0.663	92.5	121.7	0.760	25	635
total	4,730	7.4	−−	−−	94.6	112.6	0.840	50	635

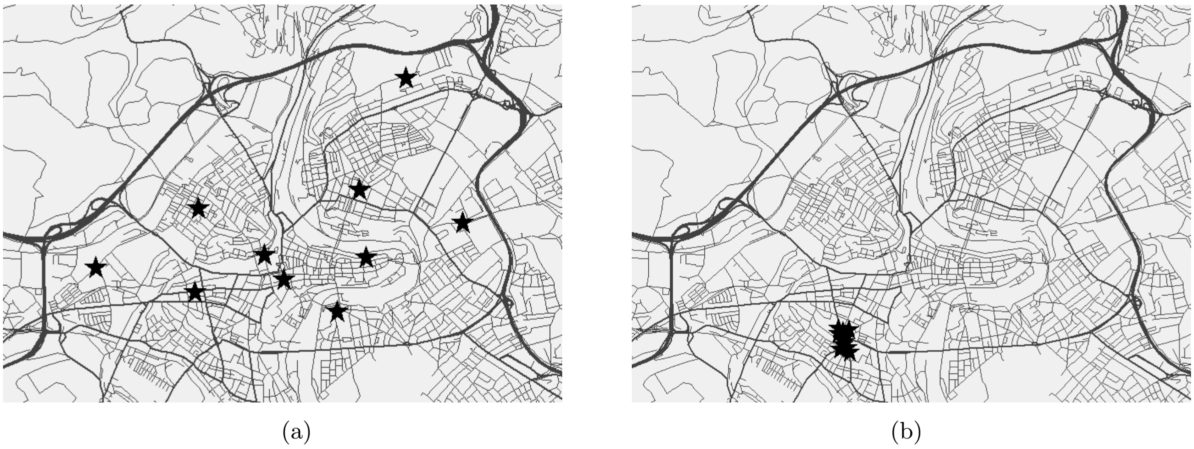
In Experiment 3 we partition the location set according to geographic characteristics into five subsets as depicted in Figure 4.8(b). The results are depicted in Table 4.9 and comply with the values in Experiment 1. However, the distribution of measured values differs between the location subsets. The largest parts of gross visits are produced in the city center and the southeast of Bern. As the average visits per entity for visit class $vc = 1$ are similar for all location subsets, the difference in gross visits must be caused by the number of visiting entities. This is confirmed by the entity coverage, which is much higher in the city center and the southeast of Bern than in the other parts. However, this result is not surprising as about one third of the test persons live each in the city center and southeast of Bern and are likely to produce most of their visits within their neighborhood. The number of average visits per location shows clearly that places in the center and southeast are most attractive. The average number of visits per location allows a direct comparison of per-site performance between location subsets because the quantity is independent of the size of the location sets. In Example 3 average visits per entity and gross visits contain the same information as our location subsets are of equal size. Experiment 3 also shows the typical human mobility pattern of movement between the outskirts and center of a city. The entity coverage of Bern city center is much higher than the percentage of test persons that live in the city center. Consequently, the center attracts further people from all over Bern.

Experiment 4 demonstrates the behavior of visit potential with respect to the spread of locations. In Experiment 3 we already calculated visit potential for 10 posters that are distributed over the city center. We now select 10 posters that are clustered in a small area of the city center. Figure 4.9 shows both location sets and Table 4.10 contains the experimental results. The entity coverage of the clustered campaign decreases considerably. However, the number of average visits per entity for visit class $vc = 1$ are very high. This results from the correlation

Table 4.9.: Visit potential quantities for Experiment 3 with directional partitioning of location set L_1, entity set E and $t = 7$ days

	$\boldsymbol{GV}$	$\boldsymbol{AV_E}$	$\boldsymbol{AV_E}$	$\boldsymbol{C_E}$	$\boldsymbol{AV_L}$	$\boldsymbol{AV_L}$	$\boldsymbol{C_L}$	$\lvert\boldsymbol{L}\rvert$	$\lvert\boldsymbol{E}\rvert$
	$vc = 0$	$vc = 0$	$vc = 1$	$vc = 1$	$vc = 0$	$vc = 1$	$vc = 1$		
L_C	1,314	2.1	4.1	0.504	131.4	187.7	0.700	10	635
L_{NW}	447	0.7	6.0	0.117	44.7	49.7	0.900	10	635
L_{NE}	626	1.0	4.5	0.217	62.6	89.4	0.700	10	635
L_{SW}	840	1.3	6.9	0.191	84.0	93.3	0.900	10	635
L_{SE}	1,503	2.4	6.4	0.372	150.3	150.3	1.000	10	635
total	4,730	7.4	−−	−−	94.6	112.6	0.840	50	635

of movement in geographic space. If a person visits one of the clustered locations, he or she naturally passes through the neighborhood and is thus likely to visit further locations of the location set. Note, however, that positive correlations between locations do not arise from geographic proximity alone. For example, consider two parallel streets of a city. The streets may be close in space, however, a person will travel either along the one or the other road. The number of average visits per location shows that locations in L_2 are higher frequented than in L_C, which leads to higher gross visits in L_2. In summary, Experiment 4 shows that a high number of visits does not automatically imply high entity coverage and vice versa. Instead, entity coverage also depends on the choice of location from a geographic point of view.


(a)

(b)

Figure 4.9.: (a) Dispersed campaign in city center of Bern (b) clustered campaign in city center of Bern

Table 4.10.: Visit potential quantities for Experiment 4 with location set L_C and L_2, entity set E and $t = 7$ days

	$\boldsymbol{GV}$	$\boldsymbol{AV_E}$	$\boldsymbol{AV_E}$	$\boldsymbol{C_E}$	$\boldsymbol{AV_L}$	$\boldsymbol{AV_L}$	$\boldsymbol{C_L}$	$\lvert\boldsymbol{L}\rvert$	$\lvert\boldsymbol{E}\rvert$
	$vc = 0$	$vc = 0$	$vc = 1$	$vc = 1$	$vc = 0$	$vc = 1$	$vc = 1$		
L_C	1,314	2.1	4.1	0.504	131.4	187.7	0.700	10	635
L_2	2,706	4.3	17.6	0.243	270.6	270.6	1.000	10	635

The final experiment shows the behavior of visit potential under geographic partitioning of the entity set (see Table 4.11). We divide the entity set according to the test persons' place of living using the city center as well as the cardinal directions west and east (see Figure 4.10). The location set consists of all southeastern locations L_{SE}. All quantities show the clear dominance of test persons from eastern Bern in the southeastern part of the city. Note that only about 50 percent of persons in eastern Bern (E_E) produce visits. Presumably, these people live in the southeast of Bern. For a detailed analysis, however, a further partitioning of the entity set is necessary. Again we can form quantities of the complete entity set from the partition.

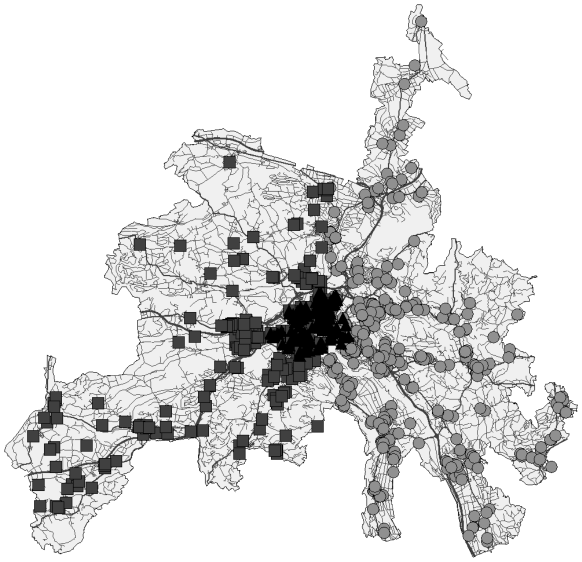


Figure 4.10.: Entity partitioning in Bern by cardinal direction

Table 4.11.: Visit potential quantities for Experiment 5 with location set L_{SO}, directional partitioning of entity set E and $t = 7$ days

	$\boldsymbol{GV}$	$\boldsymbol{AV_E}$	$\boldsymbol{AV_E}$	$\boldsymbol{C_E}$	$\boldsymbol{AV_L}$	$\boldsymbol{AV_L}$	$\boldsymbol{C_L}$	$\boldsymbol{\lvert L \rvert}$	$\boldsymbol{\lvert E \rvert}$
	$vc = 0$	$vc = 0$	$vc = 1$	$vc = 1$	$vc = 0$	$vc = 1$	$vc = 1$		
E_C	54	0.4	2.1	0.203	5.4	7.7	0.700	10	128
E_W	61	0.3	2.2	0.153	6.1	7.6	0.800	10	183
E_E	1,388	4.3	7.6	0.562	138.8	138.8	1.000	10	324
total	1,503	2.4	6.4	0.372	150.3	−−	−−	10	635

The examples in this section demonstrate that visit potential is a powerful tool to analyze the interaction between mobile entities and geographic locations. They also show that visit potential quantities complement each other and transport the full meaning of a context only in combination. For example, in outdoor adverting gross visits can be used to estimate the overall power of a campaign. However, an analysis of entity coverage and average visits per entity allows for optimal placement of posters and a high spread of information. In addition, the examples show the usefulness of location and entity partitioning, which can be used to trace mobile behavior with respect to different geographical areas and at several levels of resolution. Such partitions are applied in outdoor advertising to distinguish, for example, the visits of inhabitants and commuters.

4.6. Summary

A number of companies and research institutions apply entity-location interaction quantities in their day-to-day business. However, these quantities are not generally defined. Typically, quantities are tailored to specific applications, use context-dependent terminology and are often only informally written down. As a result, a number of quantities have evolved which are not suitable for methodological research and interdisciplinary exchange as their common background is hard to identify. In this chapter we therefore provided a formalization and application-independent notation of entity-location interaction quantities.

First, we formally defined spatiotemporal interactions between mobile entities and geographic locations and provided a mathematic concept to specify the number of visits over time and to aggregate interactions between sets of locations and sets of entities. Second, based on this concept, we defined a family of quantities called *visit potential*, which contains basic entity-location interaction quantities. We extended these quantities with respect to *visit classes* in order to disclose more characteristics of the underlying frequency distributions of k-visiting entities and k-visited locations. Based on the common background of all visit potential quantities, we were third able to analyze relationships between different quantities and to infer consistency requirements under partitioning of the location set and the entity set. Fourth, we demonstrated the general applicability of visit potential using two real-world applications, namely outdoor advertisement and bird tracking, for which we gave a precise definition of the employed entity-location interaction quantities in terms of visit potential. Finally, we illustrated typical characteristics of visit potential quantities by example calculations on real-world application data of outdoor advertising.

Given the formalization of Chapter 4 we are now able to specify entity-location interactions precisely and to analyze the relationship between a set of mobile entities and a set of geographic locations using visit potential.

5. Robust Estimation of Visit Potential under Missing Data

> *'Any conclusions [concerning the Pattern]?'*
>
> *'Either it possesses an element of irrationality itself, like living things, or it is an intelligence on such an order that some of its processes only seem irrational to lesser beings.'*
>
> (Roger Zelazny, The Great Book of Amber)

In the previous chapter we *defined* visit potential quantities and analyzed their characteristics. This chapter treats the *estimation* of visit potential quantities based on mobility data. The challenge of this estimation lies in the incompleteness of provided mobility data in real-life applications. Existing research on trajectory data mining has mostly neglected the problem of incomplete data. Outside of our work there is no publication to-date which systematically analyzes techniques for the handling of missing movement data. For example, Schönfelder and Axhausen (2001) apply survival analysis to investigate recurrent activities from travel diaries. However, the authors make the assumption that the activity data meets the preconditions of the applied algorithms. Algorithms that do not address incompleteness or do not address incompleteness correctly, inevitably lead to poor results. This is a problem in practice where typically only a small fraction of data sets are complete as, for example, in the application domain presented in Chapter 3. This chapter therefore addresses the estimation of visit potential under missing data. Missing data have been addressed in a number of publications outside of the spatiotemporal domain. We therefore begin with a review of the literature and evaluate the applicability of approaches to the spatiotemporal domain. We then adapt and apply the most promising methods to estimate visit potential and analyze their robustness with respect to different missing data mechanisms and a varying percentage of missing data.

We perform our experiments on mobility data of the German outdoor advertising application. Note, however, that our experiments focus only on one part of the modeling process within the application, namely the treatment of missing data, and do not incorporate the full complexity of the actual model. The obtained results therefore do not correspond to poster performance as provided by the outdoor advertising business sector.

This chapter is organized as follows. Section 5.1 introduces patterns and types of missing data and analyzes the presented application data with respect to these properties. Section 5.2 provides a comprehensive literature review of estimation methods for missing data and evaluates their usability and practicability with respect to visit potential and the application domain. In addition, we describe four methods in detail which we selected for evaluation. In Section 5.3 we set up the test scenario for robustness evaluation of the methods. In Section 5.4 we describe the parameter tuning process of the selected methods and in Sections 5.5 - 5.7 we evaluate the methods for a selection of visit potential quantities using the outdoor advertising test scenario. We provide a summarizing discussion and a perspective on future work in Section 5.8 and conclude the chapter with a short summary.

Before proceeding to the next sections note that this chapter combines aspects from different areas of statistics and machine learning, each having slightly different notational conventions. Some notations in this chapter are therefore overloaded. This applies especially to the usage of upper and lower case letters. We follow the common statistical notation and use upper case letters to denote random variables and lower case letters to denote specific values of such a variable. At the same time we will use upper case letter to denote (parts of) data matrices and lower case letters to denote an element of a matrix. Finally, we will use upper case letters to denote distribution functions and lower case letters to denote probability density functions. However, the respective meaning will be clear from the given context.

Excerpts of this chapter have been published with significant contribution of the author in (May et al., 2009a), (May et al., 2009b), (Hecker et al., 2010c) and (Pasquier et al., 2008).

5.1. Characteristics of Missing Data

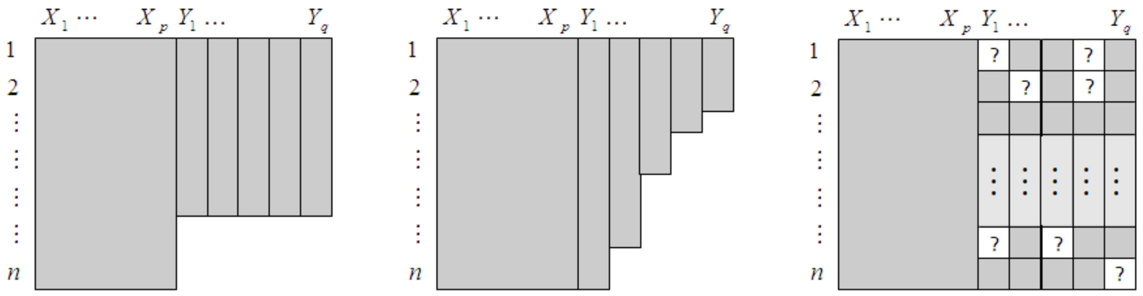
The absence of measurement data can be caused by various reasons. Therefore, the shape of missing data differs as well. The literature distinguishes, on the one hand, the *pattern* and, on the other hand, the *mechanism* of missing data. The pattern reflects, for example, the random or sequential absence of data while the mechanism captures relationships between the *missingness* and other variables of the data set. The term missingness is used in the literature to refer to a variable that indicates whether the value of an observed quantity is known or missing.

This section begins with an introduction to the patterns and mechanisms of missing data and then proceeds to the analysis of the application data with respect to these two characteristics.

5.1.1. Patterns of Missing Data

The pattern of missing data influences the choice of the missing data method that can be applied and is therefore an important characteristic of a data set. Nonresponse in surveys is generally distinguished into *unit nonresponse* and *item nonresponse* (Schafer and Graham, 2002). Unit nonresponse refers to the case when a person refuses to participate in a survey or cannot be reached by the survey request, i.e. the data are completely missing. Item nonresponse occurs if a person does not answer all questions of the survey, i.e. the data are observed only in part. This thesis focuses on item nonresponse in mobility surveys. In compliance with data mining terminology, “units” of a survey will be referred to as “entities” or “objects”, and “items” are termed “variables” or “attributes”. Further, we will denote random variables with upper case letters and realizations of the variables with lower case letters.

Item nonresponse can take several forms, and commonly the univariate, monotone and arbitrary pattern of missingness are distinguished (Schafer and Graham, 2002; Little and Rubin, 2002). Consider a data set with n objects and m attributes that can be represented by an $n \times m$ matrix. We partition the attributes into a list of independent variables $X = (X_1, \ldots, X_p)$ which are observed completely, and a list of dependent variables $Y = (Y_1, \ldots, Y_q)$ which may contain missing values, further $p + q = m$. The resulting data sets are represented as matrix $X = (x_{ij})$ with $i = 1..n$ and $j = 1..p$, respectively $Y = (y_{ik})$ with $i = 1..n$ and $k = 1..q$, where x_{ij} denotes for object i the value of variable X_j and y_{ik} denotes the value of variable Y_k. If the set of variables Y is either completely observed or completely missing for each object, the pattern is called univariate (see Figure 5.1a). Note the special case $q = 1$, which always results in a univariate pattern. A monotone pattern occurs if for all objects the following property holds: $\forall j = 1..q$ if Y_j is missing then $Y_{j+1}, \ldots, Y_q$ are missing as well (Figure 5.1b). Monotone patterns usually arise in longitudinal studies with dropout behavior. Longitudinal studies perform repeated observations of the dependent variable over a longer period of time. Dropouts refer to objects that leave the study and do not enter it again at a later point in time. For example, persons may leave a medical study because they change their place of living. Finally, an arbitrary pattern of missingness arises if intermittent missing values occur in the data (Figure 5.1c).



(a) univariate pattern

(b) monotone pattern

(c) arbitrary pattern

Figure 5.1.: Patterns of missing data; figures adopted from Schafer and Graham (2002)

We can encode the missingness of Y within a separate (multivariate) variable, which is typically named M or R. Hereby, M usually refers to an encoding scheme where *missing* values denote positive events whereas R interprets observed values, i.e. the *response*, as positive events. In this thesis we will use the latter encoding scheme. Let I_R denote an indicator function which has a value of one if the value of its argument is observed and zero if it is missing. More formally, let the term *null* encode a missing value, for any v

$$I_R(v) = \begin{cases} 1 & \text{if } v \neq \mathit{null} \\ 0 & \text{if } v = \mathit{null}. \end{cases}$$

Depending on the pattern of missingness, different definitions of R are possible. In the univariate case $R = (r_i)$ with $i = 1..n$ is a vector with one value r_i for each object. The value of r_i is equal for a all variables $Y_1, ..., Y_q$ of object i and has the value

$$r_i = I_R(y_{i1}) = \ldots = I_R(y_{iq}) \qquad \forall i = 1..n.$$

In case of a monotone pattern of missingness R is a vector that states the number of observed values per object, i.e.

$$r_i = \sum_{j=1}^{q} I(y_{ij}) \qquad \forall i = 1..n.$$

For arbitrary patterns of missingness R is a matrix which states the response of each dependent variable and object, i.e. $R = (R_1, \ldots, R_q) = (r_{ij}) = I_R(y_{ij})$ with $i = 1..n$ and $j = 1..q$.

Depending on the relationship between R and Y different *mechanisms* of missing data are distinguished in the literature which we describe in the next section.

5.1.2. Mechanisms of Missing Data

The first major works on missing data appeared in the 1970s. Rubin (Rubin, 1976) introduced a typology for missing data mechanisms and discussed their effect on the inference process. The term mechanism hereby refers to the relationship between missing data and the variables or values of variables in the considered data set, not to the actual real-world process behind the missingness. Three variants of missing data are distinguished in the literature: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR) (Little and Rubin, 2002; Schafer and Graham, 2002). The differentiation is important, because the properties of methods that treat missing data depend strongly on these dependencies.

Assuming complete knowledge Y denotes the the matrix of complete observations. Y can be partitioned into subsets Y_{obs} and Y_{mis} that contain the values for observed and unobserved parts of the data, respectively. The different mechanisms of missing data are then defined as follows, given that all objects are sampled independently from the population.

Definition 5.1.1 (Missing completely at random (MCAR)) *Missing values are missing completely at random if the missingness is independent of the data, i.e.*

$$P(R \mid X, Y) = P(R).$$

We can relax MCAR to allow a dependency between the missingness and the observed data, which results in MAR.

Definition 5.1.2 (Missing at random (MAR)) *Missing values are missing at random if the missingness depends at most on the observed data, i.e.*

$$P(R \mid X, Y) = P(R \mid X, Y_{obs}).$$

Finally, we obtain MNAR if the missingness depends on the missing values themselves and cannot be removed by conditioning on the observed values.

Definition 5.1.3 (Missing not at random (MNAR)) *Missing values are missing not at random if the missingness depends on the unobserved data, i.e. Definition 5.1.2 is violated:*

$$P(R \mid X, Y) \neq P(R \mid X, Y_{obs}).$$

For longitudinal data one further mechanism of missingness is distinguished by Little (1995): covariate-dependent missingness. The term *covariate* refers to the independent variables and means that the missingness may depend only on these completely observed variables. It is thus a stricter version of MAR, which we will call in this thesis CDMAR.

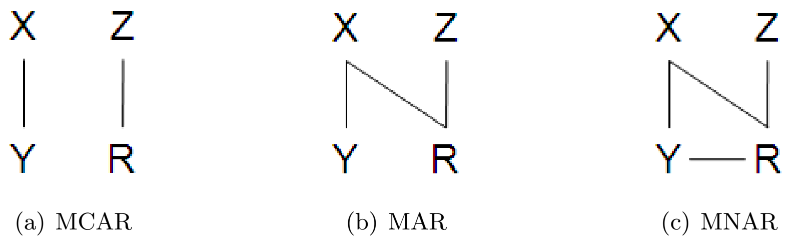
Definition 5.1.4 (Covariate-dependent missing at random (CDMAR)) *Missing values are covariate-dependent missing at random if the missingness depends at most on completely observed independent variables of the data, i.e.*

$$P(R \mid X, Y) = P(R \mid X).$$

Note that in the case of univariate missing data MAR and CDMAR coincide.

In order to give an intuitive explanation of the definitions, consider the univariate setting in which the values of Y are either completely observed or completely missing for each object. In this case Definition 5.1.2 simplifies to $P(R \mid X, Y) = P(R \mid X)$.

MCAR occurs if the probability of missing values depends neither on X nor on Y. This relationship is depicted in Figure 5.2a. Variable Z hereby denotes influences on R which, however, are independent of X and Y. If a relationship between R and X exists but R is still independent of Y, the data are defined to be MAR. MAR denotes a conditional independence of missingness given a fixed value of X (see Figure 5.2b). However, under MAR a relationship between R and Y may exist due to their mutual dependency on X. This relationship disappears



(a) MCAR (b) MAR (c) MNAR

Figure 5.2.: Graphical models of types of missing data in the univariate setting; source of figures: Schafer and Graham (2002)

once the value of X is taken into account. Finally, if the distribution of missing values depends on Y, the data are said to be MNAR (see Figure 5.2c).

For example, assume that we want to observe the daily travel distance of the German population. We recruit a representative sample of size n of the population which we ask for the sociodemographic variables gender and age, as well as their traveled distance on the previous day. For gender and age we obtain a complete observation, i.e. $X = (X_g, X_a)$. However, not all persons remember their travel distance of the previous day. Travel distance Y and response R are two vectors of length n. Let us assume that travel distance and age are related, which is a well-known result from travel surveys (see Section 2.2.3). MCAR exists if missingness depends neither on sociodemographic characteristics nor on travel distance. If missingness depends on age, for example, older persons may be less likely to recall their travel distance, the mechanism is MAR. Finally, MNAR results if missingness depends on the missing travel distance itself, i.e. a relationship between R and Y remains even if X has been taken into account. For example, all distances above or below a certain threshold could have been deleted due to plausibility reasons.

We can extend the example to a longitudinal setting with a monotone or arbitrary pattern of missingness by observing travel distance for q days, i.e. $Y = (Y_1, \dots, Y_q)$ and $R = (R_1, \dots, R_q)$. In this setting completely observed variables as well as partially observed variables carry information. If we require CDMAR, missingness may still depend only on socio-demography. However, for MAR missingness may depend in addition to socio-demography also on any recorded value of travel distance of a given entity.

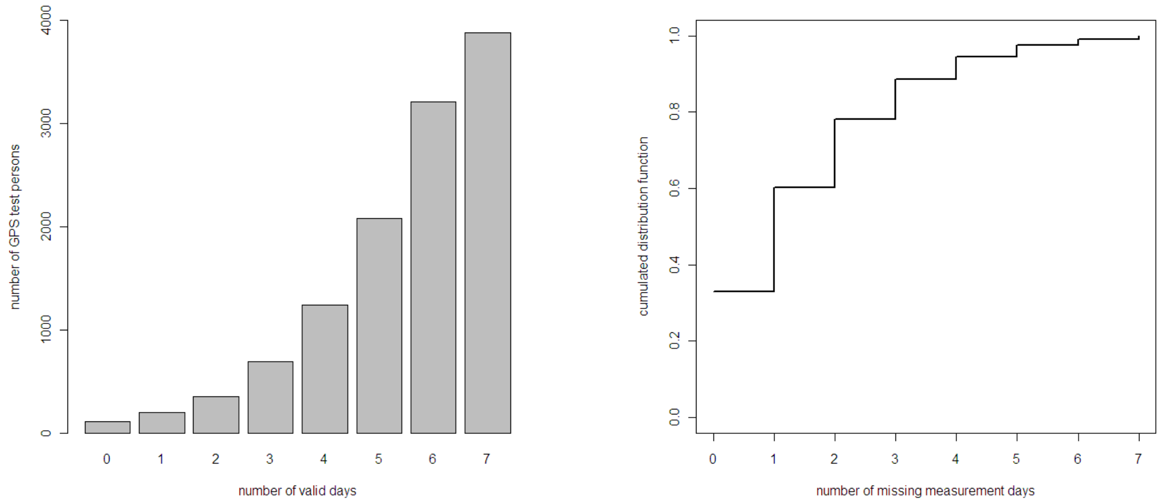
Depending on the goal of a study, missing data mechanisms have different implications (Little and Rubin, 2002). If the interest lies in the conditional distribution of the partly observed variables Y given the completely observed variables X, an analysis is only unbiased if the data are CDMAR. Returning to our above example, this means that we can estimate the travel distance for sociodemographic groups directly from the data sample. However, if the interest lies in the marginal distribution of Y (i.e. we are interested in the average travel distance of the *whole* population) CDMAR is not sufficient. In this situation only MCAR assures unbiased results. However, the observed data may be used to reduce the bias if data are not MCAR. If CDMAR or MAR dependencies are known, missing data algorithms can be applied that mitigate the induced bias by conditioning on the respective variables. It is therefore important to know which mechanisms of missingness exist in partially observed data.

5.1.3. Analyzing Patterns of Missingness in Mobility Data

In the following we will analyze the German application data set introduced in Section 3.2 in order to determine patterns and mechanisms of missing data. The German mobility survey relies on both GPS technology and computer assisted telephone interviews (CATI) to record movement data. However, CATI interviews have been conducted only for a single day per test person, and therefore lack a temporal structure. In all further analyses we will therefore consider only the GPS part of the mobility data set.

Missing measurement days in the GPS data arise due to various reasons. People may forget to carry or to charge the device. Devices may be defective or people may simply tire of the study and drop out early. The German mobility study follows a layout that allows to detect such complete missing measurement days by a follow-up survey. Short gaps are closed by using routing algorithms. However, missing trips within a day cannot be detected and are disregarded in this thesis. The study uses a *day* as chosen unit of measurement. Given a survey duration of seven days in Germany, we obtain the variable $Y = (Y_j \mid j = 1 \dots 7)$ to encode the mobility per day. Analogously, we can specify the variable for missingness as follows: $R = (R_j \mid j = 1 \dots 7)$. Two kinds of patterns of missingness occur in the data. First, missing measurements due to early drop out or due to defective devices lead to monotone patterns of missingness, i.e. if one day is missing, then all following days are missing as well. Second, single missing measurements arise due to forgetting, not loading or switching off the devices which lead to an arbitrary pattern of missing data.

In total, the data set contains 11,770 GPS test persons. Only one third of these persons provide seven valid measurement days, however, 99.1% provide at least one valid measurement day. Figure 5.3 shows the distribution of available measurement days per person and the cumulative distribution function of missing measurement days per person. An analysis of the data for monotone and arbitrary patterns of missingness showed that only 14.7% of the test persons with at least one and at most six missing measurement days followed a monotone pattern of missingness. For the majority of test persons the missingness pattern was arbitrary. On the whole, the data set therefore shows an arbitrary pattern of missingness.



(a)

(b)

Figure 5.3.: Missing measurements in German GPS mobility survey (a) distribution of valid measurement days per person (b) cumulative distribution function of missing measurement days per person

Note that a valid measurement day does not necessarily correspond to the presence of mobility data. If a test person has stayed at home for the whole day, we have full information about the mobility on this day although no movement data may be available. We refer to such a case as *immobility*.

5.1.4. Analysing Missing Data Mechanisms in Mobility Data

While the pattern of missingness is easily determined, it is hard to identify the mechanism of missing data. For MCAR several tests have been proposed (Little, 1988; Park and Davis, 1993). However, the distinction between MAR and MNAR is not straightforward. In general, a distinction is not possible unless additional knowledge about the data or the surveying process is known (Little and Rubin, 2002; Schafer and Graham, 2002). One major indication for MNAR is that the distribution of the observed values differs from the known shape of distribution. If, for example, a variable follows a normal distribution, however, possesses an asymmetric shape in the data sample, it is likely that the mechanism of missingness is not MAR (Little and Rubin, 2002). In addition, knowledge about the surveying process helps to identify the mechanism of missingness. For example, Murray and Findlay (1988) describe a longitudinal study on drugs against hypertension. Patients whose blood pressure exceeded a certain threshold were naturally withdrawn from the study and received a different treatment. In this case the mechanism of missingness is MAR because blood pressure was recorded before drop-out. A different situation arises if measurements are rejected because they exceed or fall below a certain threshold, e.g. due to plausibility reasons. In this case the missingness depends on the rejected value itself and results in a MNAR mechanism. If no information about the censoring mechanism or the distribution of the data is available, it is often assumed that the mechanism of missingness is MAR. This assumption is reasonable for many real-world scenarios. However, the robustness of applied algorithm should be assured as the degree of bias in results may depend on several factors, including the hight of missing data, the implementation of the missing data mechanism, the provided independent variables and the estimated statistical quantity (e.g. mean, variance, standard error), as shown by Collins et al. (2001).

Further, as already stated in Section 5.1.2, the analysis goal determines how detailed dependencies between the mobility characteristic and the missingness have to be tested. For statements on the whole data set it suffices to insure the independence between both variables by itself. However, if certain characteristics shall be evaluated, for example, for sociodemographic subgroups, the independence between mobility characteristic and missingness has to be assured for each subgroup as well. Therefore, the level of detail during evaluation influences the analysis of missing data mechanism as well.

In the German mobility study, the mobility information of interest is the daily number of visits of a test persons to poster sites. However, the number of visits depends on the selected poster campaign. Depending on how many locations are chosen and where they are situated, the number of visits varies. We therefore need a substitute that is proportional to the number of poster passages for an average poster campaign. We chose as substitute the daily distance that a person travels. Clearly, the more a person travels outside, the higher is the probability that he or she will see a poster. We determine the average daily travel distance for each person from their available number of measurement days, i.e.

$$D = (d_i) = \frac{\sum_{j=1}^{7} d_{ij}}{r_i} \quad \forall i = 1..n$$

with d_{ij} the traveled distance of person i on day j in kilometers and $r_i \in \{1, ..., 7\}$ the number of valid measurement days for person i. We know the number of valid measurement days, i.e. the response of each test person, due to a follow-up survey. As the average travel distance covers only available measurement days, it forms only an approximation of the true average travel distance per person. However, we are in the comfortable situation to provide a reference value for each test person independently of the number of valid measurement days.

For the dependency analysis we discretize the travel distance and measurement response in three groups each. For travel distance, we form the groups according to quantiles of the lowest, middle and highest one third of travel distances, i.e.

$$\textit{travel group}\ (d_i) = \begin{cases} \textit{low} & \text{if } d_i < Q_{0.33}(D), \\ \textit{medium} & \text{if } Q_{0.33}(D) \leq d_i < Q_{0.66}(D), \\ \textit{high} & \text{if } Q_{0.66}(D) \leq d_i. \end{cases}$$

For measurement response, we form the following groups:

$$\textit{response group}\ (r_i) = \begin{cases} \textit{low} & \text{if } r_i \in \{0,1,2\}, \\ \textit{medium} & \text{if } r_i \in \{3,4,5\}, \\ \textit{high} & \text{if } r_i \in \{6,7\}. \end{cases}$$

We perform chi-square tests in order to detect dependencies between variables. In the first place we are interested in the relationship between the independent variables and either travel group or response group. However, the relationships between the independent variables are also interesting in order to reduce complexity later on. In addition, we are interested in the relationship between travel group and response group under conditioning on the independent variables. This information is important for two reasons. First, conditioning is necessary in order to evaluate the data separately for sociodemographic groups. This, however, may induce a dependency between travel and response group and thus bias results. Second, if we detect a dependency between travel and response group in the first place, conditioning offers the possibility to reduce the dependency and may thus improve results.

In the first analysis we evaluate the dependency between any two sociodemographic variables, travel group and response group. The results are depicted in Table 5.1. Unfortunately our data set shows a dependency between travel group and response group, i.e. our data are at least CDMAR. If we assume a level of statistical significance of $\alpha = 0.05$, all variables with the exception of travel group are independent of the response group. Further, all variables with exception to response group and occupation are independent of the travel group. The dependency between our independent variables varies, however, it should be noted that householder, occupation and education show very strong dependencies to the other independent variables.

Table 5.1.: P-Values of chi-square tests between all sociodemographic variables, travel group and response group for test persons in Hamburg, Germany

	gender	**age group**	**educa-tion**	**occu-pation**	**house-holder**	**travel group**	**resp. group**
gender	0	0.709	0.079	0.004	≤ 0.001	0.159	0.212
age group	0.709	0	≤ 0.001	≤ 0.001	≤ 0.001	0.264	0.895
education	0.079	≤ 0.001	0	*≤ 0.001	≤ 0.001	0.593	* 0.407
occupation	0.004	≤ 0.001	*≤ 0.001	0	≤ 0.001	0.001	* 0.944
householder	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001	0	0.118	0.073
travel group	0.159	0.264	0.593	0.001	0.118	0	≤ 0.001
resp. group	0.212	0.895	* 0.407	* 0.944	0.073	≤ 0.001	0

* approximation may be incorrect due to small cell counts

In the second analysis we test the dependency between travel group and response group while conditioning on each of the independent variables. I.e., we perform a chi-square test of independence between travel and response group given all test persons with the same value of a variable. As mentioned above, this step is necessary because the application requires information about sociodemographic groups. Besides, the conditioning may reduce the dependency between travel and response group. The results are shown in Table 5.2. Again we obtain dependencies for each variable in at least one value. This means that we also need to test combinations of independent variables.

Table 5.2.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables gender, age group, education, occupation and householder for test persons in Hamburg, Germany

	male	**female**
gender	0.030	0.001

* approximation may be incorrect due to small cell counts

	14 - 29	**30 - 49**	**$\geq$ 50**
age group	*0.064	0.042	*0.007

* approximation may be incorrect due to small cell counts

	in school	**secondary general school**	**intermediate secondary school**	**high school / university**
education	$\leq$ 0.001	*0.188	*0.146	*0.903

* approximation may be incorrect due to small cell counts

	in training	**employed**	**retired**	**unemployed**
occupation	$\leq$ 0.001	*0.300	*0.073	*0.117

* approximation may be incorrect due to small cell counts

	yes	**no**
householder	$\leq$ 0.001	*0.016

* approximation may be incorrect due to small cell counts

In the third analysis we therefore perform chi-square tests between travel and response group for any combination of value-pairs of two variables. As the conditioning reduces the number of data points in the analysis strongly, not all analyses could be conducted. We performed chi-square tests only for groups with at least 30 test persons. However, even for these groups correct approximation cannot be completely guaranteed due to small cell counts. The complete results are given in the appendix C.1. In this place we restrict the tables to the most important results. Tables 5.3 and 5.4 show all combinations that provide independence between travel group and response group for a level of statistical significance of $\alpha = 0.03$. Both groups contain age group as a variable for conditioning, which is plausible as age is a strong differentiator for travel behavior. Under the assumption that our results may be safely interpreted, we can use each pair of the depicted variables for conditioning. The preferred choice is the pair (age group, householder). First, it has the highest minimum of statistical significance for all value groups. Second, it consists of a comparably small number of groups which increases the number of cases per group and thus the stability of results.

Table 5.3.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables age group and occupation for test persons in Hamburg, Germany

	occupation			
age group	**in training**	**employed**	**retired**	**unemployed**
14 - 29	*0.298	*0.264	NA	NA
30 - 49	NA	*0.033	NA	NA
$\geq$ 50	NA	*0.087	*0.078	NA

* approximation may be incorrect due to small cell counts

Table 5.4.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables age group and householder for test persons in Hamburg, Germany

	householder	
age group	**yes**	**no**
14 - 29	*0.058	*0.152
30 - 49	*0.096	*0.109
$\geq$ 50	*0.117	*0.043

* approximation may be incorrect due to small cell counts

5.2. Literature Review and Evaluation of Methods for Handling Missing Data

5.2.1. Overview of Missing Data Methods

In this section we give an overview of methods that handle missing data based on Little and Rubin (2002) as well as Schafer and Graham (2002). In general, the following five main categories of missing-data methods are distinguished: *case deletion*, *reweighting*, *single imputation*, *maximum likelihood* and *multiple imputation*. The first three categories contain older methods that may be applied very easily, but are in general inferior to multiple imputation and maximum likelihood approaches. In addition to theses approaches, we will introduce *survival analysis*. Survival analysis, also known as *event history analysis*, is a branch of statistics that investigates the occurrence of events as they take place over time. The field is inherently connected to missing data and fits with the event character of entity-location interactions.

Case deletion. Case deletion denotes methods that discard objects with incomplete data. Two types of case deletion exist: *complete-case analysis* and *available-case analysis*. Complete-case analysis uses only objects for which all variables are present. Available-case analysis determines objects with complete data in respect to the analysis task at hand. If, for example, the average of one variable shall be determined, all objects with values for this variable are considered, independent of possible missingness in other variables. - Case deletion requires missingness to be MCAR in order to guarantee unbiased results. In this case the remaining objects can be considered as random subsample of the original data set and, given representativity of the original sample, are representative for the full population as well. Case deletion is a very simple method which may be applied for small amounts of missing data. However, with increasing missingness, case deletion becomes inefficient as a lot of data is wasted.

Reweighting. Each object in a sample represents a certain number of objects in the population. The exact number depends upon the design weight appended to each unit. The idea of reweighting is that additional weights can be applied in order to adjust for nonresponse in the data sample. In this way nonresponse appears to be part of the sample design. - Reweighting can be used in MCAR as well as in specific MAR-situations. It can be easily applied to univariate and monotone patterns of missingness, however, is cumbersome for arbitrary patterns because a different set of weights must be computed for each variable.

Single imputation. During single imputation each missing value is filled in with a plausible value, resulting in a complete data set that can be analyzed using standard methods. Schafer and Graham (2002) distinguish four types of single imputation: *imputing unconditional means*, *imputing from unconditional distributions*, *imputing conditional means* and *imputing from a conditional distribution*. The first two methods require MCAR conditions while the latter two methods work under MAR, as their names let already assume. *Imputing unconditional means* or *mean substitution* simply replaces each missing value by the average of the observed variable. The disadvantage of this method is that it greatly underestimates the variance of the variable. This can be avoided by imputing from unconditional distributions, which aims to preserve the distribution of a variable. One method of this type of single imputation is *hot deck imputation*. Hot deck imputation substitutes missing values by randomly drawing from the observed values of the variable. Imputation from *conditional means* or *distributions* base on a regression from variables with known values to variables with unknown values. In the case of conditional mean imputation a missing value is directly replaced by the regression prediction. In the case of imputation from a conditional distribution, missing values are replaced by a value that is randomly drawn from the conditional distribution. In case of a linear model, such a value may be achieved by adding white noise to the regression value before substituting. This process counteracts the reduction of variance as well as the distortion of covariances. Imputation from conditional means or a conditional distribution can be applied without effort to univariate and monotone patterns of missingness. For monotone patters usually a step-wise regression is performed, substituting first missing values of one variable and then starting a new regression for the next variable. For arbitrary patterns the conditional distribution may take a complex shape, and by this also the sampling from it.

Multiple imputation. Similar to single imputation, multiple imputation replaces missing values with plausible values. However, the process is repeated multiple times, hence the name multiple imputation. In each simulation round missing values are replaced first and then the data set is analyzed by standard methods. In order to conduct *proper* multiple imputation the missing values have to be replaced by repeated draws from the posterior predictive distribution of the missing values. This means to perform independent draws of the missing values as well as of the model parameters on which the drawing of missing values is based. In this way the imputed data reflect sample variability as well as well as uncertainty about the model. If missing values are repeatedly drawn from a single estimate of model parameters, only the sample variability is reflected. This approach is therefore also called *improper* multiple imputation (Rubin, 1987). The combination of the several imputations for scalar parameters along with their standard error can be estimated as follows. Let Y denote the variable of interest, then $\widehat{Y}^{(j)}$ denotes the estimate of Y obtained in round j of the imputation with $j = 1..m$. The resulting estimate is the average of the m estimates:

$$\overline{Y} = \frac{1}{m} \sum_{j=1}^{m} \widehat{Y}^{(j)}. \tag{5.1}$$

In order to estimate the overall standard error of Y, the average within-imputation variance $\overline{W}$ as well as the between-imputation variance B have to be considered. Hereby the average within-imputation variance $\overline{W}$ takes the form

$$\overline{W} = \frac{1}{m} \sum_{j=1}^{m} W^{(j)}. \tag{5.2}$$

where $W^{(j)}$ is the squared standard error of estimate $\widehat{Y}$ in round j of the imputation. The between-imputation variance B is

$$B = \frac{1}{m-1} \sum_{j=1}^{m} \left[\widehat{Y}^{(j)} - \overline{Y}\right]^2. \tag{5.3}$$

The overall standard error T then takes the form

$$T = \sqrt{\overline{W} + (1 + \frac{1}{m}) \cdot B}. \tag{5.4}$$

Schafer and Graham (2002) state that for many practical applications a number of 20 imputations is sufficient to remove noise from the estimate and other statistical summaries as significance levels or probability values. Typically one set of imputations is generated and re-used for later analyses.

The most important part of multiple imputation is the creation of missing values. Multiple imputation can be motivated from a Bayesian perspective, therefore it is important to select the right model and prior distribution of parameters. It is often assumed that the data follow a multivariate normal distribution. Graham and Schafer (1999) showed that also highly non-normal variables can be imputed with very good performance using a normal distribution. The main task of the imputation model is to preserve important features as, for example, the mean, variance or correlation of variables with missing values. It does not serve to describe structural or causal relationships in the data itself. In addition, imputation methods for categorical data or mixed continuous and categorical data exist. They are described in Schafer (1997), and we will review one specific method of it in the next section. Multiple imputation typically assumes that the data are MAR, however, approaches for MNAR have also been developed. Depending on the imputation technique, uniform, monotone or arbitrary patterns of missing data can be treated.

Maximum likelihood. Maximum likelihood approaches rely on a model for the observed and unobserved data and determine those parameters which show the highest likelihood or posterior distribution under the model given the data. In difference to multiple imputation, maximum likelihood approaches treat missing values *within* the modeling process and not in a separate step prior to data analysis.

If the data are fully observed, maximum likelihood takes the following form. Let Y denote the data, which may be of scalar, vector or matrix form. Let further θ denote a vector of parameters that specify the distribution of Y under a given model. For example, assuming a multivariate distribution, θ would contain the mean and covariance information of the distribution. The goal of analysis is to find those parameters that are most likely given the data, i.e. that maximize

$$P(\theta|Y) = \frac{P(Y|\theta) \cdot P(\theta)}{P(Y)} \tag{5.5}$$

where $P(\cdot)$ denotes the probability density function. The denominator is a constant in the above equation and may therefore be neglected when searching for the most likely value of θ. The likelihood function is obtained by assuming further that the prior distribution of θ is uniform, i.e. $P(\theta_i) = P(\theta_j) \;\; \forall i, j$:

$$L(\theta|Y) = P(\theta|Y) \propto P(Y|\theta). \tag{5.6}$$

The maximum likelihood then denotes the value of θ for which equation 5.6 obtains its maximum value.

Given missing values in the data, the conditional distribution of the observed data Y_{obs} given θ is obtained by integrating over the missing data Y_{mis}:

$$P(Y_{obs}|\theta) = \int P(Y_{obs}, Y_{mis}|\theta) \; dY_{mis}. \tag{5.7}$$

However, Equation 5.7 only provides the correct likelihood

$$L(\theta|Y_{obs}) \propto P(Y_{obs}|\theta) \tag{5.8}$$

if the data are MAR. Little and Rubin (2002) therefore call the likelihood in Equation 5.6 also the likelihood "ignoring the missing-data mechanism".

With exception to a few problems, the maximum likelihood in Equation 5.8 cannot be formulated in closed form. Its computation therefore requires iterative steps for which typically the Expectation Maximization (EM) algorithm as proposed in (Dempster et al., 1977) is used.

Maximum likelihood methods typically perform a factorization of the log-likelihood for uniform and monotone patterns of missing data in order to obtain posterior distributions of simpler form. For arbitrary patterns of missing data such a factorization does often not exist or the factors of the factorization are not distinct. Then again, iterative computation techniques as EM can be applied.

Survival Analysis. Survival analysis (or event history analysis) investigates the occurrence of events as they take place over time. More precisely, survival analysis considers the individual time from an initiating event to an event of interest for a group of objects, also called *survival time* (Aalen et al., 2008; Kleinbaum and Klein, 2005). Such events denote, for example, the occurrence of some disease in a clinical study or the failure of a device in quality control. In practice the analysis of survival times is inherently connected to missing data because the time to event may be smaller than the surveying period. Assume, for example, that we monitor the time until relapse of some disease after accomplishment of a new treatment. During the monitoring period some patients will encounter a relapse, others will not. However, we do not know whether the latter persons possibly encounter a relapse after the monitoring period ends. The observation data is therefore incomplete. In survival analysis missing measurements are also termed *censored* data. Depending on whether the unknown time of event lies before or after the monitoring period, the data are called *left* or, respectively, *right* censored. If both situations apply the data are *interval* censored. In the following we will introduce important concepts of survival analysis along with the most popular estimation techniques.

Let $T \geq 0$ be a random variable of survival time, i.e. the time until an event occurs. We will denote the distribution function of T with

$$F(t) = P(T \leq t) \tag{5.9}$$

and the probability density function of T with

$$f(t) = \frac{d}{dt}F(t). \tag{5.10}$$

Two very basic concepts of event history analysis are the *survival function* and the *hazard function*. The survival function $S(t)$ states the probability that the specified event occurs later than some time t. More formally,

$$S(t) = P(T > t). \tag{5.11}$$

In general, it is assumed that $S(0) = 1$. The survival function decreases monotonically over time and often approaches zero as time increases. However, $S(t)$ may also converge to a number within the interval $(0, 1)$ for $T \to \infty$ as events do not need to happen during the lifetime of an individual. The following relationship holds clearly for the survival function:

$$S(t) = 1 - F(t). \tag{5.12}$$

The second basic concept in survival analysis is the *hazard function* or *hazard rate* $h(t)$. It specifies the instantaneous rate that an event occurs at a specific point in time. More formally, the hazard function is defined as the following limit:

$$h(t) = \lim_{\Delta t \to \infty} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{5.13}$$

In contrast to the survival function which states an unconditional probability, the hazard function represents a rate based on a conditional probability and can assume any non-negative value. The survival function and the hazard function are closely related and can be transformed into each other. First, the hazard function can be expressed using the survival function as follows:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to \infty} \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)} = -\frac{\frac{d}{dt}S(t)}{S(t)} \\ &= -\frac{d}{dt} ln(S(t)). \end{aligned} \tag{5.14}$$

Second, we can resolve Equation 5.14 by integration for the survival component:

$$S(t) = exp\left\{ -\int_0^t h(s)ds \right\}. \tag{5.15}$$

The integral of $h(t)$ is also called the *cumulative hazard function*.

$$H(t) = \int_0^t h(s)ds. \tag{5.16}$$

It can be used to simplify Equation 5.15 to

$$S(t) = exp\left\{ -H(t) \right\}. \tag{5.17}$$

The three most well-know techniques of survival analysis are the Nelson-Aalen estimator, the Kaplan-Meier estimator and the Cox regression model. Nelson-Aalen is a non-parametric method and provides an estimate of the cumulative hazard rate $H(t)$. Kaplan-Meier (Kaplan and Meier, 1958) is also a nonparametric method and estimates the survival function $S(t)$. It adapts to missing data by calculating conditional probabilities between two consecutive events. Finally, the Cox regression model (Cox, 1972) is a semiparametric model for the estimation of the hazard function $h(t)$. Similar to ordinary regression models it estimates the influence that a change in the independent variables causes on the dependent variable. One major assumption of Cox regression is hereby that the hazard rates of any two objects develop proportional over time.

In general, methods in survival analysis assume monotone patterns of missing data, although the censoring may occur at the end as well as at the beginning of the observation period. Due to conditioning (or the regression components), survival methods are able to handle MAR mechanisms.

5.2.2. Preconditions for Modeling Visit Potential

The modeling process and treatment of missing data is closely connected to the pattern and mechanism of missing data, the analysis goal and application requirements. In this thesis we will formulate preconditions for the evaluation of visit potential quantities and instantiate the data and application requirements using the outdoor advertising scenario. We believe that the following preconditions are reasonable in a number of other application scenarios, but, of course, they will not be the preferred choice everywhere.

Our analyses in Section 5.1.3 showed that the mobility data possess an arbitrary pattern of missing measurement days. This means that algorithms have to comply with intermittent missing values. However, in the case of the outdoor advertisement application this restriction is mitigated by the fact that visit potential quantities are calculated for average days. Therefore, we may permute measurement days not only to anticipate an uneven mixture of weekdays in the data sample but also to generate a monotone pattern where all missing values occur at the end of the surveying period.

With respect to the mechanism of missing data our analyses in Section 5.1.4 implied a MAR dependency. Therefore, the selected methods must be able to compensate MAR mechanisms. In the outdoor advertising scenario the compensation of MAR can take place by conditioning on sociodemographic variables, which means that the missing data methods must be able to handle numeric as well as categorical variables.

Our analysis goal is the estimation of visit potential quantities. As visit potential relies on count statistics, two possibilities for estimation exist. First, we can directly estimate visit potential quantities from the given (incomplete) data set. Second, we can impute the missing values and subsequently calculate visit potential on the completed data set. In the first case the methods have to ensure a non-negative domain for gross visits and average visits and values in the interval $[0, 1]$ for coverage. In the second case the imputed values have to conform to count statistics, i.e. they have to be non-negative integer values.

For the general acceptance of performance indicators in outdoor advertisement it is important that the applied methods are transparent and understandable. In practice performance values of poster sites are only accepted if they can be explained in a rational and comprehensible way. A second very important aspect in practice is the impartiality of the applied methods. As the business models of a whole industry depend on the performance indicators, it must be ensured that the methods cannot be influenced by one of the partners or by a third party. Therefore parameter free methods are preferred for the application. Finally, poster performance has to be evaluated for a large number of possibly very large poster sets and target groups. On thee one side, any combination of poster locations and target population may be selected, which leads to an exponential number of evaluation candidates. On the other side, nationwide campaigns may be formed that include thousands of posters and test persons. Therefore the applied methods have to be scalable.

In the remaining section we describe four algorithms in greater detail which we selected for experimental evaluation according to the above criteria. All of the methods are state-of-the-art methods in their respective fields or are especially applicable to the application domain. In addition, we selected the methods such that they cover a broad spectrum of modeling techniques. Three of the four methods are imputation techniques and one method comes from the area of survival analysis. Imputation techniques have the advantage that they complete the missing values of the data set, and further evaluation can be conducted using standard analysis techniques. This is especially helpful when different quantities have to be estimated from the data. However, they perform an intermediary step. In contrast, the selected survival technique derives the quantity directly from the data without materialization of missing values. The method reduces the estimation problem from several values to a single value. However, this is also a problem because different methods have to be applied in order to estimate further quantities, which is likely to cause inconsistencies.

The first method that we selected is the General Location Model (GLM) for mixed data. It provides a model for for the joint distribution of categorical and continuous variables and can be embedded in a multiple imputation framework. Second, we selected support vector methods as representative of state-of-the-art machine learning. Although a few approaches in the literature exist to handle missing values in support vector methods (Chechik et al., 2007; Pelckmans et al., 2005), standard implementations such as the packages e1071 and klaR for R do not offer the possibility to handle missing data (Dimitriadou et al., 2011; Roever et al., 2011). Therefore we designed a two-step imputation schema to apply the method. Third, we modeled the data based on a Poisson distribution and again embedded it in a multiple imputation schema. Poisson distributions have the advantage that they are designed for event data and therefore naturally cover the application domain. In addition, parameter estimation of Poisson distributions is straightforward and easy to understand. Finally, we selected Kaplan-Meier estimation from the area of survival analysis. This method directly infers entity coverage without imputing missing data. In addition, it is a parameter free method that has been designed for event data.

5.2.3. Multiple Imputation via General Location Model (MI-GLM)

The general location model (GLM) (Olkin and Tate, 1961) is a model for the joint distribution of categorical and continuous variables. It models all numeric variables under a multivariate normal distribution which is conditioned on the marginal distribution of the categorical variables. The model assumptions allow to handle arbitrary patterns of missing data and to compensate MAR mechanism based on the conditioning on categorical variables. The assumption of a multivariate normal distribution allows further to model the correlation between visits on consecutive days. However, the distribution also generates real-valued, possibly negative numbers during the imputation step. The method thus does not directly fit the application domain but requires additional data transformations. The method is parametric and the estimation process of the parameters under missing data sophisticated which are slight disadvantages for its practical application. However, the underlying data model is reasonable for our scenario and the method incorporates state-of-the-art parameter estimation and proper multiple imputation for the handling of missing values.

In the following we will introduce the GLM data model and present the inference and imputation process. The data transformation step for the treatment of real-valued results is deferred to Section 5.4, because it is not part of the original model. The following introduction is based on Schafer (1997), Chapter 9.

Data model. Let $X_1, X_2, \ldots, X_p$ denote a set of categorical variables and let $Y_1, Y_2, \ldots, Y_q$ denote a set of continuous variables. Further, let X and Y represent the $n \times p$ and $n \times q$ data matrices of the respective variables that are observed for n entities. The categorical data can be summarized in a p-dimensional contingency table with a total of m cells. Let $cell : \mathbb{N}^p \rightarrow \mathbb{N}$ denote a function that assigns each entity to a cell according to the (numerically encoded) values of its categorical variables. For all cells we count the number of entities and denote their number by $f = (f_1, f_2, \ldots, f_m)$ with $\sum_{h=1}^{m} f_h = n$.

The GLM is defined by the distribution of X and the conditional distribution of Y given X. The first part of the model is described by a multinomial distribution on the cell counts f, i.e.

$$f \sim M(n, \pi) \tag{5.18}$$

with $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$, $\sum_{h=1}^{m} \pi_h = 1$, the cell probabilities corresponding to f. The second part of the model defines the relationship between the variables $Y_1, Y_2, \ldots, Y_q$ by a multivariate normal distribution for each cell $h = 1..m$. Hereby, the means of $Y_1, Y_2, \ldots, Y_q$ can differ between the cells, however, the covariance is assumed to be equal for all cells. More formally, let $y_i = (y_{i1}, y_{i2}, \ldots, y_{iq})$ denote the continuous values of entity i with $i = 1..n$, let $\mu_h = (\mu_{h1}, \mu_{h2}, \ldots, \mu_{hq})$ denote the mean vector for cell h and let Σ denote the $q \times q$ covariance matrix, then we assume the following distribution for y_i:

$$y_i \sim N(\mu_h, \Sigma) \quad \text{with} \quad cell(x_i) = h. \tag{5.19}$$

If we combine both model parts, we can denote the parameters of GLM by

$$\theta = (\pi, \mu, \Sigma). \tag{5.20}$$

Hereby, $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$ contains the mean vectors of all cells. The GLM possesses $m - 1$ free parameters for the cell distribution of the first part of the model and $m \cdot q + q \cdot (q + 1)/2$ free parameters for the mean vectors and covariance of the second part. Note that the number of entities should to a considerable degree be larger than the number of free parameters in order for GLM to work well. If this is not the case, a restricted version of GLM can be used which reduces the number of free parameters by further modeling assumptions.

We can now formulate the likelihood of the complete data set as the product of the likelihood for the multinomial and the normal likelihoods as follows

$$L(\, \theta \mid (X, Y)\,) \;\propto\; L(\pi \mid X)\; L(\mu, \Sigma \mid X, Y). \tag{5.21}$$

The two factors of the likelihood are

$$L(\pi \mid X) \;\propto\; \prod_{h=1}^{m} (\pi_h)^{f_h} \quad \text{and} \tag{5.22}$$

$$L(\mu, \Sigma \mid X, Y) \;\propto\; |\Sigma|^{-\frac{n}{2}} exp \left\{ -\frac{1}{2} \sum_{h=1}^{m} \sum_{i \,|\, cell(x_i) = h} (y_i - \mu_h)^T \Sigma^{-1} (y_i - \mu_h) \right\}. \tag{5.23}$$

We can calculate the maximum likelihood (ML) estimates of the two factors in Equation 5.21 separately because the parameters of the factors are distinct. For the multinomial model in Equation 5.22 the ML estimate is

$$\hat{\pi}_h = \frac{f_h}{n} \qquad \forall h = 1..m. \tag{5.24}$$

The maximum likelihood estimators for the parameters in Equation 5.23 are

$$\hat{\mu}_h = \frac{1}{f_h} \sum_{i\,|\,cell(x_i)=h} y_i \qquad \forall h = 1..m \quad \text{and} \tag{5.25}$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{h=1}^{m} \sum_{i\,|\,cell(x_i)=h} (y_i - \hat{\mu}_h)(y_i - \hat{\mu}_h)^T. \tag{5.26}$$

The estimates μ_h are simply the average values per cell and Σ summarizes the deviations of the entities from their cell mean.

Treatment of missing data. We have to be able to model the joint distribution of any subset of incompletely observed variables given the remaining variables in order to form the predictive distribution for each entity with missing values. Let us denote the observed and missing parts of X and Y by X_{obs}, X_{mis}, Y_{obs} and Y_{mis}. Similarly, we will denote the observed and missing values of a single entity by $x_{i(obs)}$, $x_{i(mis)}$, $y_{i(obs)}$ and $y_{i(mis)}$. The joint predictive distribution for a given entity $i = 1..n$ then takes the following form

$$P(\, x_{i(mis)},\, y_{i(mis)} \mid x_{i(obs)},\, y_{i(obs)},\, \theta\,). \tag{5.27}$$

The joint predictive distribution is characterized by the conditional distribution of the cell of the entity given its observed data

$$P(\, cell(x_i) = h \mid x_{i(obs)},\, y_{i(obs)},\, \theta\,) \tag{5.28}$$

and the conditional normal distribution of the entity's missing numeric values given its cell and observed numeric data

$$P(\, y_{i(mis)} \mid cell(x_i) = h,\, y_{i(obs)},\, \theta\,). \tag{5.29}$$

The first distribution in Equation 5.28 can be derived from the joint density of $cell(x_i)$ and y_i under the general location model, which is

$$P(\, cell(x_i) = h,\, y_i \mid \theta\,) \;\propto\; \pi_h \,|\,\Sigma\,|^{-\frac{1}{2}}\, exp\left\{ -\frac{1}{2}\,(y_i - \mu_h)^T \Sigma^{-1} (y_i - \mu_h) \right\}. \tag{5.30}$$

The conditional distribution is then as follows

$$P(\, cell(x_i) = h \mid y_i,\, \theta) = \frac{P(\, cell(x_i) = h,\, y_i \mid \theta)}{P(\, y_i \mid \theta\,)} \tag{5.31}$$

$$= \frac{\pi_h \; exp\left\{ -\frac{1}{2}\,(y_i - \mu_h)^T \Sigma^{-1} (y_i - \mu_h)\,\right\}}{\sum_{h'=1}^{m} exp\left\{ -\frac{1}{2}\,(y_i - \mu_{h'})^T \Sigma^{-1} (y_i - \mu_{h'})\,\right\}} \tag{5.32}$$

$$\propto \; exp(\, \delta_{h,i}\,). \tag{5.33}$$

The term in Equation 5.33 is also known as the *linear discriminant function* of y_i with respect to μ_h and has the following form

$$\delta_{h,i} = \mu_h^T \Sigma^{-1} y_i \;-\; \frac{1}{2} \mu_h^T \Sigma^{-1} \mu_h \;+\; log\, \pi_h. \tag{5.34}$$

So far, Equations 5.31-5.33 and 5.34 apply if the data are completely observed. In order to account for missing data let us assume first that $X_1, \ldots, X_p$ are missing completely and a subset of $Y_1, \ldots, Y_q$ are missing. The conditional probability in Equations 5.31-5.33 is then obtained by integrating over all possible values of $y_{i(mis)}$. This results in

$$P(\, cell(x_i) = h \mid y_{i(obs)},\, \theta) \propto \; exp(\, \delta^*_{h,i}\,). \tag{5.35}$$

Hereby, $\delta^*_{h,i}$ is a linear discriminant which relies only on the observed data $y_{i(obs)}$. It has the form

$$\delta^*_{h,i} \;=\; {\mu^*_{h,i}}^T \; {\Sigma^*_i}^{-1} \; y_{i(obs)} \;-\; \frac{1}{2} \; {\mu^*_{h,i}}^T \; {\Sigma^*_i}^{-1} \; \mu^*_{h,i} \;+\; log(\pi_h) \tag{5.36}$$

where $\mu^*_{h,i}$ and Σ^*_i denote the subvector and square submatrix of μ_h and Σ which correspond to the observed values of y_i, respectively.

Let us now assume that subsets of $X_1, \ldots, X_p$ as well as of $Y_1, \ldots, Y_q$ are missing. We then have to account for the given information in X_{obs} when calculating the conditional cell information. Basically, the additional information restricts the cell of entity i to a specific subset of cells, which we will denote with H_i. Thus, we can determine the conditional probability of the cells by normalizing the probability over the specific subset of cells. All remaining cells obtain probability zero, i.e.

$$P(\, cell(x_i) = h \mid x_{i(obs)},\, y_{i(obs)},\, \theta\,) = \begin{cases} \frac{exp(\delta^*_{h,i})}{\sum_{h' \in H_i} exp(\delta^*_{h',i})} & \text{if } h \in H_i \\ 0 & \text{else} \end{cases} . \tag{5.37}$$

The second part of the predictive distribution in Equation 5.27, the conditional normal distribution of the entity's missing numeric values given its cell and observed numeric data (Equation 5.29), is a multivariate normal distribution that can be obtained from the complete data distribution

$$P(\, y_i \mid cell(x_i) = h,\, \theta\,) \;\sim\; N(\mu_h,\, \Sigma) \tag{5.38}$$

by partitioning y_i into $y_{i(obs)}$ and $y_{i(mis)}$. Let $\mu_{h(obs)}$ and $\mu_{h(mis)}$ denote the mean vectors of $y_{i(obs)}$ and $y_{i(mis)}$, respectively, with $\mu = (\, \mu_{h(obs)},\, \mu_{h(mis)}\,)$. In addition, the covariance matrix is partitioned as follows

$$\Sigma = \begin{bmatrix} \Sigma_{(obs)(obs)} & \Sigma_{(obs)(mis)} \\ \Sigma_{(mis)(obs)} & \Sigma_{(mis)(mis)} \end{bmatrix}. \tag{5.39}$$

The marginal distributions of $y_{i(obs)}$ and $y_{i(mis)}$ are again normal with

$$y_{i(obs)} \sim N(\mu_{h(obs)}, \Sigma_{(obs)(obs)}), \tag{5.40}$$

$$y_{i(mis)} \sim N(\mu_{h(mis)}, \Sigma_{(mis)(mis)}). \tag{5.41}$$

In addition, the conditional distribution of $y_{i(mis)}$ given $y_{i(obs)}$ is known and has the following form

$$P(\, y_{i(mis)} \mid y_{i(obs)},\, cell(x_i) = h,\, \theta\,) \;\sim\; N(\mu'_h,\, \Sigma') \tag{5.42}$$

with

$$\mu'_h = \mu_{h(mis)} + \Sigma_{(mis)(obs)}\Sigma^{-1}_{(obs)(obs)}(y_{i(obs)} - \mu_{h(obs)}), \tag{5.43}$$

$$\Sigma' = \Sigma_{(mis)(mis)} - \Sigma_{(mis)(obs)}\Sigma^{-1}_{(obs)(obs)}\Sigma_{(obs)(mis)}. \tag{5.44}$$

The new parameters μ'_h and Σ' can be obtained from μ_h and Σ by applying the *sweep* operator (Beaton, 1964) on the positions of the observed variables $y_{i(obs)}$.

In general the maximum likelihood estimates of GLM cannot be obtained in closed form when data are missing. Therefore, the EM algorithm as described in Little and Schluchter (1985) is applied to the model. During the M-step the maximum likelihood estimates $\hat{\pi}$, $\hat{\mu}$ and $\hat{\Sigma}$ are calculated. The estimates can be obtained as given in Equations 5.24 - 5.26, however relying on the expected values under the current model where data are missing. In the E-step we have to substitute all missing values by their expected value under the observed data and current model parameters. This step relies on the predictive distributions as stated in Equations 5.37 and 5.42. In practice instead of computing the missing values themselves, only the sufficient statistics of the data are computed under the observed values and current model parameters.

In order to embed GLM in a proper multiple imputation schema, we have to perform multiple draws of the posterior predictive distribution. Hereby, each draw corresponds to the independent drawing of the parameters as well as of the missing values. In order to achieve the independent draw of parameters, we can adapt the EM algorithm as described above to a data augmentation algorithm. Data augmentation is an iterative process which consists of two steps. During the I-step missing data are randomly imputed given the current model parameters. In the P-step new parameters are drawn from the posterior distribution of the parameters given the observed and imputed data. Instead of generating expected values for the missing data, data augmentation performs a random draw from the predictive distributions. In addition, the parameters in each iteration are not maximum likelihood estimates but are also randomly drawn from their posterior distribution. After a sufficiently large number of iterations, the obtained parameters may be considered random draws from their posterior distribution.

In practice this means that we use the maximum likelihood estimates of θ which we obtain from EM as input to the data augmentation. After several iterations of the I- and P-step, we obtain new parameters, which we use to generate a single imputation of the missing values. The data augmentation process is repeated until all imputations are generated. Subsequently, we can calculate visit potential quantities for each imputation and average the results. Algorithm 1 summarizes this process.

Algorithm 1: Multiple imputation via GLM (MI-GLM)

```
Input:
$X = (X_1, \dots, X_p)$ // data set of completely observed variables
$Y = (Y_1, \dots, Y_q)$ // data set of partially observed variables
$m$    // number of multiple imputations
$s$    // number of data augmentation steps
Output:
visit potential quantities for data set $X, Y$

calculate maximum likelihood estimate $\hat{\theta}$ // perform EM
for $i = 1$ to $m$ do
    $\theta_{tmp} = \hat{\theta}$
    for $j = 1$ to $s$ do // perform data augmentation
        impute missing values based on $\theta_{tmp}$
        draw new parameters $\theta_{tmp}$ from their posterior distribution given the observed
        and imputed values
    end
    impute missing values based on $\theta_{tmp}$
    calculate visit potential quantities on observed and imputed values
end
average visit potential quantities over $m$ imputations
```

5.2.4. Single Imputation via Support Vector Regression (SI-SVR)

Support vector methods have continuously developed since their first introduction by Boser et al. (1992) and are very competitive machine learning techniques. When deciding on a representative state-of-the-art machine learning technique for the evaluation of visit potential under missing data, we therefore selected support vector regression (SVR) as introduced by Drucker et al. (1997). As the standard SVR implementations such as the R packages e1071 (relying on LIBSVM) and klaR (relying on SVMlight) do not offer the possibility to handle missing data (Dimitriadou et al., 2011; Roever et al., 2011), we designed a two-step single imputation schema to apply the method. The schema has been designed as a general framework for machine learning methods and allows for an easy exchange of the base learner. It is not especially adapted to SVR, and therefore may not lead to optimal results that could be obtained by a specialized SVR. Our imputation schema is able to handle arbitrary patterns of missing data and, by the inclusion of further independent variables, SVR has the possibility to compensate MAR dependencies. Similar to GLM, SVR predicts real-valued, possibly negative numbers. Therefore, an additional data transformation step must be applied. Support vector methods are also known to require careful parameter tuning, which is also a disadvantage for the practical use of the method. However, SVR is a state-of-the-art machine learning method, which has been successfully applied in a number of different application domains.

We begin this section with an introduction of support vector regression which is based on the tutorial of Smola and Schölkopf (1998). Subsequently, we present our imputation schema. Similar to GLM, we will postpone the treatment of real-valued results until Section 5.4.

Learning algorithm. Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathbb{R}$ denote a training data set where $\mathcal{X}$ denotes the space of the independent variables, for example, $\mathcal{X} = \mathbb{R}^d$ with $d \in \mathbb{Z}$ and $d \geq 1$. The aim of ε-support vector regression is to find a function $f(x) : \mathcal{X} \to \mathbb{R}$ which a) shows a deviation of at most ε from the actual values y_i for $i = 1..n$ and b) is as flat as possible. *Flat* hereby refers to a restriction of the complexity of the function in order to avoid overfitting. In the case of a linear function, f has the following form

$$f(x) = \langle w, x \rangle + b \qquad \text{with } w \in \mathcal{X}, b \in \mathbb{R}. \tag{5.45}$$

Hereby, $\langle \cdot, \cdot \rangle$ denotes the inner product. We can formulate a convex optimization problem for SVR as follows

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}||w||^2 \\ &\text{subject to} && \begin{cases} y_i - \langle w, x_i \rangle - b & \leq & \varepsilon \\ \langle w, x_i \rangle + b - y_i & \leq & \varepsilon \end{cases} . \end{aligned} \tag{5.46}$$

The term $||w||^2$ denotes the squared norm of w. It is a regularization term that reduces the complexity of the function by minimization. However, Equation 5.46 has the disadvantage that it assumes that a function f exists which predicts the data with an error of at most ε, i.e. $|y_i - f(x_i)| \leq \varepsilon$ for all $i = 1..n$. Often, however, this is not the case. We can then relax the optimization problem by introducing slack variables ξ, ξ^* similar to a soft-margin approach. This leads to the following optimization problem

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}||w||^2 + C \textstyle\sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &\text{subject to} && \begin{cases} y_i - \langle w, x_i \rangle - b & \leq & \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq & 0 \end{cases} . \end{aligned} \tag{5.47}$$

The constant C determines the cost of prediction errors that are larger than ε. Equations 5.47 are typically solved using the dual formulation. This has the advantage that non-linear functions can be easily integrated into the framework later on. The transformation of the primal (Equations 5.47) to the dual can be accomplished using Lagrange multipliers, which yields

$$\begin{aligned} L \; := \; & \tfrac{1}{2}||w||^2 + C \textstyle\sum_{i=1}^{n} (\xi_i + \xi_i^*) - \sum_{i=1}^{n} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \textstyle\sum_{i=1}^{n} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\ & - \textstyle\sum_{i=1}^{n} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b). \end{aligned} \tag{5.48}$$

L denotes the Lagrangian and $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ are the dual variables or Lagrange multipliers, which have to be non-negative, i.e. $\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0$. In order to solve the dual, the partial derivatives of L with respect to the primal variables w, b, ξ_i, ξ_i^* must be derived and set to zero. The derivatives are

$$\begin{array}{lclcl} \frac{\partial}{\partial b} L & = & \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) & = & 0 \\ \frac{\partial}{\partial w} L & = & w - \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) x_i & = & 0 \\ \frac{\partial}{\partial \xi} L & = & C - \alpha_i - \eta_i & = & 0 \\ \frac{\partial}{\partial \xi^*} L & = & C - \alpha_i^* - \eta_i^* & = & 0. \end{array} \tag{5.49}$$

In order to obtain the dual optimization problem, the derivatives must be substituted into Equation 5.48. Further details on the solution of the dual can be found in Smola and Schölkopf (1998). Note, however, that the derivative in the second line of Equation 5.49 can be rewritten as

$$w = \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) x_i. \tag{5.50}$$

It implies that w is a linear combination of the input data. In combination with Equation 5.45 we obtain

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b. \tag{5.51}$$

From Equation 5.51 we can predict the value of the dependent variable of some data instance x without explicitly computing w. The calculation relies only on the inner product of x and the training instances. More specifically, the calculation relies only on those training instances for which $|f(x_i) - y_i| \geq \varepsilon$, the so-called support vectors.

Equation 5.51 allows in addition to extend SVR to non-linear functions by applying some mapping function $\Phi : \mathcal{X} \to \mathcal{F}$ which transforms instances of input space $\mathcal{X}$ to some higher dimensional input space $\mathcal{F}$. The inner product now takes the form $\langle \Phi(x_i), \Phi(x) \rangle$, which can be considered as similarity function of the instances in the transformed space $\mathcal{F}$. However, the kernel trick allows to avoid the explicit computation of the transformation and the inner product in $\mathcal{F}$. If a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies Mercer's theorem, i.e. as long as it is positive definite, a transformation exists such that the value of the kernel function computed for two instances is equal to the inner product of the instances in the transformed feature space (Mercer, 1909). Typical examples of such kernel functions are

$$\begin{array}{ll} \text{polynomial:} & k(x_i, x_j) = (\ \gamma \langle x_i, x_j \rangle\ )^d,\ \gamma > 0, \\ \text{radial basis:} & k(x_i, x_j) = exp(\ -\gamma \, ||x_i - x_j||^2\ ),\ \gamma > 0, \\ \text{sigmoid:} & k(x_i, x_j) = tanh(\ \gamma \langle x_i, x_j \rangle\ ). \end{array} \tag{5.52}$$

In summary, when applying SVR we have to select an appropriate kernel function along with its parameterization and a cost C for prediction errors. Previous to our experiments we therefore tested different parameterizations of SVR which are described in more detail in Section 5.4.2.

Treatment of missing data. In our experiments we apply SVR to predict missing values of the variable $Y = (Y_1, \ldots, Y_q)$. We hereby iteratively predict missing values of one variable Y_j using the remaining $(q-1)$ variables $(Y_k \mid k = 1..q,\ k \neq j)$ and possibly additional sociodemographic variables X as independent variables. The training data set then comprises all entities for which Y_j is observed. However, the variables Y_k may also contain missing values, which cannot be treated with standard SVR. We therefore perform a secondary imputation step prior to application of SVR during which we temporarily impute missing values for variables Y_k with $k = 1..q,\ k \neq j$. The secondary imputation is simply a mean substitution where we replace missing values with average values. We performed different variants of mean substitution, which we call *vertical* and *horizontal* mean substitution. The terms vertical and horizontal indicate the direction in the data matrix over which the average is formed, i.e. over a column or a row. During vertical mean substitution (VMS) we replace missing values of Y_k with an average of the same variable, possibly subject to conditioning on sociodemographic characteristics of the entity of interest. During horizontal mean substitution (HMS) we form averages for Y_k over all observed values of $Y_1, \ldots, Y_q$ for a given entity. We tested both types of mean substitution during parameter tuning of SI-SVR. The results are given in Section 5.4.2.

We perform both imputation steps of SI-SVR independently for all variable $Y_1, \ldots, Y_q$, leading to the approach depicted in Algorithm 2.

Algorithm 2: Single imputation via SVR (SI-SVR)

```
Input:
$X = (X_1, \ldots, X_p)$ // data set of completely observed variables
$Y = (Y_1, \ldots, Y_q)$ // data set of partially observed variables
$\theta$ // SVR parameterization
$\varphi$ // mean substitution parameterization
Output:
visit potential quantities for data set $(X, Y)$
for $j = 1$ to $q$ do
    // determine set of independent partially missing variables
    $Y_{1..q \setminus j} = (Y_k \mid k = 1..q,\ k \neq j)$

    // temporarily impute missing values
    $Y_{1..q \setminus j\ (mis)} = applyMeanSubstitution(X, Y, \varphi)$

    // determine training and prediction data set
    $(X, Y_{1..q \setminus j}, Y_j)_{train} = \{\ (x_{i1}, \ldots, x_{ip}, y_{i1}, \ldots, y_{iq}) \mid y_{ij}$ observed, $i = 1..n\ \}$
    $(X, Y_{1..q \setminus j})_{predict} = \{\ (x_{i1}, \ldots, x_{ip}, y_{ik}) \mid y_{ij}$ missing, $i = 1..n\ \}$

    // train SVR and impute missing values
    $h = trainSVR(\ \theta,\ (X, Y_{1..q \setminus j}, Y_j,\ )_{train}\ )$
    $Y_j^* = (\ Y_{j(obs)},\ applySVR(\ h,\ (X, Y_{1..q \setminus j})_{predict}\ )\ )$
end
calculate visit potential quantities on observed and imputed values $Y^* = (Y_1^*, \ldots, Y_q^*)$
```

5.2.5. Multiple Imputation from a Conditional Poisson Distribution (MI-Poisson)

In this scenario we assume that the movements of a person are correlated over days (see also Section 2.2.3) and that therefore also the number of visits are correlated over days. We represent the number of daily visits of an entity as Poisson distribution with parameter λ and assume that λ does not change over time. Of course, this assumption is not completely true in practice because it is known that movement varies between different days of the week (e.g. workdays vs. weekend). A constant parameterization of the Poisson distribution cannot reflect this difference. However, as we evaluate visit potential for one week in our application, the distribution of visits *within* a week may be safely ignored.

For the estimation of λ we rely on all observed measurement days of an entity, which may be in random order. Therefore, the method can handle arbitrary patterns of missing data. As we further estimate λ individually for each entity, we perform a conditioning on the finest level of granularity. Consequently, MI-Poisson is able to handle MAR mechanisms. Further, Poisson is a discrete distribution for non-negative values and thus perfectly fits the domain of our application. Poisson is a parametric distribution as its shape is determined by the parameter λ, however the possibility of external influences is small, as we directly estimate λ from the observed data. Finally, the model behind Poisson is well-known and easily understandable, which is a positive aspect for its application in practice.

Data model. The Poisson distribution is a discrete probability distribution which models the probability that a number of events occur in a given period of time. The model hereby assumes that the probability of an event is very small within a given small time span Δt. It is therefore also called the distribution of *rare* events. More formally, let X be a random variable which denotes the number of events within time span t. We then divide t into a number of very small time intervals $\Delta t_i, i = 1..n$ for which we assume that

1. an event occurs at most once within Δt_i,
2. the probability of an event is proportional to the length of Δt_i, i.e. $P(\text{event in } \Delta t_i) = \lambda \cdot \Delta t_i$,
3. the occurrence of an event in any two time intervals Δt_i and Δt_j with $i \neq j$ is independent.

As the parameter λ determines the frequency at which events occur, it is also known as *intensity rate*. When we normalize the length of time interval t to one, we obtain the Poisson distribution function which has the following form

$$f(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} \cdot e^{-\lambda} & x \in \mathbb{N}_0 \\ 0 & \text{else.} \end{cases} \tag{5.53}$$

The expected value and variance of X under Poisson are

$$E(X) = Var(X) = \lambda. \tag{5.54}$$

Due to the assumptions about the counting process that underlies the Poisson distribution, random variables that are Poisson distributed have the following two properties (Fahrmeir et al., 2010). Let X and Y be two independent random variables that are drawn from two Poisson distributions, i.e. $X \sim Po(\lambda)$ and $Y \sim Po(\mu)$. The sum of both variables is again Poisson distributed, i.e.

$$X + Y \sim Po(\lambda + \mu). \tag{5.55}$$

Let us further assume that $X \sim Po(\lambda)$ denotes the number of events in the unit time interval. If we observe with variable Z the number of events in a time interval of length t, Z is again Poisson distributed with parameter λt, i.e. $Z \sim Po(\lambda t)$.

Treatment of missing data. For the estimation of missing values we have designed a multiple imputation schema which estimates the parameter λ individually for each entity based on its observed measurements. Subsequently, the visits of missing measurement days are imputed by randomly drawing from a Poisson distribution with parameter λ. The process is repeated several times in order to decrease the influence of extreme samples. Note that for all imputation rounds we derive λ from the observed data only. This means that λ is constant over all imputation rounds and we perform an improper multiple imputation. Note also that λ is the only parameter that influences the results of imputation. It is determined without reference to additional (e.g. sociodemographic) variables. However, as we determine λ separately for each entity, the conditioning takes effect at the finest level of granularity. Algorithm 3 shows the proceeding for MI-Poisson.

Algorithm 3: Multiple imputation from a conditional Poisson distribution (MI-Poisson)

```
Input:
$Y = (y_{ij}),\ i = 1..n,\ j = 1..q$ // data set of partially observed variables
$m$   // number of multiple imputations
Output:
visit potential quantities for data set $Y$
$\Lambda = (\lambda_1, \ldots, \lambda_n)$ with $\lambda_i = avg(y_{i\cdot(obs)})$ // calculate $\lambda$ for each entity
for $i = 1$ to $m$ do
    impute missing values based on $\Lambda$
    calculate visit potential quantities on observed and imputed values
end
average visit potential quantities over $m$ imputations
```

5.2.6. Kaplan-Meier Estimation (KM)

Kaplan-Meier (Kaplan and Meier, 1958) is a well-known method from the field of survival analysis. It is a direct technique which estimates the survival function from the data without replacement of missing values. The method hereby adapts to changes in the sample size by calculating conditional probabilities between the occurrence of consecutive events. Kaplan-Meier is designed for right censored data and thus requires a monotonic pattern of missingness. This is a clear constraint on the application of Kaplan-Meier, however, as stated in Section 5.2.2 we are able to generate such a pattern from the data because the outdoor advertising application allows permutation of measurement days. Kaplan-Meier assumes that the event of interest and the missing data are independent of each other, which applies to MCAR and MAR. In case of MAR the method handles the missingness mechanisms by conditioning on categoric variables. Kaplan-Meier has been designed for the analysis of event data and the resulting survival function can easily be turned into the visit potential quantity coverage as we will show below. It is a non-parametric method and easy to understand, which are both advantages from the application side. However, the strength of Kaplan-Meier as direct method has at the same time the disadvantage that it can only be used to estimates a single visit potential quantity. In consequence, if it is required to derive all visit potential quantities, we have to apply several methods and thus have to make an additional effort to ensure consistency among the results.

Kaplan-Meier does not make an assumption about the distribution of the data. We therefore directly begin with the treatment of missing measurements. The following introduction of Kaplan-Meier is based on Aalen et al. (2008) and Kleinbaum and Klein (2005) where also further details on the method can be found.

Treatment of missing data. Let T denote a random variable that states the survival time of an object, i.e. the time until the occurrence of the event of interest. Recall that the function

$$S(t) = P(T > t) \tag{5.56}$$

is called the survival function and denotes the probability that the specified event occurs later than some time t. For a given data set, Kaplan-Meier analyzes at which times t_i events occur (with $t_0 = 0$) and determines the following variables

- r_i - number of objects at risk shortly before time t_i,
- v_i - number of events at time t_i,
- c_i - number of dropouts in time interval $(t_{i-1}, t_i]$.

In our application setting the objects are mobile entities and an event denotes a visit to the location set. More precisely, depending on the visit class vc, an event denotes the vc-th visit of an entity to the location set. The term dropout refers to entities that leave the survey permanently, i.e. a dropout of an entity in our application occurs at the end of its last day with measurements *after* generating the right-censored data set. Objects at risk are entities which are exposed to the critical event at a given point in time, i.e. their measurement day is still observed and the event has not occurred to them yet. The number of objects at risk for time point t_j is measured slightly before t_j, i.e. the set includes also those objects for which an event at t_j occurs. The number of objects at risk at time t_{j+1} results from the previous objects at risk reduced by objects with an event at t_j as well as by the objects that drop out in the preceding time interval, i.e. $r_{i+1} = r_i - v_i - c_i$. Note that for time moment $t_0 = 0$ it is generally assumed that neither an event nor a dropout occurs, i.e. $v_0 = c_0 = 0$, and r_0, r_1 represent both the whole data sample.

Kaplan-Meier adapts to differing sample sizes by calculating conditional probabilities between two consecutive events. Objects that drop out of the study between two events are assumed to survive until the next event occurs and are then removed. The conditional probability p_i to survive time point t_i given that t_{i-1} has been survived is calculated as

$$p_i = P(T > t_i \mid T > t_{i-1}) = \frac{r_i - v_i}{r_i}. \tag{5.57}$$

Given the conditional probabilities p_i, the total probability to survive some time point t_k is

$$S(t_k) = P(T > t_k) = \prod_{i=1}^{k} P(T > t_i \mid T > t_{i-1}) = \prod_{i=1}^{k} p_i \tag{5.58}$$

The transformation from survival probability to coverage is straightforward. So far, $S(t)$ states the probability that entities do not visit any location (vc-times) within the location set until t. Consequently, entity coverage is given by the probability of the complimentary event

$$F(t) = P(t \leq T) = 1 - S(t). \tag{5.59}$$

Note that Kaplan-Meier generates survival probabilities only for time moments where critical events occur. Between two consecutive time moments the survival function takes on a constant value, i.e. $S(t) = S(t_j)$ for $t \in [t_j, t_{j+1})$. This means also that the survival function remains constant after the end of the surveying period. Possibly the survival function levels off even earlier if the set of objects at risk is reduced to zero by dropout before the end of the surveying period. In this case the model lacks information to continue the survival function and we will overestimate the value of $S(t)$ if we assume it to remain constant over time. Therefore we complemented Kaplan-Meier with a log-logistic regression model which we applied whenever the last observed critical event had lain before the time moment of interest. Further details on the model and a comparison of results with and without the regression model can be found in section 5.4.4.

Algorithm 4 summarizes the estimation of entity coverage with Kaplan-Meier, possibly subject to our extension with log-logistic regression.

Algorithm 4: Kaplan-Meier estimation (KM)

```
Input:
$X = (X_1, \ldots, X_p)$ // data set of completely observed variables
$Y = (Y_1, \ldots, Y_q)$ // data set of partially observed variables
$t$ // time moment of interest for visit potential quantity
Output:
visit potential quantity entity coverage for data set $(X, Y)$
calculate survival function $S(t)$ from data set $(X, Y)$ with Kaplan-Meier
if last critical event observed before $t$ then // extend Kaplan-Meier
    apply log-logistic regression model to predict survival function at $t$
end
$C_E = 1 - S(t)$ // transform survival prob. into entity coverage
```

5.3. Experimental Set-up

5.3.1. Test Scenario

We conducted our experiments on the GPS data set of the German audience measurement study as introduced in Section 3.2.2. However, as the usage of the complete data set would have resulted in very long computation times we restricted the analysis to a subset of the data. For our experiments we selected the city of Hamburg, i.e. the universal set of entities consists of all GPS participants in Hamburg and the universal set of locations consists of all poster sites in Hamburg. Note that the selection of a single city instead of distributing entities and locations randomly over Germany is necessary in order to concentrate visits and to obtain a reasonable level of visit potential. As most individual movements take place locally, a distributed location set would decrease the probability of a visit strongly. We chose the city of Hamburg because it offers a comparably large set of test persons and possesses a complex city structure. Although we selected only a single city for the experiments, the results can be expected to generalize because the subset comprises typical movement behavior.

Note that even though we use data from the outdoor advertising application in our experiments, we focus only on one part of the modeling process. The obtained results are therefore not directly comparable to the actual values used in the application.

One problem of evaluating missing data methods is that given data with missing values the true value of any derived quantity is unknown and thus an evaluation of missing data methods is impossible. We therefore use only test persons which are completely observed and introduce artificial missingness into the data. This allows us on the one hand, to control the mechanism of missing data and, on the other hand, to vary the amount of missingness. The mobility data provides up to seven measurement days, however, a restriction to test persons with seven complete measurement days reduces the data set considerably. Therefore, we decrease the number of observed measurement days to five and evaluate visit potential for $t = 5$. For our experiments we selected all test persons with at least five observed measurement days and contracted the trajectory data set to the first five observed days of each of these persons. The reduced entity set contains 393 of the original 548 test persons in Hamburg.

In Section 4.4.1 we showed how visit potential can be used to define precisely poster performance indicators. The most important visit potential quantities in this context are gross visits, average visits per entity for visit class $vc = 1$ and entity coverage for visit class $vc = 1$. We will therefore conduct our experiments with respect to these three quantities. Note that we will shorten the names of the tested visit potential quantities to gross visits, average visits and entity coverage because only the entity perspective is applied in the scenario and therefore the quantities cannot be mixed up.

For a given entity set visit potential varies according to the size of the location set. Clearly, the more posters a campaign contains, the higher is the chance that a test person passes a poster of the campaign. We will therefore vary the size of the location set in order to test the performance of missing data methods at different levels of visit potential. In particular we will conduct our experiments for location sets of size 25, 50, 100, 250 and 500.

In order to obtain stable results, we test each missing data method on 30 different location sets of the same size. The location sets are sampled at the beginning of the experiments and are the same for each method.

A detailed parameterization of all experiments is given in Appendix B.

5.3.2. Error Measurement

We measure the performance of each missing data method using mean error, relative mean error and root mean squared error. The mean error expresses the bias of a method in absolute numbers while the relative mean error relates the bias to the value of the measured variable. The root mean squared error contains the bias as well as the variance of an estimation method, however, expressed in units of the analyzed variable. We will refer to these errors as *basic* errors. More precisely, the errors are defined as following. Note that we refrain from including the data set and variable in the parameterization of the basic errors and only specify the name of the missing data method in order to reduce the notation to essentials.

Definition 5.3.1 (Mean error) *Let y_i with $i = 1..n$ denote the true values of some variable Y and $\hat{y}_i$ denote the estimated or predicted values of the variable by some method m. The mean error is defined as*

$$me(m) = \frac{\sum_{i=1}^{n} \hat{y}_i - y_i}{n}.$$

Definition 5.3.2 (Relative mean error) *Let y_i with $i = 1..n$ denote the true values of some variable Y and $\hat{y}_i$ denote the estimated or predicted values of the variable by some method m. The relative mean error is defined as*

$$rme(m) = \frac{\sum_{i=1}^{n} \frac{\hat{y}_i - y_i}{y_i}}{n}.$$

Definition 5.3.3 (Root mean squared error) *Let y_i with $i = 1..n$ denote the true values of some variable Y and $\hat{y}_i$ denote the estimated or predicted values of the variable by some method m. The root mean squared error is defined as*

$$rmse(m) = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}}.$$

As we conduct our experiments for different rates of missing data as well as for different sizes of the location set, we form three further errors that aggregate the results of the mean error, relative mean error and root mean squared error over all parameterizations. We will call these aggregated errors *compound* errors.

Definition 5.3.4 (Average absolute compound error) *Let $err(m)_{ij}$ denote an arbitrary basic error of some method m measured for location size s_i and missing data rate r_j with $i = 1..n$, $j = 1..m$. We define the average absolute compound error then as*

$$aace(err, m) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \mid err(m)_{ij} \mid}{n \cdot m}.$$

The main purpose of the compound error is to provide a single error value for a series of tests and thus to ease the complexity of method comparison. We choose the absolute values of the basic errors so that possible positive and negative biases for different parameterizations do not cancel each other out. The average value instead of a sum of errors was selected in order to retain the relation to the true value of the evaluated variable.

In summary, our *basic errors* are formed for a given rate of missing data and a given location set size over 30 experiments with different poster campaigns. Our *compound errors* aggregate the basic errors for 10 (during parameter tuning only 6) different rates of missing data and 5 location set sizes, i.e. the compound error summarizes 1,500 (respectively 900) experiments.

5.3.3. Generation of Artificial Missing Data

In order to evaluate the robustness of the selected methods for missing data, we implemented the mechanisms MCAR, CDMAR and MAR (see Section 5.1.2). Further, we varied the rate of missing data. Note that we use the term rate to refer to the proportion of partially observed entities, i.e. the proportion of the entity set with at least one missing measurement day. The term does not refer to the proportion of missing measurement days in total. The reason for our definition is that a completely random introduction of missing measurement days according to a given rate, i.e. each day has a given probability to be missing, may lead to the deletion of all measurement days of a test person. This, however, reduces the size of the entity set, which falsifies the rate of missingness for the remaining entities and increases the standard error of results. Therefore we follow a strategy which first selects a group of persons according to a given missingness rate. Second, one or more measurement days of each person are deleted. However, at least one measurement day is retained. Table 5.5 shows the corresponding expected percentage of missing measurement days for the applied rates partially observed entities.

For MCAR the group of persons in the first step is chosen randomly. The number of deleted measurement days within the second step is also determined at random.

For CDMAR we select different proportions of persons with missing data within different sociodemographic groups. Hereby it is important that the sociodemographic variable influences the mobile behavior. Else, the connection between mobility behavior and missingness would still be at random.

Finally, we introduce a version of MAR where the selection of persons depends solely on their mobile behavior. In this case only an inclusion of mobility information can help to reduce the bias.

5.4. Parameter Tuning

This section contains the parameter optimization for the missing data methods described in Section 5.2.3 - 5.2.6. Most of the analyzed parameters concern the data transformation for MI-

Table 5.5.: Rates of partially observed entities and corresponding expected percentage of missing measurement days

rate of partially observed entities	**expected rate of missing measurement days**
0.1	0.05
0.2	0.10
0.3	0.15
0.4	0.20
0.5	0.25
0.6	0.30
0.7	0.35
0.8	0.40
0.9	0.45
1.0	0.50

GLM and SI-SVR where we have to ensure a non-negative integer output. Further we analyzed the number of imputation iterations in order to find a good trade-off between accuracy and computation time. In these cases overfitting is not an issue. For SI-SVR the parameter tuning includes the selection of an appropriate kernel function which, however, is a necessary step for the application of support vector methods.

We compare errors for the visit potential quantities gross visits, average visits and entity coverage. However, for evaluation we will concentrate on the visit potential quantity gross visits because it is the most basic quantity to describe the interaction between the entity and location set. We decided for this visit potential quantity because if the gross visits already show a large error, other derived quantities are not likely to be more reliable. The exact parameterizations of each preliminary experiment can be found in Appendix B.2.

5.4.1. Parameterization of Multiple Imputation via General Location Model (MI-GLM)

As stated earlier, the general location model (GLM) assumes a multivariate normal distribution of the numeric variables and therefore produces real-valued results. Our model, however, bases on non-negative integer values. Therefore, additional transformations are necessary in order to adapt MI-GLM or respectively its outcome to the required domain. We try to avoid negative values by a log-transformation of the daily number of visits. Non-integer values after imputation are rounded based on a probabilistic procedure. In addition, we also test the number of required imputation rounds.

Preliminary Experiment 1. We perform a log-transformation in order to avoid negative number of visits in the results. More precisely, the log-transformation has the following form

$$y'_{ij} = ln(y_{ij} + a) \tag{5.60}$$

where y_{ij} denotes the number of visits of person i on day j and y'_{ij} denotes the transformed value. The transformation requires an additional additive term a, because y_{ij} may be zero, leading to an invalid transformation. The inverse transformation of the imputed values is given by

$$y_{ij} = e^{y'_{ij}} - a. \tag{5.61}$$

The first preliminary experiment tests which value of a leads to the best performance of MI-GLM. Note that after inverse transformation y_{ij} may still result in negative values due to the additive term. In such cases all values below zero were set to zero. The detailed parameterization of Experiment 1 is given in Appendix B.2.1.

Table 5.6.: Preliminary Experiment 1, MI-GLM with different log-transformations

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM no log	344.0	0.4	0.105	0.141	0.054	0.198	360.6	0.5	0.106
MI-GLM log(y+0.10)	550.9	1.5	0.025	0.103	0.114	0.042	595.4	1.7	0.028
MI-GLM log(y+0.50)	63.9	0.6	0.061	0.019	0.093	0.110	120.9	0.7	0.062
MI-GLM log(y+0.75)	35.1	0.7	0.069	0.018	0.098	0.126	102.8	0.8	0.071
MI-GLM log(y+1.00)	31.5	0.8	0.075	0.023	0.099	0.136	100.2	0.9	0.076
MI-GLM log(y+1.50)	38.4	0.8	0.081	0.034	0.098	0.150	102.4	0.9	0.082
MI-GLM log(y+2.00)	42.2	0.8	0.085	0.042	0.096	0.158	105.9	0.9	0.087
MI-GLM log(y+2.50)	46.6	0.8	0.088	0.048	0.094	0.164	109.7	0.8	0.089
MI-GLM log(y+5.00)	83.1	0.7	0.095	0.071	0.083	0.178	132.3	0.7	0.096

Table 5.6 shows for each parameterization the average absolute compound error (*aace*) for the basic errors mean error, relative mean error and root mean squared error. Remember that *aace* averages the basic errors over five location set sizes with each 30 different location sets and six different rates of artificially induced missingness. The mean error states the bias of the parameterizations while the relative mean error states the bias in proportion to the true value of the respective visit potential quantity. The root mean squared error combines the bias as well as the variance of the parameterizations. For example, without log-transformation GLM has an average absolute compound mean error of 344.0 for gross visits, which amounts on average to 14.1% of the true value of gross visits. The average compound root mean squared error is 360.6, which means that the variance is comparably small with respect to the bias.

The parameterization with smallest $aace(me,\ GV_E\,)$ is log-transformation with $a = 1.0$. This parameterization shows not only the smallest bias but also the smallest error variance as stated by $aace(rmse,\ GV_E\,)$. The *aace* for average visits and entity coverage is not optimal for $a = 1.0$, however, the error for this parameterization is not considerably higher than for others. Therefore, we will conduct all major experiments of MI-GLM with log-transformation and an additive term of $a = 1.0$.

Preliminary Experiment 2. The handling of non-integer values is important because ignorance or simple rounding may lead to underestimation in case of small numbers of visits. Assume, for example, that all persons show 0.99 visits. Given visit class $vc = 1$ and ignoring the problem of non-integer values, the number of gross visits would be zero because no person reaches at least one visit. Ordinary rounding solves this problem in general, however, it fails for smaller numbers of visits. If we decrease the total number of visits of each person to 0.49, the same problem arises again. Be aware that GLM estimates multivariate normal distributions. The predominant rounding of small numbers of visits to zero therefore decreases the mean of the distribution. Figure 5.4 illustrates this behavior. It shows the mean error of the quantity gross visits when estimated with MI-GLM without rounding, with normal rounding and with probabilistic rounding for all location sets of size $|L| = 100$. The mean error is given as difference between the gray line, representing the average visits over 30 location sets, and the black dashed line of each alternative.

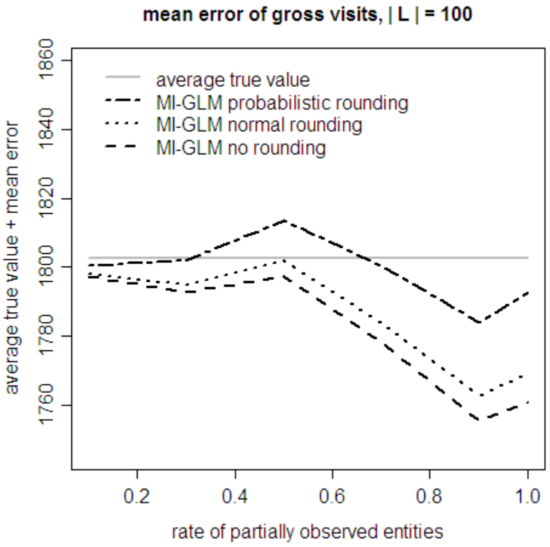


Figure 5.4.: Mean error of gross visits for GLM without rounding, with normal rounding and with probabilistic rounding for location set size $|L| = 100$

We therefore decided for a probabilistic interpretation of the digits after the decimal point. Let $dec(y_{ij})$ denote the decimal part of y_{ij} with $0 \leq dec(y_{ij}) < 1$. With probability $dec(y_{ij})$ variable y_{ij} will assume the next higher integer value and with probability $1 - dec(y_{ij})$ it will assume the integer part of y_{ij}. Naturally, the probabilistic evaluation introduces a variance into the data. In the second preliminary experiment we therefore tested how many iterations of probabilistic rounding are required for stable results. Table 5.7 shows that the results are nearly identical over all versions. The stability is probably a result of several factors, among them the averaging over 30 different location sets and the imputation iterations of MI-GLM itself. For all major experiments we selected a value of 10 iterations for probabilistic rounding. Details on the parameterization of Experiment 2 are given in Appendix B.2.2.

Table 5.7.: Preliminary Experiment 2, MI-GLM with different numbers of probabilistic rounding

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM 1x prob. r.	39.5	0.8	0.074	0.023	0.103	0.136	99.3	0.9	0.076
MI-GLM 10x prob. r.	39.5	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076
MI-GLM 20x prob. r.	39.6	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076
MI-GLM 50x prob. r.	39.6	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076
MI-GLM 100x prob. r.	39.6	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076
MI-GLM 200x prob. r.	39.6	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076
MI-GLM 500x prob. r.	39.6	0.8	0.074	0.023	0.102	0.136	99.2	0.9	0.076

Preliminary Experiment 3. Finally, we tested how many imputation iterations are required for MI-GLM. Schafer and Graham (2002) noted that 20 iterations are sufficient to remove noise from the estimate and other statistical summaries in many practical applications. In our third preliminary experiment we therefore varied the number of imputation rounds between 5 and 50 and compared results. Details on the parameterization of Experiment 3 are given in Appendix B.2.3.

Table 5.8.: Preliminary Experiment 3, MI-GLM with different numbers of imputation rounds

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM 5 imp.	34.9	0.8	0.075	0.022	0.102	0.137	102.7	0.9	0.076
MI-GLM 10 imp.	34.1	0.8	0.074	0.022	0.102	0.136	98.0	0.9	0.076
MI-GLM 15 imp.	34.3	0.8	0.074	0.022	0.102	0.136	97.5	0.9	0.076
MI-GLM 20 imp.	35.7	0.8	0.074	0.022	0.102	0.136	97.9	0.9	0.076
MI-GLM 25 imp.	35.8	0.8	0.074	0.022	0.102	0.136	97.0	0.9	0.076
MI-GLM 30 imp.	34.9	0.8	0.074	0.022	0.102	0.136	96.0	0.9	0.076
MI-GLM 40 imp.	34.6	0.8	0.074	0.022	0.102	0.136	96.1	0.9	0.076
MI-GLM 50 imp.	34.4	0.8	0.074	0.022	0.102	0.136	95.8	0.9	0.076

The *aace* over different numbers of imputations shows little variation in Table 5.8. Only $aace(rmse, \cdot)$ for gross visits improves steadily for increasing numbers of imputations. However, the increase is comparably small considering that the computation time increases linear with the number of imputations. Table 5.9 shows the time that was necessary to complete all experiments behind each version. We decided for 15 rounds of imputations for future experiments with MI-GLM.

Table 5.9.: Computation time of each version of Preliminary Experiment 3 in seconds

5 imp.	10 imp.	15 imp.	20 imp.	25 imp.	30 imp.	40 imp.	50 imp.
745	1,462	2,177	2,884	3,588	4,290	5,705	7,105

5.4.2. Parameterization of Single Imputation via Support Vector Regression (SI-SVR)

As already mentioned in Section 5.2.4, SI-SVR iteratively predicts missing data values for each measurement day. We applied mean substitution to temporarily fill missing values of the independent variables. In the first parameterization experiment for SI-SVR we therefore tested different variants of mean substitution. Second, we tested different kernels and also evaluated the usefulness of log-transformation because, similar to GLM, SVR produces real-valued results. Third, we refined the parameterization for log-transformation and tested different additive terms. Finally, we tested the number of iterations of probabilistic rounding.

Preliminary Experiment 4. Mean substitution is typically performed over the values of a single variable which we will call vertical mean substitution (VMS). As different sociodemographic groups possess different mobile behaviors, we also allow for stratification according to sociodemographic characteristics during VMS. In addition, we perform mean stratification across the number of daily visits per entity which we call horizontal mean stratification (HMS). This approach is motivated by the fact that movements of a person correlate over time (see

Section 2.2.3), and their number of daily visits with a given location set is therefore also likely to be similar over time.

More formally, let variable Y_j with $j = 1..m$ denote the number of daily visits between an entity and a location set on day j. Further, let y_{ij} denote the number of daily visits of entity e_i $(i = 1..n)$ on day j with location set L. For the stratification of VMS we tested the three sociodemographic variables gender (X_g), age group (X_a) and occupation (X_o) as they show a high dependency to travel group. We did not include the variable householder as it possesses a high dependency to the other three variables (see Section 5.1.4). The selected sociodemographic variables are categorical and possess the following domains.

$$X_g \in \{\text{male, female}\}$$
$$X_a \in \{14 - 29 \text{ years}, 30 - 49 \text{ years}, \geq 50 \text{ years}\}$$
$$X_o \in \{\text{in training (pupil, apprentice, student), employed, retired, unemployed}\}$$

We formed stratifications for each variable itself as well as for the combinations (gender, age group) and (gender, occupation). We did not consider combinations with age group and occupation because both variables are highly correlated (see Section 5.1.4). We denote a stratification with the letter S and the stratifying variables as set, e.g. $S = \{X_g\}$ denotes a stratification according to gender and $S = \{\}$ denotes no stratification. We denote a single stratum of S with s. Each stratum consists of an ordered list of values of the independent variables. Further, we define an indicator function I_S which tests whether an entity e belongs to a given stratum s, i.e.

$$I_S(s, e) = \begin{cases} 1 & if\ (x_g, x_a, x_o) = s \\ 0 & else \end{cases} \quad . \tag{5.62}$$

Given a stratification S and a stratum s we can calculate the average value of some variable Y_j for stratum s as

$$\overline{y}_j^s = \frac{\sum_{i=1,\, I_{obs}(y_{ij})}^{n} \; y_{ij} \cdot I_S(e_i, s)}{\sum_{i=1,\, I_{obs}(y_{ij})}^{n} \; I_S(e_i, s)} \tag{5.63}$$

Hereby, the average is formed only over entities where variable Y_j is observed. This is indicated by the boolean function $I_{obs}(y_{ij})$.

For HMS we average the observed number of daily visits per entity, i.e.

$$\overline{y}_i = \frac{\sum_{j=1,\, I_{obs}(y_{ij})}^{m} \; y_{ij}}{\sum_{j=1,\, I_{obs}(y_{ij})}^{m} \; 1}. \tag{5.64}$$

All other parameters for the experiment are given in Appendix B.2.4. Table 5.10 shows the results of the experiment. The average absolute compound mean error of gross visits and average visits is very similar for all versions of VMS. For HMS the aace improves. However, the improvement in mean error is accompanied by an increase in root mean squared error, which is greater for HMS than for VMS. As there is no obvious interpretation of the results, we will select averaging across persons without stratification, averaging across persons with stratification by gender and occupation (both variables show a high dependency to travel group) and averaging across different measurements of the same person for further parameter tuning.

Table 5.10.: Preliminary Experiment 4, SI-SVR with different mean substitution methods for temporary filling of missing data

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
SI-SVR $S = \{\ \}$	222.9	1.1	0.032	0.077	0.099	0.043	492.9	1.8	0.037
SI-SVR $S = \{X_g\}$	223.2	1.1	0.032	0.076	0.099	0.043	490.8	1.8	0.037
SI-SVR $S = \{X_a\}$	224.9	1.1	0.032	0.078	0.100	0.044	498.1	1.9	0.038
SI-SVR $S = \{X_o\}$	227.0	1.1	0.032	0.075	0.097	0.043	497.7	1.8	0.037
SI-SVR $S = \{X_g, X_a\}$	222.3	1.1	0.032	0.074	0.096	0.044	506.9	1.9	0.038
SI-SVR $S = \{X_g, X_o\}$	223.9	1.1	0.032	0.074	0.094	0.044	494.5	1.8	0.038
SI-SVR HMS	189.2	0.6	0.031	0.086	0.086	0.040	625.8	2.0	0.037

Preliminary Experiment 5. Given the above selected averaging methods, we next tested different kernel functions as well as log-transformation for the reduction of negative values. We tested SI-SVR with polynomial kernel and radial basis function. The detailed parameterizations are given in Appendix B.2.5. The results are shown in Table 5.11. Clearly, the combination of a polynomial kernel and log-transformation does not work. For the other combinations of kernel functions and log-transformation, a polynomial kernel without log-transformation achieved the best results in all three cases of mean substitution. Note that the results of polynomial kernel differ from the results in Preliminary Experiment 4, which were also obtained using a polynomial kernel and the same parameterizations otherwise. The only differences in the experiments arise due to random determination of missing values. As SI-SVR shows comparably large errors, it is natural that the error values between different experiments show greater differences as well. However, the differences also indicate an instability of the method, as our results are averaged over 900 individual runs.

Table 5.11.: Preliminary Experiment 5, SI-SVR with different kernel functions, with and without log-transformation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
SI-SVR ra, a1	476.5	2.0	0.034	0.159	0.174	0.047	485.3	2.1	0.041
SI-SVR ra+log, a1	557.5	1.8	0.041	0.195	0.149	0.073	566.1	1.8	0.045
SI-SVR po, a1	215.6	1.2	0.032	0.081	0.105	0.043	496.9	1.9	0.038
SI-SVR po+log, a1	$8.4E^{16}$	$7.8E^{14}$	0.042	$1.7E^{14}$	$2.3E^{14}$	0.075	$4.6E^{17}$	$4.3E^{15}$	0.046
SI-SVR ra, a2	465.2	2.0	0.036	0.158	0.174	0.050	474.1	2.0	0.041
SI-SVR ra+log, a2	548.2	1.8	0.038	0.194	0.151	0.069	556.7	1.8	0.042
SI-SVR po, a2	203.5	1.2	0.032	0.075	0.101	0.044	485.0	1.8	0.038
SI-SVR po+log, a2	$4.0E^{11}$	$3.6E^{9}$	0.041	$7.8E^{8}$	$1.0E^{9}$	0.075	$2.2E^{12}$	$2.0E^{10}$	0.046
SI-SVR ra, a3	229.4	1.3	0.030	0.076	0.098	0.039	247.0	1.3	0.037
SI-SVR ra+log, a3	308.0	0.4	0.048	0.109	0.032	0.080	322.6	0.5	0.050
SI-SVR po, a3	162.3	0.7	0.031	0.089	0.093	0.040	600.0	2.0	0.038
SI-SVR po+log, a3	$1.8E^{7}$	$8.7E^{4}$	0.048	$1.5E^{4}$	$1.7E^{4}$	0.079	$9.4E^{7}$	$4.6E^{5}$	0.050

ra: radial kernel   po: polynomial kernel
ra+log: radial kernel with log-transf.   po+log: polynomial kernel with log-transf.
a1: $S = \{\ \}$ a2: $S = \{X_g, X_o\}$ a3: avg. per entity

Preliminary Experiment 6. So far, SI-SVR used only the contact values as independent variables for prediction. However, as sociodemographic characteristics influence mobile behavior as well, we next added the variables gender and occupation to the data set. Both variables show a strong dependency to average daily travel distance and are therefore most likely to improve our results further. The outcome for the parameterization (see Appendix B.2.6) as in Preliminary Experiment 5 is shown in Table 5.12. Surprisingly, only HMS can improve on further sociodemographic variables. The improvement applies to both, mean error and root mean squared error. This experiment shows that the type of mean substitution as well as the provision of informative sociodemographic variables has great influence on the performance of SI-SVR. For all further experiments we will therefore select SI-SVR with polynomial kernel, without log-transformation and with averaging per person.

Table 5.12.: Preliminary Experiment 6, SI-SVR with different kernel functions, with and without log-transformation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
SI-SVR ra, a1	466.9	2.0	0.038	0.162	0.172	0.056	477.1	2.0	0.042
SI-SVR rad+log, a1	558.3	1.9	0.041	0.198	0.159	0.074	567.7	1.9	0.044
SI-SVR po, a1	208.3	1.1	0.035	0.070	0.089	0.052	349.8	1.5	0.040
SI-SVR po+log, a1	$6.4E^5$	$5.1E^3$	0.039	$1.3E^3$	$1.7E^3$	0.070	$3.5E^6$	$2.8E^4$	0.042
SI-SVR ra, a2	464.2	2.0	0.038	0.161	0.173	0.055	474.5	2.0	0.042
SI-SVR ra+log, a2	555.9	1.9	0.039	0.197	0.158	0.070	565.7	1.9	0.042
SI-SVR po, a2	218.4	1.2	0.035	0.073	0.090	0.051	349.7	1.5	0.039
SI-SVR po+log, a2	$2.0E^5$	$1.2E^3$	0.039	$2.4E^2$	$3.0E^2$	0.070	$1.1E^6$	$6.4E^3$	0.042
SI-SVR ra, a3	289.6	1.4	0.028	0.099	0.108	0.040	305.3	1.4	0.034
SI-SVR ra+log, a3	375.9	0.6	0.053	0.133	0.051	0.087	389.5	0.7	0.054
SI-SVR po, a3	61.0	0.6	0.027	0.033	0.074	0.043	264.8	1.1	0.032
SI-SVR po+log, a3	$1.0E^5$	$5.8E^2$	0.057	$1.1E^2$	$1.4E^2$	0.094	$5.2E^5$	$2.9E^3$	0.058

ra: radial kernel po: polynomial kernel
ra+log: radial kernel with log-transf. po+log: polynomial kernel with log-transf.
a1: $S = \{\ \}$ a2: $S = \{X_g, X_o\}$ a3: HMS

Preliminary Experiment 7. Finally, we tested the number of required iterations of probabilistic rounding for SI-SVR. Details on the parameterization can be found Appendix B.2.7 and the results are given in Table 5.13 and contain no surprises. Similar to MI-GLM we will select 10 iterations of probabilistic rounding for all further experiments.

5.4.3. Parameterization of Multiple Imputation from a Conditional Poisson Distribution (MI-Poisson)

In case of Poisson estimation we apply a two-step procedure as descried in Section 5.2.5. First, we estimate the parameter λ for each entity from its observed measurements. Second, we randomly draw values from a Poisson distribution with parameter λ for the missing data of an entity. In this preliminary experiment we determine how many simulations of random draws are necessary in order to obtain stable results.

Preliminary Experiment 8. In this experiment we perform Poisson estimation with 1, 10, 20, 50, 100, 200 and 500 simulations of repeated draws (see Appendix B.2.8). Each simulation

Table 5.13.: Preliminary Experiment 7, SI-SVR with different numbers of probabilistic rounding

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
SI-SVR 1x prob. r.	187.6	0.6	0.032	0.085	0.084	0.041	646.0	2.0	0.041
SI-SVR 10x prob. r.	187.3	0.6	0.031	0.085	0.084	0.040	646.4	2.0	0.038
SI-SVR 20x prob. r.	187.1	0.6	0.031	0.085	0.084	0.040	645.9	2.0	0.038
SI-SVR 50x prob. r.	187.2	0.6	0.031	0.085	0.084	0.041	646.0	2.0	0.038
SI-SVR 100x prob. r.	187.1	0.6	0.031	0.085	0.084	0.040	645.9	2.0	0.038
SI-SVR 200x prob. r.	187.2	0.6	0.031	0.085	0.084	0.040	646.0	2.0	0.038
SI-SVR 500x prob. r.	187.2	0.6	0.031	0.085	0.084	0.040	646.0	2.0	0.038

generates one complete data set by filling missing values with random draws from a Poisson distribution that is parameterized with the estimated λ of the respective entity. The data set is evaluated and the results are subsequently averaged over all simulations. Table 5.14 shows the average absolute compound error. As expected, the average compound root mean squared error decreases with increasing number of simulations. However, the decrease levels off after 50 simulations. As the error values for the different numbers of simulation are very close and additional simulations increase the running time, we decided to conduct all following experiments of MI-Poisson with 30 simulations.

Table 5.14.: Preliminary Experiment 8, MI-Poisson with differing numbers of simulations

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
parameterization	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-Poisson 1 sim.	14.2	1.6	0.094	0.005	0.200	0.155	102.1	1.7	0.095
MI-Poisson 10 sim.	16.6	1.6	0.094	0.006	0.199	0.155	100.1	1.7	0.095
MI-Poisson 20 sim.	16.3	1.6	0.094	0.006	0.199	0.155	100.0	1.7	0.095
MI-Poisson 50 sim.	16.0	1.6	0.094	0.006	0.199	0.155	99.4	1.7	0.095
MI-Poisson 100 sim.	16.1	1.6	0.094	0.006	0.199	0.155	99.5	1.7	0.095
MI-Poisson 200 sim.	16.2	1.6	0.094	0.006	0.199	0.155	99.4	1.7	0.095
MI-Poisson 500 sim.	16.2	1.6	0.094	0.006	0.199	0.155	99.5	1.7	0.095

5.4.4. Parameterization of Kaplan-Meier Estimation (KM)

Kaplan-Meier has been designed for event data and therefore fulfills the structural requirements of our application. In addition, it is a parameter-free method. However, as it sequentially removes entities that drop out early or that experience a critical event, the problem may arise that all entities from the sample are removed before the time moment in question is reached. In this case the survival function remains at the same level of the last critical event. If the empty set of objects at risk is caused by dropout, the survival function lacks information about how to continue and will consequently overestimate the true survival value in the remaining time. Therefore, subsequent to Kaplan-Meier we performed a regression of the survival function for the time moment of interest if the last critical event had occurred before that.

Preliminary Experiment 9. In this experiment we compare the original version of Kaplan-Meier with the extended version. In the original version we simply repeat the survival value of the last day with a critical event until the time moment in question is reached. In the extended version we predict the value of the requested time span under the assumption that the survival function can be modeled using the log-logistic distribution function. In fact, the log-logistic model is well-know in survival analysis as it allows for flexible hazard functions that may either decrease over time or first increase and then decrease over time (Kleinbaum and Klein, 2005). When applied directly to the survival function $S(t)$, the log-logistic model has the following form:

$$S(t) = \frac{1}{1 + \lambda t^{\alpha}} \qquad \text{with } \lambda > 0, \alpha > 0. \tag{5.65}$$

The parameters λ and α determine the shape of the function. Hereby, λ scales the function in height while α determines the shape (and consequently the hazard) of the function. We use nonlinear (weighted) least-squares estimation (function *nls* of R-toolkit) in order to estimate the parameters of function 5.65. For further details on the experiment see Appendix B.2.9.

Figure 5.5 shows the mean error and root mean squared error for different rates of entities with missing data and location set size $|L| = 100$. Remember that we can estimate only entity coverage with Kaplan-Meier. The original and extended version of Kaplan-Meier possess the same results for a rate of entities with missing data below one. If $r = 1$, no data for the last day is available, and Kaplan-Meier clearly underestimates entity coverage. The subsequent application of our predictive model, however, compensates this effect. The comparison of the compound error over all location set sizes in Table 5.15 confirms that the extended model improves the result of Kaplan-Meier. In our further experiments we will therefore use Kaplan-Meier complemented with a log-logistic regression.

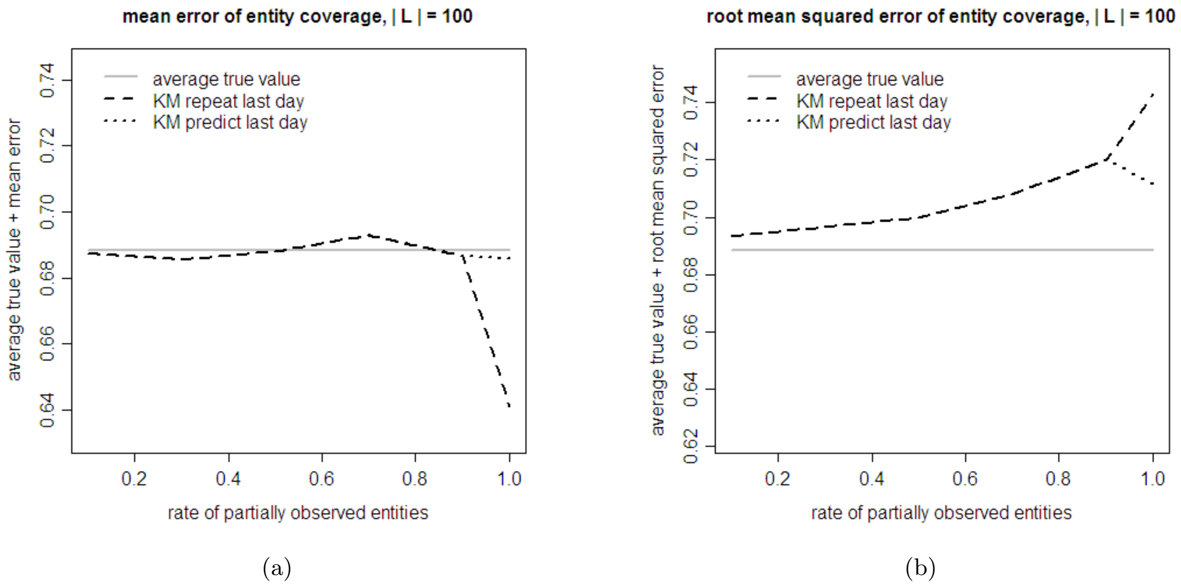


Figure 5.5.: (a) Mean error and (b) root mean squared error (right) of entity coverage for KM with repetition and prediction of the last day in case less than five days with events for location set size $|L| = 100$

Table 5.15.: Preliminary Experiment 9, Kaplan-Meier with repetition and prediction of the last day in case of fewer days with events

	aace(me, ·)	aace(rme, ·)	aace(rmse, ·)
parameterization	C_E	C_E	C_E
KM repeat last day	0.009	0.016	0.021
KM predict last day	0.003	0.005	0.017

5.5. Robustness Test under MCAR

After selecting and adapting methods to the estimation of visit potential under missing data, this section and the following two sections test the performance of the methods. We will test how robust the methods perform with respect to different mechanisms and amounts of missing data. This section addresses the most basic mechanism of missing data, namely missing completely at random (MCAR).

5.5.1. Test Scenario

In this section we test the performance of multiple imputation via general location model (MI-GLM), single imputation via support vector regression (SI-SVR), multiple imputation from a conditional Poisson distribution (MI-Poisson) and Kaplan-Meier (KM) as described in Section 5.2.3 - 5.2.6 under MCAR mechanism. Our aim is to compare the general performance of the methods against each other, and we therefore apply the most basic mechanism of missing data. We observe the behavior of the methods for different sets of independent variables. First we provide only the daily number of visits. Second, we provide additional sociodemographic variables. Furthermore, we compare the computation time of the methods.

All methods are parameterized according to the parameter tuning described in Section 5.4. Details on parameterization and results of each experiment can be found in Appendices B.3, C.2 and C.3. Note that we performed experiments only on a selection of sociodemographic variables (gender, age group, occupation, gender & age group, gender & occupation) due to combinatorial multiplicity. The presented experiments in this section are again a selection of all performed experiments because the display of complete results would overwhelm this chapter. However, the results of the tested variables lead to similar conclusions, and the shown experiments present a generally observed tendency.

We induce missing data randomly into the data set as described in Section 5.3.3. Remember that the rate corresponds to the proportion of *entities* with at least one missing measurement day. We increase the proportion of partially observed entities from 0.1 to 1.0 in steps of 0.1. In addition, we test the algorithms for location sets of size 25, 50, 100, 250 and 500 in order to obtain different levels of the number of visits in the data set. Finally, we conducted each parameterizations with 30 different poster campaigns.

5.5.2. MCAR without Sociodemographic Variables

The first experiment uses no sociodemographic information. All algorithms rely only on the number of daily visits for calculation. Table 5.16 shows the average absolute compound error (*aace*) for the basic errors mean error (*me*), relative mean error (*rme*) and root mean squared error (*rmse*).

Remember that the Kaplan-Meier provides only entity coverage, therefore errors for gross visits and average visits are missing. If we compare the compound errors for gross visits across

Table 5.16.: Experiment 1, MCAR mechanism without sociodemographic variables

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	35.7	0.8	0.070	0.021	0.098	0.128	97.5	0.9	0.071
SI-SVR	171.5	0.6	0.030	0.065	0.073	0.039	585.8	1.9	0.036
MI-Poisson	14.5	1.5	0.089	0.005	0.184	0.146	102.4	1.6	0.090
KM	–	–	0.003	–	–	0.006	–	–	0.017

the first three methods, we see that MI-Poisson has the smallest mean error, followed by MI-GLM with about twice the mean error. The mean error of SI-SVR is comparably large against the other two methods. MI-Poisson possesses also the smallest relative mean error, however, its root mean squared error is slightly above of MI-GLM. This means that MI-Poisson has the smallest bias, which, however, is compensated by the variation of the method if compared with MI-GLM. The root mean squared error of SI-SVR lies again considerably higher than for MI-Poisson and MI-GLM. For average visits, the *aacm* of mean error and relative mean error of MI-GLM and SI-SVR are nearly identical, however, SI-SVR possesses a higher variance than MI-GLM. The compound error of MI-Poisson is about twice the size of the MI-GLM error. For entity coverage KM clearly performs best. It is followed by SI-SVR, MI-GLM and, finally, MI-Poisson.

Similar to our evaluation during parameter tuning, we attribute the highest importance to the estimation of gross visits. Gross visits is the most basic visit potential quantity, and it is used for the calculation of average visits and entity coverage (see Section 4.2.2). Therefore, we believe that if gross visits are not estimated correctly, the estimation of the other visit potential quantities is not reliable as well. This is clearly the case for SI-SVR. Although, SI-SVR provides smaller errors for entity coverage than MI-Poisson and MI-GLM, the results are most likely unreliable. A second indication for the unreliability of SI-SVR is its high root mean squared error for gross visits and average visits when compared to MI-Poisson and MI-GLM. For the remaining evaluations we therefore take a conservative point of view when we interpret the results of SI-SVR. It is not clear why MI-SVR performs worse than the other methods, especially as support vector methods have proved to work well in many other domains. Two directions for improvement are a more sophisticated adaptation of the SVR and an improvement of the imputation schema. Currently the imputations are performed independently for each variable. A combined inference similar to Gibbs sampling in a dependency network (Heckerman et al., 2001) might improve results.

The relative mean error helps to assess the height of the observed mean error. With exception to the estimation of gross visits with MI-Poisson and the estimation of entity coverage with KM all methods have a bias larger than (on average) one percent for the estimation of visit potential quantities. This is an unexpected result. If we take a closer look, the distribution of relative mean error for MI-Poisson and MI-GLM reveals that the estimation of gross visits is easier for both methods than the estimation of average visits and entity coverage. This behavior is plausible because gross visits rely only on a summarization of visits for all entities, while average visits and entity coverage also require a correct distribution of the visits across entities.

Figures 5.6 - 5.8 depict the results of Experiment 1 before the aggregation of the basic errors. The figures show the average value of gross visits, average visits and entity coverage for a given location set size and rate of partially observed entities. Each picture contains a single location set size and shows the behavior for an increasing rate of missing entities. The mean error is given as difference between the gray horizontal line, representing the average visit potential

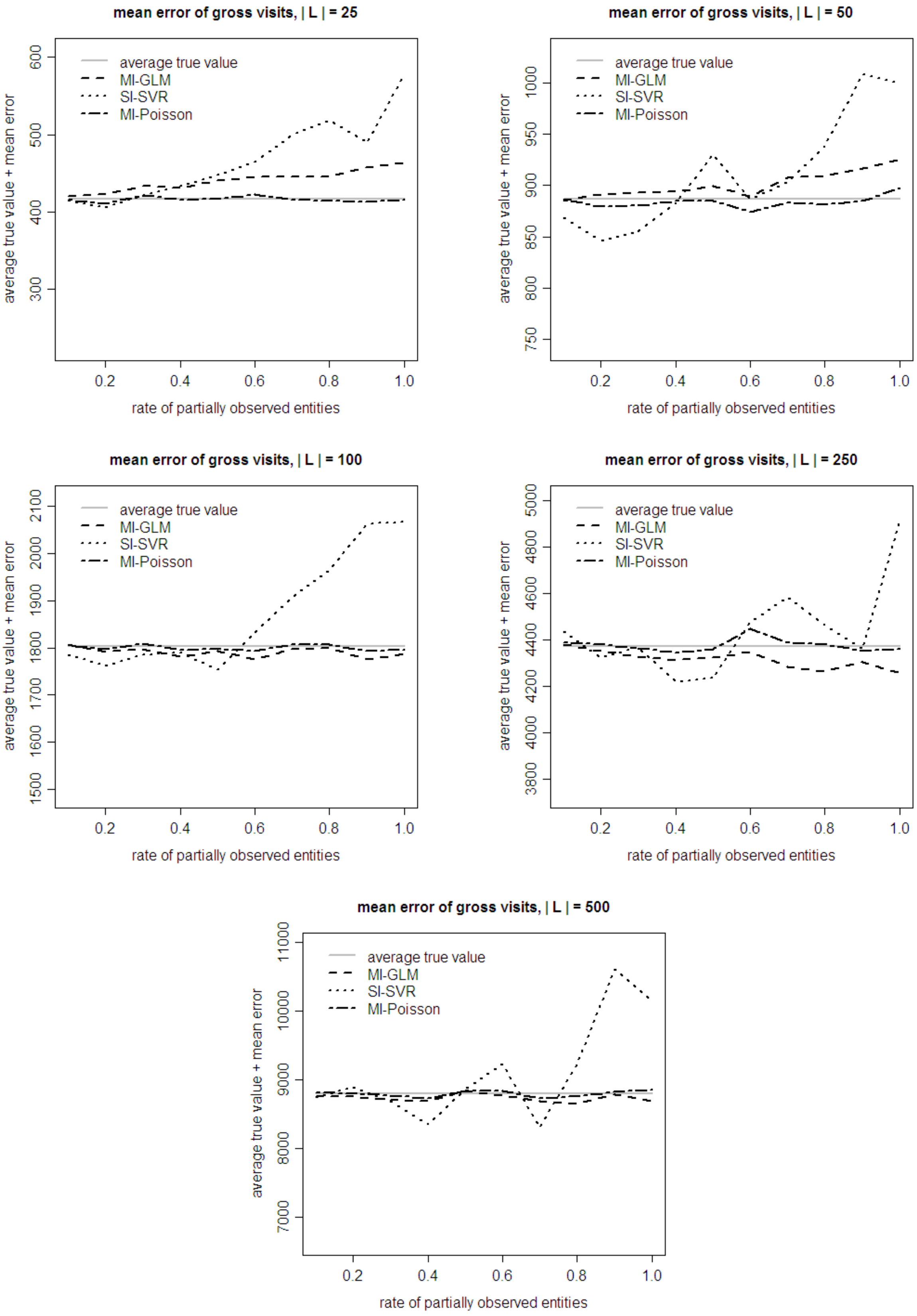


Figure 5.6.: Experiment 1, MCAR mechanism without sociodemographic variables: mean error of gross visits for location set sizes $|L| \in \{25, 50, 100, 250, 500\}$

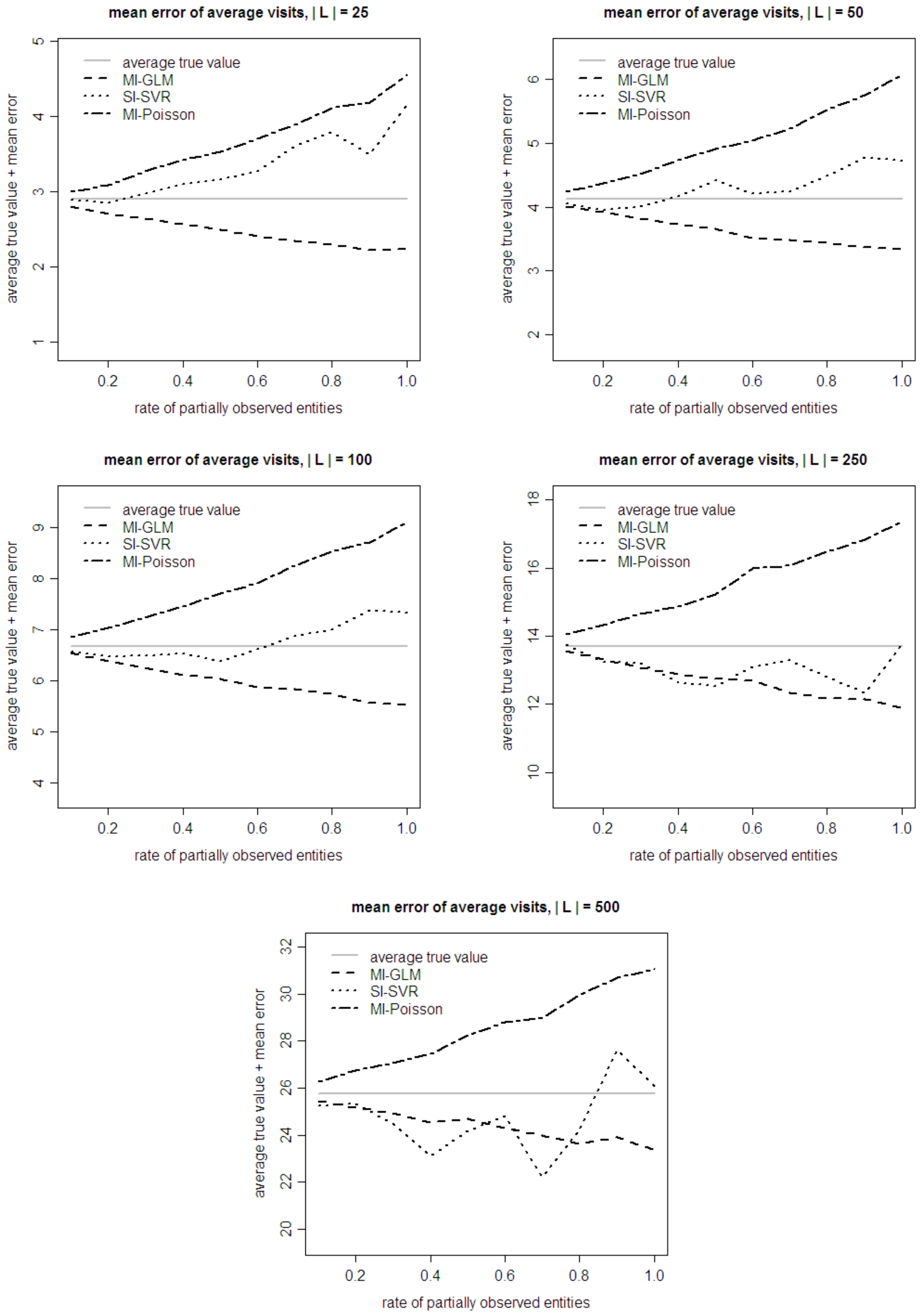


Figure 5.7.: Experiment 1, MCAR mechanism without sociodemographic variables: mean error of average visits for location set sizes $|L| \in \{25, 50, 100, 250, 500\}$

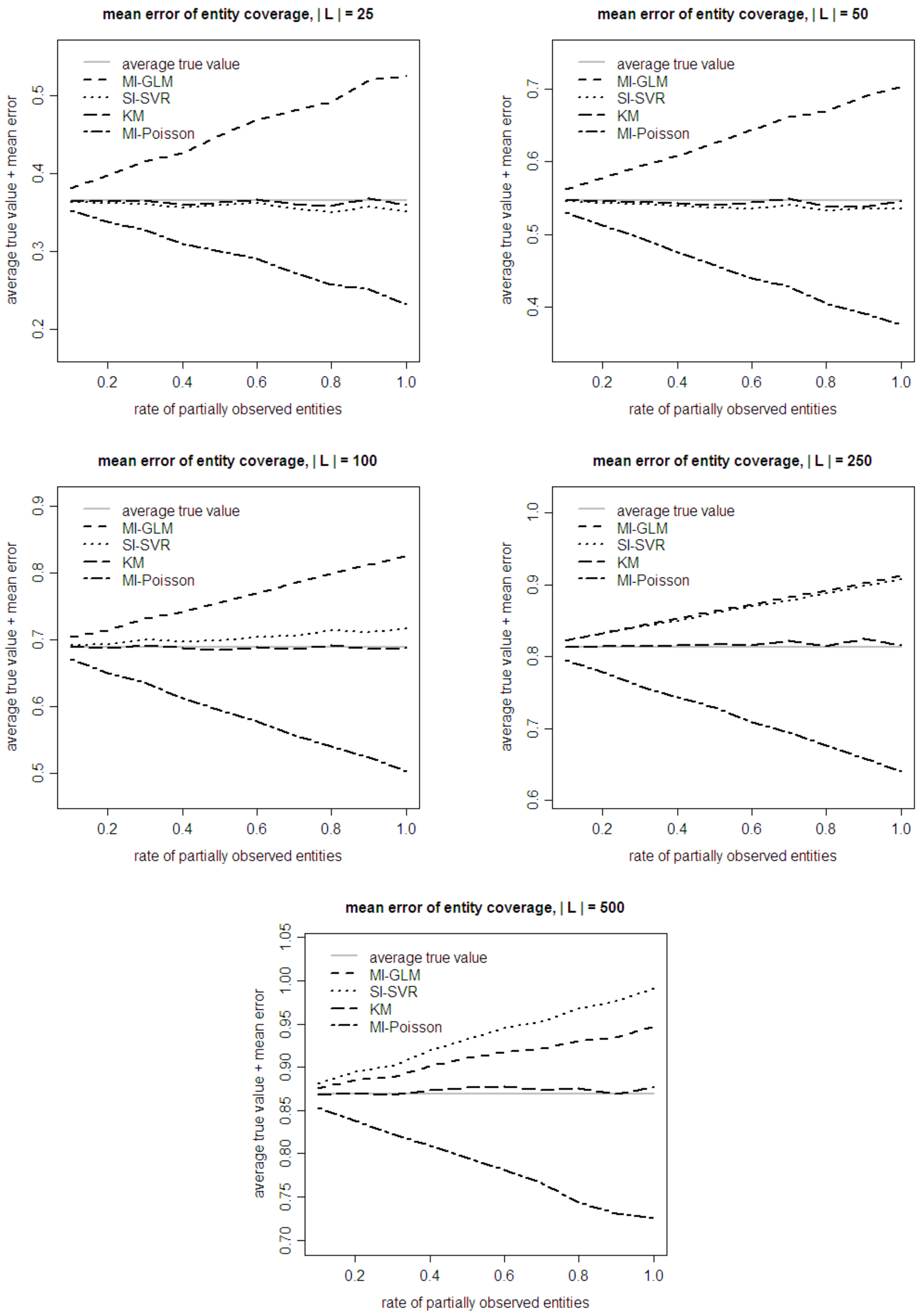


Figure 5.8.: Experiment 1, MCAR mechanism without sociodemographic variables: mean error of entity coverage for location set sizes $|L| \in \{25, 50, 100, 250, 500\}$

over 30 location sets, and the black dashed line of each alternative, representing the average estimated values. Note that the pictures are adapted to the respective levels of visit potential, i.e. the scales differ between different sizes of location sets for the same visit potential quantity. The respective disaggregated results for each figure can be found in Appendix C.2, Tables C.12 - C.14. The detailed results for root mean squared error are listed in Appendix C.3 in Tables C.51 - C.53.

The figures show two relationships between the mean error and the parameterization of the experiments. First, for all visit potential quantities the mean error increases with an increasing rate of missing data. Second, for the quantities gross visits and average visits the mean error increases with an increasing size of the location set. The first observation is plausible, because the less observations are available the more imputations are necessary. However, if the methods are slightly biased, the bias will add up. An exception here is Kaplan-Meier, which remains unbiased. The second observation becomes plausible when we set the mean error in proportion to the true value of gross and average visits. As the values of the quantities increase with larger location sets, the height of the error increases as well. When we compared the relative mean error for both quantities across location set sizes, we observed a constant or even decreasing error with increasing location set size.

For gross visits the pictures in Figure 5.6 show two additional effects. The most obvious one is the instable behavior of SI-SVR. Its mean error contains very irregular fluctuations, which is one further reason for a conservative interpretation of the results of SI-SVR. The second observation is the changing direction of the mean error when we observe MI-GLM for different location set sizes. While MI-GLM overestimates gross visits for small location sets, it underestimates gross visits for large location sets. The reason for this behavior is not obvious. It may be connected to the log-transformation of the data. However, it suggests that although the root mean squared error of MI-GLM is smaller than for MI-Poisson, the unbiased estimate of MI-Poisson may be preferred.

For average visits and entity coverage we can observe an opposite behavior of the methods. For average visits the behavior of SI-SVR is again very instable and turns from an overestimation for small location sets to an increasing underestimation for large location sets. MI-GLM constantly underestimates the true value for average visits and MI-Poisson shows a permanent overestimation. This behavior is reversed when we examine entity coverage. Here, SI-SVR has for small location sets a slight underestimation, which turns into an overestimation for large location sets. MI-GLM constantly overestimates entity coverage and MI-Poisson constantly underestimates entity coverage. The opposite behavior of mean error for average visits and entity coverage can be explained by the relationship of the three visits potential quantities as stated in Equation 4.12. Consider, for example, the method MI-GLM, which has only a small bias for gross visits. If average visits are underestimated, entity coverage must be overestimated in order to balance the mean error of gross visits. Practically, it means that MI-GLM spreads visits across more entities than actually have contact to the location set while MI-Poisson concentrates visits on fewer entities. For MI-GLM this behavior seems to stem from an underestimation of the correlation across the visits for different days. The behavior of Poisson can be explained as follows. Whenever an entity possesses zero visits on all *observed* days, the estimated λ will be zero and consequently all imputed values for this entity. However, in reality visits may have taken place on the missing measurement days. In contrast, if all days with zero visits of an entity are *missing*, there remains still a positive probability that a missing value may be substituted by a zero value. Thus, a concentration of visits takes place on certain entities. In total, however, the number of visits is unbiased because the average number of visits per entity is a true estimate of λ and the total set of entities acts stabilizing in the statistical sense.

In summary, MI-Poisson and MI-GLM are the best methods to estimate gross visits under

MCAR. While MI-GLM possesses the smaller root mean squared error, MI-Poisson provides an unbiased estimate. KM is the best method to estimate entity coverage. For average visits the evaluation is not clear. The method SI-SVR shows very instable results, especially for gross visits. It is therefore likely that also the other quantities estimated over SI-SVR are not reliable.

We made similar observations about the behavior of MI-GLM, SI-SVR, MI-Poisson and KM in our other experiments. Therefore, we will address these behaviors again but concentrate on the aim of the experiments.

5.5.3. MCAR with Sociodemographic Variables

In the next experiments we provide sociodemographic variables for the missing data methods. This information is used by MI-GLM, SI-SVR and KM for conditioning. MI-Poisson does not evaluate sociodemographic information, because the Poisson distributions are estimated separately for each entity (see Section 5.2.5). We have evaluated the following socio-demography: gender (Experiment 2), age group (Experiment 3) and occupation (Experiment 4) as well as the combinations gender & age group (Experiment 5) and gender & occupation (Experiment 6). The compound errors of the experiments are shown in Tables 5.17 - 5.21. Details on the mean error and root mean squared error for individual location set sizes and rates of missing data can be found in Appendix C.2 in Tables C.15 - C.29 and in Appendix C.3 in Tables C.54 - C.68, respectively.

Table 5.17.: Experiment 2, MCAR mechanism, with sociodemographic variable gender

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	40.9	0.8	0.071	0.024	0.098	0.129	101.3	0.9	0.072
SI-SVR	79.9	0.5	0.023	0.047	0.079	0.037	290.6	1.1	0.029
MI-Poisson	18.2	1.5	0.089	0.007	0.185	0.146	102.8	1.6	0.090
KM	–	–	0.003	–	–	0.006	–	–	0.017

Table 5.18.: Experiment 3, MCAR mechanism, with sociodemographic variable age group

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	34.1	0.8	0.071	0.023	0.097	0.131	94.9	0.8	0.072
SI-SVR	71.0	0.6	0.024	0.040	0.069	0.035	259.0	1.0	0.029
MI-Poisson	16.2	1.5	0.089	0.005	0.184	0.145	101.5	1.6	0.090
KM	–	–	0.002	–	–	0.004	–	–	0.017

Table 5.19.: Experiment 4, MCAR mechanism, with sociodemographic variable occupation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	31.1	0.8	0.071	0.023	0.096	0.130	95.4	0.8	0.072
SI-SVR	63.2	0.5	0.022	0.036	0.059	0.033	285.0	1.1	0.028
MI-Poisson	15.4	1.5	0.088	0.005	0.184	0.145	98.0	1.6	0.090
KM	–	–	0.004	–	–	0.007	–	–	0.018

Table 5.20.: Experiment 5, MCAR mechanism, with sociodemographic variables gender and age group

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	30.0	0.8	0.071	0.022	0.096	0.130	97.2	0.8	0.072
SI-SVR	55.9	0.6	0.027	0.037	0.077	0.042	265.1	1.1	0.031
MI-Poisson	17.0	1.5	0.089	0.006	0.185	0.146	100.6	1.6	0.090
KM	–	–	0.004	–	–	0.007	–	–	0.017

Table 5.21.: Experiment 6, MCAR mechanism, with sociodemographic variables gender and occupation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E 4	AV_E	C_E
MI-GLM	16568.5	46.3	0.071	3.188	2.883	0.132	90056.5	251.2	0.073
SI-SVR	64.7	0.6	0.025	0.029	0.061	0.039	247.2	1.1	0.030
MI-Poisson	15.3	1.5	0.089	0.006	0.185	0.146	102.5	1.6	0.090
KM	–	–	0.004	–	–	0.008	–	–	0.018

The main improvements are achieved by SI-SVR. However, the remaining error of SI-SVR is still above the error of the other methods. For MI-GLM the calculation of gross visits improves when adding age group, occupation or gender & age group. However, the improvement is comparably small and, due to the worsening of results when adding variable gender, it is likely that the results are partially caused by random effects. The small change in error of MI-GLM is plausible when considering that we induce missing data completely at random and that the previous estimation of gross visits has been comparably close. However, we would have expected that also average visits and entity coverage improve with additional sociodemographic variables. MI-GLM shows a very high error when adding gender & occupation, which was a surprise. When we look at the detailed results in Appendix C.2, Tables C.27 - C.29 and in Appendix C.3, Tables C.66 - C.68 we see that the large increase in error occurs only for large location sets and high rates of missing data. On closer examination it turned out that this behavior is connected to the log-transformation of MI-GLM. In some cases the imputed values are comparably high, and the inverse exponential transformation amplifies the effect. The performance of KM remains at the same high level as before. Note that the changes in the error of MI-Poisson denote random effects, which are caused by the random generation of artificially missing measurement days.

When we compare the improvement of SI-SVR over the different sociodemographic variables, we see that the variables have different impacts. If only a single sociodemographic variable is added, variable occupation achieves the best results, as may have been expected from the dependency analysis in Section 5.1.4. In combination with gender no further improvement occurs. However, the combination of gender and age group again improves results. This behavior may again be an indication that the selected imputation schema is not sufficient. As a single additional variable leads to great improvement, a joint inference process in the first imputation step might improve results as well.

In summary, the experiments show that additional sociodemographic variables improve results only for SI-SVR and MI-GLM. The height of improvement is connected to the predictive strength of the variables. While the improvement of SI-SVR is very strong as already observed during parameter tuning, it is small for MI-GLM under MCAR. In addition, the experiments showed that a log-transformation of the data in order to avoid negative results may cause extreme results. The general strategy to apply real-numbered methods and to transform the data accordingly in order to make the methods applicable to event data should therefore be reassessed and applied only within limits.

5.5.4. Computation Time

In many real-world applications computation time is one crucial factor that decides about the applicability of a method. We therefore tested the computation time of the selected missing data methods. The timing was taken for a parameterization similar to Experiment 1, i.e. under MCAR without sociodemographic variables. As MI-GLM, SI-SVR and MI-Poisson possess iteration parameters that influence computation time, we performed the experiments once for a single iteration and once for the selected number of iterations. Table 5.22 shows the mean computation time of one run of each method averaged over 1500 experiments as performed for each evaluation.

Table 5.22.: Computation time

time in s	method
1.54	GLM, 15x imputation, 10 x probabilistic rounding
0.84	GLM, 15x imputation, 1 x probabilistic rounding
0.10	GLM, 1x imputation, 1 x probabilistic rounding
0.90	SVM, 10x probabilistic rounding
0.85	SVM, 1x probabilistic rounding
0.13	KM
0.32	Poisson, 30x imputation
0.04	Poisson, 1x imputation

All methods are fast enough to be conducted in real-time. KM is the fastest method, followed by MI-Poisson if we assume realistic parameterizations. The two simple models clearly possess a runtime advantage over the two more complex models. KM is about 7 times faster than SI-SVR and about 12 times faster than MI-GLM. MI-Poisson is still about 5 times faster than SI-SVR and about 5 times faster than MI-GLM. Another reason for the time advantage of KM is that it works directly on the observed data and does not have to impute missing values. If we reduce the other methods to a single imputation and rounding step, MI-GLM becomes very competitive in computation time although it possesses a complex model. The computation time of SI-SVR changes only little because SI-SVR consists of a single imputation step itself. Remember that we train a SVR for each measurement day, so the model actually consists of five successive training and prediction steps, which contributes to the high computation time. From the two time measurements of SI-SVR we can conclude that probabilistic rounding costs about 0.04 seconds. Naturally, the difference of probabilistic rounding for MI-GLM is higher because it has to be applied to all 15 imputations.

In summary, simple and direct missing data methods have a clear advantage in computation time. The costs for data transformations are visible and add up if applied to a multiple imputation schema.

5.6. Robustness Test under CDMAR

In the previous section we tested the performance of the selected missing data methods under a missing completely at random (MCAR) mechanism. In this section we will introduce a dependency between sociodemographic variables and the response variable, leading to covariate-dependent missing at random (CDMAR). We will test how robust the methods perform when biasing the missing data mechanism using a variable with weak or with strong dependencies to the movement behavior.

5.6.1. Test Scenario

In this section we test the performance of multiple imputation via general location model (MI-GLM), single imputation via support vector regression (SI-SVR), multiple imputation from a conditional Poisson distribution (MI-Poisson) and Kaplan-Meier (KM) as described in Section 5.2.3 - 5.2.6 under CDMAR mechanism. We hereby induce systematically missing values for entities with certain sociodemographic attributes. We start with a variable that shows only a weak connection to travel group. Afterwards, we proceed to a variable with high dependency to travel group. The aim of the experiments is to determine how strong CDMAR can influence the performance of the algorithms in our application setting and to what degree the methods can compensate the effect if the according variable is available for conditioning. All methods are parameterized according to the parameter tuning described in Section 5.4. Details on parameterization and results of each experiment can be found in Appendices B.4, C.2 and C.3.

We induce missing data targeted to entities with certain sociodemographic attributes as described in Section 5.3.3. Remember that the rate corresponds to the proportion of *entities* with at least one missing measurement day of the respective sociodemographic group. We increase this rate from 0.1 to 1.0 in steps of 0.1. For the remaining entities we apply a constant rate of missing data of 0.5. In consequence, entities of the selected sociodemographic group are over-represented at the beginning of an experiment and underrepresented at the end of an experiment. Note that due to the different censoring schema the total number of missing measurement days for a given rate differs between the experiments under CDMAR. The reason for this is that the selected attributes have different shares in the data sample and we therefore apply the rate to different proportions of entities. For the same reason, the number of missing measurement days deviates also from the experiments under MCAR. We tested all algorithms again for location sets of sizes 25, 50, 100, 250 and 500 and conducted each parameterization with 30 different poster campaigns.

5.6.2. CDMAR for Variable with Weak Dependency

As analyzed in Section 5.1.4, the variable gender shows only a weak dependency to travel group. We therefore selected this variable for the first analysis of CDMAR mechanism and chose the attribute gender=”female” to define the group of entities with a varying rate of missing data. Table 5.23 shows the aggregated results of the experiment. Details on the mean error and root mean squared error for individual location set sizes and rates of missing data can be found in Appendix C.2 in Tables C.30 - C.32 and in Appendix C.3 in Tables C.69 - C.71, respectively. If we compare the results with Experiment 1 (see Table 5.16), we see that with exception of SI-SVR only small differences occur, which may be attributed to random effects. The behavior of SI-SVR results from the lower total proportion of missing data. As about 25 percent of the entities always keep five measurement days, extreme errors at high rates of missingness are avoided. KM and MI-Poisson still obtain very small errors, from which we may conclude that missingness related to independent variables with lose connection to the dependent variable has only little influence on results.

Table 5.23.: Experiment 7, CDMAR mechanism on gender (female), without sociodemographic variables

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	39.1	0.8	0.067	0.022	0.098	0.124	95.7	0.8	0.069
SI-SVR	102.0	0.7	0.028	0.047	0.067	0.037	546.2	1.9	0.034
MI-Poisson	13.9	1.4	0.086	0.005	0.168	0.140	99.6	1.5	0.087
KM	–	–	0.003	–	–	0.005	–	–	0.015

5.6.3. CDMAR for Variable with Strong Dependency

In this section we perform CDMAR for the sociodemographic variable with the highest dependency to travel group: occupation. More specifically, we chose the attribute occupation=”employed” to define the group of entities with a varying rate of missing data. In Experiment 8 we do not provide any sociodemographic variables to the algorithms while in Experiment 9 we provide variable occupation for conditioning. The results are given in Tables 5.24 and 5.25, respectively. Further details on the mean error and root mean squared error for individual location set sizes and rates of missing data can be found in Appendix C.2 in Tables C.33 - C.38 and in Appendix C.3 in Tables C.72 - C.77, respectively.

Table 5.24.: Experiment 8, CDMAR mechanism on occupation (employed), without sociodemographic variables

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	60.4	0.8	0.066	0.025	0.099	0.122	106.5	0.9	0.068
SI-SVR	127.8	0.7	0.027	0.055	0.074	0.036	496.7	1.7	0.033
MI-Poisson	16.7	1.4	0.085	0.006	0.169	0.139	102.0	1.4	0.086
KM	–	–	0.005	–	–	0.008	–	–	0.015

Table 5.25.: Experiment 9, CDMAR mechanism on occupation (employed), with sociodemographic variable occupation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	37.6	0.8	0.068	0.022	0.098	0.124	95.2	0.8	0.069
SI-SVR	65.6	0.6	0.022	0.027	0.053	0.032	280.3	1.1	0.027
MI-Poisson	18.4	1.4	0.085	0.006	0.168	0.140	101.5	1.5	0.086
KM	–	–	0.003	–	–	0.005	–	–	0.016

Clearly, the errors increase under CDMAR based on occupation. Now KM also shows a slight increase in error for entity coverage. However, the provision of occupation for conditioning reverses the effect completely for KM. MI-GLM and SI-SVR improve also in Experiment 9 and are able to compensate the CDMAR mechanism. The behavior of MI-Poisson for gross visits is a random effect, as MI-Poisson does not rely on sociodemographic variables and performs the evaluation of both experiments under the same condition. In comparison with MI-GLM and SI-SVR, the error for gross visits is still small. This is plausible because MI-Poisson imputes visits separately for each entity based on the available visits of the entity. If the assumption of correlated visits over days as well as the assumed model is correct, MI-Poisson only has to face statistical variation during the calculation of gross visits.

5.7. Robustness Test under MAR

In this section we finally test how robust the missing data methods perform to varying degrees of missing data under a missing at random (MAR) mechanism, i.e. where missingness may relate to any of the observed data. This means that we bias the missing data mechanism to remove measurements preferable of test persons that possess a high (low) travel distance.

5.7.1. Test Scenario

In this section we test the performance of multiple imputation via general location model (MI-GLM), single imputation via support vector regression (SI-SVR), multiple imputation from a conditional Poisson distribution (MI-Poisson) and Kaplan-Meier (KM) as described in Section 5.2.3 - 5.2.6 under MAR mechanism. We hereby induce missing values systematically for entities with certain travel behavior. Note that our variable travel group does not actually correspond to the evaluated number of visits but is used as a substitute because the number of visits varies with each location set (see also Section 5.1.4).

The aim of our experiments under MAR is twofold. First, we want to estimate how well a MAR mechanism can be compensated by other (related) sociodemographic variables. We therefore start the experiments similarly to the previous section without sociodemographic variables and include the variables occupation and travel group in the later experiments. Second, we want to explore the impact of censoring measurements of opposite travel groups. All methods are again parameterized according to the parameter tuning described in Section 5.4. Details on parameterization and results of each experiment can be found in Appendices B.5, C.2 and C.3.

We induce missing data targeted to entities with certain travel behavior as described in Section 5.3.3. For the selected group we increase the proportion of entities with missing data from 0.1 to 1.0 in steps of 0.1. For the remaining entities we apply a constant rate of missing data of 0.5. In Section 5.7.2 we selected entities with the attribute travel group="high" to define the group of entities with a varying rate of missing data. In consequence, entities with high travel group are over-represented at the beginning of an experiment and underrepresented at the end of an experiment. In Section 5.7.3 we vary the rate of missing data for travel group= "low", which has the reverse effect on the data. We tested all algorithms for location sets of sizes 25, 50, 100, 250 and 500 and conducted each parameterizations with 30 different poster campaigns.

5.7.2. Test of Compensation by Sociodemographic Variables for MAR

In this section we perform MAR for the travel group with the highest mobility. In Experiment 10 we do not provide any sociodemographic variables to the algorithms, in Experiment 11 we provide variable occupation and in Experiment 12 we provide variable travel group for conditioning. The aggregated results are given in Tables 5.26, 5.27 and 5.28, respectively. The detailed results about the mean error and root mean squared error for individual location set sizes and rates of missing data can be found in Appendix C.2 in Tables C.39 - C.47 and in Appendix C.3 in Tables C.78 - C.86, respectively.

Similar to Experiment 8 the errors in Experiment 10 increase for all methods with exception to MI-Poisson. However, the increase is stronger than before. This behavior is expected

Table 5.26.: Experiment 10, MAR mechanism on travel group (high), without sociodemographic variables

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	87.5	0.8	0.064	0.033	0.098	0.118	127.6	0.9	0.065
SI-SVR	152.3	0.7	0.028	0.062	0.082	0.037	539.2	1.9	0.033
MI-Poisson	17.3	1.4	0.083	0.007	0.164	0.137	107.2	1.4	0.084
KM	–	–	0.008	–	–	0.014	–	–	0.016

because the daily visits depend stronger on the average number of daily traveled kilometers than on the occupation of a person.

When we add the sociodemographic variable occupation, the results do not improve with exception of SI-SVR. This is surprising as travel group and occupation are dependent on each other according to the analyses in section 5.1.4. The improvement of SI-SVR results most likely from the additional sociodemographic information, however, does not compensate the MAR effect.

Table 5.27.: Experiment 11, MAR mechanism on travel group (high), with sociodemographic variable occupation

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	85.5	0.8	0.064	0.032	0.098	0.118	126.9	0.9	0.065
SI-SVR	72.9	0.5	0.022	0.033	0.057	0.033	309.0	1.2	0.027
MI-Poisson	14.2	1.4	0.084	0.005	0.162	0.137	104.8	1.4	0.085
KM	–	–	0.007	–	–	0.013	–	–	0.016

The results improve when we add travel group as independent variable. KM provides again unbiased results. The mean error of gross visits for MI-GLM nearly halves, however, does not reach the level in Experiments 1 and 9. One explanation for this difference may be the difference in the level of missing data between the experiments. However, it is also possible that MI-GLM is not able to compensate the effect of MAR completely. The results of SI-SVR do not improve, which is surprising because the information about travel group are more appropriate for prediction than occupation.

In summary, our results show that it is important to test missing data mechanisms not only for dependencies with independent variables but whenever possible also for dependencies with the variable of interest itself, because in the second case the influence on results is much stronger. In the mobility application we have the possibility to perform such a test because we assume a strong correlation between the measurements of an entity on successive days. The experiments show further that in case of a MAR dependency on partially observed variables, it is better to supply the original information (e.g. to provide an additional variable on the mobile behavior) than to rely on sociodemographic variables that are related to mobile behavior. Finally, the experiment shows the robustness of MI-Poisson. Due to the estimation of individual Poisson distributions, the method resulted in unbiased gross visits under MAR without use of any independent variables.

Table 5.28.: Experiment 12, MAR mechanism on travel group (high), with sociodemographic variable travel group

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	46.2	0.8	0.064	0.023	0.097	0.118	99.4	0.8	0.065
SI-SVR	72.6	0.6	0.025	0.033	0.062	0.037	281.0	1.2	0.030
MI-Poisson	15.7	1.3	0.084	0.006	0.161	0.138	105.8	1.4	0.085
KM	–	–	0.002	–	–	0.004	–	–	0.014

5.7.3. Variation of Attribute Values for MAR

In this section we perform MAR for the travel group with the lowest mobility and do not provide any sociodemographic variables for conditioning. The results of Experiment 13 are given in Table 5.29. Details about the mean error and root mean squared error for individual location set sizes and rates of missing data can be found in Appendix C.2 in Tables C.48 - C.50 and in Appendix C.3 in Tables C.87 - C.89.

Again, the errors are distinct, however, they are smaller than in Experiment 12. In Figure 5.9 we show the basic error of entity coverage for Experiments 10 (left) and 13 (right) for location sets of size 100. Note that we restricted both figures to the missing data method KM in order to depict only the bias related to MAR. The figures show the opposite development of the bias when censoring persons with high or low mobility. While in Experiment 10 persons with high travel group are overrepresented at the beginning of the experiment and underrepresented at the end, the order is reversed for Experiment 13. Accordingly, entity coverage is overestimated for low rates of missing data in Experiment 10 and underestimated for high rates of missing data. For Experiment 13 the effect is reversed. These findings are similar to our experiments in (May et al., 2009a) based on data from the Swiss outdoor advertising application.

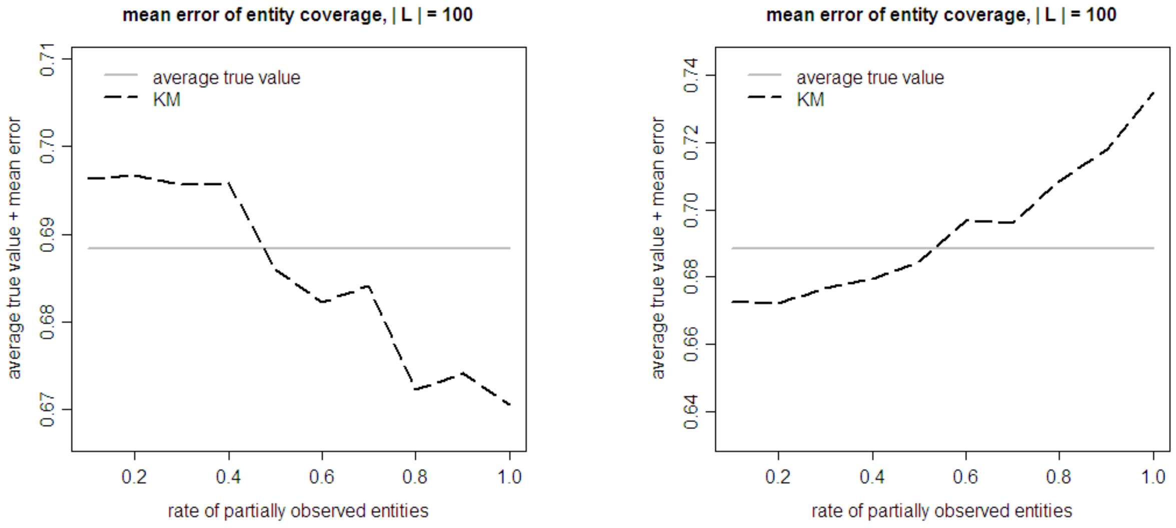
Table 5.29.: Experiment 13, MAR mechanism on travel group (low), without sociodemographic variables

	aace(me, ·)			aace(rme, ·)			aace(rmse, ·)		
Method	GV_E	AV_E	C_E	GV_E	AV_E	C_E	GV_E	AV_E	C_E
MI-GLM	61.2	0.8	0.069	0.029	0.096	0.126	107.0	0.8	0.070
SI-SVR	68.5	0.7	0.030	0.036	0.066	0.040	519.8	1.8	0.036
MI-Poisson	12.9	1.4	0.084	0.006	0.160	0.137	102.2	1.4	0.085
KM	–	–	0.014	–	–	0.023	–	–	0.021

5.8. Discussion

Altogether, the methods reacted to different experimental situations as may have been expected. Introducing dependencies between movement related variables and the censoring mechanism resulted in a decrease of performance. The availability of the respective variables for conditioning had a compensating effect. In addition, large rates of missing data generally lead to higher errors than small rates of missing data. However, we also noticed some unexpected behaviors of the methods which we will discuss in this section.

First, the bad performance of SI-SVR was not expected when designing the experiments. We can think of two main directions to explain the behavior. On the one hand, the parameter-



(a)

(b)

Figure 5.9.: Mean error of entity coverage for KM under MAR mechanism on (a) travel group = ”high” and (b) travel group = “low” for location set size $|L| = 100$

ization of the SVR was probably not optimal. We tested only two kernel methods and used the provided default parameterizations of the software packages. Therefore, a more sophisticated tuning of the base learner will probably improve results. On the other hand, the imputation schema offers room for improvement. So far we performed two independent imputation steps and the imputations for each variable were independent as well. We can overcome this limitation by allowing a combined inference process, for example, by using Gibbs sampling as applied in dependency networks (Heckerman et al., 2001). In addition to improving SI-SVR, one further option is to test the robustness of advanced support vector methods that handle missing data directly as, for example, proposed by Chechik et al. (2007) and Pelckmans et al. (2005).

Second, MI-Poisson proved to perform very robust against CDMAR and MAR mechanisms of missing data. Relying only on the daily number of visits MI-Poisson compensated both mechanisms. This result achieved MI-Poisson because it estimated a Poisson distribution for the number of daily visits separately for each entity. The method therefore performed a conditioning on the smallest level of resolution, i.e. per entity. In practice this characteristic is of great advantage because it means that the method can be applied independently of an analysis of the mechanisms of missingness. It is even applicable if the appropriate variables for conditioning are not available (assuming a mechanism of missingness of at most MAR).

Third, none of the methods yielded good results for the estimation of average visits. Compared to the estimation of gross visits the increased error is plausible because gross visits rely only on a summarization of visits for all entities, while average visits also require a correct distribution of the visits across entities, which makes the estimation task harder.

Fourth, for MI-GLM and MI-Poisson we observed an opposite development of the bias when estimating average visits and entity coverage. MI-GLM showed an underestimation of average visits and an overestimation of entity coverage while MI-Poisson showed an overestimation of average visits and an underestimation of entity coverage. This means that MI-GLM spread visits across more entities than actually had contact to the location set while MI-Poisson concentrated visits on fewer entities. For GLM this behavior seems to stem from an underestimation of the correlation across the visits for different days while the behavior of Poisson is related to the positive probability of zero-visits as described in Section 5.5.2.

Fifth, we observed in experiments with MI-GLM that the often-used strategy of log-transformation in order to obtain non-negative results from real-valued methods may cause extreme values. This effect became visible only in a few situations and therefore leaves an impression of incalculability. In practice a method with such an instability cannot or only very limited be applied. Therefore, future research for the estimation of visit potential from incomplete mobility data should prefer event-based methods.

Finally, we did not expect that different visit potential quantities yield different error relations between the missing data methods. For gross visits MI-Poisson achieved the best results, for entity coverage KM performed best. However, none of the methods performed equally well for all three types of visit potential quantities. In practice this means that the estimation of visit potential from incomplete mobility data is cumbersome because it requires the implementation and maintenance of different methods as well as excessive testing in order to ensure coherent results between the methods.

5.9. Summary

This chapter addressed the estimation of visit potential from incomplete mobility data sets. In particular, our goal was to observe the robustness of missing data methods under various conditions in our experiments. We set up three test scenarios for the missing data mechanisms MCAR, CDMAR and MAR using the mobility data of the German outdoor advertising application. In each scenario we evaluated the performance of MI-GLM, SI-SVR, MI-Poisson and KM for different location set sizes and different proportions of entities with missing data. In the first scenario we applied a MCAR mechanism and tested the performance when providing only the daily number of visits as well as when providing additional sociodemographic variables. In the second scenario we applied a CDMAR mechanism where we introduced missing values targeted to specific sociodemographic groups. We then observed the performance of the four missing data methods once based only on the visit information and once based on additional sociodemographic variables. Finally, we applied a MAR mechanism where we eliminated values of entities with high or low daily travel distance. Again we tested the performance of the methods with and without sociodemographic variables as well as the variable used for censoring.

Under MCAR we observed that only MI-Poisson and KM achieved unbiased results for gross visits and entity coverage, respectively. For average visits none of the methods yielded good results. The addition of sociodemographic variables under MCAR improved especially the performance of SI-SVR. However, SI-SVR still remained with the highest error. As already stated in the discussion section, different possibilities for improvement of SI-SVR exist, which will have to be tested in future work.

In the CDMAR scenario, as may have been expected, a missingness mechanism related to independent variables with lose connection to the dependent variables showed only little influence on results while a missingness mechanism related to an independent variable with strong connection to the dependent variables resulted in increased errors. All methods compensated the induced bias when the according sociodemographic variable was added. An exception hereby was MI-Poisson because it compensated CDMAR already without additional sociodemographic variables.

In the MAR scenario the error increased stronger than under CDMAR, which was expected because the daily visits depend stronger on the average number of daily traveled kilometers than, for example, on the occupation of a person. When we added the sociodemographic variable occupation, the results did not improve. This was surprising because travel group and occupation are dependent on each other according to the analyses in Section 5.1.4. Only when we added travel group as independent variable the bias decreased. Again, MI-Poisson yielded unbiased gross visits for all experiments under MAR without additional variables.

One unexpected result of our experiments was that none of the methods was able to yield unbiased results for all three types of visit potential quantities. Thus, the currently best strategy to estimate visit potential from incomplete data is to apply MI-Poisson for the estimation of gross visits, KM for the estimation of coverage, and to derive average visits from its relationship to the other two quantities. This approach, however, requires to ensure consistency of results between the quantities which applies not only to a single parameterization of the entity set, location set and visit class but also between the results of related entity and location sets, over time and between different visit classes. This makes it very hard to apply the methods in practice and requires further research.

To conclude, Chapter 5 is the first systematic approach to analyze missing data methods for their applicability to mobility data and the estimation of visit potential quantities. We provide a customization of missing data methods to the domain, a test scenario for different missing data mechanisms and extensive experiments that show the strengths and weaknesses of the applied methods. Researchers and practitioners in this domain will now be able to select objectively an appropriate missing data method and to pay attention to possible shortcomings of the method.

6. Conclusion

> *Perfection is attained not when there is no longer anything to add, but when there is no longer anything to take away.*
>
> (Antoine de Saint-Exupéry; Wind, Sand and Stars)

6.1. Summary

Every day people interact with the environment by passing or visiting geographic locations. The knowledge about such interactions is invaluable for a number of applications as, for example, outdoor advertisement which is the central application scenario of this thesis. In consequence, companies as well as public institutions are interested in the evaluation of entity-location interactions. However, they face two problems. First, no uniform terminology and systematic definition of entity-location interactions exists, which makes the specification of application problems, methodological research and interdisciplinary exchange of results difficult. Second, although modern positioning technologies, such as the Global Positioning System (GPS), Global System for Mobile Communications (GSM) or Radio Frequency Identification (RFID), offer a rich source of movement information, they have one common disadvantage: they cannot guarantee complete observation. Thus, the underlying data from which entity-location interaction quantities are typically derived, inherently contain missing data. This problem has to be addressed in order to obtain correct results. However, the current literature does not systematically address the question of handling missing movement data. In this thesis we therefore investigated the following two questions:

1. How can entity-location interactions and interaction quantities be defined?
2. How can entity-location interaction quantities be estimated from incomplete mobility data for applications under real-world conditions?

We addressed the first research question in Chapter 4 and provided a formal definition of entity-location interactions as well as a mathematic concept and context-independent terminology to specify entity-location interaction quantities. We defined a family of basic entity-location interaction quantities which we called *visit potential* and analyzed the relationships between the defined quantities. Furthermore, we demonstrated the applicability of the quantities by formalizing two real-world problems and by providing example calculations using the outdoor advertising application scenario.

We addressed the second research question in Chapter 5 by providing the first systematic approach to analyze missing data methods for their applicability to mobility data. We set up a comprehensive test scenario to evaluate the effect of different mechanisms and degrees of missing data for the estimation of visit potential. More specific, we evaluated the performance of four state-of-the-art missing data methods under the mechanisms missing completely at random (MCAR), covariate-dependent missing at random (CDMAR) and missing at random

(MAR). The methods we applied were multiple imputation via general location model (MI-GLM), single imputation via support vector regression (SI-SVR), multiple imputation from a conditional Poisson distribution (MI-Poisson) and Kaplan-Meier estimation (KM), which we adapted according to the requirements of visit potential. We performed all experiments on real-world application data from the German outdoor advertisement scenario.

Our experiments showed that only two methods were able to obtain unbiased results under the simplest missing data mechanism MCAR, namely MI-Poisson for the estimation of gross visits and KM for the estimation of entity coverage. MI-Poisson turned out to be very robust against the mechanisms CDMAR and MAR, where it was the only method that yielded unbiased results without adding further independent variables for conditioning. However, against our expectations none of the methods was able to obtain unbiased estimates for all three types of visit potential quantities.

6.2. Discussion

How relevant are our results for further exploitation in research and practice and what limits are attached to them? Before proceeding to the first part of the question, let us address the limits of our results. The work of this thesis is certainly limited to a specific part of spatiotemporal data analysis and mobility mining, i.e. the evaluation of entity-location interactions. It cannot answer the general question how to treat missing data during the analysis of mobility patterns or when evaluating the relevance of mobility clusters. It is also limited in the addressed type of missing data. Throughout this thesis we assume that only complete measurement days are missing. However, in practice the problem of missing trips within a day arises as well. This problem is insofar complicated as it requires the detection of missing trips in the first place. Also the evaluated missing data methods allow room for further improvement. As discussed in Section 5.8 more advanced imputation schemes and parameterizations are possible. Last but not least we have limited the applied methods to handle missing data *after* the evaluation of visits. Knowing only a short measurement period, this approach is reasonable. However, with the growing interest in long term mobility studies, the repetitive behavior of human movements may allow for a direct estimation of the missing movement trajectories.

Regarding the relevance of our work for further research and in practice, the historic development of this thesis may give some insights. The estimation of poster performance indicators from incomplete mobility data is a true applied problem in two commercial projects at Fraunhofer IAIS with the Swiss and German outdoor advertising industry. This thesis abstracts the problem setting and formalizes it in an application independent way. Such a formalization - possibly related to different problems - did not exist before. The formalization is, however, essential in order to set the problem into perspective with related work and to analyze it systematically. We therefore see our formalization as pioneering work which opens a number of challenging applied questions to the scientific community. In addition, our framework is very general and extensible which makes it applicable to a potentially large number of scenarios. The second part of our thesis evaluates methods for the estimation of visit potential from missing data. It is the first systematic approach provided for the domain of mobility data analysis. Certainly, the results need further improvement, but they already provide a good direction for practical application and for further research. We therefore see the value of this thesis in bridging the gap between practice and science, providing the scientific community with a formal framework and challenging research questions, and allowing the application side to profit from current and future work.

6.3. Future Work

One clear direction of future work is to improve missing data methods for the estimation of visit potential. As we have already discussed this point several times we will use this section to highlight another research direction of future work. Over the past years the yearning for large mobility surveys, possibly providing mobility data in real-time, has become very strong. On the one hand, obtaining such a data set is technically feasible because positioning technologies have long found their way into personal life (e.g. navigation systems, mobile phones). On the other hand, such a data set would provide a rich source of information for many community and commercial services. For example, visit potential could be applied to estimate the number of visitors at touristic sights or shopping locations. Also the regularity of visiting behavior could be analyzed. However, a comprehensive mobility data set infringes the privacy of individuals. Mobility data contain very sensitive information and it does not suffice to simply remove personal identifiers from the data in order to anonymize them (Monreale et al., 2010). Therefore, researchers in the mobility domain are challenged to develop privacy preserving data mining methods. This challenge applies to visit potential as well. We therefore see one major future research direction in the development of privacy preserving methods for the estimation of visit potential from distributed real-time data sources. The development of such methods is challenging because they have to provide (probabilistic) guarantees for the protection of sensitive information. However, the respect of human rights, including privacy and informational self-determination, is at the heart of any scientific and economic progress.

A. Mathematical Foundations

A.1. Point-Set Topology

Point-set topology (also general topology) establishes the basic concepts of topology. It studies the general aspects of continuity in spaces.

The first part of this section provides general definitions of relevant concepts in point-set topology. The second part adapts the given definitions to the special case of metric spaces. All definitions in Section A.1 are based on Willard (2004) if not otherwise specified.

A.1.1. Topological Spaces

Definition A.1.1 (Topology) *A topology on a set X is a collection of subsets τ which satisfy the following conditions:*

1. *the $\emptyset$ and X belong to τ,*
2. *any union of elements of τ belongs to τ,*
3. *any finite intersection of elements of τ belongs to τ.*

Definition A.1.2 (Topological space) *The ordered pair (X, τ) of a set X and a topology τ on X is called a topological space.*

Definition A.1.3 (Open set) *Given a topological space (X, τ), the sets of the topology τ are called open sets.*

Definition A.1.4 (Closed set) *Given a topological space (X, τ) and a subset $E \subset X$, E is closed if $X - E$ is open.*

Definition A.1.5 (Closure) *Given a topological space X and a subset $E \subset X$, the closure of E in X is the intersection of all closed sets in X that contain E:*

$$\overline{E} = \bigcap \{K \subset X \mid K \textit{ is closed and } E \subset K\}.$$

The closure of a set E is thus the smallest closed set which contains E.

A.1.2. Metric Spaces

The above definitions can be specialized to define, respectively to be applicable to, metric spaces. Metric spaces form the underlying mathematical concept to specify the position and distance of objects in geographic space, because coordinate-based reference systems along with an appropriate distance function are metric spaces.

Definition A.1.6 (Metric space) *A metric space is an ordered pair (M, d) of a set M and a metric (or distance function) d with $d : M \times M \to \mathbb{R}$, so that for any $x, y, z \in M$ holds:*

1. $d(x, y) \geq 0$ *(non-negativity),*

2. $d(x, x) = 0 \Leftrightarrow x = y$ *(identity of indiscernibles) ,*

3. $d(x, y) = d(y, x)$ *(symmetry),*

4. $d(x, y) + d(y, z) \geq d(x, z)$ *(triangle inequality).*

Definition A.1.7 (ϵ-Disk) *Given a metric space (M, d) and a point $x \in M$, the ϵ-disk about x for $\epsilon > 0$ is defined as*

$$U(x, \epsilon) = \{y \in M \mid d(x, y) < \epsilon\}.$$

Definition A.1.8 (Open set) *A subset E of a metric space (M, d) is open iff for each $x \in E$ an ϵ-disc $U(x, \epsilon)$ about x exists which is contained in E.*

The definitions of closed sets and the closure of a set in metric space are equal to the definitions in topological space.

Given a function between two metric spaces, the continuity of the function on a given subset of the domain is defined as follows (Shirali and Vasudeva, 2006).

Definition A.1.9 (Continuous function between metric spaces) *Given two metric spaces (X, d_X) and (Y, d_Y) and a subset $A \subseteq X$. A function $f : A \to Y$ is said to be continuous at $a \in A$, if for every $\varepsilon > 0$, there exists some $\delta > 0$ such that*

$$d_Y(f(x), f(a)) < \varepsilon \text{ whenever } x \in A \text{ and } d_X(x, a) < \delta.$$

If f is continuous at every point of A, then it is said to be continuous on A.

B. Parameterization Details of Experiments

B.1. General Setup

All experiments were conducted with the following parameters.

Entity selection:

- $\mathcal{E}$ = {residents of Hamburg with at least five valid measurement days}
- $|E| = 393$

Location selection:

- $\mathcal{L}$ = {poster locations in Hamburg}
- $|L| \in \{25, 50, 100, 250, 500\}$
- All location sets are randomly drawn from the universal location set $\mathcal{L}$ without repetition.
- For each location set size 30 different location sets were drawn and evaluated.

Time span and visit class:

- time span $t = 5$
- visit class $vc = 1$

In each experiment we measured the visit potential quantities gross visits (GV_E), average visits (AV_E) and entity coverage (C_E).

All experiments were conducted using the R toolkit version 2.11.1 (R Development Core Team, 2010). For implementation of GLM we used package *mix* (Schafer, 2010), for SVR we used package *e1071* (Dimitriadou et al., 2010), which relies on the library *LIBSVM* (Chang and Lin, 2001), and for Kaplan-Meier we used package *survival* (Therneau, 2009).

B.2. Preliminary Experiments

Artificial missing data:

- mechanism of missing data = MCAR
- proportion of entities with missing data $r \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$

B.2.1. Preliminary Experiment 1

This experiment tested MI-GLM with different log-transformations and was conducted with

- 20 imputation rounds,
- log transformation with additive term $a \in \{0.1, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 5.0\}$,
- all values below zero after retransformation set to zero,
- 20 iterations of probabilistic rounding,
- no additional socio-demography.

B.2.2. Preliminary Experiment 2

This experiment tested MI-GLM with different iteration numbers of probabilistic rounding and was conducted with

- 20 imputation rounds,
- log transformation with additive term $a = 1.0$,
- all values below zero after retransformation set to zero,
- respectively 1, 10, 20, 50, 100, 200, 500 iterations of probabilistic rounding,
- no additional socio-demography.

B.2.3. Preliminary Experiment 3

This experiment tested MI-GLM with different numbers of imputation rounds and was conducted with

- respectively 5, 10, 15, 20, 25, 30, 40 and 50 imputation rounds,
- log transformation with additive term $a = 1.0$,
- all values below zero after retransformation set to zero,
- 20 iterations of probabilistic rounding,
- no additional socio-demography.

B.2.4. Preliminary Experiment 4

This experiment tested SI-SVR with different mean substitution strategies for temporarily filling of missing values of independent variables and was conducted with

- polynomial kernel,
- no log transformation,
- all values below zero set to zero,
- 20 iterations of probabilistic rounding,
- VMS without stratification, VMS with stratification by gender (X_g), by age (X_a), by occupation (X_o), by gender and age (X_g, X_a), by gender and occupation (X_g, X_o), HMS,
- no additional socio-demography.

B.2.5. Preliminary Experiment 5

This experiment tested SI-SVR with different kernel functions, with and without log-transformation with

- polynomial kernel, radial kernel,
- no log transformation, log transformation with additive term $a = 1.0$,
- all values below zero set to zero,
- 20 iterations of probabilistic rounding,
- VMS without stratification, VMS with stratification by gender and occupation (X_g, X_o), HMS,
- no additional socio-demography.

B.2.6. Preliminary Experiment 6

This experiment tested SI-SVR with different kernel functions, with and without log-transformation as above, however, with additional sociodemographic information and was conducted with

- polynomial kernel, radial kernel,
- no log transformation, log transformation with additive term $a = 1.0$,
- all values below zero set to zero,
- 20 iterations of probabilistic rounding,
- VMS without stratification, VMS with stratification by gender and occupation (X_g, X_o), HMS,
- additional sociodemographic variables gender and occupation.

B.2.7. Preliminary Experiment 7

This experiment tested SI-SVR with different iteration numbers of probabilistic rounding and was conducted with

- polynomial kernel,
- no log transformation,
- all values below zero set to zero,
- respectively 1, 10, 20, 50, 100, 200, 500 iterations of probabilistic rounding,
- HMS,
- no additional socio-demography.

B.2.8. Preliminary Experiment 8

This experiment tested MI-Poisson estimation with different numbers of simulations for each missing value.

- 1, 10, 20, 50, 100, 200 and 500 simulations.

B.2.9. Preliminary Experiment 9

This experiment tested Kaplan-Meier with different methods for the prediction of survival probability given that no entity survived until the time span in question and was conducted with

- prediction by repetition of last survival probability and prediction based on a log-logistic regression model.

B.3. Experiments under MCAR

All experiments under MCAR were conducted with the optimal parameterization as obtained from parameter tuning. The method for the induction of artificial missing data was the same for all experiments under MCAR.

Parameterization of compared methods:

- multiple imputation via general location model (MI-GLM)
    - log-transformation, $\alpha = 1$, remaining negative values are set to zero
    - 10 rounds of probabilistic rounding
    - 15 rounds of imputation
- single imputation via support vector regression (SI-SVR)
    - polynomial kernel
    - HMS in order to temporarily fill missing values of independent variables
    - no log-transformation, remaining negative values are set to zero
    - 10 rounds of probabilistic rounding
- multiple imputation from a conditional Poisson distribution (MI-Poisson)
    - 30 rounds of imputation
- Kaplan-Meier (KM)
    - prediction of the last day based on a regression log-logistic model

Artificial missing data:

- mechanism of missing data = MCAR
- proportion of entities with missing data $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$

B.3.1. Experiment 1

This experiment compared the missing data methods under MCAR without additional sociodemographic information.

B.3.2. Experiment 2

This experiment compared the missing data methods under MCAR with the additional sociodemographic variable gender.

B.3.3. Experiment 3

This experiment compared the missing data methods under MCAR with the additional sociodemographic variable age group.

B.3.4. Experiment 4

This experiment compared the missing data methods under MCAR with the additional sociodemographic variable occupation.

B.3.5. Experiment 5

This experiment compared the missing data methods under MCAR with the additional sociodemographic variables gender and age group.

B.3.6. Experiment 6

This experiment compared the missing data methods under MCAR with the additional sociodemographic variables gender and occupation.

B.4. Experiments under CDMAR

All experiments under CDMAR were conducted with the optimal parameterization as obtained from parameter tuning. However, the induction of artificial missing data was applied to different sociodemographic groups between the experiments.

Parameterization of compared methods:

- multiple imputation via general location model (MI-GLM)
  - log-transformation, $\alpha = 1$, remaining negative values are set to zero
  - 10 rounds of probabilistic rounding
  - 15 rounds of imputation
- single imputation via support vector regression (SI-SVR)
  - polynomial kernel
  - HMS in order to temporarily fill missing values of independent variables
  - no log-transformation, remaining negative values are set to zero
  - 10 rounds of probabilistic rounding
- multiple imputation from a conditional Poisson distribution (MI-Poisson)
  - 30 rounds of imputation
- Kaplan-Meier (KM)
  - prediction of the last day based on a regression log-logistic model

Artificial missing data:

- mechanism of missing data = CDMAR
- varying proportion of entities with missing data according to a given sociodemographic attribute $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- constant proportion of entities with missing data for all other entities $r_c = 0.5$

B.4.1. Experiment 7

This experiment compared the missing data methods under CDMAR. The varying proportion of entities with missing data was applied to all entities with variable gender=”female” (56.0% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted without additional sociodemographic information for the missing data methods.

B.4.2. Experiment 8

This experiment compared the missing data methods under CDMAR. The varying proportion of entities with missing data was applied to all entities with variable occupation=”employed” (55.7% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted without additional sociodemographic information for the missing data methods.

B.4.3. Experiment 9

This experiment compared the missing data methods under CDMAR. The varying proportion of entities with missing data was applied to all entities with variable occupation=”employed” (55.7% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted with the additional sociodemographic variable occupation for the missing data methods.

B.5. Experiments under MAR

All experiments under MAR were conducted with the optimal parameterization as obtained from parameter tuning. However, the induction of artificial missing data differed between the experiments.

Parameterization of compared methods:

- multiple imputation via general location model (MI-GLM)
    - log-transformation, $\alpha = 1$, remaining negative values are set to zero
    - 10 rounds of probabilistic rounding
    - 15 rounds of imputation
- single imputation via support vector regression (SI-SVR)
    - polynomial kernel
    - HMS in order to temporarily fill missing values of independent variables
    - no log-transformation, remaining negative values are set to zero
    - 10 rounds of probabilistic rounding
- multiple imputation from a conditional Poisson distribution (MI-Poisson)
    - 30 rounds of imputation
- Kaplan-Meier (KM)
    - prediction of the last day based on a regression log-logistic model

Artificial missing data:

- mechanism of missing data = MAR
- varying proportion of entities with missing data according to a given sociodemographic attribute $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
- constant proportion of entities with missing data for all other entities $r_c = 0.5$

B.5.1. Experiment 10

This experiment compared the missing data methods under MAR. The varying proportion of entities with missing data was applied to all entities with variable travel group=”high” (36.1% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted without additional sociodemographic information for the missing data methods.

B.5.2. Experiment 11

This experiment compared the missing data methods under MAR. The varying proportion of entities with missing data was applied to all entities with variable travel group="high" (36.1% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted with the additional sociodemographic variable occupation for the missing data methods.

B.5.3. Experiment 12

This experiment compared the missing data methods under MAR. The varying proportion of entities with missing data was applied to all entities with variable travel group="high" (36.1% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted with the additional variable travel group for the missing data methods.

B.5.4. Experiment 13

This experiment compared the missing data methods under MAR. The varying proportion of entities with missing data was applied to all entities with variable travel group="low" (28.8% of the sample). The remaining entities were censored according to the constant rate r_c. The experiment was conducted without additional sociodemographic information for the missing data methods.

C. Analysis Results

C.1. Tables of Dependency Analysis

This section contains the detailed results of the dependency analysis between travel group and response group for all pairs of sociodemographic variables.

Table C.1.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables gender and age group for test persons in Hamburg, Germany

	age group		
gender	**14 - 29**	**30 - 49**	**$\geq$ 50**
male	*0.774	*0.317	*0.080
female	*0.022	*0.144	*0.061

* approximation may be incorrect due to small cell counts

Table C.2.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables gender and education for test persons in Hamburg, Germany

	education			
gender	**in school**	**secondary general school**	**intermediate secondary school**	**high school / university**
male	NA	*0.240	*0.040	*0.376
female	NA	*0.306	*0.009	*0.600

* approximation may be incorrect due to small cell counts

Table C.3.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables gender and occupation for test persons in Hamburg, Germany

	occupation			
gender	**in training**	**employed**	**retired**	**unemployed**
male	*0.776	*0.061	*0.086	NA
female	*0.164	0.011	*0.476	*0.364

* approximation may be incorrect due to small cell counts

Table C.4.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables gender and householder for test persons in Hamburg, Germany

	householder	
gender	**yes**	**no**
male	*0.015	*0.655
female	0.044	*0.016

* approximation may be incorrect due to small cell counts

Table C.5.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables age group and education for test persons in Hamburg, Germany

	education			
age group	**in school**	**secondary general school**	**intermediate secondary school**	**high school / university**
14 - 29	NA	NA	*0.142	*0.061
30 - 49	NA	*0.518	*0.002	*0.963
$\geq$ 50	NA	*0.227	*0.531	*0.017

* approximation may be incorrect due to small cell counts

Table C.6.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables age group and occupation for test persons in Hamburg, Germany

	occupation			
age group	**in training**	**employed**	**retired**	**unemployed**
14 - 29	*0.298	*0.264	NA	NA
30 - 49	NA	*0.033	NA	NA
$\geq$ 50	NA	*0.087	*0.078	NA

* approximation may be incorrect due to small cell counts

Table C.7.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables age group and householder for test persons in Hamburg, Germany

	householder	
age group	**yes**	**no**
14 - 29	*0.058	*0.152
30 - 49	*0.096	*0.109
$\geq$ 50	*0.117	*0.043

* approximation may be incorrect due to small cell counts

Table C.8.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables education and occupation for test persons in Hamburg, Germany

	occupation			
education	**in training**	**employed**	**retired**	**unemployed**
in school	NA	NA	NA	NA
secondary general school	NA	*0.466	*0.069	NA
intermediate secondary school	NA	*0.000	*0.695	NA
high school / university	*0.161	*0.728	NA	NA

* approximation may be incorrect due to small cell counts

Table C.9.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables education and householder for test persons in Hamburg, Germany

	householder	
education	**yes**	**no**
in school	NA	NA
secondary general school	*0.330	*0.232
intermediate secondary school	*0.022	*0.018
high school / university	*0.021	*0.938

* approximation may be incorrect due to small cell counts

Table C.10.: P-Values of chi-square tests between travel group and response group under conditioning on the sociodemographic variables occupation and householder for test persons in Hamburg, Germany

	householder	
occupation	**yes**	**no**
in training	*0.411	NA
employed	0.002	*0.088
retired	*0.625	*0.131
unemployed	NA	*0.134

* approximation may be incorrect due to small cell counts

C.2. Mean Error of Experiments under MCAR, CDMAR and MAR

This section contains the detailed results of the presented experiments under MCAR, CDMAR and MAR. Tables C.12 - C.29 contain the mean error for Experiments 1 - 6 testing MCAR, Tables C.30 - C.38 contain the mean error for Experiments 7 - 9 testing CDMAR and finally Tables C.39 - C.50 contain the mean error for Experiments 10 - 13 testing MAR.

To each experiment belong three tables showing the visit potential quantities gross visits, average visits and entity coverage. The mean error is formed over 30 repetition of each parameterization. Each part of a table shows the mean error for one tested method. Each row contains results for a certain location set size and columns show results for different rates of missing data.

In order to set the errors in proportion to the true value of the visit potential quantities, table C.11 first shows the values of gross visits, average visits and entity coverage when measured on the complete data set. The values are averages over 30 test location sets per location set size.

Table C.11.: Average true values of visit potential quantities (calculated on complete data) per location set size

visit potential quantity	$\|L\| = 25$	$\|L\| = 50$	$\|L\| = 100$	$\|L\| = 250$	$\|L\| = 500$
GV_E	416.500	886.500	1802.500	4371.933	8791.933
AV_E	2.896	4.121	6.664	13.690	25.751
C_E	0.365	0.546	0.688	0.813	0.869

Table C.12.: Mean error of gross visits under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	4.149	6.197	17.006	14.969	24.934	29.460	29.650	29.032	40.828	46.950
$\|L\|$ =50	-0.700	4.544	6.760	7.370	12.958	3.054	21.010	22.536	30.115	38.996
$\|L\|$ =100	4.600	-11.689	-5.408	-20.680	-9.862	-27.828	-5.213	-2.488	-26.494	-14.486
$\|L\|$ =250	10.228	-16.906	-44.170	-57.038	-48.289	-21.782	-87.258	-105.375	-65.351	-113.902
$\|L\|$ =500	-30.610	-40.418	-94.397	-100.684	48.340	-24.250	-101.461	-150.084	-3.552	-99.421

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	-2.363	-10.120	5.687	16.890	31.330	47.857	83.333	102.600	73.577	160.977
$\|L\|$ = 50	-17.717	-40.673	-31.367	-3.380	44.187	0.263	16.200	52.227	121.527	112.370
$\|L\|$ =100	-17.527	-41.670	-18.017	-12.263	-48.113	30.353	105.607	161.180	260.730	263.953
$\|L\|$ =250	66.773	-45.777	-1.457	-152.047	-136.403	103.977	213.223	95.057	-11.513	536.343
$\|L\|$ =500	-49.657	102.957	-108.180	-443.853	71.173	430.730	-479.603	418.470	1813.207	1363.003

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-1.077	-5.661	4.991	-0.553	0.502	5.919	-1.417	-1.619	-3.450	-0.262
$\|L\|$ =50	-0.868	-6.550	-6.164	-1.399	-1.842	-12.846	-2.926	-5.368	-1.738	11.297
$\|L\|$ =100	2.178	-5.412	6.038	-8.159	-4.724	-8.577	5.250	4.666	-10.138	-6.962
$\|L\|$ =250	16.893	7.420	-4.839	-27.419	-12.441	77.520	18.650	10.234	-18.542	-11.239
$\|L\|$ =500	14.601	10.107	-29.480	-57.036	40.173	48.424	-60.861	-39.896	26.830	51.637

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.13.: Mean error of average visits under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.098	-0.197	-0.254	-0.335	-0.406	-0.487	-0.545	-0.598	-0.669	-0.662
$\|L\|$ =50	-0.118	-0.205	-0.297	-0.393	-0.467	-0.615	-0.638	-0.678	-0.747	-0.782
$\|L\|$ =100	-0.128	-0.283	-0.412	-0.555	-0.622	-0.798	-0.834	-0.930	-1.099	-1.146
$\|L\|$ =250	-0.131	-0.384	-0.620	-0.815	-0.948	-0.992	-1.344	-1.508	-1.544	-1.816
$\|L\|$ =500	-0.295	-0.585	-0.830	-1.217	-1.059	-1.437	-1.767	-2.101	-1.827	-2.400

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.004	-0.045	0.079	0.197	0.261	0.375	0.712	0.892	0.597	1.257
$\|L\|$ =50	-0.072	-0.168	-0.115	0.043	0.290	0.086	0.118	0.362	0.656	0.604
$\|L\|$ =100	-0.103	-0.197	-0.176	-0.137	-0.292	-0.044	0.203	0.341	0.718	0.664
$\|L\|$ =250	0.058	-0.456	-0.474	-1.059	-1.160	-0.605	-0.411	-0.891	-1.343	0.069
$\|L\|$ =500	-0.504	-0.439	-1.244	-2.644	-1.585	-0.931	-3.551	-1.542	1.851	0.326

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.099	0.188	0.371	0.520	0.631	0.800	0.991	1.206	1.280	1.653
$\|L\|$ =50	0.124	0.246	0.396	0.610	0.794	0.922	1.116	1.409	1.638	1.954
$\|L\|$ =100	0.189	0.375	0.575	0.795	1.035	1.248	1.595	1.854	2.044	2.437
$\|L\|$ =250	0.362	0.629	0.969	1.169	1.527	2.295	2.392	2.796	3.139	3.655
$\|L\|$ =500	0.521	0.989	1.328	1.713	2.500	3.036	3.253	4.207	4.937	5.293

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.14.: Mean error of entity coverage under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.016	0.032	0.051	0.061	0.083	0.104	0.116	0.126	0.155	0.160
$\|L\|$ =50	0.016	0.031	0.047	0.062	0.079	0.098	0.115	0.124	0.143	0.157
$\|L\|$ =100	0.015	0.026	0.043	0.054	0.066	0.081	0.096	0.110	0.124	0.136
$\|L\|$ =250	0.010	0.020	0.030	0.040	0.051	0.059	0.070	0.079	0.089	0.100
$\|L\|$ =500	0.007	0.016	0.019	0.033	0.042	0.049	0.053	0.061	0.066	0.079

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.002	-0.003	-0.004	-0.009	-0.006	-0.003	-0.011	-0.015	-0.008	-0.014
$\|L\|$ =50	-0.001	-0.003	-0.004	-0.008	-0.010	-0.011	-0.006	-0.014	-0.011	-0.011
$\|L\|$ =100	0.004	0.004	0.011	0.009	0.011	0.016	0.017	0.025	0.022	0.028
$\|L\|$ =250	0.010	0.019	0.029	0.037	0.048	0.057	0.065	0.075	0.086	0.094
$\|L\|$ =500	0.012	0.026	0.033	0.050	0.064	0.076	0.084	0.099	0.108	0.122

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.013	-0.027	-0.037	-0.056	-0.065	-0.075	-0.093	-0.108	-0.114	-0.133
$\|L\|$ =50	-0.016	-0.035	-0.051	-0.071	-0.089	-0.107	-0.118	-0.142	-0.156	-0.171
$\|L\|$ =100	-0.018	-0.039	-0.053	-0.076	-0.094	-0.111	-0.131	-0.148	-0.164	-0.186
$\|L\|$ =250	-0.018	-0.034	-0.054	-0.069	-0.083	-0.104	-0.118	-0.136	-0.154	-0.173
$\|L\|$ =500	-0.016	-0.031	-0.045	-0.060	-0.073	-0.087	-0.103	-0.125	-0.138	-0.144

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.001	-0.001	0.000	-0.006	-0.003	0.002	-0.005	-0.007	0.002	-0.005
$\|L\|$ =50	-0.000	-0.001	-0.002	-0.003	-0.006	-0.002	0.003	-0.009	-0.009	-0.001
$\|L\|$ =100	0.001	-0.002	0.003	-0.001	-0.004	0.001	-0.003	0.003	-0.002	-0.000
$\|L\|$ =250	0.000	0.002	0.001	0.003	0.004	0.003	0.009	0.002	0.012	0.003
$\|L\|$ =500	-0.000	0.002	-0.001	0.005	0.008	0.009	0.005	0.007	0.000	0.008

Table C.15.: Mean error of gross visits under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	3.868	12.055	18.992	17.100	20.693	27.623	37.944	37.191	38.986	58.170
$\|L\| = 50$	0.719	2.487	9.412	16.389	14.193	22.462	18.513	21.650	42.410	32.838
$\|L\| = 100$	1.194	1.080	-11.942	-9.200	-15.394	-5.054	-16.040	-14.713	7.680	-29.794
$\|L\| = 250$	-2.311	-32.259	-72.178	-47.112	-34.702	-32.616	-130.680	-132.373	-101.049	-73.127
$\|L\| = 500$	-12.980	-19.857	-76.886	-71.122	-127.620	-104.264	-102.601	-225.936	-52.165	-25.506

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.253	-2.117	19.010	7.547	69.453	55.797	67.570	75.780	77.707	91.257
$\|L\| = 50$	-8.800	-28.440	-4.117	23.847	-20.567	17.567	54.163	59.117	120.240	69.567
$\|L\| = 100$	-14.083	-2.623	5.300	14.517	12.083	26.360	28.677	21.290	204.820	83.830
$\|L\| = 250$	-52.690	-131.510	-85.187	-120.143	-129.840	4.480	-142.650	79.487	135.463	255.273
$\|L\| = 500$	-29.063	-89.753	-133.280	-215.207	-233.623	-66.940	116.267	-52.130	188.830	448.387

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	-0.763	-0.738	5.670	-4.223	-5.086	5.898	7.791	1.518	-5.053	11.383
$\|L\| = 50$	-0.310	-6.202	5.533	13.158	2.456	4.628	-0.400	-9.918	14.167	0.747
$\|L\| = 100$	-2.016	2.210	10.720	11.508	-3.217	14.629	-21.579	9.968	39.874	-21.318
$\|L\| = 250$	-2.242	-19.532	-18.238	-21.156	-11.789	44.450	-86.803	-1.787	-26.781	25.569
$\|L\| = 500$	20.092	43.993	24.267	-59.774	-39.059	-60.480	30.627	-63.979	17.628	47.829

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\| =100$	–	–	–	–	–	–	–	–	–	–
$\|L\| =250$	–	–	–	–	–	–	–	–	–	–
$\|L\| =500$	–	–	–	–	–	–	–	–	–	–

Table C.16.: Mean error of average visits under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	-0.105	-0.176	-0.247	-0.348	-0.422	-0.490	-0.511	-0.582	-0.639	-0.646
$\|L\| = 50$	-0.108	-0.215	-0.305	-0.398	-0.473	-0.559	-0.621	-0.698	-0.746	-0.794
$\|L\| = 100$	-0.135	-0.265	-0.413	-0.530	-0.646	-0.731	-0.840	-0.957	-0.985	-1.185
$\|L\| = 250$	-0.181	-0.431	-0.722	-0.773	-0.907	-1.073	-1.448	-1.559	-1.632	-1.733
$\|L\| = 500$	-0.288	-0.469	-0.846	-1.069	-1.443	-1.526	-1.813	-2.360	-2.048	-2.188

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.008	0.023	0.198	0.185	0.726	0.550	0.756	0.800	0.843	0.938
$\|L\| = 50$	0.006	-0.030	0.084	0.251	0.140	0.337	0.665	0.696	0.965	0.819
$\|L\| = 100$	-0.023	0.062	0.134	0.205	0.258	0.396	0.422	0.399	1.148	0.666
$\|L\| = 250$	-0.234	-0.486	-0.414	-0.536	-0.620	-0.257	-0.636	0.016	0.140	0.365
$\|L\| = 500$	-0.321	-0.638	-0.900	-1.373	-1.585	-1.259	-0.872	-1.533	-0.971	-0.378

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.093	0.222	0.362	0.472	0.608	0.812	1.051	1.170	1.388	1.657
$\|L\| = 50$	0.137	0.270	0.433	0.609	0.766	0.939	1.138	1.329	1.605	1.831
$\|L\| = 100$	0.161	0.387	0.621	0.834	1.047	1.399	1.537	1.869	2.380	2.401
$\|L\| = 250$	0.294	0.579	0.872	1.202	1.564	2.107	2.132	2.953	3.271	3.547
$\|L\| = 500$	0.479	1.003	1.425	1.808	2.217	2.919	3.662	3.967	4.979	5.581

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\| =100$	–	–	–	–	–	–	–	–	–	–
$\|L\| =250$	–	–	–	–	–	–	–	–	–	–
$\|L\| =500$	–	–	–	–	–	–	–	–	–	–

Table C.17.: Mean error of entity coverage under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.017	0.034	0.051	0.066	0.082	0.102	0.117	0.131	0.145	0.169
$\|L\| = 50$	0.015	0.031	0.050	0.069	0.080	0.101	0.110	0.127	0.152	0.155
$\|L\| = 100$	0.015	0.029	0.041	0.055	0.067	0.082	0.092	0.109	0.123	0.135
$\|L\| = 250$	0.010	0.020	0.031	0.039	0.051	0.062	0.069	0.077	0.088	0.102
$\|L\| = 500$	0.009	0.014	0.022	0.030	0.038	0.044	0.055	0.063	0.070	0.078

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	-0.001	-0.005	-0.007	-0.014	-0.019	-0.017	-0.026	-0.025	-0.028	-0.027
$\|L\| = 50$	-0.006	-0.014	-0.013	-0.017	-0.030	-0.032	-0.047	-0.047	-0.044	-0.056
$\|L\| = 100$	-0.003	-0.007	-0.012	-0.016	-0.021	-0.029	-0.030	-0.031	-0.033	-0.034
$\|L\| = 250$	0.004	0.005	0.009	0.010	0.013	0.016	0.012	0.014	0.016	0.025
$\|L\| = 500$	0.008	0.013	0.018	0.026	0.032	0.038	0.042	0.050	0.054	0.058

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	-0.012	-0.027	-0.037	-0.054	-0.067	-0.076	-0.092	-0.104	-0.121	-0.126
$\|L\| = 50$	-0.018	-0.037	-0.049	-0.063	-0.084	-0.099	-0.118	-0.138	-0.147	-0.168
$\|L\| = 100$	-0.017	-0.037	-0.055	-0.073	-0.094	-0.115	-0.135	-0.147	-0.169	-0.188
$\|L\| = 250$	-0.017	-0.036	-0.052	-0.069	-0.085	-0.101	-0.123	-0.144	-0.161	-0.164
$\|L\| = 500$	-0.014	-0.028	-0.043	-0.062	-0.072	-0.094	-0.105	-0.121	-0.139	-0.151

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.000	-0.001	0.002	-0.003	-0.004	0.002	-0.003	0.002	-0.009	0.013
$\|L\| = 50$	-0.001	-0.002	0.002	0.005	-0.001	0.001	-0.004	-0.004	-0.002	0.007
$\|L\| = 100$	0.001	0.001	-0.000	0.001	0.001	-0.001	-0.005	0.002	-0.001	-0.002
$\|L\| = 250$	0.001	0.000	0.004	0.003	0.006	0.005	0.003	0.002	-0.002	0.009
$\|L\| = 500$	0.002	0.001	0.001	0.001	0.004	0.001	0.009	0.012	0.011	0.007

Table C.18.: Mean error of gross visits under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	3.986	10.073	15.315	18.396	24.804	26.404	34.317	39.198	46.436	47.707
$\lvert L\rvert = 50$	2.745	10.802	10.809	10.040	18.497	23.436	33.586	17.877	34.455	43.755
$\lvert L\rvert = 100$	2.633	0.069	-6.541	-3.160	5.732	-30.323	3.058	-5.261	-16.574	3.906
$\lvert L\rvert = 250$	-20.542	-39.880	-35.757	-47.138	-44.259	-74.531	-114.018	-48.856	-129.565	-43.821
$\lvert L\rvert = 500$	-7.854	7.268	12.069	4.373	-28.434	-86.279	-0.770	-128.481	-104.415	178.027

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	2.810	7.900	9.650	50.133	22.910	41.657	65.043	45.050	62.587	95.213
$\lvert L\rvert = 50$	-7.593	-11.203	4.183	-7.717	5.497	9.473	74.773	18.500	103.427	103.680
$\lvert L\rvert = 100$	33.377	-21.027	-7.337	-7.877	-16.957	14.810	0.560	104.847	50.910	151.507
$\lvert L\rvert = 250$	-52.620	-71.530	-46.263	-131.747	-139.360	34.297	22.583	62.153	81.770	133.467
$\lvert L\rvert = 500$	-123.250	-139.293	-56.660	-164.590	-122.403	-259.927	-44.357	-71.913	301.750	330.480

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	0.403	2.579	-2.237	3.752	2.386	-1.288	3.396	0.203	3.052	-1.776
$\lvert L\rvert = 50$	-3.507	5.356	1.993	-4.053	7.393	3.360	6.486	-12.058	11.939	2.258
$\lvert L\rvert = 100$	-1.007	2.023	8.852	11.999	10.523	-15.608	13.797	10.472	-16.342	6.161
$\lvert L\rvert = 250$	-23.634	-1.547	2.388	-17.602	-2.698	0.313	-4.353	26.619	-60.187	11.187
$\lvert L\rvert = 500$	-14.974	22.024	62.591	51.913	58.387	-53.437	20.041	-84.166	2.471	111.604

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert = 50$	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert =100$	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert =250$	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert =500$	–	–	–	–	–	–	–	–	–	–

Table C.19.: Mean error of average visits under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	-0.090	-0.199	-0.268	-0.344	-0.423	-0.502	-0.538	-0.580	-0.624	-0.692
$\|L\| = 50$	-0.116	-0.199	-0.290	-0.394	-0.467	-0.542	-0.595	-0.722	-0.767	-0.780
$\|L\| = 100$	-0.127	-0.262	-0.403	-0.519	-0.572	-0.823	-0.811	-0.943	-1.076	-1.119
$\|L\| = 250$	-0.223	-0.455	-0.604	-0.786	-0.973	-1.116	-1.404	-1.366	-1.715	-1.601
$\|L\| = 500$	-0.225	-0.443	-0.598	-0.831	-1.177	-1.539	-1.500	-2.054	-2.250	-1.589

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.052	0.077	0.148	0.480	0.252	0.421	0.633	0.480	0.641	0.879
$\|L\| = 50$	-0.005	-0.001	0.123	0.105	0.219	0.260	0.626	0.408	0.800	0.888
$\|L\| = 100$	0.132	-0.051	0.022	0.014	0.075	0.219	0.160	0.523	0.426	0.705
$\|L\| = 250$	-0.253	-0.398	-0.396	-0.691	-0.815	-0.220	-0.352	-0.386	-0.359	-0.309
$\|L\| = 500$	-0.598	-0.936	-0.896	-1.421	-1.569	-2.097	-1.753	-2.063	-1.259	-1.387

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	0.121	0.221	0.339	0.512	0.632	0.749	0.942	1.141	1.378	1.521
$\|L\| = 50$	0.121	0.287	0.437	0.594	0.794	0.964	1.163	1.313	1.658	1.957
$\|L\| = 100$	0.181	0.375	0.608	0.803	1.081	1.234	1.559	1.815	2.142	2.466
$\|L\| = 250$	0.217	0.550	0.930	1.242	1.617	1.991	2.400	2.850	3.156	3.759
$\|L\| = 500$	0.438	0.977	1.543	2.026	2.474	2.837	3.472	3.954	4.843	6.016

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\| =100$	–	–	–	–	–	–	–	–	–	–
$\|L\| =250$	–	–	–	–	–	–	–	–	–	–
$\|L\| =500$	–	–	–	–	–	–	–	–	–	–

Table C.20.: Mean error of entity coverage under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	0.015	0.035	0.051	0.067	0.087	0.103	0.118	0.132	0.150	0.167
$\lvert L\rvert = 50$	0.017	0.034	0.049	0.064	0.083	0.099	0.115	0.128	0.150	0.161
$\lvert L\rvert = 100$	0.014	0.028	0.042	0.056	0.067	0.083	0.097	0.111	0.125	0.141
$\lvert L\rvert = 250$	0.009	0.020	0.031	0.040	0.053	0.057	0.069	0.080	0.089	0.098
$\lvert L\rvert = 500$	0.007	0.016	0.022	0.029	0.039	0.046	0.054	0.061	0.072	0.076

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	-0.004	-0.003	-0.009	-0.011	-0.011	-0.014	-0.016	-0.017	-0.023	-0.021
$\lvert L\rvert = 50$	-0.004	-0.007	-0.012	-0.019	-0.023	-0.027	-0.031	-0.040	-0.035	-0.045
$\lvert L\rvert = 100$	-0.001	-0.003	-0.006	-0.005	-0.014	-0.017	-0.015	-0.014	-0.023	-0.014
$\lvert L\rvert = 250$	0.005	0.011	0.015	0.017	0.024	0.020	0.026	0.034	0.037	0.043
$\lvert L\rvert = 500$	0.008	0.019	0.025	0.033	0.043	0.049	0.059	0.068	0.076	0.084

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	-0.015	-0.024	-0.040	-0.052	-0.064	-0.076	-0.088	-0.103	-0.116	-0.128
$\lvert L\rvert = 50$	-0.018	-0.033	-0.051	-0.071	-0.084	-0.101	-0.117	-0.138	-0.151	-0.174
$\lvert L\rvert = 100$	-0.019	-0.036	-0.055	-0.070	-0.093	-0.112	-0.126	-0.144	-0.172	-0.184
$\lvert L\rvert = 250$	-0.017	-0.031	-0.051	-0.070	-0.086	-0.103	-0.122	-0.136	-0.161	-0.173
$\lvert L\rvert = 500$	-0.016	-0.030	-0.043	-0.058	-0.071	-0.091	-0.101	-0.123	-0.137	-0.155

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert = 25$	-0.002	0.002	-0.002	-0.002	0.000	-0.000	0.000	0.002	-0.001	0.009
$\lvert L\rvert = 50$	0.000	0.001	0.001	-0.003	0.001	0.000	0.004	-0.001	-0.002	0.001
$\lvert L\rvert = 100$	0.000	0.001	0.001	0.003	-0.002	0.003	0.002	-0.001	-0.002	0.003
$\lvert L\rvert = 250$	0.001	0.002	0.003	0.002	0.006	0.002	0.006	0.004	-0.000	0.000
$\lvert L\rvert = 500$	-0.000	0.002	0.001	0.002	0.005	0.003	0.010	0.008	0.005	0.002

Table C.21.: Mean error of gross visits under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	5.477	13.268	9.781	20.827	22.072	26.946	33.139	39.771	44.744	54.416
$\|L\|$ =50	1.458	7.378	0.315	15.798	19.095	19.439	31.022	39.912	36.602	41.063
$\|L\|$ =100	4.501	-7.138	4.252	-4.160	4.825	1.794	-4.102	-3.631	2.313	23.606
$\|L\|$ =250	-33.130	-51.960	-39.754	-26.278	-51.210	-78.668	-84.141	-40.518	-60.979	-79.595
$\|L\|$ =500	-1.683	-28.411	-87.318	-15.367	22.254	-95.138	-107.828	-12.183	-76.486	-19.242

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-3.213	9.063	3.360	17.527	51.147	43.463	32.630	60.730	64.730	80.177
$\|L\|$ =50	-18.113	-24.607	-25.010	-15.910	-16.877	-19.390	-20.020	7.713	68.167	80.433
$\|L\|$ =100	-11.450	-22.283	-23.717	-48.977	-46.320	60.593	-38.110	-19.837	-30.367	41.817
$\|L\|$ =250	-33.397	-87.613	-29.240	-102.233	-105.360	-107.473	-100.160	143.357	193.380	-20.280
$\|L\|$ =500	-61.883	-72.917	-117.070	-214.703	-197.393	-41.480	-290.970	10.350	128.780	95.453

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-1.458	7.003	-1.376	5.804	-2.723	5.721	4.048	3.010	-2.250	-0.306
$\|L\|$ =50	-2.774	-1.208	-8.914	5.671	-5.381	-5.161	3.360	17.616	-0.217	-0.637
$\|L\|$ =100	7.660	-4.169	11.807	0.704	7.037	6.411	17.813	-0.903	-2.219	23.416
$\|L\|$ =250	-21.082	-36.233	11.164	-11.733	-11.308	-6.844	-24.869	75.630	3.727	-5.273
$\|L\|$ =500	6.053	-13.936	-24.626	3.672	48.918	-23.307	-56.980	65.386	-94.256	60.086

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.22.: Mean error of average visits under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.097	-0.188	-0.298	-0.342	-0.421	-0.464	-0.511	-0.569	-0.648	-0.650
$\|L\|$ =50	-0.109	-0.208	-0.346	-0.364	-0.461	-0.556	-0.633	-0.657	-0.752	-0.827
$\|L\|$ =100	-0.111	-0.280	-0.375	-0.521	-0.589	-0.714	-0.830	-0.943	-0.992	-1.041
$\|L\|$ =250	-0.259	-0.499	-0.628	-0.753	-0.933	-1.160	-1.328	-1.343	-1.534	-1.732
$\|L\|$ =500	-0.238	-0.523	-0.892	-0.927	-1.028	-1.548	-1.751	-1.683	-2.038	-2.173

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.002	0.062	0.079	0.194	0.493	0.457	0.417	0.650	0.680	0.836
$\|L\|$ =50	-0.058	-0.070	-0.043	0.042	0.066	0.083	0.085	0.260	0.623	0.705
$\|L\|$ =100	-0.026	-0.055	-0.064	-0.138	-0.130	0.271	-0.074	0.027	0.038	0.270
$\|L\|$ =250	-0.175	-0.473	-0.355	-0.709	-0.693	-0.797	-0.848	-0.188	-0.113	-0.710
$\|L\|$ =500	-0.448	-0.713	-1.002	-1.575	-1.667	-1.522	-2.356	-1.773	-1.445	-1.805

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.098	0.238	0.332	0.507	0.607	0.844	1.024	1.236	1.345	1.549
$\|L\|$ =50	0.124	0.265	0.377	0.607	0.772	0.933	1.120	1.446	1.565	1.824
$\|L\|$ =100	0.213	0.371	0.614	0.829	1.020	1.252	1.585	1.860	2.137	2.543
$\|L\|$ =250	0.226	0.447	0.939	1.152	1.666	2.013	2.295	3.116	3.222	3.641
$\|L\|$ =500	0.441	0.857	1.331	1.959	2.678	3.030	3.260	4.288	4.524	5.522

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.23.: Mean error of entity coverage under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.017	0.037	0.050	0.068	0.083	0.096	0.112	0.131	0.155	0.165
$\|L\|$ =50	0.015	0.034	0.050	0.063	0.082	0.099	0.121	0.133	0.149	0.169
$\|L\|$ =100	0.013	0.027	0.043	0.057	0.069	0.083	0.096	0.112	0.121	0.137
$\|L\|$ =250	0.009	0.021	0.031	0.042	0.049	0.059	0.070	0.080	0.090	0.101
$\|L\|$ =500	0.008	0.015	0.022	0.031	0.038	0.045	0.052	0.060	0.066	0.078

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.002	0.000	-0.007	-0.008	-0.013	-0.015	-0.020	-0.023	-0.021	-0.026
$\|L\|$ =50	-0.004	-0.006	-0.010	-0.015	-0.018	-0.022	-0.022	-0.029	-0.036	-0.037
$\|L\|$ =100	-0.002	-0.003	-0.003	-0.004	-0.005	-0.005	-0.008	-0.011	-0.016	-0.012
$\|L\|$ =250	0.004	0.012	0.016	0.024	0.022	0.029	0.034	0.038	0.043	0.040
$\|L\|$ =500	0.009	0.017	0.023	0.034	0.039	0.050	0.056	0.065	0.065	0.076

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.013	-0.022	-0.039	-0.050	-0.065	-0.079	-0.092	-0.107	-0.117	-0.128
$\|L\|$ =50	-0.018	-0.034	-0.051	-0.067	-0.088	-0.103	-0.115	-0.134	-0.151	-0.168
$\|L\|$ =100	-0.018	-0.038	-0.054	-0.076	-0.089	-0.106	-0.126	-0.150	-0.168	-0.184
$\|L\|$ =250	-0.017	-0.032	-0.050	-0.065	-0.090	-0.105	-0.121	-0.139	-0.154	-0.171
$\|L\|$ =500	-0.014	-0.029	-0.045	-0.061	-0.077	-0.094	-0.103	-0.118	-0.138	-0.149

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.000	0.004	-0.001	-0.000	-0.002	-0.004	-0.011	-0.006	-0.001	0.018
$\|L\|$ =50	-0.001	0.000	-0.001	-0.003	-0.004	-0.001	0.007	-0.005	0.003	0.024
$\|L\|$ =100	-0.000	-0.001	0.002	0.001	0.002	0.001	0.003	0.001	-0.009	0.006
$\|L\|$ =250	-0.000	0.002	0.002	0.006	0.002	0.004	0.006	0.007	0.001	0.005
$\|L\|$ =500	0.001	0.001	0.002	0.003	0.003	0.002	0.003	0.002	0.001	0.008

Table C.24.: Mean error of gross visits under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	3.898	6.952	16.630	17.994	25.334	29.607	33.606	34.828	44.530	56.904
$\|L\|$ =50	2.737	4.709	11.488	16.464	19.430	18.573	17.349	27.318	21.102	42.568
$\|L\|$ =100	-1.267	3.823	-6.556	-13.320	-13.181	20.157	-13.600	-1.657	-29.707	7.176
$\|L\|$ =250	-4.472	-46.678	-38.185	-63.627	-20.650	-69.992	-81.785	-109.120	-61.459	-33.094
$\|L\|$ =500	-15.554	4.814	-68.931	0.846	26.391	-10.242	-43.080	-64.255	32.397	142.817

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	15.213	17.603	17.760	24.007	32.363	39.697	43.020	63.487	82.670	90.353
$\|L\|$ =50	-6.180	-12.080	-4.670	-17.733	2.350	41.230	57.700	56.260	11.523	41.510
$\|L\|$ =100	13.957	10.753	-25.120	7.367	-40.333	62.747	12.363	36.227	82.430	78.787
$\|L\|$ =250	-8.197	-106.353	-86.647	-80.137	-21.563	11.800	73.033	-97.797	127.843	38.220
$\|L\|$ =500	-72.340	-20.543	-226.037	-138.810	-229.790	36.647	-42.113	-183.093	-90.950	58.060

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	1.353	-1.561	3.071	3.150	5.626	2.303	3.601	-6.176	0.432	10.363
$\|L\|$ =50	0.603	-2.531	1.247	-1.548	7.171	2.111	-0.367	-6.648	-15.947	-4.697
$\|L\|$ =100	2.039	9.493	8.693	-2.386	-13.353	29.706	-10.044	12.793	-40.580	4.849
$\|L\|$ =250	7.644	-33.309	-15.973	-35.967	40.248	-6.161	8.028	-64.593	-11.510	-39.697
$\|L\|$ =500	0.612	58.499	-86.857	0.838	86.142	51.539	5.092	-23.822	17.880	43.041

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.25.: Mean error of average visits under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.102	-0.210	-0.264	-0.355	-0.407	-0.488	-0.535	-0.591	-0.647	-0.651
$\|L\|$ =50	-0.109	-0.225	-0.277	-0.352	-0.451	-0.546	-0.631	-0.698	-0.768	-0.811
$\|L\|$ =100	-0.149	-0.230	-0.393	-0.543	-0.642	-0.685	-0.849	-0.895	-1.107	-1.102
$\|L\|$ =250	-0.174	-0.460	-0.576	-0.827	-0.872	-1.166	-1.335	-1.513	-1.534	-1.607
$\|L\|$ =500	-0.276	-0.383	-0.872	-0.915	-1.002	-1.273	-1.675	-1.850	-1.802	-1.725

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.137	0.194	0.199	0.266	0.357	0.488	0.535	0.759	0.891	1.015
$\|L\|$ =50	0.008	0.027	0.125	0.112	0.245	0.475	0.630	0.690	0.538	0.673
$\|L\|$ =100	0.066	0.114	0.022	0.157	0.048	0.401	0.272	0.394	0.629	0.650
$\|L\|$ =250	-0.106	-0.464	-0.440	-0.495	-0.415	-0.334	-0.208	-0.756	-0.168	-0.487
$\|L\|$ =500	-0.479	-0.524	-1.394	-1.430	-1.805	-1.180	-1.768	-2.321	-2.302	-2.041

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.115	0.213	0.341	0.500	0.666	0.792	0.983	1.121	1.377	1.731
$\|L\|$ =50	0.132	0.268	0.451	0.606	0.792	0.978	1.165	1.363	1.573	1.855
$\|L\|$ =100	0.186	0.425	0.609	0.796	0.991	1.407	1.499	1.902	2.010	2.463
$\|L\|$ =250	0.301	0.511	0.893	1.208	1.718	1.905	2.380	2.609	3.106	3.402
$\|L\|$ =500	0.441	1.086	1.230	1.934	2.774	3.123	3.594	4.052	4.872	5.539

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.26.: Mean error of entity coverage under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.016	0.035	0.051	0.067	0.084	0.103	0.117	0.131	0.154	0.168
$\|L\|$ =50	0.017	0.034	0.047	0.062	0.080	0.096	0.111	0.131	0.141	0.165
$\|L\|$ =100	0.015	0.026	0.040	0.055	0.068	0.087	0.094	0.106	0.123	0.140
$\|L\|$ =250	0.010	0.019	0.028	0.040	0.051	0.061	0.071	0.078	0.089	0.101
$\|L\|$ =500	0.008	0.014	0.024	0.032	0.038	0.044	0.056	0.060	0.069	0.078

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.004	-0.006	-0.007	-0.012	-0.015	-0.021	-0.024	-0.031	-0.030	-0.036
$\|L\|$ =50	-0.005	-0.011	-0.018	-0.025	-0.029	-0.036	-0.041	-0.047	-0.057	-0.055
$\|L\|$ =100	-0.002	-0.008	-0.012	-0.013	-0.020	-0.017	-0.024	-0.026	-0.029	-0.035
$\|L\|$ =250	0.005	0.008	0.010	0.015	0.021	0.023	0.026	0.028	0.033	0.038
$\|L\|$ =500	0.009	0.016	0.026	0.036	0.041	0.045	0.059	0.066	0.075	0.081

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.013	-0.026	-0.036	-0.051	-0.064	-0.077	-0.090	-0.106	-0.117	-0.132
$\|L\|$ =50	-0.017	-0.035	-0.053	-0.071	-0.084	-0.104	-0.120	-0.139	-0.158	-0.172
$\|L\|$ =100	-0.018	-0.038	-0.055	-0.074	-0.093	-0.111	-0.130	-0.149	-0.171	-0.184
$\|L\|$ =250	-0.016	-0.035	-0.052	-0.072	-0.084	-0.100	-0.119	-0.140	-0.152	-0.168
$\|L\|$ =500	-0.015	-0.030	-0.048	-0.060	-0.077	-0.089	-0.106	-0.120	-0.137	-0.150

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.001	0.000	0.001	0.000	-0.001	0.000	-0.001	-0.008	0.005	0.029
$\|L\|$ =50	0.000	-0.000	-0.002	-0.005	-0.002	-0.002	-0.007	0.002	-0.008	0.017
$\|L\|$ =100	0.001	-0.001	0.000	0.001	-0.002	0.005	-0.002	-0.002	-0.001	0.013
$\|L\|$ =250	0.001	0.000	0.000	0.003	0.004	0.009	0.004	0.003	0.002	0.010
$\|L\|$ =500	0.001	-0.000	0.002	0.005	0.001	0.005	0.006	0.004	0.007	0.012

Table C.27.: Mean error of gross visits under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	6.004	11.703	16.778	20.480	32.769	35.832	36.920	39.112	39.749	67.342
$\|L\|$ =50	1.260	7.946	13.571	13.314	25.245	26.639	28.489	31.634	64.601	62.504
$\|L\|$ =100	-2.677	-3.039	-10.791	1.170	7.009	22.843	-32.665	7.676	19.689	99.892
$\|L\|$ =250	-18.661	-34.685	-16.670	-48.073	-40.049	-49.812	3740.923	25.272	534118.077	1485.890
$\|L\|$ =500	18.063	-26.441	-36.682	0.817	-21.329	27018.021	941.362	104235.518	3120.050	152639.501

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	15.837	5.057	8.727	1.203	13.117	16.693	36.947	45.420	14.450	85.253
$\|L\|$ =50	-1.570	5.120	-11.980	-32.343	32.990	16.580	19.320	-27.763	16.077	52.257
$\|L\|$ =100	-19.927	-29.450	5.013	-61.923	-41.923	15.777	-73.793	-8.187	32.573	45.263
$\|L\|$ =250	-20.013	-113.377	-35.183	22.860	-134.293	-36.723	68.613	84.523	-93.137	-151.337
$\|L\|$ =500	-65.737	-159.310	-115.480	-258.953	-339.050	-258.570	-152.457	-132.967	-160.780	-40.000

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	1.633	2.654	4.276	-0.820	5.236	3.338	-4.756	-5.449	-7.417	8.908
$\|L\|$ =50	0.252	4.698	4.786	-7.782	11.653	0.564	-4.564	-5.860	-3.644	10.518
$\|L\|$ =100	-0.004	6.354	4.144	8.052	-14.257	16.050	-28.306	17.270	-0.627	11.973
$\|L\|$ =250	-0.564	-34.801	22.317	19.342	-16.123	7.972	-17.501	32.881	-49.073	-45.106
$\|L\|$ =500	2.398	-15.103	14.756	28.820	-31.523	11.266	-12.867	115.081	-64.826	-17.661

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.28.: Mean error of average visits under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	-0.093	-0.194	-0.257	-0.331	-0.377	-0.462	-0.523	-0.608	-0.654	-0.611
$\lvert L\rvert$ =50	-0.108	-0.195	-0.292	-0.405	-0.433	-0.526	-0.609	-0.675	-0.666	-0.724
$\lvert L\rvert$ =100	-0.124	-0.262	-0.436	-0.492	-0.593	-0.647	-0.896	-0.894	-0.962	-0.825
$\lvert L\rvert$ =250	-0.210	-0.424	-0.547	-0.776	-0.924	-1.097	9.619	-1.155	1498.319	2.661
$\lvert L\rvert$ =500	-0.208	-0.537	-0.822	-0.869	-1.105	73.745	1.127	287.576	6.580	412.166

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.140	0.068	0.135	0.141	0.231	0.268	0.479	0.579	0.437	0.927
$\lvert L\rvert$ =50	0.028	0.098	0.039	-0.007	0.333	0.319	0.387	0.175	0.433	0.655
$\lvert L\rvert$ =100	-0.045	-0.058	0.082	-0.151	-0.020	0.208	-0.063	0.176	0.354	0.334
$\lvert L\rvert$ =250	-0.123	-0.494	-0.339	-0.224	-0.762	-0.544	-0.159	-0.286	-0.815	-1.020
$\lvert L\rvert$ =500	-0.478	-0.970	-1.120	-1.735	-2.073	-2.159	-2.097	-2.211	-2.451	-2.245

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.120	0.227	0.368	0.503	0.667	0.780	0.959	1.079	1.373	1.655
$\lvert L\rvert$ =50	0.129	0.296	0.439	0.552	0.818	0.966	1.128	1.302	1.538	1.940
$\lvert L\rvert$ =100	0.184	0.397	0.607	0.793	1.045	1.387	1.492	1.951	2.176	2.489
$\lvert L\rvert$ =250	0.285	0.556	1.038	1.319	1.570	1.930	2.481	2.931	3.099	3.647
$\lvert L\rvert$ =500	0.452	0.870	1.364	2.156	2.291	2.888	3.515	4.475	4.646	5.448

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ = 25	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ = 50	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =100	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =250	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =500	–	–	–	–	–	–	–	–	–	–

Table C.29.: Mean error of entity coverage under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.018	0.037	0.051	0.067	0.086	0.105	0.118	0.139	0.150	0.172
$\|L\|$ =50	0.016	0.032	0.051	0.068	0.081	0.098	0.115	0.130	0.152	0.162
$\|L\|$ =100	0.012	0.027	0.044	0.055	0.070	0.083	0.092	0.110	0.125	0.140
$\|L\|$ =250	0.009	0.019	0.030	0.040	0.051	0.061	0.064	0.080	0.093	0.098
$\|L\|$ =500	0.009	0.016	0.025	0.030	0.037	0.046	0.052	0.062	0.068	0.076

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.003	-0.004	-0.009	-0.015	-0.015	-0.017	-0.024	-0.026	-0.037	-0.031
$\|L\|$ =50	-0.004	-0.010	-0.012	-0.019	-0.022	-0.031	-0.035	-0.039	-0.044	-0.049
$\|L\|$ =100	-0.003	-0.005	-0.007	-0.009	-0.014	-0.015	-0.022	-0.022	-0.022	-0.017
$\|L\|$ =250	0.003	0.008	0.014	0.018	0.021	0.026	0.022	0.033	0.033	0.035
$\|L\|$ =500	0.010	0.018	0.028	0.035	0.040	0.052	0.060	0.067	0.074	0.079

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.013	-0.024	-0.038	-0.054	-0.065	-0.076	-0.094	-0.102	-0.122	-0.127
$\|L\|$ =50	-0.016	-0.034	-0.050	-0.068	-0.084	-0.103	-0.119	-0.134	-0.149	-0.171
$\|L\|$ =100	-0.019	-0.036	-0.056	-0.071	-0.098	-0.113	-0.134	-0.150	-0.170	-0.184
$\|L\|$ =250	-0.017	-0.038	-0.053	-0.068	-0.086	-0.099	-0.127	-0.138	-0.157	-0.177
$\|L\|$ =500	-0.015	-0.030	-0.042	-0.064	-0.074	-0.087	-0.106	-0.119	-0.138	-0.153

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.000	0.001	-0.000	-0.003	-0.002	-0.001	-0.001	0.005	-0.009	0.038
$\|L\|$ =50	-0.000	-0.002	0.002	0.002	-0.002	-0.003	-0.001	-0.002	-0.001	0.024
$\|L\|$ =100	-0.001	-0.000	0.001	0.002	-0.003	0.000	-0.004	-0.003	0.002	0.018
$\|L\|$ =250	0.000	-0.001	0.001	0.001	0.005	0.006	-0.006	0.002	0.010	0.006
$\|L\|$ =500	0.001	0.002	0.004	-0.000	0.002	0.005	0.005	0.008	0.005	0.012

Table C.30.: Mean error of gross visits under CDMAR on gender (female) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	7.866	8.058	15.368	24.518	22.007	24.392	24.836	32.525	33.623	34.380
$\|L\|$ =50	-0.037	7.232	8.082	12.051	16.166	17.482	24.951	12.941	39.467	36.380
$\|L\|$ =100	-9.324	-17.052	-18.694	-23.024	-28.632	-7.449	-12.553	-12.432	10.743	-6.418
$\|L\|$ =250	-74.888	-86.439	-53.928	-78.411	-33.276	-55.261	-76.944	-80.699	-79.604	-85.472
$\|L\|$ =500	-139.225	-124.080	-82.574	-29.920	-84.428	-101.501	10.353	-86.546	36.576	7.420

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	7.313	17.537	19.210	46.113	26.947	28.047	21.040	39.697	65.163	107.657
$\|L\|$ =50	-31.750	-24.277	6.933	-5.233	3.500	38.310	30.400	80.270	68.460	91.240
$\|L\|$ =100	-16.937	-77.883	-57.527	1.910	-51.623	148.777	73.597	-12.527	15.337	83.643
$\|L\|$ =250	-64.950	-156.220	-188.297	49.670	-25.183	-111.147	-340.460	135.847	480.720	236.157
$\|L\|$ =500	-218.543	-104.130	-246.343	-243.523	255.727	-221.007	42.723	76.857	46.663	557.537

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.488	-2.654	-0.322	7.846	2.973	-0.081	-7.841	-1.818	2.514	-0.759
$\|L\|$ =50	3.747	4.470	10.011	-5.811	-2.134	-2.868	1.741	-8.238	12.928	5.317
$\|L\|$ =100	11.294	-8.391	0.727	-17.793	-8.702	1.150	-8.121	-19.469	0.038	-22.178
$\|L\|$ =250	-36.330	-21.508	-5.870	7.930	30.404	3.518	-35.371	-27.711	-19.086	-8.469
$\|L\|$ =500	-20.811	-31.584	-0.263	52.810	23.228	-65.353	47.623	-8.077	64.623	-1.413

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.31.: Mean error of average visits under CDMAR on gender (female) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.269	-0.334	-0.363	-0.360	-0.425	-0.461	-0.490	-0.514	-0.553	-0.564
$\|L\|$ =50	-0.309	-0.317	-0.375	-0.423	-0.475	-0.517	-0.553	-0.636	-0.593	-0.657
$\|L\|$ =100	-0.368	-0.469	-0.526	-0.600	-0.706	-0.690	-0.788	-0.829	-0.799	-0.923
$\|L\|$ =250	-0.655	-0.786	-0.766	-0.949	-0.856	-1.053	-1.178	-1.274	-1.327	-1.411
$\|L\|$ =500	-0.997	-1.015	-1.077	-1.074	-1.302	-1.515	-1.330	-1.686	-1.411	-1.667

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.091	0.166	0.168	0.380	0.242	0.276	0.254	0.364	0.505	0.873
$\|L\|$ =50	-0.134	-0.067	0.081	0.040	0.059	0.267	0.175	0.445	0.360	0.469
$\|L\|$ =100	-0.138	-0.363	-0.312	-0.081	-0.332	0.393	0.089	-0.206	-0.160	0.079
$\|L\|$ =250	-0.599	-0.980	-1.147	-0.509	-0.808	-1.191	-1.892	-0.600	0.295	-0.462
$\|L\|$ =500	-1.560	-1.365	-2.000	-2.223	-0.961	-2.476	-1.968	-2.016	-2.257	-1.103

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.309	0.362	0.433	0.579	0.638	0.705	0.789	0.885	1.040	1.171
$\|L\|$ =50	0.385	0.497	0.617	0.670	0.751	0.889	0.942	1.071	1.251	1.348
$\|L\|$ =100	0.563	0.617	0.748	0.855	0.955	1.191	1.244	1.468	1.662	1.730
$\|L\|$ =250	0.708	0.912	1.215	1.410	1.753	1.880	1.951	2.144	2.503	2.818
$\|L\|$ =500	1.154	1.456	1.834	2.247	2.485	2.563	3.267	3.361	3.822	4.087

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.32.: Mean error of entity coverage under CDMAR on gender (female) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.044	0.054	0.067	0.075	0.084	0.092	0.099	0.111	0.121	0.125
$\|L\|$ =50	0.044	0.050	0.060	0.071	0.082	0.090	0.102	0.109	0.120	0.130
$\|L\|$ =100	0.036	0.044	0.051	0.058	0.069	0.076	0.086	0.093	0.098	0.108
$\|L\|$ =250	0.026	0.032	0.037	0.045	0.048	0.056	0.061	0.067	0.071	0.076
$\|L\|$ =500	0.021	0.023	0.029	0.035	0.038	0.044	0.048	0.052	0.054	0.061

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.004	-0.005	-0.003	-0.003	-0.007	-0.009	-0.010	-0.009	-0.010	-0.012
$\|L\|$ =50	-0.003	-0.006	-0.005	-0.006	-0.007	-0.010	-0.005	-0.009	-0.006	-0.006
$\|L\|$ =100	0.008	0.008	0.011	0.009	0.015	0.014	0.018	0.016	0.021	0.022
$\|L\|$ =250	0.024	0.031	0.036	0.041	0.046	0.054	0.056	0.063	0.068	0.073
$\|L\|$ =500	0.033	0.037	0.047	0.056	0.059	0.068	0.076	0.082	0.089	0.097

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.035	-0.043	-0.047	-0.055	-0.064	-0.073	-0.083	-0.087	-0.095	-0.105
$\|L\|$ =50	-0.044	-0.056	-0.065	-0.079	-0.085	-0.098	-0.101	-0.116	-0.121	-0.132
$\|L\|$ =100	-0.050	-0.062	-0.069	-0.084	-0.089	-0.104	-0.111	-0.130	-0.137	-0.148
$\|L\|$ =250	-0.046	-0.055	-0.067	-0.075	-0.087	-0.098	-0.107	-0.115	-0.129	-0.140
$\|L\|$ =500	-0.039	-0.049	-0.058	-0.065	-0.074	-0.084	-0.094	-0.101	-0.107	-0.119

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.001	-0.001	0.003	0.002	-0.001	-0.005	-0.005	-0.003	-0.005	-0.011
$\|L\|$ =50	0.001	-0.003	-0.001	-0.003	0.001	-0.003	-0.001	-0.005	-0.003	-0.005
$\|L\|$ =100	0.001	-0.000	0.001	-0.003	0.002	-0.001	0.002	-0.003	-0.003	-0.008
$\|L\|$ =250	0.002	0.003	0.002	0.002	0.001	0.004	-0.000	0.002	-0.003	-0.005
$\|L\|$ =500	0.002	0.002	0.003	0.005	0.002	0.004	0.004	0.005	0.003	0.004

Table C.33.: Mean error of gross visits under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	17.725	17.490	18.064	20.757	23.778	21.559	20.119	33.218	26.459	23.437
$\|L\|$ =50	25.417	16.236	15.526	12.894	24.337	26.063	16.762	20.300	5.429	-0.707
$\|L\|$ =100	18.052	6.183	9.342	11.280	-2.338	-12.706	-46.099	-14.476	-53.109	-86.683
$\|L\|$ =250	17.271	5.894	-19.699	-62.013	-55.696	-75.777	-91.195	-170.725	-136.468	-170.409
$\|L\|$ =500	90.985	57.287	-66.349	-80.300	-132.162	-151.024	-199.471	-259.388	-241.012	-289.814

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	8.407	2.330	12.713	21.727	41.953	47.603	84.697	55.350	58.467	151.877
$\|L\|$ =50	-18.813	-9.393	-13.130	23.243	17.527	24.530	34.390	-2.643	166.673	98.860
$\|L\|$ =100	-52.813	-83.343	-14.280	-88.657	-4.323	22.717	2.653	55.870	30.667	135.303
$\|L\|$ =250	-177.747	-231.263	-221.333	67.083	-4.263	-109.590	-59.337	68.640	103.663	602.957
$\|L\|$ =500	-332.880	-374.713	-195.543	-139.503	-247.690	-25.350	585.957	656.370	516.027	283.010

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	1.259	-0.308	-0.580	0.562	2.909	0.066	-5.528	7.500	7.031	-4.036
$\|L\|$ =50	4.190	-4.047	5.933	1.552	14.443	10.057	13.030	12.050	13.049	-1.358
$\|L\|$ =100	1.048	-9.353	4.847	6.942	9.941	0.906	-19.508	2.984	-26.829	-22.562
$\|L\|$ =250	1.244	-7.453	-11.830	-8.241	10.608	-14.293	50.262	-50.437	21.383	43.478
$\|L\|$ =500	27.914	27.768	-83.847	9.029	-98.314	-32.400	54.449	-11.740	25.878	-29.327

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.34.: Mean error of average visits under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.232	-0.276	-0.349	-0.372	-0.416	-0.453	-0.486	-0.507	-0.566	-0.613
$\|L\|$ =50	-0.237	-0.304	-0.362	-0.427	-0.460	-0.499	-0.563	-0.614	-0.671	-0.741
$\|L\|$ =100	-0.357	-0.430	-0.464	-0.527	-0.610	-0.713	-0.840	-0.800	-0.927	-1.093
$\|L\|$ =250	-0.495	-0.612	-0.714	-0.934	-0.927	-1.056	-1.162	-1.462	-1.405	-1.539
$\|L\|$ =500	-0.572	-0.734	-1.141	-1.185	-1.443	-1.513	-1.802	-1.971	-2.022	-2.261

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.080	0.062	0.131	0.221	0.365	0.454	0.777	0.473	0.519	1.213
$\|L\|$ =50	-0.085	0.015	-0.022	0.210	0.118	0.137	0.247	0.004	0.935	0.584
$\|L\|$ =100	-0.331	-0.420	-0.175	-0.471	-0.154	-0.052	-0.095	0.039	0.014	0.402
$\|L\|$ =250	-1.080	-1.312	-1.304	-0.520	-0.739	-1.089	-1.009	-0.676	-0.630	0.790
$\|L\|$ =500	-2.185	-2.400	-2.055	-1.882	-2.374	-1.792	-0.317	-0.206	-0.691	-1.510

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.321	0.396	0.450	0.560	0.647	0.787	0.814	0.934	1.031	1.109
$\|L\|$ =50	0.411	0.477	0.617	0.663	0.816	0.894	1.012	1.064	1.288	1.321
$\|L\|$ =100	0.513	0.627	0.823	0.923	1.056	1.149	1.200	1.473	1.583	1.663
$\|L\|$ =250	0.876	1.068	1.182	1.327	1.653	1.740	2.156	2.079	2.518	2.774
$\|L\|$ =500	1.416	1.615	1.776	2.222	2.076	2.776	3.067	3.197	3.702	3.730

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.35.: Mean error of entity coverage under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.048	0.054	0.067	0.074	0.084	0.089	0.094	0.111	0.116	0.123
$\|L\|$ =50	0.050	0.054	0.063	0.071	0.084	0.093	0.098	0.110	0.110	0.118
$\|L\|$ =100	0.046	0.050	0.055	0.064	0.068	0.077	0.079	0.087	0.088	0.095
$\|L\|$ =250	0.034	0.039	0.041	0.047	0.048	0.053	0.057	0.062	0.065	0.067
$\|L\|$ =500	0.029	0.031	0.033	0.034	0.038	0.038	0.044	0.044	0.048	0.052

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.001	-0.005	-0.003	-0.006	-0.007	-0.013	-0.012	-0.006	-0.010	-0.013
$\|L\|$ =50	-0.000	-0.005	-0.005	-0.005	-0.006	-0.004	-0.010	-0.004	-0.013	-0.016
$\|L\|$ =100	0.015	0.012	0.013	0.015	0.014	0.014	0.011	0.015	0.009	0.009
$\|L\|$ =250	0.033	0.038	0.040	0.045	0.045	0.048	0.053	0.054	0.059	0.062
$\|L\|$ =500	0.045	0.048	0.054	0.054	0.061	0.062	0.070	0.072	0.076	0.083

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.036	-0.044	-0.049	-0.058	-0.065	-0.078	-0.083	-0.084	-0.091	-0.103
$\|L\|$ =50	-0.047	-0.059	-0.067	-0.074	-0.084	-0.092	-0.101	-0.106	-0.123	-0.134
$\|L\|$ =100	-0.049	-0.062	-0.073	-0.081	-0.091	-0.101	-0.111	-0.124	-0.140	-0.144
$\|L\|$ =250	-0.049	-0.060	-0.067	-0.073	-0.086	-0.094	-0.102	-0.115	-0.123	-0.130
$\|L\|$ =500	-0.042	-0.049	-0.064	-0.068	-0.074	-0.088	-0.088	-0.097	-0.107	-0.112

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.001	-0.001	0.002	-0.000	-0.001	-0.009	-0.007	-0.002	-0.008	-0.012
$\|L\|$ =50	0.003	-0.001	0.000	-0.001	0.001	0.002	-0.003	-0.001	-0.014	-0.014
$\|L\|$ =100	0.008	0.003	0.001	0.004	0.001	-0.001	-0.004	-0.005	-0.018	-0.017
$\|L\|$ =250	0.006	0.007	0.004	0.007	0.002	-0.001	-0.002	-0.002	-0.008	-0.009
$\|L\|$ =500	0.008	0.007	0.004	0.002	0.003	-0.002	0.004	-0.003	-0.003	-0.002

Table C.36.: Mean error of gross visits under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	15.836	17.860	16.955	21.972	25.558	23.635	25.374	26.032	24.282	31.194
$\|L\|$ =50	15.241	24.393	13.622	21.002	8.905	25.065	13.279	21.494	7.392	13.550
$\|L\|$ =100	7.710	0.682	0.872	5.962	4.954	-4.931	-33.504	-38.260	-50.192	-47.980
$\|L\|$ =250	19.124	-30.117	-50.472	-25.056	-22.245	-93.454	-76.705	-50.443	-101.203	-87.520
$\|L\|$ =500	37.948	-57.817	-52.554	-56.874	-85.454	-2.625	-99.018	-141.062	-132.288	-68.148

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.570	2.447	8.663	26.187	13.707	36.530	32.277	100.950	4.093	35.840
$\|L\|$ =50	-23.960	-17.193	-12.033	-20.903	-5.660	-1.117	22.553	-2.213	37.510	-0.203
$\|L\|$ =100	-26.670	-35.483	-47.470	-17.353	-49.853	5.153	22.640	-7.353	40.740	79.460
$\|L\|$ =250	-31.667	-64.840	-101.580	-87.200	-7.080	38.860	-111.880	26.173	-32.350	7.500
$\|L\|$ =500	-141.710	-312.117	-285.387	-180.707	-327.470	62.850	-19.100	-245.097	-249.193	210.443

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.530	-0.950	-3.089	2.043	-3.418	-4.446	1.069	2.544	-7.353	2.830
$\|L\|$ =50	-1.133	4.859	-2.956	-3.969	-3.527	10.836	-0.142	16.217	1.534	-3.211
$\|L\|$ =100	0.689	2.076	-15.113	3.950	7.341	17.049	-24.087	-7.793	-10.631	3.179
$\|L\|$ =250	38.881	9.307	-37.737	-0.322	1.413	-21.721	25.812	39.424	-16.379	38.361
$\|L\|$ =500	64.831	-45.526	-15.580	-24.390	-59.249	92.754	0.147	-56.586	-44.969	120.529

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.37.: Mean error of average visits under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.247	-0.293	-0.348	-0.363	-0.390	-0.470	-0.504	-0.557	-0.588	-0.588
$\|L\|$ =50	-0.264	-0.303	-0.399	-0.419	-0.487	-0.497	-0.588	-0.607	-0.693	-0.698
$\|L\|$ =100	-0.356	-0.437	-0.501	-0.527	-0.617	-0.685	-0.826	-0.875	-0.947	-0.998
$\|L\|$ =250	-0.525	-0.691	-0.802	-0.818	-0.889	-1.093	-1.126	-1.126	-1.319	-1.344
$\|L\|$ =500	-0.633	-1.030	-1.068	-1.200	-1.246	-1.143	-1.584	-1.628	-1.731	-1.687

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.038	0.074	0.132	0.273	0.236	0.377	0.372	0.883	0.180	0.434
$\|L\|$ =50	-0.012	-0.014	0.052	0.010	0.141	0.143	0.311	0.149	0.402	0.226
$\|L\|$ =100	-0.036	-0.069	-0.112	0.024	-0.142	0.062	0.159	0.056	0.284	0.361
$\|L\|$ =250	-0.315	-0.495	-0.597	-0.628	-0.485	-0.262	-0.821	-0.492	-0.707	-0.604
$\|L\|$ =500	-1.092	-1.717	-1.719	-1.626	-2.003	-1.037	-1.507	-2.161	-2.291	-1.217

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.295	0.365	0.440	0.583	0.645	0.689	0.806	0.934	0.934	1.175
$\|L\|$ =50	0.401	0.474	0.544	0.630	0.776	0.899	0.981	1.126	1.225	1.356
$\|L\|$ =100	0.555	0.679	0.732	0.968	1.021	1.183	1.231	1.403	1.577	1.783
$\|L\|$ =250	0.966	1.071	1.165	1.449	1.672	1.757	2.107	2.316	2.439	2.814
$\|L\|$ =500	1.553	1.550	1.833	2.078	2.251	2.893	2.891	3.165	3.552	4.245

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.38.: Mean error of entity coverage under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.048	0.058	0.066	0.074	0.082	0.093	0.103	0.113	0.118	0.126
$\|L\|$ =50	0.047	0.059	0.068	0.076	0.079	0.092	0.100	0.109	0.115	0.121
$\|L\|$ =100	0.042	0.048	0.056	0.061	0.072	0.076	0.082	0.087	0.091	0.099
$\|L\|$ =250	0.036	0.037	0.041	0.047	0.052	0.052	0.057	0.063	0.066	0.070
$\|L\|$ =500	0.026	0.030	0.032	0.037	0.035	0.040	0.046	0.044	0.049	0.054

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.006	-0.006	-0.009	-0.012	-0.014	-0.014	-0.014	-0.015	-0.018	-0.019
$\|L\|$ =50	-0.013	-0.009	-0.013	-0.015	-0.021	-0.019	-0.023	-0.021	-0.028	-0.029
$\|L\|$ =100	-0.007	-0.007	-0.007	-0.009	-0.005	-0.005	-0.008	-0.010	-0.013	-0.008
$\|L\|$ =250	0.013	0.018	0.018	0.022	0.029	0.024	0.030	0.035	0.037	0.039
$\|L\|$ =500	0.024	0.029	0.032	0.040	0.038	0.043	0.052	0.053	0.058	0.065

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.035	-0.041	-0.050	-0.060	-0.068	-0.074	-0.079	-0.087	-0.094	-0.103
$\|L\|$ =50	-0.049	-0.054	-0.065	-0.075	-0.088	-0.092	-0.105	-0.110	-0.124	-0.136
$\|L\|$ =100	-0.053	-0.063	-0.073	-0.086	-0.089	-0.098	-0.115	-0.122	-0.135	-0.145
$\|L\|$ =250	-0.047	-0.057	-0.070	-0.078	-0.088	-0.096	-0.104	-0.111	-0.125	-0.132
$\|L\|$ =500	-0.043	-0.054	-0.059	-0.067	-0.075	-0.080	-0.088	-0.100	-0.109	-0.112

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.002	0.001	-0.004	-0.004	-0.007	-0.003	0.002	-0.001	-0.003	0.005
$\|L\|$ =50	-0.006	0.000	0.001	-0.001	-0.005	0.001	-0.000	0.007	-0.001	0.005
$\|L\|$ =100	-0.004	-0.005	-0.002	-0.004	0.002	0.005	0.005	0.003	-0.003	-0.001
$\|L\|$ =250	0.005	0.002	-0.000	0.002	0.005	0.002	0.004	0.005	0.008	0.001
$\|L\|$ =500	-0.000	0.001	0.001	0.004	0.000	0.004	0.007	0.003	0.010	0.001

Table C.39.: Mean error of gross visits under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	34.793	32.064	28.392	31.114	21.661	17.710	17.819	16.829	13.338	2.291
$\|L\|$ =50	44.995	34.755	37.582	33.295	21.218	13.516	-2.491	-12.939	-7.832	-24.593
$\|L\|$ =100	63.623	45.007	41.136	10.005	-25.056	-6.345	-41.836	-83.916	-101.335	-121.297
$\|L\|$ =250	101.980	62.216	37.888	-22.904	-50.804	-115.770	-134.745	-172.340	-300.938	-287.230
$\|L\|$ =500	238.520	115.962	70.431	-14.694	6.565	-178.202	-274.769	-346.395	-429.781	-425.417

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	13.550	11.770	21.107	21.300	12.413	33.760	53.790	84.607	65.583	83.477
$\|L\|$ =50	-34.340	-26.943	-18.787	14.157	30.597	18.357	37.623	125.817	128.617	207.987
$\|L\|$ =100	8.287	-107.623	-69.727	-16.540	-84.663	31.140	-20.103	127.400	143.187	162.990
$\|L\|$ =250	-117.423	-277.760	-29.320	-38.000	-87.760	-107.370	-6.440	162.857	337.413	719.900
$\|L\|$ =500	-279.617	-280.323	-110.703	-200.783	134.727	298.727	19.163	330.993	1155.693	1103.127

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	1.126	3.296	-2.401	3.443	-4.228	-0.854	-0.088	5.199	-0.624	-9.374
$\|L\|$ =50	0.323	-2.900	11.434	8.477	5.560	19.094	1.891	9.618	14.424	6.127
$\|L\|$ =100	7.407	-1.304	14.996	-1.170	-22.661	3.853	2.797	-21.249	-6.370	-17.558
$\|L\|$ =250	23.878	-16.840	18.202	4.048	17.323	5.660	40.748	39.921	-55.568	33.646
$\|L\|$ =500	36.639	-40.921	-6.692	-50.311	63.891	-47.747	-74.884	10.917	36.298	29.917

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.40.: Mean error of average visits under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.289	-0.317	-0.368	-0.371	-0.421	-0.446	-0.483	-0.487	-0.509	-0.539
$\|L\|$ =50	-0.308	-0.357	-0.365	-0.412	-0.444	-0.493	-0.555	-0.606	-0.602	-0.630
$\|L\|$ =100	-0.334	-0.435	-0.459	-0.590	-0.685	-0.621	-0.760	-0.872	-0.966	-1.044
$\|L\|$ =250	-0.431	-0.592	-0.643	-0.837	-0.942	-1.125	-1.213	-1.305	-1.664	-1.620
$\|L\|$ =500	-0.380	-0.706	-0.947	-1.084	-1.055	-1.662	-1.927	-2.086	-2.308	-2.346

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.049	0.062	0.187	0.168	0.162	0.333	0.481	0.769	0.642	0.887
$\|L\|$ =50	-0.200	-0.159	-0.095	0.054	0.183	0.139	0.259	0.750	0.734	1.196
$\|L\|$ =100	-0.148	-0.574	-0.432	-0.234	-0.418	0.002	-0.175	0.431	0.488	0.590
$\|L\|$ =250	-1.075	-1.601	-0.839	-0.859	-1.032	-1.022	-0.750	-0.270	0.359	1.480
$\|L\|$ =500	-2.430	-2.410	-1.999	-2.206	-1.326	-0.927	-1.764	-0.782	1.421	1.223

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.367	0.443	0.483	0.574	0.608	0.711	0.759	0.882	0.938	0.977
$\|L\|$ =50	0.488	0.533	0.671	0.696	0.773	0.909	0.891	1.055	1.113	1.192
$\|L\|$ =100	0.781	0.787	0.904	0.912	0.973	1.144	1.193	1.244	1.369	1.433
$\|L\|$ =250	1.329	1.265	1.554	1.469	1.689	1.675	1.904	1.988	1.884	2.273
$\|L\|$ =500	2.056	1.924	2.286	2.133	2.583	2.444	2.488	2.800	3.024	3.143

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.41.: Mean error of entity coverage under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.074	0.075	0.080	0.084	0.083	0.084	0.091	0.091	0.091	0.084
$\lvert L\rvert$ =50	0.074	0.075	0.078	0.082	0.080	0.083	0.083	0.085	0.087	0.080
$\lvert L\rvert$ =100	0.062	0.066	0.067	0.071	0.068	0.068	0.070	0.066	0.071	0.072
$\lvert L\rvert$ =250	0.046	0.049	0.047	0.048	0.050	0.049	0.051	0.050	0.049	0.049
$\lvert L\rvert$ =500	0.037	0.036	0.040	0.037	0.038	0.041	0.041	0.039	0.039	0.041

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.006	0.002	-0.002	-0.001	-0.008	-0.011	-0.011	-0.015	-0.017	-0.027
$\lvert L\rvert$ =50	0.005	0.004	0.001	0.001	-0.004	-0.008	-0.011	-0.016	-0.017	-0.025
$\lvert L\rvert$ =100	0.019	0.020	0.018	0.018	0.010	0.011	0.010	0.003	0.004	-0.000
$\lvert L\rvert$ =250	0.046	0.049	0.046	0.047	0.048	0.044	0.045	0.046	0.041	0.041
$\lvert L\rvert$ =500	0.060	0.059	0.061	0.060	0.061	0.063	0.066	0.061	0.062	0.065

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	-0.040	-0.046	-0.054	-0.058	-0.067	-0.072	-0.076	-0.082	-0.090	-0.099
$\lvert L\rvert$ =50	-0.058	-0.064	-0.070	-0.075	-0.083	-0.089	-0.096	-0.106	-0.109	-0.120
$\lvert L\rvert$ =100	-0.070	-0.073	-0.077	-0.083	-0.095	-0.099	-0.104	-0.115	-0.119	-0.127
$\lvert L\rvert$ =250	-0.068	-0.071	-0.080	-0.078	-0.086	-0.087	-0.093	-0.096	-0.107	-0.110
$\lvert L\rvert$ =500	-0.061	-0.064	-0.072	-0.071	-0.073	-0.080	-0.083	-0.084	-0.088	-0.092

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.012	0.007	0.004	0.002	-0.003	-0.007	-0.007	-0.012	-0.017	-0.028
$\lvert L\rvert$ =50	0.012	0.009	0.006	0.005	-0.001	-0.007	-0.005	-0.012	-0.017	-0.029
$\lvert L\rvert$ =100	0.008	0.008	0.007	0.007	-0.002	-0.006	-0.004	-0.016	-0.014	-0.018
$\lvert L\rvert$ =250	0.007	0.009	0.002	0.005	0.003	-0.000	0.001	-0.001	-0.008	-0.011
$\lvert L\rvert$ =500	0.008	0.005	0.007	0.001	0.003	0.004	0.004	0.000	-0.002	-0.001

Table C.42.: Mean error of gross visits under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	31.899	34.627	22.932	24.624	25.614	20.698	15.261	14.237	9.938	13.928
$\|L\|$ =50	45.233	44.653	37.862	23.407	19.610	9.821	6.434	-13.304	-21.289	-23.787
$\|L\|$ =100	59.920	44.751	14.772	16.098	-11.533	-38.700	-43.823	-62.302	-87.562	-92.983
$\|L\|$ =250	94.318	78.547	40.173	-22.760	-56.503	-100.456	-111.124	-203.298	-214.987	-265.130
$\|L\|$ =500	219.810	193.582	91.804	71.397	-32.766	-95.562	-242.898	-408.389	-327.489	-474.605

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	5.483	1.847	4.050	-1.103	1.980	47.800	99.807	17.907	43.890	72.227
$\|L\|$ =50	-22.057	10.227	-12.390	-31.553	-28.133	-18.577	-14.833	26.433	38.563	7.650
$\|L\|$ =100	-68.583	-74.247	-59.313	-9.073	-24.277	-55.530	-46.430	20.973	17.907	52.777
$\|L\|$ =250	-94.997	-138.117	-70.843	-232.550	-13.390	-111.107	-30.033	142.557	212.980	0.580
$\|L\|$ =500	-182.200	-288.843	-234.743	-323.613	-175.743	-137.007	31.583	75.380	121.020	91.307

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.648	-0.452	-6.194	-5.023	-2.511	4.373	-2.750	-1.966	-6.870	4.722
$\|L\|$ =50	-1.628	6.414	1.274	-6.714	-7.978	2.309	10.044	-2.239	0.973	0.281
$\|L\|$ =100	-6.478	-5.824	-14.759	21.486	-3.876	-15.530	-5.884	5.467	6.866	5.884
$\|L\|$ =250	-7.236	-11.232	9.521	-29.599	-25.941	-7.912	29.378	-30.974	80.254	-15.851
$\|L\|$ =500	7.501	-13.397	-51.642	16.396	-3.774	29.681	-9.839	-82.257	61.024	-19.433

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.43.: Mean error of average visits under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.317	-0.331	-0.392	-0.382	-0.411	-0.447	-0.484	-0.498	-0.511	-0.528
$\|L\|$ =50	-0.320	-0.326	-0.389	-0.442	-0.448	-0.496	-0.535	-0.589	-0.631	-0.658
$\|L\|$ =100	-0.388	-0.434	-0.513	-0.555	-0.637	-0.738	-0.778	-0.820	-0.886	-0.938
$\|L\|$ =250	-0.450	-0.501	-0.635	-0.846	-0.962	-1.111	-1.112	-1.448	-1.428	-1.565
$\|L\|$ =500	-0.443	-0.537	-0.828	-0.801	-1.180	-1.369	-1.786	-2.200	-2.044	-2.467

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.004	0.016	0.090	0.078	0.136	0.474	0.883	0.348	0.598	0.779
$\|L\|$ =50	-0.089	0.100	0.001	-0.056	0.015	0.094	0.133	0.417	0.494	0.354
$\|L\|$ =100	-0.311	-0.303	-0.187	-0.028	-0.025	-0.114	-0.032	0.263	0.309	0.462
$\|L\|$ =250	-0.735	-0.839	-0.662	-1.073	-0.449	-0.716	-0.393	0.040	0.414	-0.235
$\|L\|$ =500	-1.689	-2.011	-1.842	-2.015	-1.624	-1.566	-1.009	-0.870	-0.763	-0.948

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.341	0.406	0.468	0.553	0.613	0.730	0.755	0.855	0.914	1.078
$\|L\|$ =50	0.474	0.578	0.595	0.648	0.736	0.859	0.980	0.989	1.088	1.147
$\|L\|$ =100	0.673	0.745	0.826	0.985	1.006	1.043	1.196	1.404	1.412	1.521
$\|L\|$ =250	1.197	1.292	1.430	1.382	1.515	1.665	1.940	1.764	2.264	2.165
$\|L\|$ =500	1.933	2.085	2.104	2.478	2.574	2.710	2.686	2.615	3.083	3.005

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.44.: Mean error of entity coverage under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.075	0.080	0.080	0.079	0.085	0.087	0.089	0.090	0.087	0.095
$\|L\|$ =50	0.076	0.077	0.082	0.081	0.080	0.081	0.085	0.081	0.082	0.086
$\|L\|$ =100	0.067	0.066	0.063	0.069	0.068	0.069	0.072	0.069	0.067	0.071
$\|L\|$ =250	0.046	0.046	0.047	0.049	0.050	0.051	0.049	0.054	0.050	0.049
$\|L\|$ =500	0.037	0.038	0.038	0.035	0.038	0.039	0.039	0.037	0.040	0.040

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.003	-0.001	-0.007	-0.011	-0.013	-0.014	-0.017	-0.024	-0.027	-0.027
$\|L\|$ =50	-0.002	-0.006	-0.006	-0.012	-0.019	-0.023	-0.026	-0.032	-0.037	-0.039
$\|L\|$ =100	0.006	0.003	-0.004	-0.001	-0.006	-0.010	-0.014	-0.018	-0.025	-0.027
$\|L\|$ =250	0.028	0.026	0.027	0.022	0.025	0.023	0.018	0.023	0.014	0.014
$\|L\|$ =500	0.042	0.043	0.042	0.039	0.040	0.042	0.039	0.039	0.038	0.043

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.038	-0.046	-0.055	-0.062	-0.066	-0.070	-0.077	-0.085	-0.092	-0.096
$\|L\|$ =50	-0.057	-0.063	-0.068	-0.078	-0.086	-0.092	-0.100	-0.107	-0.114	-0.119
$\|L\|$ =100	-0.065	-0.071	-0.081	-0.081	-0.091	-0.098	-0.106	-0.118	-0.118	-0.126
$\|L\|$ =250	-0.066	-0.072	-0.075	-0.079	-0.085	-0.089	-0.096	-0.098	-0.102	-0.113
$\|L\|$ =500	-0.060	-0.066	-0.070	-0.075	-0.079	-0.080	-0.083	-0.087	-0.088	-0.093

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.013	0.009	0.003	-0.002	-0.001	-0.004	-0.008	-0.013	-0.018	-0.021
$\|L\|$ =50	0.012	0.009	0.009	0.003	-0.002	-0.007	-0.008	-0.016	-0.021	-0.020
$\|L\|$ =100	0.011	0.008	0.002	0.004	-0.001	-0.004	-0.006	-0.013	-0.014	-0.015
$\|L\|$ =250	0.005	0.004	0.005	0.004	0.004	0.003	-0.003	0.000	-0.003	-0.006
$\|L\|$ =500	0.007	0.006	0.005	0.000	0.001	0.002	0.001	-0.000	-0.000	0.001

Table C.45.: Mean error of gross visits under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	27.576	25.376	25.754	25.070	19.024	19.108	21.556	16.271	20.740	8.107
$\lvert L\rvert$ =50	34.380	29.529	21.439	26.450	17.692	12.143	17.439	-14.082	-3.640	-8.615
$\lvert L\rvert$ =100	21.356	24.486	26.714	2.971	-5.421	-4.830	-27.392	-46.843	-69.795	-58.897
$\lvert L\rvert$ =250	9.580	-22.408	-51.041	-31.918	-39.985	-67.326	-124.089	-152.980	-126.105	-172.219
$\lvert L\rvert$ =500	-10.515	4.102	-56.862	10.696	-87.224	-131.278	-126.785	-85.410	-120.638	-198.323

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	6.437	11.200	50.110	28.233	12.560	0.677	8.837	23.403	80.043	58.443
$\lvert L\rvert$ =50	-19.597	-13.373	-14.833	-19.653	-9.157	80.267	11.080	-25.417	8.273	79.003
$\lvert L\rvert$ =100	-50.063	5.103	-50.683	-2.040	-34.447	-6.540	-0.407	-26.017	7.430	-1.740
$\lvert L\rvert$ =250	-141.877	-181.640	-196.567	-101.553	-133.887	-121.280	26.563	91.040	90.360	243.593
$\lvert L\rvert$ =500	-150.670	-185.367	-331.440	-18.323	-311.343	-126.253	-124.720	14.293	224.390	71.343

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	3.006	-4.203	-2.343	-2.584	-5.677	-2.040	-7.450	-5.696	8.293	-3.508
$\lvert L\rvert$ =50	3.193	1.646	-5.152	2.689	-7.558	1.264	10.137	-19.757	-5.837	6.140
$\lvert L\rvert$ =100	-6.800	-0.116	11.411	0.102	-0.109	9.190	-8.442	-7.991	-19.764	-2.004
$\lvert L\rvert$ =250	-21.809	-31.409	-24.906	22.099	26.800	-1.533	-12.814	-21.086	-4.297	-20.389
$\lvert L\rvert$ =500	54.807	24.388	-63.494	55.529	-14.533	-66.542	-10.572	38.627	35.399	-57.977

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ = 25	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ = 50	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =100	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =250	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =500	–	–	–	–	–	–	–	–	–	–

Table C.46.: Mean error of average visits under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.300	-0.351	-0.358	-0.381	-0.435	-0.458	-0.458	-0.519	-0.515	-0.556
$\|L\|$ =50	-0.325	-0.339	-0.388	-0.414	-0.454	-0.497	-0.489	-0.629	-0.601	-0.625
$\|L\|$ =100	-0.470	-0.468	-0.482	-0.561	-0.605	-0.622	-0.720	-0.799	-0.906	-0.872
$\|L\|$ =250	-0.678	-0.787	-0.904	-0.900	-0.871	-0.995	-1.165	-1.255	-1.197	-1.379
$\|L\|$ =500	-0.976	-0.942	-1.182	-1.046	-1.360	-1.495	-1.410	-1.394	-1.485	-1.723

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.063	0.098	0.420	0.298	0.214	0.118	0.235	0.385	0.805	0.743
$\|L\|$ =50	-0.031	0.059	0.065	0.031	0.135	0.570	0.284	0.199	0.354	0.755
$\|L\|$ =100	-0.235	0.026	-0.177	0.052	-0.028	0.094	0.161	0.082	0.248	0.263
$\|L\|$ =250	-0.961	-1.049	-1.112	-0.848	-0.852	-0.789	-0.293	-0.085	-0.053	0.456
$\|L\|$ =500	-1.729	-1.770	-2.196	-1.359	-2.179	-1.651	-1.623	-1.311	-0.606	-1.016

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.379	0.376	0.490	0.556	0.605	0.654	0.716	0.836	0.949	1.024
$\|L\|$ =50	0.495	0.579	0.635	0.690	0.742	0.835	0.971	0.925	1.043	1.229
$\|L\|$ =100	0.636	0.790	0.913	0.938	1.066	1.165	1.189	1.313	1.316	1.443
$\|L\|$ =250	1.161	1.243	1.290	1.503	1.662	1.777	1.821	1.864	1.984	2.012
$\|L\|$ =500	2.178	2.197	2.062	2.562	2.268	2.421	2.763	2.924	3.021	2.968

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.47.: Mean error of entity coverage under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.069	0.074	0.076	0.080	0.083	0.087	0.089	0.096	0.101	0.095
$\|L\|$ =50	0.070	0.069	0.071	0.079	0.079	0.083	0.085	0.088	0.090	0.091
$\|L\|$ =100	0.061	0.062	0.065	0.064	0.066	0.068	0.072	0.073	0.077	0.078
$\|L\|$ =250	0.044	0.045	0.047	0.051	0.047	0.050	0.050	0.051	0.052	0.055
$\|L\|$ =500	0.033	0.033	0.036	0.038	0.039	0.040	0.037	0.041	0.041	0.041

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.002	-0.001	-0.006	-0.012	-0.012	-0.015	-0.020	-0.022	-0.021	-0.032
$\|L\|$ =50	-0.008	-0.015	-0.018	-0.016	-0.023	-0.023	-0.028	-0.039	-0.039	-0.043
$\|L\|$ =100	0.005	-0.001	-0.001	-0.006	-0.010	-0.013	-0.016	-0.019	-0.023	-0.026
$\|L\|$ =250	0.033	0.031	0.032	0.033	0.027	0.026	0.023	0.022	0.020	0.018
$\|L\|$ =500	0.047	0.044	0.045	0.047	0.047	0.047	0.045	0.048	0.044	0.043

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.040	-0.046	-0.054	-0.061	-0.066	-0.070	-0.078	-0.085	-0.084	-0.098
$\|L\|$ =50	-0.057	-0.066	-0.076	-0.077	-0.088	-0.091	-0.099	-0.110	-0.113	-0.123
$\|L\|$ =100	-0.063	-0.073	-0.079	-0.085	-0.094	-0.100	-0.107	-0.116	-0.120	-0.123
$\|L\|$ =250	-0.067	-0.073	-0.074	-0.077	-0.084	-0.093	-0.098	-0.100	-0.104	-0.107
$\|L\|$ =500	-0.063	-0.066	-0.070	-0.074	-0.072	-0.081	-0.085	-0.085	-0.088	-0.095

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.000	0.001	-0.002	-0.001	-0.001	-0.000	-0.004	-0.002	-0.001	-0.002
$\|L\|$ =50	-0.001	-0.005	-0.006	0.000	-0.003	0.000	-0.003	-0.004	-0.004	-0.004
$\|L\|$ =100	0.005	0.001	0.001	-0.001	-0.003	-0.001	-0.003	-0.001	0.004	-0.000
$\|L\|$ =250	0.002	0.002	0.004	0.007	0.001	0.000	-0.000	-0.003	-0.001	0.002
$\|L\|$ =500	-0.001	-0.001	0.002	0.005	0.004	0.002	-0.003	0.002	0.002	0.001

Table C.48.: Mean error of gross visits under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	1.043	6.979	18.104	21.608	23.685	28.550	30.747	36.236	40.642	46.496
$\|L\|$ =50	-21.524	-12.412	-4.310	0.867	26.920	27.271	30.485	39.317	45.640	52.242
$\|L\|$ =100	-65.666	-42.475	-26.444	-19.230	-17.464	22.628	2.435	38.834	39.327	58.885
$\|L\|$ =250	-123.313	-158.637	-131.378	-81.538	-32.725	-45.548	7.362	38.944	85.585	85.597
$\|L\|$ =500	-278.640	-250.622	-172.998	-100.486	-78.966	-1.496	114.711	97.606	127.331	195.658

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	11.463	72.850	70.813	28.193	40.537	17.727	5.260	29.903	17.297	30.140
$\|L\|$ =50	-14.820	-16.507	89.880	40.683	9.657	55.270	-17.647	13.507	-7.797	-6.587
$\|L\|$ =100	0.063	4.860	75.493	-93.563	-29.240	-44.093	-48.437	79.160	-50.707	-36.863
$\|L\|$ =250	-63.627	-189.080	-99.727	115.557	45.253	-185.517	-148.807	57.557	-101.427	53.297
$\|L\|$ =500	-87.413	60.460	-157.247	-292.990	37.107	-254.857	-9.730	57.460	161.247	-185.800

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-4.347	1.061	4.024	6.236	-0.033	-6.469	-6.498	-2.010	-4.741	-2.230
$\|L\|$ =50	-1.331	-0.802	-1.160	4.380	15.218	0.133	-7.730	-1.071	-3.334	-7.941
$\|L\|$ =100	9.726	16.219	9.569	-5.097	-15.979	6.071	-20.207	17.962	-17.804	-7.136
$\|L\|$ =250	37.637	-26.808	-27.230	31.348	18.689	-32.417	-2.770	32.082	35.780	-1.159
$\|L\|$ =500	0.044	3.332	-18.914	-0.083	3.483	1.763	63.794	18.124	-26.281	-55.303

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.49.: Mean error of average visits under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.338	-0.350	-0.349	-0.401	-0.417	-0.423	-0.486	-0.481	-0.518	-0.515
$\|L\|$ =50	-0.350	-0.393	-0.408	-0.450	-0.434	-0.517	-0.555	-0.550	-0.598	-0.609
$\|L\|$ =100	-0.517	-0.494	-0.519	-0.569	-0.644	-0.632	-0.736	-0.706	-0.798	-0.825
$\|L\|$ =250	-0.689	-0.928	-0.941	-0.908	-0.886	-1.054	-1.008	-1.036	-1.050	-1.156
$\|L\|$ =500	-1.203	-1.276	-1.272	-1.204	-1.259	-1.262	-1.132	-1.376	-1.440	-1.420

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.213	0.702	0.659	0.229	0.334	0.199	0.047	0.209	0.090	0.160
$\|L\|$ =50	0.091	0.031	0.537	0.299	0.075	0.276	-0.080	0.050	-0.076	-0.079
$\|L\|$ =100	0.035	0.037	0.260	-0.421	-0.211	-0.360	-0.362	0.018	-0.485	-0.502
$\|L\|$ =250	-0.489	-0.973	-0.825	-0.248	-0.599	-1.393	-1.425	-0.960	-1.547	-1.222
$\|L\|$ =500	-1.003	-0.819	-1.685	-2.295	-1.525	-2.627	-2.176	-2.280	-2.190	-3.304

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.510	0.580	0.628	0.602	0.646	0.652	0.614	0.690	0.687	0.751
$\|L\|$ =50	0.640	0.648	0.735	0.764	0.826	0.782	0.780	0.890	0.906	0.940
$\|L\|$ =100	0.833	0.942	0.964	0.993	1.020	1.114	1.132	1.248	1.213	1.321
$\|L\|$ =250	1.232	1.278	1.326	1.539	1.744	1.592	1.855	2.069	2.174	2.137
$\|L\|$ =500	1.669	1.907	2.050	2.315	2.511	2.649	3.054	3.114	3.259	3.449

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.50.: Mean error of entity coverage under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.048	0.056	0.067	0.079	0.084	0.090	0.105	0.109	0.121	0.126
$\|L\|$ =50	0.036	0.049	0.056	0.067	0.083	0.097	0.105	0.112	0.126	0.132
$\|L\|$ =100	0.031	0.037	0.047	0.056	0.066	0.081	0.086	0.098	0.111	0.123
$\|L\|$ =250	0.019	0.027	0.034	0.041	0.050	0.058	0.066	0.074	0.085	0.092
$\|L\|$ =500	0.014	0.019	0.027	0.032	0.037	0.045	0.052	0.059	0.065	0.071

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.016	-0.014	-0.011	-0.004	-0.007	-0.008	-0.001	-0.001	0.003	0.004
$\|L\|$ =50	-0.021	-0.015	-0.017	-0.011	-0.005	-0.001	0.001	0.003	0.006	0.005
$\|L\|$ =100	-0.004	-0.003	0.003	0.008	0.011	0.021	0.019	0.028	0.034	0.040
$\|L\|$ =250	0.018	0.024	0.032	0.036	0.046	0.053	0.063	0.072	0.082	0.090
$\|L\|$ =500	0.026	0.035	0.045	0.053	0.059	0.071	0.079	0.091	0.098	0.107

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.058	-0.060	-0.061	-0.058	-0.067	-0.072	-0.069	-0.072	-0.074	-0.077
$\|L\|$ =50	-0.074	-0.075	-0.084	-0.083	-0.083	-0.087	-0.091	-0.097	-0.099	-0.106
$\|L\|$ =100	-0.073	-0.080	-0.084	-0.091	-0.097	-0.097	-0.107	-0.102	-0.111	-0.116
$\|L\|$ =250	-0.061	-0.074	-0.077	-0.077	-0.089	-0.090	-0.097	-0.102	-0.105	-0.110
$\|L\|$ =500	-0.053	-0.060	-0.066	-0.072	-0.077	-0.081	-0.086	-0.092	-0.100	-0.108

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	-0.014	-0.012	-0.008	0.000	-0.003	-0.002	0.007	0.011	0.016	0.022
$\|L\|$ =50	-0.023	-0.015	-0.017	-0.010	0.002	0.008	0.012	0.014	0.030	0.030
$\|L\|$ =100	-0.016	-0.016	-0.012	-0.009	-0.004	0.008	0.008	0.020	0.029	0.046
$\|L\|$ =250	-0.011	-0.010	-0.006	-0.002	0.003	0.010	0.013	0.021	0.034	0.045
$\|L\|$ =500	-0.008	-0.006	-0.002	0.000	0.001	0.009	0.013	0.019	0.025	0.036

C.3. Root Mean Squared Error of Experiments under MCAR, CDMAR and MAR

This section contains the detailed results of the presented experiments under MCAR, CDMAR and MAR. Tables C.51 - C.68 contain the root mean squared error for Experiments 1 - 6 testing MCAR, Tables C.69 - C.77 contain the root mean squared error for Experiments 7 - 9 testing CDMAR and finally Tables C.78 - C.89 contain the root mean squared error for Experiments 10 - 13 testing MAR.

To each experiment belong three tables showing the visit potential quantities gross visits, average visits and entity coverage. The root mean squared error is formed over 30 repetition of each parameterization. Each part of a table shows the root mean squared error for one tested method. Each row contains results for a certain location set size and columns show results for different rates of missing data.

Table C.51.: Root mean squared error of gross visits under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	9.127	11.906	20.812	23.691	33.950	34.848	33.563	37.443	48.165	53.642
$\|L\|$ =50	14.057	19.070	29.885	23.397	30.530	34.798	36.616	47.710	42.510	66.482
$\|L\|$ =100	20.271	30.775	36.111	52.162	74.412	58.691	57.429	64.844	78.011	89.722
$\|L\|$ =250	41.852	72.227	87.224	130.884	95.692	152.021	177.326	172.987	207.210	179.166
$\|L\|$ =500	89.017	150.224	168.895	210.424	163.455	314.646	285.105	369.012	335.849	259.336

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	17.644	22.266	34.316	68.902	83.427	101.501	154.573	193.050	116.605	239.067
$\|L\|$ =50	27.622	56.876	74.825	123.993	258.322	163.846	135.306	167.589	291.770	245.856
$\|L\|$ =100	65.504	133.040	153.297	197.548	229.866	380.037	428.754	470.126	816.601	634.070
$\|L\|$ =250	573.011	483.868	477.532	564.245	415.573	711.477	667.020	784.300	716.423	874.139
$\|L\|$ =500	447.016	1516.058	1260.917	774.042	1359.001	2189.114	1090.343	1924.581	3759.279	2614.168

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	9.222	12.376	17.436	20.906	20.697	20.179	24.819	29.872	30.082	24.448
$\|L\|$ =50	13.408	26.862	27.492	35.505	37.106	39.550	35.131	45.185	40.650	55.396
$\|L\|$ =100	28.856	36.133	44.481	57.787	73.988	51.963	68.679	73.921	88.374	115.595
$\|L\|$ =250	60.948	99.751	97.075	116.474	129.346	209.623	162.328	185.527	207.659	207.139
$\|L\|$ =500	98.091	164.484	188.557	197.421	190.088	337.948	303.399	307.744	395.546	253.913

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.52.: Root mean squared error of average visits under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.115	0.213	0.270	0.361	0.431	0.512	0.581	0.628	0.705	0.708
$\|L\|$ =50	0.133	0.220	0.324	0.421	0.483	0.642	0.660	0.701	0.777	0.808
$\|L\|$ =100	0.151	0.306	0.441	0.584	0.663	0.825	0.859	0.973	1.137	1.197
$\|L\|$ =250	0.193	0.432	0.673	0.874	0.983	1.083	1.420	1.577	1.645	1.892
$\|L\|$ =500	0.392	0.728	0.946	1.314	1.188	1.674	1.909	2.302	2.041	2.520

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.107	0.174	0.255	0.514	0.584	0.750	1.144	1.503	0.849	1.694
$\|L\|$ =50	0.116	0.238	0.332	0.566	1.192	0.735	0.635	0.836	1.360	1.121
$\|L\|$ =100	0.235	0.497	0.606	0.683	0.846	1.334	1.433	1.696	2.850	2.140
$\|L\|$ =250	1.807	1.560	1.516	1.963	1.668	2.173	1.782	2.326	2.435	1.988
$\|L\|$ =500	1.381	4.338	3.725	3.214	3.890	5.804	4.435	5.120	8.721	5.682

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.111	0.216	0.396	0.539	0.665	0.820	1.020	1.243	1.308	1.673
$\|L\|$ =50	0.140	0.261	0.409	0.632	0.823	0.949	1.141	1.440	1.656	2.004
$\|L\|$ =100	0.204	0.393	0.606	0.827	1.065	1.272	1.632	1.890	2.084	2.491
$\|L\|$ =250	0.399	0.684	1.044	1.229	1.595	2.379	2.448	2.889	3.235	3.753
$\|L\|$ =500	0.616	1.099	1.468	1.827	2.578	3.238	3.392	4.316	5.111	5.368

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	−	−	−	−	−	−	−	−	−	−
$\|L\|$ = 50	−	−	−	−	−	−	−	−	−	−
$\|L\|$ =100	−	−	−	−	−	−	−	−	−	−
$\|L\|$ =250	−	−	−	−	−	−	−	−	−	−
$\|L\|$ =500	−	−	−	−	−	−	−	−	−	−

Table C.53.: Root mean squared error of entity coverage under MCAR without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.018	0.033	0.052	0.063	0.086	0.105	0.117	0.128	0.157	0.163
$\|L\|$ =50	0.017	0.033	0.049	0.064	0.081	0.099	0.116	0.125	0.146	0.160
$\|L\|$ =100	0.017	0.027	0.044	0.055	0.068	0.082	0.097	0.112	0.125	0.138
$\|L\|$ =250	0.011	0.022	0.031	0.042	0.052	0.061	0.072	0.080	0.090	0.101
$\|L\|$ =500	0.008	0.017	0.021	0.034	0.043	0.049	0.054	0.062	0.067	0.079

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.006	0.009	0.011	0.018	0.019	0.017	0.021	0.027	0.025	0.031
$\|L\|$ =50	0.005	0.011	0.013	0.016	0.019	0.021	0.019	0.026	0.030	0.034
$\|L\|$ =100	0.008	0.008	0.014	0.014	0.021	0.023	0.025	0.031	0.030	0.038
$\|L\|$ =250	0.011	0.021	0.030	0.039	0.049	0.059	0.067	0.076	0.087	0.095
$\|L\|$ =500	0.013	0.027	0.034	0.052	0.065	0.077	0.085	0.100	0.109	0.123

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.014	0.028	0.038	0.058	0.067	0.076	0.094	0.109	0.115	0.135
$\|L\|$ =50	0.017	0.036	0.053	0.072	0.090	0.108	0.118	0.143	0.157	0.172
$\|L\|$ =100	0.020	0.040	0.053	0.077	0.096	0.111	0.132	0.149	0.166	0.187
$\|L\|$ =250	0.019	0.035	0.055	0.070	0.085	0.105	0.119	0.137	0.155	0.174
$\|L\|$ =500	0.017	0.032	0.047	0.060	0.074	0.088	0.104	0.126	0.138	0.145

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.005	0.008	0.010	0.016	0.017	0.018	0.014	0.023	0.031	0.029
$\|L\|$ =50	0.004	0.009	0.011	0.014	0.018	0.018	0.020	0.024	0.045	0.028
$\|L\|$ =100	0.006	0.007	0.009	0.010	0.017	0.020	0.018	0.020	0.038	0.030
$\|L\|$ =250	0.005	0.009	0.008	0.011	0.016	0.017	0.018	0.023	0.035	0.013
$\|L\|$ =500	0.004	0.006	0.007	0.010	0.013	0.014	0.016	0.021	0.026	0.019

Table C.54.: Root mean squared error of gross visits under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	7.179	15.676	23.944	25.908	27.309	34.573	42.482	42.069	43.555	63.868
$\|L\|$ =50	12.237	18.703	22.476	29.310	26.992	46.566	45.311	36.976	59.661	49.439
$\|L\|$ =100	23.143	26.503	39.108	49.557	58.669	58.513	53.597	62.792	67.883	101.100
$\|L\|$ =250	44.940	63.738	91.229	118.132	109.433	137.458	176.896	193.483	198.672	214.441
$\|L\|$ =500	91.159	136.268	170.624	202.092	247.242	274.155	304.775	326.639	387.336	358.699

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	16.851	25.234	51.193	48.106	218.492	104.019	107.133	112.197	126.393	109.454
$\|L\|$ =50	23.029	46.385	87.222	156.600	70.514	107.641	174.641	164.026	251.838	155.021
$\|L\|$ =100	56.521	134.536	189.798	198.812	362.462	163.881	182.152	125.188	387.924	186.763
$\|L\|$ =250	81.724	164.230	224.808	276.493	295.565	406.081	357.005	417.357	615.576	635.718
$\|L\|$ =500	238.868	313.556	422.090	578.358	599.606	646.006	758.700	852.607	1044.885	1454.489

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	8.656	12.177	18.917	19.122	23.119	23.652	27.499	26.462	24.911	29.104
$\|L\|$ =50	12.509	18.713	30.661	41.272	34.393	46.113	50.013	40.519	55.841	47.420
$\|L\|$ =100	25.319	37.846	49.982	57.326	71.589	68.871	68.149	59.332	99.793	79.924
$\|L\|$ =250	52.107	75.614	92.192	117.362	111.689	164.886	173.796	134.115	210.174	206.879
$\|L\|$ =500	102.793	200.682	201.560	183.415	265.530	250.290	353.343	264.048	331.760	439.398

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.55.: Root mean squared error of average visits under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.114	0.203	0.267	0.380	0.443	0.518	0.547	0.622	0.665	0.690
$\|L\|$ =50	0.122	0.229	0.324	0.418	0.490	0.595	0.650	0.722	0.782	0.818
$\|L\|$ =100	0.165	0.294	0.439	0.557	0.671	0.757	0.867	0.987	1.016	1.228
$\|L\|$ =250	0.233	0.467	0.749	0.848	0.967	1.147	1.500	1.602	1.716	1.827
$\|L\|$ =500	0.371	0.588	0.989	1.225	1.560	1.703	2.010	2.458	2.282	2.403

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.097	0.164	0.438	0.411	1.802	0.783	0.981	1.053	1.103	1.061
$\|L\|$ =50	0.116	0.185	0.426	0.806	0.373	0.588	1.055	0.990	1.452	1.017
$\|L\|$ =100	0.223	0.528	0.688	0.767	1.386	0.697	0.764	0.660	1.725	0.926
$\|L\|$ =250	0.296	0.588	0.769	0.937	1.058	1.163	1.147	1.293	1.882	1.748
$\|L\|$ =500	0.754	1.096	1.469	2.025	2.224	2.227	2.288	2.914	3.027	3.825

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.111	0.233	0.373	0.490	0.644	0.832	1.079	1.188	1.419	1.686
$\|L\|$ =50	0.146	0.285	0.460	0.632	0.790	0.965	1.165	1.352	1.630	1.861
$\|L\|$ =100	0.193	0.420	0.644	0.867	1.095	1.427	1.566	1.896	2.435	2.431
$\|L\|$ =250	0.335	0.619	0.924	1.258	1.605	2.163	2.208	2.991	3.347	3.633
$\|L\|$ =500	0.574	1.168	1.581	1.899	2.362	3.035	3.860	4.070	5.122	5.766

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.56.: Root mean squared error of entity coverage under MCAR with sociodemographic variable gender

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.018	0.035	0.052	0.067	0.084	0.104	0.119	0.134	0.147	0.172
$\lvert L\rvert$ =50	0.016	0.033	0.051	0.070	0.081	0.102	0.111	0.129	0.154	0.157
$\lvert L\rvert$ =100	0.016	0.030	0.041	0.057	0.069	0.083	0.093	0.110	0.124	0.136
$\lvert L\rvert$ =250	0.012	0.022	0.032	0.040	0.052	0.064	0.070	0.078	0.089	0.103
$\lvert L\rvert$ =500	0.010	0.015	0.023	0.031	0.039	0.045	0.056	0.064	0.070	0.079

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.007	0.010	0.013	0.021	0.027	0.024	0.033	0.031	0.037	0.039
$\lvert L\rvert$ =50	0.009	0.016	0.020	0.021	0.034	0.036	0.054	0.053	0.053	0.061
$\lvert L\rvert$ =100	0.006	0.011	0.017	0.022	0.027	0.035	0.037	0.036	0.039	0.042
$\lvert L\rvert$ =250	0.007	0.010	0.012	0.015	0.020	0.026	0.023	0.029	0.027	0.033
$\lvert L\rvert$ =500	0.010	0.015	0.020	0.029	0.035	0.042	0.046	0.052	0.058	0.062

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.014	0.028	0.038	0.056	0.068	0.078	0.093	0.105	0.122	0.128
$\lvert L\rvert$ =50	0.019	0.038	0.050	0.064	0.085	0.100	0.120	0.139	0.149	0.168
$\lvert L\rvert$ =100	0.018	0.038	0.056	0.074	0.095	0.116	0.136	0.148	0.170	0.189
$\lvert L\rvert$ =250	0.019	0.038	0.052	0.070	0.086	0.103	0.124	0.146	0.162	0.164
$\lvert L\rvert$ =500	0.015	0.030	0.044	0.063	0.073	0.095	0.106	0.122	0.140	0.151

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.006	0.007	0.008	0.012	0.016	0.017	0.023	0.024	0.034	0.026
$\lvert L\rvert$ =50	0.006	0.009	0.012	0.013	0.014	0.016	0.023	0.028	0.035	0.024
$\lvert L\rvert$ =100	0.005	0.009	0.008	0.011	0.017	0.016	0.021	0.024	0.038	0.021
$\lvert L\rvert$ =250	0.006	0.008	0.007	0.010	0.013	0.020	0.019	0.025	0.036	0.018
$\lvert L\rvert$ =500	0.005	0.006	0.007	0.011	0.011	0.018	0.019	0.024	0.034	0.015

Table C.57.: Root mean squared error of gross visits under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	9.049	16.020	19.343	24.470	28.794	31.130	40.296	47.358	51.287	55.249
$\|L\|$ =50	11.068	20.308	23.022	28.282	35.578	41.489	45.557	42.499	55.969	79.241
$\|L\|$ =100	21.137	27.258	41.206	35.027	62.368	66.716	57.390	61.561	73.520	86.762
$\|L\|$ =250	53.913	71.089	86.642	113.423	115.029	123.324	190.722	155.844	198.837	153.730
$\|L\|$ =500	62.882	111.346	135.630	160.976	200.959	242.357	227.882	362.143	323.889	416.997

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	22.836	35.861	47.181	118.389	43.101	91.855	124.425	73.700	109.515	150.457
$\|L\|$ =50	34.155	35.798	73.012	88.824	63.916	114.263	188.397	127.696	233.280	194.676
$\|L\|$ =100	232.339	82.329	180.640	153.783	112.370	266.103	145.329	269.048	183.224	328.561
$\|L\|$ =250	120.845	143.342	210.004	231.854	257.413	409.214	393.512	461.135	683.523	474.027
$\|L\|$ =500	159.660	229.790	474.413	398.131	550.743	518.197	798.906	700.122	898.880	913.168

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	12.105	15.488	17.238	22.043	16.788	16.796	28.569	29.239	26.746	30.531
$\|L\|$ =50	11.432	20.668	22.536	29.641	38.763	40.414	35.213	48.900	47.903	66.856
$\|L\|$ =100	27.779	39.896	47.251	61.461	66.385	67.005	78.797	83.908	85.653	95.245
$\|L\|$ =250	75.249	77.742	94.994	119.500	128.044	155.556	173.033	170.143	182.933	189.533
$\|L\|$ =500	89.147	132.747	192.953	205.511	247.978	270.267	276.316	356.225	296.759	410.706

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.58.: Root mean squared error of average visits under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.112	0.222	0.282	0.364	0.453	0.530	0.577	0.612	0.664	0.720
$\|L\|$ =50	0.131	0.215	0.307	0.419	0.492	0.570	0.612	0.754	0.800	0.823
$\|L\|$ =100	0.153	0.283	0.433	0.539	0.618	0.858	0.837	0.971	1.101	1.162
$\|L\|$ =250	0.275	0.490	0.643	0.835	1.033	1.177	1.472	1.459	1.795	1.669
$\|L\|$ =500	0.306	0.562	0.693	0.947	1.322	1.675	1.618	2.224	2.424	1.871

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.152	0.289	0.373	0.894	0.383	0.733	1.007	0.640	0.861	1.310
$\|L\|$ =50	0.186	0.168	0.366	0.427	0.389	0.616	1.056	0.743	1.370	1.125
$\|L\|$ =100	0.849	0.299	0.644	0.592	0.458	0.991	0.657	1.058	0.786	1.242
$\|L\|$ =250	0.436	0.512	0.751	0.908	1.130	1.271	1.249	1.345	2.117	1.348
$\|L\|$ =500	0.681	1.102	1.637	1.771	2.180	2.451	2.848	2.757	2.600	2.685

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.138	0.244	0.363	0.541	0.645	0.770	0.959	1.164	1.409	1.558
$\|L\|$ =50	0.138	0.307	0.451	0.609	0.820	0.994	1.181	1.335	1.685	2.006
$\|L\|$ =100	0.208	0.397	0.626	0.841	1.108	1.266	1.600	1.861	2.181	2.522
$\|L\|$ =250	0.311	0.601	0.988	1.286	1.667	2.064	2.492	2.913	3.222	3.858
$\|L\|$ =500	0.515	1.069	1.625	2.107	2.580	2.959	3.581	4.100	4.957	6.170

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.59.: Root mean squared error of entity coverage under MCAR with sociodemographic variable age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.016	0.036	0.052	0.069	0.089	0.104	0.120	0.134	0.153	0.170
$\|L\|$ =50	0.018	0.036	0.050	0.065	0.084	0.100	0.117	0.130	0.153	0.162
$\|L\|$ =100	0.015	0.030	0.043	0.057	0.069	0.084	0.098	0.112	0.126	0.142
$\|L\|$ =250	0.011	0.021	0.031	0.042	0.054	0.058	0.070	0.081	0.090	0.099
$\|L\|$ =500	0.008	0.017	0.023	0.031	0.039	0.047	0.055	0.062	0.073	0.077

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.009	0.010	0.012	0.020	0.022	0.023	0.022	0.029	0.035	0.037
$\|L\|$ =50	0.008	0.012	0.016	0.023	0.027	0.031	0.037	0.044	0.043	0.054
$\|L\|$ =100	0.005	0.010	0.013	0.012	0.023	0.024	0.026	0.026	0.032	0.030
$\|L\|$ =250	0.007	0.013	0.018	0.022	0.027	0.025	0.032	0.040	0.041	0.049
$\|L\|$ =500	0.009	0.020	0.027	0.035	0.044	0.050	0.060	0.069	0.077	0.085

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.016	0.026	0.041	0.054	0.066	0.077	0.089	0.105	0.118	0.130
$\|L\|$ =50	0.019	0.034	0.052	0.072	0.085	0.102	0.118	0.139	0.152	0.175
$\|L\|$ =100	0.019	0.038	0.055	0.071	0.094	0.113	0.127	0.145	0.173	0.185
$\|L\|$ =250	0.018	0.033	0.052	0.071	0.087	0.104	0.123	0.137	0.162	0.174
$\|L\|$ =500	0.017	0.031	0.045	0.060	0.072	0.092	0.102	0.124	0.138	0.156

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.007	0.008	0.007	0.016	0.016	0.017	0.024	0.027	0.034	0.028
$\|L\|$ =50	0.006	0.009	0.010	0.014	0.017	0.016	0.017	0.029	0.033	0.025
$\|L\|$ =100	0.004	0.009	0.010	0.013	0.021	0.018	0.024	0.027	0.035	0.027
$\|L\|$ =250	0.004	0.007	0.010	0.012	0.013	0.014	0.021	0.020	0.026	0.019
$\|L\|$ =500	0.004	0.006	0.009	0.008	0.011	0.014	0.021	0.021	0.034	0.014

Table C.60.: Root mean squared error of gross visits under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	8.631	16.656	17.014	27.080	29.096	33.371	36.492	50.772	48.817	60.052
$\|L\|$ =50	12.609	17.206	28.081	24.715	35.894	41.173	45.953	53.756	53.140	61.851
$\|L\|$ =100	21.812	28.025	40.484	43.736	51.655	67.282	48.371	57.140	83.975	66.513
$\|L\|$ =250	77.985	88.108	81.967	85.797	113.941	156.461	138.410	175.958	164.364	209.547
$\|L\|$ =500	71.576	85.399	166.308	176.316	227.716	268.484	321.444	272.928	349.506	327.316

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	12.373	29.544	40.931	48.497	191.790	83.062	63.716	130.901	150.786	145.008
$\|L\|$ =50	21.005	36.452	53.192	56.279	58.012	94.216	84.851	114.852	223.657	184.831
$\|L\|$ =100	55.962	191.756	135.913	147.404	144.926	534.821	165.603	185.846	202.914	244.332
$\|L\|$ =250	180.241	189.412	336.632	287.187	367.503	436.014	371.345	417.358	751.823	368.891
$\|L\|$ =500	163.478	461.809	497.369	482.758	581.915	1069.177	1032.404	621.085	910.406	888.680

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	8.488	15.535	20.826	19.457	22.677	21.960	26.739	33.679	27.567	31.954
$\|L\|$ =50	15.894	23.732	29.225	26.406	37.067	40.096	40.012	56.274	45.542	44.190
$\|L\|$ =100	32.752	30.503	57.233	51.803	68.311	66.672	69.545	82.439	93.998	77.348
$\|L\|$ =250	92.993	91.099	110.094	94.097	119.984	144.245	109.810	180.440	134.787	198.022
$\|L\|$ =500	94.615	135.632	179.563	215.444	241.830	235.677	364.291	309.101	319.424	308.856

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.61.: Root mean squared error of average visits under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.111	0.209	0.325	0.370	0.456	0.505	0.544	0.601	0.684	0.672
$\|L\|$ =50	0.128	0.228	0.373	0.388	0.487	0.580	0.658	0.685	0.776	0.856
$\|L\|$ =100	0.139	0.299	0.405	0.542	0.621	0.751	0.847	0.970	1.024	1.066
$\|L\|$ =250	0.322	0.538	0.668	0.807	0.982	1.243	1.382	1.425	1.619	1.808
$\|L\|$ =500	0.332	0.599	0.986	1.079	1.256	1.662	1.937	1.867	2.204	2.361

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.086	0.182	0.309	0.375	1.458	0.690	0.605	1.057	1.273	1.262
$\|L\|$ =50	0.084	0.136	0.224	0.261	0.285	0.447	0.458	0.612	1.166	1.056
$\|L\|$ =100	0.199	0.687	0.458	0.563	0.502	2.007	0.596	0.672	0.719	0.861
$\|L\|$ =250	0.584	0.705	1.109	1.080	1.327	1.521	1.449	1.230	2.148	1.293
$\|L\|$ =500	0.623	1.486	1.719	2.052	2.271	3.374	3.554	2.410	2.819	3.079

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.117	0.255	0.350	0.533	0.632	0.866	1.051	1.273	1.367	1.574
$\|L\|$ =50	0.141	0.294	0.411	0.621	0.804	0.968	1.144	1.486	1.588	1.844
$\|L\|$ =100	0.244	0.391	0.651	0.854	1.053	1.292	1.619	1.895	2.175	2.579
$\|L\|$ =250	0.353	0.530	1.001	1.193	1.715	2.080	2.331	3.186	3.277	3.708
$\|L\|$ =500	0.535	0.973	1.438	2.100	2.785	3.133	3.437	4.394	4.649	5.615

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.62.: Root mean squared error of entity coverage under MCAR with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.018	0.038	0.051	0.070	0.084	0.098	0.113	0.133	0.157	0.168
$\|L\|$ =50	0.017	0.035	0.051	0.065	0.083	0.101	0.122	0.134	0.150	0.171
$\|L\|$ =100	0.015	0.028	0.044	0.058	0.070	0.084	0.097	0.113	0.123	0.139
$\|L\|$ =250	0.010	0.022	0.032	0.043	0.050	0.060	0.071	0.081	0.091	0.102
$\|L\|$ =500	0.009	0.016	0.023	0.032	0.039	0.047	0.053	0.060	0.067	0.079

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.005	0.008	0.013	0.016	0.020	0.024	0.026	0.033	0.030	0.039
$\|L\|$ =50	0.007	0.011	0.014	0.021	0.023	0.030	0.029	0.036	0.044	0.047
$\|L\|$ =100	0.007	0.009	0.010	0.013	0.014	0.017	0.017	0.023	0.028	0.026
$\|L\|$ =250	0.007	0.014	0.019	0.027	0.026	0.031	0.038	0.042	0.046	0.045
$\|L\|$ =500	0.010	0.018	0.024	0.035	0.040	0.052	0.057	0.066	0.068	0.078

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.014	0.024	0.041	0.052	0.067	0.080	0.094	0.109	0.118	0.130
$\|L\|$ =50	0.019	0.035	0.052	0.068	0.089	0.105	0.116	0.135	0.152	0.169
$\|L\|$ =100	0.019	0.038	0.055	0.077	0.090	0.107	0.127	0.151	0.169	0.185
$\|L\|$ =250	0.018	0.033	0.051	0.066	0.091	0.106	0.121	0.140	0.155	0.172
$\|L\|$ =500	0.015	0.031	0.046	0.062	0.079	0.095	0.104	0.120	0.138	0.149

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.005	0.008	0.010	0.013	0.015	0.023	0.019	0.025	0.039	0.036
$\|L\|$ =50	0.006	0.010	0.009	0.013	0.017	0.025	0.023	0.028	0.037	0.035
$\|L\|$ =100	0.006	0.008	0.011	0.013	0.013	0.016	0.027	0.020	0.040	0.027
$\|L\|$ =250	0.004	0.006	0.009	0.012	0.014	0.016	0.019	0.026	0.034	0.022
$\|L\|$ =500	0.004	0.007	0.007	0.009	0.011	0.017	0.017	0.021	0.029	0.018

Table C.63.: Root mean squared error of gross visits under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	8.196	13.523	23.162	24.333	30.274	35.409	40.510	39.632	50.899	65.652
$\|L\|$ =50	8.108	21.881	26.064	31.197	32.448	37.086	42.184	49.181	39.830	63.380
$\|L\|$ =100	20.547	28.189	36.326	51.924	49.555	61.027	53.694	65.112	66.948	83.734
$\|L\|$ =250	51.317	87.253	97.115	128.262	117.545	135.551	173.450	174.166	152.419	168.148
$\|L\|$ =500	119.848	103.857	178.475	174.105	200.357	270.343	259.633	340.730	321.728	407.901

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	68.677	78.141	60.798	58.161	69.617	91.681	111.157	126.076	152.848	136.646
$\|L\|$ =50	22.241	40.829	79.246	66.491	69.850	131.785	198.642	215.785	122.639	119.838
$\|L\|$ =100	134.426	112.782	95.054	161.670	116.530	237.918	312.383	183.717	330.787	306.289
$\|L\|$ =250	178.946	150.611	214.285	374.119	340.577	573.191	498.401	379.333	379.782	454.029
$\|L\|$ =500	198.415	340.270	386.294	571.458	441.434	997.792	640.283	630.912	567.788	923.652

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	12.008	12.171	20.091	20.480	23.959	18.963	30.672	24.374	27.608	35.965
$\|L\|$ =50	12.027	26.757	26.821	35.588	29.252	41.115	45.788	53.211	46.667	49.780
$\|L\|$ =100	27.500	41.962	37.043	52.977	59.460	74.073	71.838	79.141	72.030	92.630
$\|L\|$ =250	55.829	91.896	93.022	127.205	140.188	145.015	158.111	137.561	185.223	168.015
$\|L\|$ =500	130.193	161.270	188.037	215.732	226.329	323.832	298.322	305.989	306.561	371.491

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.64.: Root mean squared error of average visits under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.117	0.226	0.290	0.387	0.428	0.518	0.570	0.623	0.692	0.691
$\|L\|$ =50	0.117	0.254	0.303	0.374	0.469	0.571	0.661	0.733	0.802	0.860
$\|L\|$ =100	0.171	0.256	0.418	0.587	0.675	0.722	0.892	0.923	1.132	1.156
$\|L\|$ =250	0.235	0.516	0.645	0.893	0.942	1.233	1.430	1.581	1.610	1.683
$\|L\|$ =500	0.442	0.520	0.996	1.022	1.171	1.469	1.836	2.029	1.987	2.008

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.506	0.637	0.494	0.446	0.573	0.811	0.959	1.155	1.296	1.246
$\|L\|$ =50	0.111	0.198	0.400	0.306	0.422	0.747	1.107	1.321	0.859	0.897
$\|L\|$ =100	0.488	0.433	0.380	0.659	0.483	0.969	1.114	0.752	1.476	1.272
$\|L\|$ =250	0.559	0.578	0.760	1.189	1.114	1.734	1.555	1.404	1.025	1.557
$\|L\|$ =500	0.711	1.155	1.674	2.031	2.112	3.034	2.540	2.867	2.673	3.155

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.138	0.233	0.367	0.516	0.691	0.811	1.016	1.144	1.406	1.758
$\|L\|$ =50	0.143	0.289	0.463	0.633	0.812	1.002	1.195	1.415	1.598	1.882
$\|L\|$ =100	0.208	0.446	0.631	0.834	1.028	1.441	1.530	1.934	2.030	2.497
$\|L\|$ =250	0.345	0.571	0.958	1.282	1.785	1.959	2.442	2.656	3.176	3.451
$\|L\|$ =500	0.581	1.188	1.325	2.069	2.860	3.266	3.724	4.162	4.989	5.669

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.65.: Root mean squared error of entity coverage under MCAR with sociodemographic variables gender and age group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|=25$	0.017	0.036	0.052	0.069	0.085	0.105	0.119	0.132	0.157	0.170
$\|L\|=50$	0.018	0.036	0.048	0.063	0.082	0.099	0.113	0.132	0.143	0.167
$\|L\|=100$	0.016	0.027	0.042	0.056	0.069	0.088	0.096	0.107	0.125	0.141
$\|L\|=250$	0.011	0.021	0.030	0.041	0.052	0.063	0.072	0.079	0.090	0.102
$\|L\|=500$	0.009	0.014	0.025	0.033	0.039	0.045	0.057	0.061	0.069	0.078

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|=25$	0.006	0.011	0.013	0.018	0.021	0.027	0.031	0.038	0.038	0.047
$\|L\|=50$	0.008	0.014	0.022	0.028	0.033	0.041	0.046	0.052	0.061	0.059
$\|L\|=100$	0.007	0.011	0.016	0.018	0.025	0.025	0.031	0.034	0.038	0.040
$\|L\|=250$	0.007	0.010	0.014	0.018	0.024	0.026	0.030	0.033	0.039	0.045
$\|L\|=500$	0.011	0.017	0.027	0.037	0.043	0.047	0.061	0.068	0.076	0.082

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|=25$	0.014	0.027	0.038	0.052	0.066	0.079	0.092	0.108	0.119	0.134
$\|L\|=50$	0.018	0.036	0.055	0.072	0.085	0.106	0.121	0.140	0.159	0.172
$\|L\|=100$	0.019	0.039	0.056	0.075	0.094	0.112	0.130	0.150	0.172	0.185
$\|L\|=250$	0.017	0.036	0.054	0.073	0.085	0.101	0.120	0.141	0.153	0.168
$\|L\|=500$	0.015	0.031	0.049	0.062	0.078	0.091	0.107	0.121	0.138	0.151

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|=25$	0.003	0.009	0.010	0.012	0.013	0.015	0.019	0.028	0.038	0.041
$\|L\|=50$	0.006	0.010	0.013	0.014	0.014	0.026	0.023	0.031	0.039	0.032
$\|L\|=100$	0.006	0.008	0.010	0.010	0.014	0.023	0.020	0.027	0.040	0.026
$\|L\|=250$	0.005	0.006	0.010	0.011	0.014	0.018	0.017	0.025	0.026	0.019
$\|L\|=500$	0.004	0.005	0.009	0.011	0.011	0.014	0.020	0.022	0.026	0.021

Table C.66.: Root mean squared error of gross visits under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	7.803	15.401	20.609	25.899	35.479	39.659	42.464	42.313	49.627	73.585
$\|L\|$ =50	13.623	20.059	24.554	28.345	36.570	39.416	47.000	48.406	114.771	102.125
$\|L\|$ =100	18.121	34.644	46.150	47.678	48.325	78.433	85.588	76.213	81.349	283.849
$\|L\|$ =250	59.736	77.984	83.547	99.913	131.311	118.181	20958.391	183.928	2921136.155	4302.572
$\|L\|$ =500	79.280	143.307	180.624	185.333	215.425	147874.290	5478.737	570351.914	16997.650	812536.585

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	72.277	42.890	40.296	36.577	58.799	60.580	79.279	98.713	51.997	181.160
$\|L\|$ =50	36.906	97.864	47.343	69.298	149.370	161.367	226.166	78.727	155.665	154.021
$\|L\|$ =100	38.708	64.962	312.611	106.190	214.192	219.617	155.167	165.682	189.011	230.799
$\|L\|$ =250	166.265	166.870	252.692	514.318	297.182	354.441	694.398	463.012	352.207	357.454
$\|L\|$ =500	210.768	272.938	478.022	507.041	479.386	630.611	830.854	490.393	579.187	666.024

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	11.497	15.785	17.206	18.796	15.823	19.922	18.990	23.005	35.773	30.476
$\|L\|$ =50	16.255	27.983	23.486	28.471	30.158	38.302	51.723	45.288	47.542	45.231
$\|L\|$ =100	22.279	43.991	65.263	53.653	49.360	82.515	93.854	76.960	85.441	113.191
$\|L\|$ =250	72.137	88.708	105.811	149.521	132.913	132.348	150.508	148.665	160.748	195.032
$\|L\|$ =500	125.532	160.949	225.692	297.147	215.960	270.682	224.708	414.523	262.276	340.616

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.67.: Root mean squared error of average visits under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.104	0.208	0.276	0.352	0.401	0.492	0.553	0.638	0.685	0.652
$\|L\|$ =50	0.124	0.213	0.315	0.426	0.465	0.552	0.636	0.694	0.754	0.791
$\|L\|$ =100	0.154	0.291	0.467	0.524	0.626	0.704	0.927	0.943	1.006	1.159
$\|L\|$ =250	0.267	0.470	0.595	0.844	1.001	1.167	59.376	1.268	8201.937	11.520
$\|L\|$ =500	0.312	0.649	0.962	1.046	1.281	410.511	14.873	1582.716	45.742	2205.130

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.546	0.294	0.307	0.297	0.503	0.561	0.674	0.883	0.595	1.476
$\|L\|$ =50	0.181	0.450	0.218	0.284	0.731	0.802	1.190	0.361	0.857	0.980
$\|L\|$ =100	0.132	0.211	1.154	0.361	0.836	0.844	0.463	0.591	0.852	0.896
$\|L\|$ =250	0.538	0.621	0.790	1.701	1.140	1.189	2.006	1.422	1.343	1.418
$\|L\|$ =500	0.771	1.158	1.748	2.144	2.285	2.670	3.110	2.589	2.841	2.967

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.150	0.248	0.385	0.524	0.693	0.804	0.972	1.115	1.417	1.680
$\|L\|$ =50	0.154	0.315	0.456	0.571	0.843	0.990	1.161	1.330	1.577	1.966
$\|L\|$ =100	0.201	0.418	0.657	0.822	1.076	1.421	1.535	1.998	2.219	2.558
$\|L\|$ =250	0.353	0.612	1.080	1.412	1.648	1.990	2.572	2.977	3.182	3.715
$\|L\|$ =500	0.628	1.007	1.515	2.386	2.380	2.998	3.607	4.618	4.739	5.571

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.68.: Root mean squared error of entity coverage under MCAR with sociodemographic variables gender and occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\vert L\vert$ =25	0.018	0.038	0.052	0.069	0.088	0.106	0.120	0.140	0.153	0.175
$\vert L\vert$ =50	0.017	0.033	0.052	0.070	0.083	0.099	0.117	0.132	0.154	0.164
$\vert L\vert$ =100	0.013	0.027	0.045	0.057	0.071	0.085	0.094	0.111	0.126	0.141
$\vert L\vert$ =250	0.010	0.020	0.032	0.041	0.052	0.062	0.065	0.081	0.093	0.099
$\vert L\vert$ =500	0.010	0.017	0.026	0.032	0.038	0.047	0.053	0.063	0.068	0.077

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\vert L\vert$ =25	0.006	0.009	0.014	0.020	0.022	0.023	0.031	0.032	0.046	0.043
$\vert L\vert$ =50	0.008	0.013	0.016	0.023	0.026	0.036	0.041	0.046	0.050	0.053
$\vert L\vert$ =100	0.007	0.009	0.014	0.015	0.025	0.019	0.028	0.029	0.029	0.028
$\vert L\vert$ =250	0.006	0.010	0.017	0.021	0.025	0.029	0.027	0.038	0.036	0.040
$\vert L\vert$ =500	0.011	0.019	0.028	0.037	0.041	0.053	0.062	0.069	0.075	0.080

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\vert L\vert$ =25	0.014	0.025	0.039	0.055	0.066	0.077	0.096	0.103	0.124	0.130
$\vert L\vert$ =50	0.018	0.035	0.051	0.070	0.085	0.104	0.120	0.135	0.150	0.171
$\vert L\vert$ =100	0.020	0.037	0.057	0.072	0.099	0.114	0.135	0.151	0.170	0.185
$\vert L\vert$ =250	0.018	0.039	0.054	0.069	0.087	0.100	0.128	0.139	0.158	0.178
$\vert L\vert$ =500	0.016	0.031	0.043	0.065	0.075	0.087	0.107	0.120	0.139	0.154

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\vert L\vert$ =25	0.005	0.007	0.011	0.013	0.014	0.016	0.022	0.026	0.031	0.051
$\vert L\vert$ =50	0.007	0.007	0.010	0.014	0.018	0.016	0.023	0.027	0.034	0.038
$\vert L\vert$ =100	0.006	0.007	0.010	0.012	0.020	0.017	0.019	0.026	0.034	0.030
$\vert L\vert$ =250	0.005	0.006	0.009	0.009	0.015	0.014	0.023	0.026	0.039	0.022
$\vert L\vert$ =500	0.004	0.007	0.010	0.010	0.011	0.013	0.019	0.022	0.019	0.019

Table C.69.: Root mean squared error of gross visits under CDMAR on gender (female) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	13.047	16.969	19.639	28.324	29.429	32.430	32.119	39.728	37.590	41.743
$\|L\|$ =50	18.055	25.131	24.251	29.777	32.424	29.875	42.696	41.460	52.367	51.796
$\|L\|$ =100	34.910	46.180	50.189	56.567	55.074	63.316	71.997	49.776	66.005	46.340
$\|L\|$ =250	100.161	134.774	105.438	126.030	89.300	114.882	178.125	134.727	144.285	167.856
$\|L\|$ =500	214.702	224.130	182.044	225.602	193.337	187.472	192.712	275.382	235.217	381.768

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	37.780	69.193	60.173	133.929	65.765	70.110	76.135	86.083	149.514	176.186
$\|L\|$ =50	46.862	67.271	128.317	176.087	149.929	161.803	207.791	339.885	185.690	270.446
$\|L\|$ =100	175.239	162.620	172.612	393.861	280.918	735.298	476.651	343.315	341.739	497.562
$\|L\|$ =250	600.551	363.971	368.379	883.815	640.667	459.723	504.330	1146.662	1826.365	1385.415
$\|L\|$ =500	507.903	1524.693	747.660	658.344	1791.602	1117.420	1564.751	1610.604	1140.896	2225.583

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	13.147	19.186	18.228	20.119	21.637	22.073	28.755	25.213	28.698	27.939
$\|L\|$ =50	29.550	32.913	33.628	32.941	36.635	26.163	46.038	47.089	47.864	46.702
$\|L\|$ =100	47.215	52.588	67.100	60.978	57.360	72.015	59.691	80.531	72.763	65.299
$\|L\|$ =250	109.624	126.550	125.702	136.562	123.975	129.462	157.879	135.336	135.597	171.948
$\|L\|$ =500	167.994	224.747	187.712	210.143	212.334	198.438	208.133	310.240	294.969	370.831

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.70.: Root mean squared error of average visits under CDMAR on gender (female) without sociodemographic variables

$\|L\|$ =25	0.288	0.368	0.382	0.383	0.449	0.493	0.519	0.541	0.584	0.590
$\|L\|$ =50	0.327	0.342	0.395	0.442	0.497	0.541	0.586	0.651	0.617	0.692
$\|L\|$ =100	0.381	0.498	0.553	0.632	0.727	0.731	0.826	0.850	0.824	0.948
$\|L\|$ =250	0.692	0.856	0.821	1.012	0.894	1.090	1.300	1.317	1.386	1.481
$\|L\|$ =500	1.105	1.134	1.186	1.263	1.398	1.612	1.452	1.854	1.586	1.979
MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.297	0.486	0.466	1.083	0.461	0.578	0.645	0.678	0.969	1.273
$\|L\|$ =50	0.216	0.310	0.596	0.867	0.666	0.831	1.017	1.564	0.857	1.247
$\|L\|$ =100	0.653	0.652	0.657	1.419	1.007	2.571	1.693	1.244	1.141	1.725
$\|L\|$ =250	1.932	1.362	1.478	2.661	2.029	1.772	2.184	3.423	4.989	3.887
$\|L\|$ =500	2.026	4.350	2.844	2.771	4.922	3.862	4.645	4.695	3.837	5.853

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.328	0.379	0.449	0.601	0.663	0.726	0.823	0.913	1.075	1.200
$\|L\|$ =50	0.401	0.531	0.649	0.688	0.770	0.912	0.982	1.102	1.287	1.384
$\|L\|$ =100	0.580	0.646	0.787	0.889	0.982	1.221	1.271	1.506	1.690	1.760
$\|L\|$ =250	0.776	0.995	1.270	1.485	1.802	1.919	2.038	2.196	2.553	2.903
$\|L\|$ =500	1.244	1.641	1.908	2.379	2.572	2.696	3.364	3.509	3.964	4.271

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.71.: Root mean squared error of entity coverage under CDMAR on gender (female) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.045	0.056	0.068	0.076	0.085	0.094	0.101	0.113	0.122	0.127
$\|L\|$ =50	0.045	0.051	0.062	0.072	0.083	0.091	0.103	0.110	0.121	0.131
$\|L\|$ =100	0.037	0.046	0.052	0.060	0.070	0.077	0.087	0.094	0.099	0.109
$\|L\|$ =250	0.027	0.034	0.039	0.046	0.049	0.058	0.062	0.068	0.072	0.077
$\|L\|$ =500	0.021	0.024	0.030	0.036	0.038	0.044	0.049	0.052	0.055	0.062

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.010	0.013	0.014	0.016	0.019	0.019	0.020	0.022	0.023	0.025
$\|L\|$ =50	0.009	0.012	0.015	0.016	0.018	0.019	0.015	0.022	0.022	0.021
$\|L\|$ =100	0.013	0.014	0.016	0.016	0.020	0.020	0.023	0.023	0.027	0.028
$\|L\|$ =250	0.026	0.032	0.038	0.043	0.047	0.056	0.057	0.065	0.069	0.074
$\|L\|$ =500	0.034	0.038	0.047	0.056	0.060	0.069	0.077	0.083	0.090	0.097

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.036	0.044	0.049	0.057	0.066	0.074	0.085	0.088	0.097	0.107
$\|L\|$ =50	0.046	0.057	0.066	0.080	0.087	0.099	0.102	0.117	0.122	0.133
$\|L\|$ =100	0.051	0.062	0.070	0.085	0.090	0.105	0.111	0.131	0.138	0.149
$\|L\|$ =250	0.047	0.056	0.068	0.076	0.088	0.098	0.108	0.115	0.129	0.141
$\|L\|$ =500	0.040	0.050	0.059	0.066	0.075	0.085	0.094	0.102	0.108	0.120

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.008	0.009	0.012	0.015	0.011	0.017	0.021	0.016	0.021	0.023
$\|L\|$ =50	0.010	0.010	0.014	0.014	0.014	0.018	0.020	0.026	0.023	0.024
$\|L\|$ =100	0.009	0.011	0.010	0.013	0.013	0.015	0.016	0.025	0.021	0.028
$\|L\|$ =250	0.007	0.011	0.011	0.010	0.014	0.016	0.012	0.014	0.014	0.022
$\|L\|$ =500	0.005	0.006	0.008	0.011	0.010	0.012	0.015	0.015	0.017	0.020

Table C.72.: Root mean squared error of gross visits under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	22.067	21.465	25.772	27.109	30.927	28.252	28.512	37.467	33.269	36.857
$\|L\|$ =50	30.115	25.464	32.101	31.792	35.563	36.320	35.272	35.426	27.845	42.787
$\|L\|$ =100	42.681	43.009	39.677	43.130	46.111	53.113	73.728	66.248	97.519	105.685
$\|L\|$ =250	61.314	86.903	104.514	112.470	115.948	151.028	147.170	204.094	186.595	204.652
$\|L\|$ =500	182.132	188.322	182.761	178.679	304.287	280.432	318.805	327.911	359.884	389.458

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	41.036	38.956	72.416	80.280	102.668	86.807	162.589	111.827	124.143	259.266
$\|L\|$ =50	50.779	174.989	82.107	353.455	155.801	169.478	131.039	96.134	312.318	231.186
$\|L\|$ =100	113.742	146.114	279.564	221.005	235.440	291.776	415.161	372.849	363.502	446.562
$\|L\|$ =250	248.468	297.470	336.859	701.340	608.854	512.382	596.401	772.262	704.575	1849.302
$\|L\|$ =500	422.601	616.457	891.091	963.559	1193.108	1320.472	1765.571	2129.308	1695.889	1484.035

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	15.326	14.903	14.311	19.978	22.439	22.956	20.280	22.501	27.474	29.434
$\|L\|$ =50	21.555	30.091	32.882	29.956	43.633	34.386	42.797	38.858	35.968	49.794
$\|L\|$ =100	42.402	47.811	56.693	65.881	63.749	54.490	75.114	71.853	92.698	79.133
$\|L\|$ =250	84.945	112.499	111.590	106.199	131.361	159.977	160.517	149.613	149.124	195.882
$\|L\|$ =500	191.368	168.484	213.999	233.608	286.699	292.315	310.668	230.384	279.740	312.084

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.73.: Root mean squared error of average visits under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.254	0.296	0.380	0.396	0.444	0.487	0.505	0.540	0.589	0.650
$\|L\|$ =50	0.250	0.332	0.378	0.446	0.484	0.521	0.592	0.641	0.693	0.772
$\|L\|$ =100	0.378	0.452	0.490	0.542	0.637	0.736	0.868	0.829	0.975	1.121
$\|L\|$ =250	0.540	0.672	0.797	0.968	0.988	1.128	1.223	1.509	1.455	1.579
$\|L\|$ =500	0.699	0.870	1.216	1.288	1.626	1.650	1.967	2.080	2.166	2.377

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.301	0.294	0.568	0.593	0.788	0.678	1.373	0.897	0.916	1.885
$\|L\|$ =50	0.238	0.915	0.355	1.966	0.720	0.805	0.631	0.446	1.646	1.090
$\|L\|$ =100	0.515	0.587	1.038	0.840	0.826	1.027	1.457	1.255	1.309	1.558
$\|L\|$ =250	1.206	1.416	1.516	2.051	1.939	1.829	1.966	2.304	2.125	5.195
$\|L\|$ =500	2.304	2.773	3.127	3.290	4.033	3.963	4.623	5.464	4.394	4.126

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.345	0.411	0.467	0.585	0.667	0.804	0.846	0.955	1.061	1.139
$\|L\|$ =50	0.422	0.504	0.646	0.697	0.841	0.915	1.035	1.081	1.307	1.362
$\|L\|$ =100	0.542	0.646	0.856	0.947	1.088	1.162	1.232	1.519	1.642	1.702
$\|L\|$ =250	0.906	1.127	1.208	1.386	1.713	1.816	2.231	2.136	2.574	2.849
$\|L\|$ =500	1.562	1.707	1.892	2.342	2.222	2.937	3.246	3.278	3.806	3.887

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.74.: Root mean squared error of entity coverage under CDMAR on occupation (employed) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.049	0.056	0.068	0.075	0.085	0.091	0.095	0.112	0.117	0.124
$\|L\|$ =50	0.051	0.055	0.065	0.072	0.086	0.094	0.099	0.111	0.111	0.120
$\|L\|$ =100	0.047	0.051	0.057	0.065	0.069	0.078	0.080	0.088	0.089	0.096
$\|L\|$ =250	0.035	0.040	0.042	0.048	0.049	0.053	0.058	0.063	0.066	0.069
$\|L\|$ =500	0.030	0.032	0.034	0.034	0.039	0.040	0.045	0.045	0.049	0.053

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.010	0.013	0.010	0.014	0.017	0.023	0.019	0.019	0.021	0.024
$\|L\|$ =50	0.011	0.014	0.015	0.012	0.020	0.020	0.018	0.018	0.022	0.026
$\|L\|$ =100	0.017	0.017	0.018	0.021	0.018	0.021	0.016	0.023	0.020	0.016
$\|L\|$ =250	0.035	0.040	0.041	0.046	0.046	0.049	0.054	0.056	0.060	0.064
$\|L\|$ =500	0.045	0.049	0.055	0.054	0.062	0.063	0.071	0.073	0.077	0.084

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.036	0.045	0.050	0.060	0.066	0.080	0.084	0.086	0.092	0.105
$\|L\|$ =50	0.048	0.060	0.068	0.075	0.085	0.093	0.102	0.107	0.124	0.134
$\|L\|$ =100	0.050	0.063	0.074	0.082	0.091	0.102	0.112	0.124	0.141	0.144
$\|L\|$ =250	0.050	0.061	0.067	0.074	0.087	0.094	0.103	0.117	0.123	0.131
$\|L\|$ =500	0.043	0.049	0.064	0.069	0.075	0.089	0.088	0.098	0.108	0.113

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.009	0.012	0.011	0.016	0.012	0.019	0.015	0.015	0.023	0.022
$\|L\|$ =50	0.010	0.013	0.016	0.009	0.016	0.013	0.016	0.017	0.022	0.027
$\|L\|$ =100	0.010	0.012	0.012	0.014	0.010	0.014	0.015	0.021	0.028	0.027
$\|L\|$ =250	0.011	0.012	0.011	0.013	0.012	0.012	0.015	0.015	0.014	0.022
$\|L\|$ =500	0.011	0.011	0.010	0.010	0.013	0.013	0.010	0.013	0.012	0.015

Table C.75.: Root mean squared error of gross visits under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	19.934	20.568	21.795	25.308	30.381	29.471	31.265	36.211	32.382	38.969
$\lvert L\rvert$ =50	21.569	32.719	32.516	32.413	31.142	41.513	32.043	50.724	40.264	46.676
$\lvert L\rvert$ =100	31.639	35.517	47.831	47.244	41.239	53.588	76.737	71.607	78.916	85.724
$\lvert L\rvert$ =250	75.389	72.785	91.548	89.493	87.819	158.419	159.535	111.535	157.137	172.162
$\lvert L\rvert$ =500	133.385	157.734	180.130	193.009	238.699	248.661	273.708	320.875	304.819	317.140

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	24.536	29.830	61.472	60.162	77.161	145.484	75.671	323.932	58.476	70.981
$\lvert L\rvert$ =50	41.483	44.985	105.449	74.163	98.220	87.761	146.224	93.627	136.516	105.123
$\lvert L\rvert$ =100	102.408	114.622	149.933	136.130	119.742	244.261	261.368	252.435	326.383	254.466
$\lvert L\rvert$ =250	238.184	279.511	443.040	396.874	446.172	925.218	336.045	463.708	349.327	365.788
$\lvert L\rvert$ =500	236.907	396.108	504.372	480.771	492.457	764.471	828.694	679.146	710.294	857.319

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	17.376	14.980	17.580	21.249	21.009	21.841	25.459	27.176	27.077	29.495
$\lvert L\rvert$ =50	20.357	28.086	33.894	27.212	40.615	40.877	32.326	52.812	53.554	54.015
$\lvert L\rvert$ =100	41.918	46.508	52.177	51.917	40.920	70.526	98.633	77.534	70.924	78.897
$\lvert L\rvert$ =250	93.684	121.949	90.215	141.148	70.821	157.914	160.593	120.229	148.903	154.251
$\lvert L\rvert$ =500	183.131	190.868	241.741	158.003	253.291	265.432	276.262	305.506	350.826	355.101

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ = 25	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ = 50	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =100	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =250	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =500	–	–	–	–	–	–	–	–	–	–

Table C.76.: Root mean squared error of average visits under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.269	0.310	0.373	0.384	0.418	0.510	0.533	0.589	0.616	0.611
$\|L\|$ =50	0.277	0.332	0.418	0.442	0.516	0.521	0.606	0.647	0.721	0.745
$\|L\|$ =100	0.376	0.458	0.529	0.559	0.643	0.713	0.857	0.903	0.978	1.028
$\|L\|$ =250	0.567	0.733	0.837	0.865	0.924	1.167	1.206	1.185	1.371	1.427
$\|L\|$ =500	0.734	1.138	1.200	1.371	1.439	1.348	1.743	1.852	1.912	1.881

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.166	0.233	0.453	0.443	0.590	1.168	0.608	2.452	0.466	0.624
$\|L\|$ =50	0.136	0.178	0.511	0.318	0.487	0.393	0.779	0.414	0.742	0.544
$\|L\|$ =100	0.325	0.422	0.505	0.551	0.460	0.941	0.949	0.894	1.282	0.897
$\|L\|$ =250	0.821	0.916	1.592	1.302	1.469	2.918	1.277	1.442	1.200	1.201
$\|L\|$ =500	1.268	1.900	2.164	2.051	2.267	2.381	2.692	2.796	2.940	2.574

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.314	0.385	0.458	0.609	0.667	0.701	0.831	0.960	0.972	1.207
$\|L\|$ =50	0.414	0.490	0.578	0.654	0.795	0.914	1.008	1.157	1.259	1.400
$\|L\|$ =100	0.576	0.705	0.756	0.996	1.036	1.210	1.291	1.441	1.611	1.818
$\|L\|$ =250	1.007	1.121	1.210	1.507	1.691	1.839	2.179	2.381	2.514	2.893
$\|L\|$ =500	1.637	1.667	1.933	2.199	2.430	2.998	2.997	3.288	3.744	4.383

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.77.: Root mean squared error of entity coverage under CDMAR on occupation (employed) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.049	0.059	0.067	0.075	0.083	0.095	0.104	0.115	0.119	0.128
$\|L\|$ =50	0.048	0.061	0.069	0.077	0.080	0.093	0.101	0.110	0.116	0.124
$\|L\|$ =100	0.043	0.049	0.057	0.063	0.073	0.078	0.083	0.089	0.092	0.101
$\|L\|$ =250	0.037	0.038	0.042	0.048	0.053	0.053	0.058	0.064	0.067	0.072
$\|L\|$ =500	0.027	0.031	0.034	0.038	0.036	0.041	0.047	0.045	0.050	0.054

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.011	0.013	0.014	0.018	0.021	0.022	0.020	0.025	0.026	0.029
$\|L\|$ =50	0.017	0.015	0.019	0.020	0.027	0.027	0.027	0.028	0.034	0.035
$\|L\|$ =100	0.013	0.013	0.014	0.018	0.015	0.018	0.019	0.021	0.020	0.021
$\|L\|$ =250	0.016	0.021	0.022	0.026	0.031	0.028	0.033	0.037	0.040	0.042
$\|L\|$ =500	0.026	0.031	0.034	0.041	0.040	0.045	0.053	0.054	0.059	0.066

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.036	0.043	0.051	0.061	0.069	0.075	0.080	0.089	0.096	0.104
$\|L\|$ =50	0.050	0.055	0.066	0.075	0.089	0.094	0.105	0.111	0.125	0.138
$\|L\|$ =100	0.053	0.064	0.075	0.087	0.090	0.099	0.116	0.123	0.136	0.145
$\|L\|$ =250	0.048	0.058	0.071	0.079	0.089	0.096	0.105	0.112	0.126	0.133
$\|L\|$ =500	0.044	0.054	0.060	0.068	0.076	0.080	0.089	0.101	0.110	0.114

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.010	0.009	0.012	0.012	0.016	0.015	0.017	0.020	0.020	0.025
$\|L\|$ =50	0.012	0.010	0.012	0.013	0.016	0.017	0.016	0.022	0.029	0.024
$\|L\|$ =100	0.010	0.011	0.014	0.013	0.014	0.017	0.022	0.017	0.028	0.023
$\|L\|$ =250	0.009	0.010	0.013	0.012	0.013	0.013	0.015	0.017	0.027	0.016
$\|L\|$ =500	0.009	0.007	0.010	0.013	0.011	0.013	0.016	0.015	0.024	0.012

Table C.78.: Root mean squared error of gross visits under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	36.185	34.199	31.382	33.569	26.492	27.622	23.030	23.144	27.427	20.272
$\|L\|$ =50	49.282	42.739	42.038	49.114	36.250	30.037	27.287	38.976	35.555	42.219
$\|L\|$ =100	72.138	62.379	60.686	30.249	66.760	59.534	66.416	102.462	124.872	131.381
$\|L\|$ =250	135.539	96.349	78.859	91.689	107.503	159.477	180.142	207.049	322.189	324.787
$\|L\|$ =500	296.226	233.054	212.928	157.372	254.548	292.492	349.823	421.501	489.938	516.540

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	43.297	50.028	92.067	56.536	62.175	78.685	131.990	170.549	143.847	160.842
$\|L\|$ =50	50.026	58.523	82.669	108.542	204.441	96.126	151.975	351.000	369.972	461.604
$\|L\|$ =100	383.445	123.343	141.226	350.906	234.610	298.982	294.909	509.338	455.828	498.745
$\|L\|$ =250	273.701	316.518	632.361	468.475	510.502	500.850	511.483	825.436	1191.934	1526.763
$\|L\|$ =500	446.570	842.845	752.501	1087.924	1516.190	1616.844	1103.845	1850.416	2459.452	2308.180

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	16.629	18.821	17.783	21.116	17.881	23.487	21.021	23.571	33.169	30.075
$\|L\|$ =50	33.074	33.566	31.719	45.330	36.236	46.946	32.373	51.303	56.150	40.890
$\|L\|$ =100	48.827	47.539	59.427	46.634	81.589	65.494	57.769	77.466	86.224	78.132
$\|L\|$ =250	122.071	83.106	108.448	132.719	128.714	161.897	178.318	143.560	147.415	200.091
$\|L\|$ =500	196.238	209.125	228.742	164.353	315.708	317.188	315.875	350.025	297.447	278.268

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\| = 25$	–	–	–	–	–	–	–	–	–	–
$\|L\| = 50$	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.79.: Root mean squared error of average visits under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.308	0.335	0.397	0.397	0.453	0.482	0.508	0.521	0.539	0.573
$\lvert L\rvert$ =50	0.334	0.375	0.382	0.450	0.476	0.517	0.573	0.630	0.630	0.667
$\lvert L\rvert$ =100	0.349	0.460	0.484	0.612	0.722	0.653	0.800	0.906	1.014	1.071
$\lvert L\rvert$ =250	0.498	0.638	0.682	0.873	0.990	1.177	1.274	1.367	1.703	1.679
$\lvert L\rvert$ =500	0.667	0.900	1.130	1.191	1.285	1.789	2.017	2.189	2.391	2.495

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.289	0.312	0.726	0.434	0.454	0.582	0.917	1.303	1.169	1.371
$\lvert L\rvert$ =50	0.277	0.288	0.399	0.494	0.957	0.453	0.743	1.797	1.812	2.248
$\lvert L\rvert$ =100	1.396	0.616	0.597	1.302	0.862	1.080	1.036	1.806	1.625	1.794
$\lvert L\rvert$ =250	1.274	1.659	2.017	1.637	1.824	1.790	1.712	2.400	3.475	4.254
$\lvert L\rvert$ =500	2.601	3.307	2.832	3.697	4.191	4.423	3.475	5.070	6.119	5.684

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	0.393	0.461	0.501	0.601	0.623	0.725	0.784	0.903	0.959	0.992
$\lvert L\rvert$ =50	0.517	0.558	0.688	0.729	0.791	0.938	0.912	1.096	1.147	1.208
$\lvert L\rvert$ =100	0.796	0.816	0.933	0.934	1.013	1.178	1.226	1.288	1.423	1.488
$\lvert L\rvert$ =250	1.390	1.318	1.587	1.531	1.755	1.766	1.988	2.052	1.946	2.376
$\lvert L\rvert$ =500	2.149	2.031	2.406	2.212	2.767	2.552	2.620	2.974	3.185	3.287

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ = 25	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ = 50	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =100	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =250	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =500	–	–	–	–	–	–	–	–	–	–

Table C.80.: Root mean squared error of entity coverage under MAR on travel group (high) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.075	0.076	0.081	0.085	0.084	0.085	0.092	0.092	0.093	0.086
$\|L\|$ =50	0.075	0.076	0.079	0.084	0.082	0.084	0.084	0.086	0.089	0.083
$\|L\|$ =100	0.062	0.067	0.069	0.072	0.069	0.070	0.072	0.067	0.072	0.074
$\|L\|$ =250	0.047	0.050	0.049	0.049	0.051	0.050	0.052	0.051	0.050	0.050
$\|L\|$ =500	0.038	0.037	0.041	0.037	0.039	0.042	0.042	0.040	0.040	0.042

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.014	0.013	0.012	0.014	0.018	0.020	0.020	0.024	0.026	0.033
$\|L\|$ =50	0.014	0.014	0.013	0.015	0.014	0.017	0.019	0.023	0.025	0.033
$\|L\|$ =100	0.023	0.024	0.023	0.023	0.017	0.019	0.018	0.014	0.015	0.018
$\|L\|$ =250	0.047	0.050	0.048	0.048	0.049	0.046	0.047	0.048	0.043	0.043
$\|L\|$ =500	0.061	0.060	0.062	0.060	0.062	0.063	0.066	0.062	0.063	0.065

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.042	0.047	0.055	0.059	0.068	0.074	0.077	0.083	0.091	0.100
$\|L\|$ =50	0.058	0.065	0.071	0.076	0.084	0.090	0.097	0.107	0.111	0.121
$\|L\|$ =100	0.071	0.074	0.078	0.084	0.096	0.100	0.104	0.116	0.120	0.128
$\|L\|$ =250	0.069	0.072	0.081	0.079	0.087	0.089	0.094	0.097	0.108	0.111
$\|L\|$ =500	0.062	0.065	0.072	0.072	0.074	0.081	0.085	0.086	0.089	0.093

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.018	0.012	0.013	0.011	0.014	0.015	0.016	0.022	0.026	0.030
$\|L\|$ =50	0.017	0.014	0.014	0.014	0.013	0.015	0.014	0.019	0.026	0.033
$\|L\|$ =100	0.013	0.015	0.017	0.012	0.012	0.017	0.016	0.021	0.021	0.022
$\|L\|$ =250	0.014	0.013	0.014	0.011	0.013	0.012	0.013	0.015	0.017	0.017
$\|L\|$ =500	0.013	0.010	0.013	0.011	0.012	0.012	0.012	0.010	0.011	0.010

Table C.81.: Root mean squared error of gross visits under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	35.678	36.663	26.890	27.959	31.434	27.442	25.061	26.003	27.166	25.905
$\|L\|$ =50	49.064	50.391	46.123	34.992	35.575	26.567	29.426	41.623	35.191	49.893
$\|L\|$ =100	65.710	56.324	45.492	50.849	41.688	77.389	68.065	86.585	110.945	120.396
$\|L\|$ =250	125.926	100.704	99.357	98.918	106.036	164.855	172.979	251.207	248.591	288.981
$\|L\|$ =500	268.462	228.840	236.471	193.844	228.773	258.650	333.609	496.774	487.342	544.582

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	24.996	30.187	53.304	32.536	42.420	117.580	242.353	70.697	142.166	131.858
$\|L\|$ =50	38.148	135.534	74.635	67.463	78.819	79.786	122.544	236.757	218.158	103.850
$\|L\|$ =100	79.002	87.654	131.487	146.543	197.207	171.486	221.019	246.583	186.753	449.750
$\|L\|$ =250	274.185	223.077	433.049	269.861	508.170	271.051	389.804	899.479	621.005	332.418
$\|L\|$ =500	309.479	331.647	722.343	446.237	541.999	736.729	936.770	1367.630	907.848	963.605

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	15.865	17.025	21.572	21.657	20.268	22.870	26.692	25.694	28.745	25.010
$\|L\|$ =50	21.976	36.041	28.799	32.185	38.602	34.974	35.494	41.503	47.090	48.715
$\|L\|$ =100	48.534	46.586	54.488	60.772	55.401	82.414	70.744	70.227	66.831	98.813
$\|L\|$ =250	112.611	97.901	128.718	110.152	104.495	125.464	146.573	199.280	177.116	165.955
$\|L\|$ =500	184.489	149.480	208.208	231.383	226.351	288.463	263.807	368.344	386.660	320.256

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.82.: Root mean squared error of average visits under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.337	0.365	0.415	0.410	0.441	0.475	0.511	0.522	0.534	0.556
$\|L\|$ =50	0.334	0.348	0.414	0.456	0.472	0.517	0.567	0.617	0.661	0.700
$\|L\|$ =100	0.408	0.455	0.543	0.579	0.659	0.771	0.803	0.850	0.918	0.977
$\|L\|$ =250	0.505	0.548	0.688	0.907	1.004	1.191	1.203	1.521	1.483	1.605
$\|L\|$ =500	0.650	0.670	1.049	0.961	1.372	1.542	1.885	2.354	2.291	2.608

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.168	0.207	0.378	0.244	0.357	0.900	1.828	0.599	1.223	1.135
$\|L\|$ =50	0.175	0.622	0.361	0.268	0.311	0.332	0.607	1.323	1.223	0.639
$\|L\|$ =100	0.360	0.365	0.452	0.542	0.754	0.630	0.772	0.993	0.772	1.733
$\|L\|$ =250	1.112	1.027	1.447	1.194	1.586	1.072	1.291	2.612	1.839	1.063
$\|L\|$ =500	1.856	2.062	2.566	2.195	2.137	2.637	2.783	4.017	2.655	2.933

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.356	0.428	0.497	0.573	0.640	0.756	0.786	0.881	0.970	1.109
$\|L\|$ =50	0.488	0.598	0.616	0.663	0.762	0.888	1.003	1.015	1.122	1.178
$\|L\|$ =100	0.698	0.768	0.850	1.012	1.036	1.099	1.230	1.452	1.453	1.573
$\|L\|$ =250	1.239	1.332	1.470	1.425	1.571	1.734	2.007	1.846	2.325	2.232
$\|L\|$ =500	2.016	2.148	2.236	2.599	2.675	2.866	2.835	2.862	3.296	3.167

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.83.: Root mean squared error of entity coverage under MAR on travel group (high) with sociodemographic variable occupation

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.076	0.081	0.081	0.080	0.086	0.089	0.090	0.092	0.089	0.098
$\|L\|$ =50	0.077	0.078	0.083	0.083	0.082	0.083	0.087	0.083	0.085	0.087
$\|L\|$ =100	0.067	0.067	0.064	0.070	0.069	0.070	0.073	0.071	0.068	0.072
$\|L\|$ =250	0.046	0.047	0.048	0.050	0.051	0.052	0.050	0.055	0.051	0.050
$\|L\|$ =500	0.038	0.039	0.040	0.036	0.039	0.040	0.040	0.038	0.041	0.041

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.013	0.013	0.013	0.019	0.019	0.023	0.024	0.031	0.034	0.034
$\|L\|$ =50	0.010	0.015	0.012	0.019	0.024	0.029	0.032	0.037	0.042	0.043
$\|L\|$ =100	0.013	0.011	0.016	0.014	0.014	0.017	0.022	0.025	0.029	0.034
$\|L\|$ =250	0.030	0.028	0.030	0.026	0.028	0.026	0.021	0.027	0.019	0.022
$\|L\|$ =500	0.043	0.043	0.044	0.040	0.041	0.043	0.040	0.040	0.040	0.044

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.040	0.047	0.056	0.064	0.067	0.072	0.079	0.086	0.093	0.097
$\|L\|$ =50	0.058	0.065	0.069	0.079	0.087	0.093	0.101	0.107	0.115	0.120
$\|L\|$ =100	0.066	0.072	0.082	0.083	0.092	0.099	0.107	0.118	0.119	0.127
$\|L\|$ =250	0.067	0.072	0.077	0.081	0.086	0.090	0.097	0.100	0.103	0.114
$\|L\|$ =500	0.061	0.067	0.071	0.075	0.080	0.081	0.084	0.088	0.089	0.093

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.017	0.014	0.013	0.012	0.013	0.017	0.018	0.022	0.025	0.027
$\|L\|$ =50	0.015	0.014	0.016	0.018	0.014	0.017	0.019	0.023	0.027	0.027
$\|L\|$ =100	0.016	0.014	0.013	0.013	0.017	0.017	0.016	0.022	0.021	0.023
$\|L\|$ =250	0.012	0.011	0.014	0.015	0.012	0.015	0.012	0.017	0.014	0.014
$\|L\|$ =500	0.012	0.012	0.013	0.010	0.012	0.012	0.010	0.009	0.012	0.012

Table C.84.: Root mean squared error of gross visits under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	29.031	27.947	29.151	28.002	23.459	25.762	32.086	24.187	28.599	25.247
$\|L\|$ =50	40.104	38.805	28.313	37.355	29.709	32.305	32.935	42.272	38.695	36.940
$\|L\|$ =100	35.781	44.881	48.805	45.858	53.262	37.641	74.026	73.446	87.743	83.851
$\|L\|$ =250	51.261	76.578	93.226	91.564	131.336	137.856	166.234	199.788	162.942	211.046
$\|L\|$ =500	140.458	154.017	169.540	227.897	263.866	305.230	310.689	238.699	294.647	325.472

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	34.968	55.659	107.727	86.387	68.249	33.935	60.502	61.515	157.888	119.988
$\|L\|$ =50	43.188	67.626	83.911	66.587	105.488	306.800	99.461	94.474	91.432	192.826
$\|L\|$ =100	89.926	308.942	74.042	362.074	174.094	194.074	257.572	148.485	178.846	119.601
$\|L\|$ =250	161.291	206.487	247.953	207.606	252.261	309.452	580.273	492.576	453.306	861.926
$\|L\|$ =500	511.506	369.129	492.338	847.578	528.597	690.151	606.976	621.206	1049.145	715.017

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	18.609	13.605	15.726	15.887	17.980	19.078	27.929	19.428	28.441	29.763
$\|L\|$ =50	25.042	34.281	24.158	33.243	31.463	37.647	49.640	47.340	37.433	54.230
$\|L\|$ =100	41.269	53.201	48.537	61.154	64.901	60.506	77.295	68.326	70.965	73.999
$\|L\|$ =250	75.619	109.357	115.497	120.766	144.667	136.250	141.273	184.949	162.654	153.595
$\|L\|$ =500	207.716	204.583	236.744	264.428	307.572	352.472	291.964	340.676	242.595	295.655

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.85.: Root mean squared error of average visits under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.316	0.369	0.380	0.399	0.457	0.483	0.491	0.542	0.548	0.585
$\|L\|$ =50	0.341	0.362	0.404	0.439	0.476	0.516	0.517	0.659	0.633	0.649
$\|L\|$ =100	0.481	0.497	0.515	0.587	0.637	0.649	0.755	0.827	0.931	0.906
$\|L\|$ =250	0.718	0.826	0.954	0.947	0.965	1.059	1.219	1.320	1.236	1.442
$\|L\|$ =500	1.040	1.047	1.257	1.253	1.517	1.704	1.658	1.528	1.667	1.873

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.244	0.471	0.776	0.651	0.620	0.295	0.462	0.562	1.299	1.028
$\|L\|$ =50	0.177	0.313	0.407	0.258	0.502	1.567	0.575	0.492	0.597	1.150
$\|L\|$ =100	0.356	1.145	0.294	1.356	0.673	0.685	0.961	0.501	0.709	0.517
$\|L\|$ =250	1.030	1.093	1.219	1.003	1.105	1.131	1.828	1.551	1.399	2.646
$\|L\|$ =500	2.151	2.031	2.428	2.762	2.506	2.626	2.348	2.179	2.919	2.268

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.405	0.389	0.501	0.565	0.628	0.678	0.734	0.857	0.994	1.049
$\|L\|$ =50	0.510	0.599	0.650	0.712	0.764	0.857	1.006	0.947	1.068	1.254
$\|L\|$ =100	0.646	0.819	0.934	0.968	1.105	1.183	1.221	1.338	1.343	1.476
$\|L\|$ =250	1.185	1.280	1.351	1.557	1.729	1.861	1.877	1.966	2.055	2.090
$\|L\|$ =500	2.250	2.287	2.161	2.688	2.435	2.632	2.927	3.093	3.104	3.149

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.86.: Root mean squared error of entity coverage under MAR on travel group (high) with sociodemographic variable travel group

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.069	0.075	0.077	0.080	0.084	0.088	0.091	0.097	0.102	0.098
$\|L\|$ =50	0.070	0.070	0.072	0.080	0.081	0.085	0.086	0.089	0.092	0.094
$\|L\|$ =100	0.062	0.063	0.066	0.065	0.068	0.070	0.073	0.074	0.078	0.079
$\|L\|$ =250	0.045	0.046	0.048	0.052	0.049	0.051	0.051	0.052	0.053	0.056
$\|L\|$ =500	0.034	0.035	0.036	0.039	0.040	0.041	0.038	0.042	0.041	0.042

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.011	0.015	0.014	0.016	0.020	0.021	0.029	0.026	0.026	0.041
$\|L\|$ =50	0.012	0.019	0.023	0.021	0.030	0.029	0.033	0.042	0.044	0.049
$\|L\|$ =100	0.017	0.014	0.011	0.014	0.021	0.021	0.022	0.025	0.029	0.031
$\|L\|$ =250	0.036	0.032	0.034	0.036	0.032	0.028	0.026	0.026	0.025	0.025
$\|L\|$ =500	0.047	0.045	0.046	0.048	0.048	0.048	0.046	0.050	0.045	0.044

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.041	0.047	0.055	0.061	0.068	0.071	0.080	0.086	0.085	0.100
$\|L\|$ =50	0.057	0.067	0.076	0.077	0.089	0.092	0.100	0.111	0.114	0.124
$\|L\|$ =100	0.063	0.074	0.079	0.086	0.096	0.101	0.108	0.117	0.120	0.124
$\|L\|$ =250	0.068	0.074	0.075	0.078	0.085	0.094	0.098	0.101	0.104	0.108
$\|L\|$ =500	0.064	0.067	0.071	0.074	0.072	0.082	0.086	0.087	0.089	0.096

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.011	0.012	0.010	0.011	0.016	0.015	0.017	0.013	0.026	0.023
$\|L\|$ =50	0.009	0.014	0.015	0.014	0.015	0.014	0.017	0.018	0.027	0.023
$\|L\|$ =100	0.012	0.012	0.013	0.011	0.017	0.015	0.016	0.016	0.023	0.016
$\|L\|$ =250	0.011	0.010	0.012	0.015	0.014	0.010	0.011	0.017	0.010	0.013
$\|L\|$ =500	0.010	0.012	0.011	0.010	0.010	0.011	0.010	0.012	0.010	0.011

Table C.87.: Root mean squared error of gross visits under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	11.769	15.651	23.014	26.790	28.045	32.212	36.402	41.377	45.483	50.744
$\lvert L\rvert$ =50	31.506	25.557	31.290	24.294	42.222	42.539	41.054	57.357	53.849	58.290
$\lvert L\rvert$ =100	79.236	70.489	52.878	51.695	55.933	53.672	59.157	63.104	61.307	79.860
$\lvert L\rvert$ =250	149.775	183.006	176.035	136.446	99.788	120.712	95.552	90.976	133.748	136.608
$\lvert L\rvert$ =500	325.971	329.579	260.172	239.409	240.471	194.544	288.205	236.297	234.631	333.699

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	51.620	283.077	131.938	59.194	110.982	60.869	49.462	65.970	58.945	62.495
$\lvert L\rvert$ =50	96.416	101.929	470.165	160.094	120.339	231.109	151.446	145.356	85.729	98.065
$\lvert L\rvert$ =100	290.478	230.403	557.421	175.100	350.158	301.857	251.599	726.621	295.337	272.992
$\lvert L\rvert$ =250	495.967	557.331	509.242	1419.271	759.303	502.372	521.667	805.566	745.211	729.719
$\lvert L\rvert$ =500	1317.702	1284.793	1328.053	807.452	1658.977	989.427	1249.142	1574.136	1569.455	1117.052

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ =25	18.348	16.984	21.035	23.310	15.981	19.686	23.207	27.331	23.969	22.493
$\lvert L\rvert$ =50	28.278	28.787	43.565	41.629	49.073	39.406	35.506	41.533	35.658	39.572
$\lvert L\rvert$ =100	51.440	54.989	58.718	53.097	68.895	52.605	64.265	58.331	57.317	57.451
$\lvert L\rvert$ =250	112.248	122.152	132.530	151.360	107.155	122.010	109.837	116.261	140.519	154.944
$\lvert L\rvert$ =500	215.990	289.649	292.615	226.019	269.687	224.362	357.791	287.905	237.030	269.946

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\lvert L\rvert$ = 25	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ = 50	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =100	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =250	–	–	–	–	–	–	–	–	–	–
$\lvert L\rvert$ =500	–	–	–	–	–	–	–	–	–	–

Table C.88.: Root mean squared error of average visits under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.364	0.388	0.372	0.420	0.443	0.452	0.511	0.515	0.555	0.545
$\|L\|$ =50	0.376	0.406	0.446	0.478	0.463	0.542	0.576	0.578	0.615	0.629
$\|L\|$ =100	0.547	0.534	0.552	0.608	0.674	0.655	0.773	0.741	0.828	0.861
$\|L\|$ =250	0.743	0.981	1.026	0.964	0.938	1.112	1.054	1.078	1.105	1.206
$\|L\|$ =500	1.306	1.413	1.423	1.367	1.449	1.395	1.409	1.541	1.533	1.585

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.449	2.359	1.106	0.427	0.752	0.453	0.341	0.426	0.407	0.356
$\|L\|$ =50	0.477	0.486	2.215	0.815	0.546	1.052	0.761	0.655	0.396	0.435
$\|L\|$ =100	1.078	0.832	2.073	0.697	1.264	1.132	0.979	2.519	1.185	1.095
$\|L\|$ =250	1.599	1.812	1.730	4.152	2.369	1.966	2.048	2.448	2.593	2.320
$\|L\|$ =500	3.823	3.747	4.114	3.121	4.788	3.767	4.064	4.702	4.716	4.412

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.529	0.599	0.649	0.631	0.666	0.669	0.634	0.721	0.700	0.788
$\|L\|$ =50	0.659	0.663	0.761	0.788	0.854	0.802	0.799	0.916	0.932	0.964
$\|L\|$ =100	0.853	0.966	0.984	1.028	1.052	1.138	1.164	1.287	1.237	1.353
$\|L\|$ =250	1.301	1.335	1.421	1.606	1.804	1.642	1.903	2.105	2.242	2.197
$\|L\|$ =500	1.814	2.082	2.224	2.436	2.639	2.793	3.236	3.278	3.330	3.547

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ = 25	–	–	–	–	–	–	–	–	–	–
$\|L\|$ = 50	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =100	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =250	–	–	–	–	–	–	–	–	–	–
$\|L\|$ =500	–	–	–	–	–	–	–	–	–	–

Table C.89.: Root mean squared error of entity coverage under MAR on travel group (low) without sociodemographic variables

MI-GLM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.049	0.058	0.069	0.080	0.086	0.091	0.106	0.111	0.123	0.127
$\|L\|$ =50	0.038	0.050	0.058	0.069	0.084	0.098	0.107	0.113	0.126	0.133
$\|L\|$ =100	0.032	0.039	0.049	0.057	0.067	0.083	0.088	0.099	0.111	0.124
$\|L\|$ =250	0.021	0.028	0.035	0.043	0.051	0.059	0.067	0.075	0.085	0.093
$\|L\|$ =500	0.015	0.021	0.028	0.033	0.037	0.046	0.053	0.060	0.065	0.072

SI-SVR	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.019	0.020	0.017	0.015	0.018	0.016	0.018	0.019	0.020	0.020
$\|L\|$ =50	0.024	0.018	0.022	0.019	0.018	0.014	0.018	0.019	0.015	0.020
$\|L\|$ =100	0.010	0.012	0.015	0.017	0.019	0.026	0.025	0.032	0.038	0.044
$\|L\|$ =250	0.020	0.026	0.033	0.038	0.048	0.054	0.065	0.073	0.083	0.091
$\|L\|$ =500	0.027	0.036	0.046	0.054	0.059	0.072	0.080	0.091	0.098	0.107

MI-Poisson	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.059	0.062	0.062	0.060	0.068	0.073	0.070	0.074	0.075	0.078
$\|L\|$ =50	0.075	0.075	0.085	0.084	0.084	0.088	0.092	0.098	0.100	0.107
$\|L\|$ =100	0.074	0.081	0.084	0.092	0.098	0.098	0.108	0.104	0.112	0.117
$\|L\|$ =250	0.062	0.075	0.077	0.078	0.090	0.091	0.098	0.102	0.106	0.111
$\|L\|$ =500	0.054	0.061	0.066	0.072	0.078	0.081	0.088	0.093	0.101	0.108

KM	r=0.1	r=0.2	r=0.3	r=0.4	r=0.5	r=0.6	r=0.7	r=0.8	r=0.9	r=1.0
$\|L\|$ =25	0.018	0.018	0.015	0.010	0.013	0.013	0.018	0.022	0.025	0.027
$\|L\|$ =50	0.025	0.017	0.020	0.017	0.018	0.016	0.021	0.021	0.039	0.035
$\|L\|$ =100	0.018	0.020	0.017	0.017	0.013	0.017	0.019	0.027	0.034	0.053
$\|L\|$ =250	0.015	0.013	0.014	0.013	0.013	0.012	0.019	0.028	0.039	0.049
$\|L\|$ =500	0.011	0.011	0.007	0.010	0.011	0.015	0.018	0.026	0.029	0.040

Bibliography

O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer, 2008.

R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM Press, 1993.

J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154, 1984.

L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In *Proc. of the 15th annual ACM international symposium on Advances in geographic information systems (GIS'07)*, pages 1–8. ACM, 2007.

M. Andersson, J. Gudmundsson, P. Laube, and T. Wolle. Reporting leaders and followers among trajectories of moving point objects. *Geoinformatica*, 12(4):497–528, 2008.

G. Andrienko, N. Andrienko, and S. Wrobel. Visual analytics tools for analysis of movement data. *SIGKDD Explor. Newsl.*, 9(2):38–46, 2007.

G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology (VAST'09)*, pages 3–10. IEEE, 2009.

N. Andrienko, G. Andrienko, N. Pelekis, and S. Spaccapietra. Basic concepts of movement data. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 1. Springer, Berlin Heidelberg, 2008.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD 1999)*, pages 49–60. ACM Press, 1999.

A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Journal of Intelligent Data Analysis*, 7(6):541–566, 2003.

A. Appice, M. Berardi, M. Ceci, and D. Malerba. Mining and filtering multi-level spatial association rules with ares. In *15th International Symposium on Foundations of Intelligent Systems (ISMIS)*, pages 342–353. Springer, 2005.

Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma). Project documentation, 2009. http://www.agma-mmc.de/03_forschung/plakat.asp?topnav=10&subnav=199.

Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma). ma Plakat, 2011. URL `http://www.werbestatistik.ch/download.php?id=26_684c49cc`.

A. E. Beaton. *The use of special matrix operations in statistical calculus.* Research Bulletin RB-64-51. Educational Testing Servic, 1964.

R. Benetis, C. S. Jensen, G. Karciauskas, and S. Saltenis. Nearest and reverse nearest neighbor queries for moving objects. *International Journal on Very Large Data Bases (VLDB Journal)*, 15(3):229–249, 2006.

M. Benkert, J. Gudmundsson, F. Hübner, and T. Wolle. Reporting flock patterns. *Computational Geometry: Theory and Applications*, 41(3):111–125, 2008.

J. Biesterfeld, E. Ennigrou, and K. Jobmann. Neural networks for location prediction in mobile networks. In *Proc. of the International Workshop on Applications of Neural Networks to Telecommunications (IWANNT'97)*, pages 207–214, 1997.

BirdTrack, 2011. URL `http://www.birdtrack.net`. [Online; accessed March 2011].

V. Bogorny, J. Valiati, S. Camargo, P. Engel, B. Kuijpers, and L. O. Alvares. Mining maximal generalized frequent geographic patterns with knowledge constraints. In *Proc. of the 6th IEEE International Conference on Data Mining (ICDM)*, pages 813–817. IEEE Computer Society, 2006.

J. Bollmann and W. G. Koch, editors. *Lexikon der Kartographie und Geomatik (in zwei Bänden).* Spektrum Akademischer Verlag, 2001.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. of the 5th Annual Workshop on Computational learning Theory (COLT'92)*, pages 144–152, 1992.

S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk. On map-matching vehicle tracking data. In *Proc. of the 31st International Conference on Very Large Data Bases (VLDB'05)*, pages 853–864, 2005.

I. N. Bronstein, K. A. Semendjaev, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik.* Harri Deutsch, 2001.

M. Buchin, A. Driemel, M. van Kreveld, and V. Sacristán. An algorithmic framework for segmenting trajectories based on spatio-temporal criteria. In *Proc. of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS'10)*, pages 202–211. ACM, 2010.

Bundesministerium für Verkehr, Bau und Stadtentwicklung. Mobilität in Deutschland 2008, Abschlussbericht (Mobility in Germany 2008, final report), 2010. http://www.mobilitaet-in-deutschland.de.

P. A. Burrough and R. A. McDonnell. *Principles of Geographical Information Systems.* Oxford University Press, 2000.

H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. In *Proc. of the 5th IEEE International Conference on Data Mining (ICDM'05)*, pages 82–89. IEEE, 2005.

M. Celik, S. Shekhar, J. P. Rogers, J. A. Shine, and J. S. Yoo. Mixed-drove spatio-temporal co-occurence pattern mining: A summary of results. In *Proceedings of the Sixth International Conference on Data Mining (ICDM '06)*, pages 119–128. IEEE Computer Society, 2006.

M. Celik, S. Shekhar, J. P. Rogers, and J. A. Shine. Mixed-drove spatiotemporal co-occurrence pattern mining. *IEEE Trans. on Knowl. and Data Eng.*, 20(10):1322–1335, 2008.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

S. Chawla, S. Shekhar, W. Wu, and U. Ozesmi. Modelling saptial dependencies for mining geospatial data. In H. J. Miller and J. Han, editors, *Geographic Data Mining and Knowledge Discovery*, chapter 6. Taylor & Francis, London and New York, 2001.

G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller. Max-margin classification of incomplete data. In *Proc. of the 20th Annual Conference on Neural Information Processing Systems (NIPS'06)*. MIT Press, 2007.

C. Cheng, R. Jain, and E. van den Berg. Location prediction algorithms for mobile wireless systems. In B. Furht and M. Ilyas, editors, *Wireless internet handbook: technologies, standards, and application*, pages 245–263. CRC Press, 2003.

J.-P. Chilès and Pierre Delfiner. *Geostatistics - Modeling Spatial Uncertainty.* Wiley & Sons, 1999.

C. Claramunt and B. Jiang. A representation of relationships in temporal spaces. In *Innovations in GIS VII: Geocomputation*, pages 41–53. Taylor & Francis, 2000.

C. Claramunt and B. Jiang. An integrated representation of spatial and temporal relationships between evolving regions. *Geographical Systems*, 3:411–428, 2001.

A. D. Cliff and J. K. Ord. *Spatial autocorrelation.* Pion Limited, London, 1973.

A. D. Cliff and J. K. Ord. The choice of a test for spatial autocorrelation. In J. C. Davis and M. J. McCullagh, editors, *Display and Analysis fo Spatial Data*, pages 54–77. Wiley & Sons, 1975.

L. M. Collins, J. L. Schafer, and C.-M. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2): 187–220, 1972.

N. A. C. Cressie. *Statistics for Spatial Data.* Wiley & Sons, 1993.

A. P. Dempster, N. M. Lairs, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

Destatis. *Statistisches Jahrbuch 2009 - Für die Bundesrepublik Deutschland (Statistical Yearbook 2009 - For the Federal Republic of Germany).* Statistisches Bundesamt (Federal Statistical Office), Wiesbaden, 2009.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2010. URL `http://cran.r-project.org/package=e1071`. R package version 1.5-24.

E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. *Reference Manual: Package e1071*, 2011. URL `http://cran.r-project.org/web/packages/e1071/e1071.pdf`.

H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9 (NIPS'96)*, pages 155–161. MIT Press, 1997.

M. J. Egenhofer. Reasoning about binary topological relations. In O. Günther and H. J. Schek, editors, *Proc. of the 2nd International Symposium on Advances in Spatial Databases (SSD)*, pages 143–160. Springer-Verlag, 1991.

M. J. Egenhofer and J. Herring. Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical report, Department of Surveying Engineering, University of Maine, 1990.

M. J. Egenhofer, E.Clementini, and P. Di Felice. Topological relations between regions with holes. *International Journal of Geographical Information Systems (IJGIS)*, 8(2):129–142, 1994.

M. Erwig, R. H. Güting, M. Schneider, and M. Vazirgiannis. Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3): 269–296, 1999.

M. Ester, J. Sander, H.-P. Kriegel, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

Fachverband Außenwerbung e.V. Netto-Werbeeinnahmen erfassbarer Werbeträger in Deutschland, 2002-2010 (Net turnover of confirmable advertising media in Gemany, 2000-2010), 2011. URL `http://www.faw-ev.de/media/download/marktdaten/4_Nettoumsaetze_aller_Werbemedien_ab_2002.pdf`.

Fachverband Außenwerbung e.V. Press conference "ma 2007 plakat", January 16th, 2008.

L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik*. Springer, 7 edition, 2010.

L. Forlizzi, R. H. Güting, E. Nardelli, and M. Schneider. A data model and data structures for moving objects databases. In *Proc. of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*, pages 319–330. ACM, 2000.

A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression*. Wiley & Sons, 2002.

A. U. Frank. Qualitative spatial reasoning: cardinal directions as an example. *IJGIS*, 10(3): 269–290, 1996.

E. Fretzos, N. Pelekis, I. Ntoutsi, and Y. Theodoridis. Trajectory database systems. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 6. Springer, Berlin Heidelberg, 2008.

H. Fritzsch. *Die verbogene Raum-Zeit: Newton, Einstein und die Gravitation*. Piper, Munich, 2000.

R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5: 115–145, 1954.

F. Giannotti and D. Pedreschi. Mobility, data mining and privacy: A vision of convergence. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, pages 1–11. Springer, Berlin Heidelberg, 2008.

F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *Proc. of the 6th SIAM International Conference on Data Mining (SDM'06)*, pages 346–357. SIAM, 2006.

F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07)*, pages 330–339. ACM, 2007.

M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

J. W. Graham and J. L. Schafer. On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle, editor, *Statistical Strategies for Small Sample Research*, pages 147–177. Sage, 1999.

B. Guc, M. May, Y. Saygin, and C. Körner. Semantic annotation of GPS trajectories. In *Proc. of the 11th AGILE International Conference on Geographic Information Science (AGILE'08)*, 2008.

J. Gudmundsson and M. van Kreveld. Computing longest duration flocks in trajectory data. In *Proc. of the 14th annual ACM International Symposium on Advances in Geographic Information Systems (ACM GIS'06)*, pages 35–42. ACM, 2006.

J. Gudmundsson, M. Kreveld, and B. Speckmann. Efficient detection of patterns in 2D trajectories of moving points. *Geoinformatica*, 11(2):195–215, 2007.

R. H. Güting and M. Schneider. *Moving Objects Databases.* Morgan Kaufmann, 2005.

R. H. Güting, M. H. Böhlen, M. Erwig, C. S. Jensen, N. A. Lorentzos, M. Schneider, and M. Vazirgiannis. A foundation for representing and querying moving objects. *ACM Transactions on Database Systems (TODS)*, 25(1):1–42, 2000.

T. Hägerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24:7–21, 1970.

R. Haining. *Spatial Data Analysis: Theory and Practice.* Cambridge University Press, 2003.

P. Hallot and R. Billen. Life and motion configurations: A basis for spatio-temporal generalized reasoning model. In *Proc. of the ER 2008 Workshops on Advances in Conceptual Modeling: Challenges and Opportunities (ER'08)*, pages 323–333. Springer-Verlag, 2008.

D. Hecker, C. Körner, and M. May. Räumlich differenzierte Reichweiten für die Außenwerbung. In J. Strobl, T. Blaschke, and G. Griesebner, editors, *Angewandte Geoinformatik 2010, Beiträge zum 22. Symposium für Angewandte Geoinformatik (AGIT'10) Salzburg*, pages 194–203. Wichmann, 2010a.

D. Hecker, C. Körner, H. Streich, and U. Hofmann. A sensitivity analysis for the selection of business critical geodata in Swiss outdoor advertisement. In R. Purves and R. Weibel, editors, *GIScience 2010, Extended Abstracts Volume*, 2010b.

D. Hecker, H. Stange, C. Körner, and M. May. Sample bias due to missing data in mobility surveys. In *Proc. of the 2010 IEEE International Conference on Data Mining Workshops (ICDMW'10)*, pages 241–248. IEEE Computer Society, 2010c.

D. Hecker, C Körner, and M. May. Robustness analyses for repeated mobility surveys in outdoor advertising. In *Proc. of the IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM 2011)*, pages 148–153, 2011a.

D. Hecker, C Körner, and M. May. Challenges and advantages of using GPS data in outdoor advertisement. In *Proc. of the 3th Conference on Geoinformatik - Geochange*, pages 257–260. Akademische Verlagsgesellschaft, 2011b.

D. Hecker, C. Körner, H. Stange, D. Schulz, and M. May. Modeling micro-movement variability in mobility studies. In S. Geertman, W. Reinhardt, and F. Toppen, editors, *Advancing Geoinformation Science for a Changing World*, Lecture Notes in Geoinformation and Cartography, pages 121–140. Springer, 2011c.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2001.

D. Hernández. *Qualitative Representation of Spatial Knowledge*. Number 804 in Lecture Notes in Artificial Intelligence. Springer-Verlag, 1994.

J. R. Herring, editor. *OpenGIS Implementation Specification for Geographic Information - Simple feature access - Part 1: Common architecture*, 2006. Open Geospatial Consortium Inc. Number OGC 06-103r3, Version 1.2.0.

K. Hornsby and M. J. Egenhofer. Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1-2):177–194, 2002.

Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: A general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12):1472–1485, 2004.

C. J. Huijbregts. Regionalized variables and quantitative analysis of spatial data. In J. C. Davis and M. J. McCullagh, editors, *Display and Analysis fo Spatial Data*, pages 38–53. Wiley & Sons, 1975.

S.-Y. Hwang, Y.-H. Liu, J.-K. Chiu, and E.-P. Lim. Mining mobile group patterns: A trajectory-based approach. In *Proc. of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'05)*, volume 3518 of *Lecture Notes in Computer Science*, pages 713–718. Springer, 2005.

International Bureau of Weights and Measures. The international system of units (si), 2006. URL `http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf`.

International Bureau of Weights and Measures. International atomic time, 2010. URL `http://www.bipm.org/en/scientific/tai/tai.html`.

International Organization for Standardization (ISO). *ISO 8601:2004(E), Data elements and interchange formats - Information interchange - Representation of dates and times*, 2004.

P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *Proc. of the 9th International Symposium on Spatial and Temporal Databases (SSTD'05)*, pages 364–381. Springer, 2005.

J. Kang and H.-S. Yong. Mining spatio-temporal patterns in trajectory data. *Journal of Information Processing Systems*, 6(4):521–536, 2010.

E. L. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

H. A. Karimi and X. Liu. A predictive location model for location-based services. In *Proc. of the 11th ACM International Symposium on Advances in Geographic Information Systems (GIS'03)*, pages 126–133. ACM, 2003.

D. Katsaros, A. Nanopoulos, M. Karakaya, G. Yavas, O. Ulusoy, and Y. Manolopoulos. Clustering mobile trajectories for resource allocation in mobile environments. In *Proc. of the 5th International Symposium on Intelligent Data Analysis (IDA'03)*, pages 319–329. Springer, 2003.

E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An online algorithm for segmenting time series. In *Proc. of the 2001 IEEE International Conference on Data Mining (ICDM '01)*, pages 289–296. IEEE Computer Society, 2001.

D. G. Kleinbaum and M. Klein. *Survival Analysis.* Statistics for Biology and Health. Springer, 2005.

J. M. Kleinberg. Hubs, authorities, and communities. *ACM Computing Surveys*, page 5, 1999.

W. Klösgen. Subgroup discovery. In W. Klösgen and J. Zytkow, editors, *Handbook of Data Mining and Knowledge Discovery*, chapter 16.3. Oxford University Press, New York, 2002.

W. Klösgen and M. May. Spatial subgroup mining integrated in an object-relational spatial database. In *Proc. of the 6th European Conference on Data Mining ans Knowledge Discovery (PKDD)*, pages 275–286, 2002.

K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. of the 4th International Symposium on Advances in Spatial Databases (SSD '95)*, pages 47–66, London, 1995. Springer.

C. Körner, D. Hecker, M. Krause-Traudes, M. May, S. Scheider, D. Schulz, H. Stange, and S. Wrobel. Spatial data mining in practice: Principles and case studies. In C. Soares and R. Ghani, editors, *Data Mining for Business Applications.* IOS Press, 2010a.

C. Körner, D. Hecker, M. May, and S. Wrobel. Visit potential: A common vocabulary for the analysis of entity-location interactions in mobility applications. In M. Painho, M. Y. Santos, and H. Pundt, editors, *Geospatial Thinking*, Lecture Notes in Geoinformation and Cartography, pages 79–95. Springer, 2010b.

W. J. Koschnick. FOCUS-Medialexikon, 2011. URL `http://www.medialine.de/deutsch/wissen/medialexikon.html`.

M. Krause-Traudes, S. Scheider, S. Rüping, and H. Meßner. Spatial data mining for retail sales forecasting. In *Proc. of the 11th International Conference on Geographic Information Science (AGILE '08)*, 2008.

Y. Kurata and M. J. Egenhofer. The 9+-intersection for topological relations between a directed line segment and a region. In *Proc. of the Workshop on Behaviour Monitoring and Interpretation (BMI)*, volume 296 of *CEUR Workshop Proceedings*, pages 62–76. CEUR-WS.org, 2007.

K. Laasonen. Clustering and prediction of mobile user routes from cellular data. In *Proc. of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, pages 569–576. Springer, 2005.

P. Laube and S. Imfeld. Analyzing relative motion within groups of trackable moving point objects. In *Proc. of the 2nd International Conference on Geographic Information Science (GIScience'02)*, pages 132–144, London, UK, 2002. Springer-Verlag.

P. Laube, M. van Kreveld, and S. Imfeld. Finding remo - detecting relative motion patterns in geospatial lifelines. In *Proc. of 11th International Symposium on Spatial Data Handling (SDH'04)*, pages 201–214. Springer, 2004.

A. J. T. Lee, Y.-A. Chen, and W.-C. Ip. Mining frequent trajectory patterns in spatial-temporal databases. *Information Sciences*, 179(13):2218–2231, 2009.

L. Leonardi, S. Orlando, A. Raffaetà, A. Roncato, and C. Silvestri. Frequent spatio-temporal patterns in trajectory data warehouses. In *ACM Symposium on Applied Computing*, pages 1433–1440. ACM, 2009.

B. Liang and Z. J. Haas. Predictive distance-based mobility management for multidimensional PCS networks. *IEEE/ACM Transactions on Networking*, 11(5):718–732, 2003.

L. Liao, D. Fox, and H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *International Journal of Robotics Research*, 26(1):119–134, 2007.

T. Liebig, H. Stange, D. Hecker, M. May, C. Körner, and U. Hofmann. A general pedestrian movement model for the evaluation of mixed indoor-outdoor poster campaigns. In *Proc. of the Third International Workshop on Pervasive Advertising and Shopping*, 2010.

S. C. Liou and Y. M. Huang. Trajectory predictions in mobile networks. *International Journal of Information Technology (IJIT)*, 11(11):109–122, 2005.

R. J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley Series in Probability & Mathematical Statistics. John Wiley & Sons, 2002.

R. J. A. Little and M. D. Schluchter. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrica*, 72:492–512, 1985.

P. A. Longley, M. F. Goodchild, D. J Maguire, and D. W. Rhind. *Geographic Information Systems and Science*, chapter 3. Wiley & Sons, 2001a.

P. A. Longley, M. F. Goodchild, D. J Maguire, and D. W. Rhind. *Geographic Information Systems and Science*, chapter 4. Wiley & Sons, 2001b.

R. Lott. OGC abstract specification topic 2, spatial referencing by coordinates. Technical report, Open Geospatial Consortium, Inc. (OGC), 2004. URL `http://portal.opengeospatial.org/files/?artifact_id=6716`.

J. Macedo, C. Vagenot, W. Othman, N. Pelekis, E. Fretzos, B. Kuijpers, I. Ntoutsi ans S. Spaccapietra, and Y. Theodoridis. Trajectory data models. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 5. Springer, Berlin Heidelberg, 2008.

D. Malerba, M. Ceci, and A. Appice. Mining model trees from spatial data. In *Proceedings of PKDD 2005*, pages 169–180, 2005.

G. Marketos, E. Frentzos, I. Ntoutsi, N. Pelekis, A. Raffaetà, and Y. Theodoridis. Building real-world trajectory warehouses. In *Proc. of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access (MobiDE'08)*, pages 8–15. ACM, 2008.

G. Matheron. *The Theory of Regionalized Variables and Its Applications.* École Nationale Supérieure des Mines de Paris, 1971.

M. May, D. Hecker, C. Körner, S. Scheider, and D. Schulz. A vector-geometry based spatial kNN-algorithm for traffic frequency predictions. In *Proc. of the 2008 IEEE International Conference on Data Mining Workshops (ICDMW '08)*, pages 442–447. IEEE Computer Society, 2008a.

M. May, S. Scheider, R. Rösler, D. Schulz, and D. Hecker. Pedestrian flow prediction in extensive road networks using biased observational data. In *Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS '08)*, pages 1–4. ACM, 2008b.

M. May, C. Körner, D. Hecker, M. Pasquier, U. Hofmann, and F. Mende. Handling missing values in GPS surveys using survival analysis: a GPS case study of outdoor advertising. In *ADKDD '09: Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 78–84. ACM, 2009a.

M. May, C. Körner, D. Hecker, M. Pasquier, Urs Hofmann, and Felix Mende. Modelling missing values for audience measurement in outdoor advertising using GPS data. In *GI Jahrestagung*, pages 3993–4006. GI, 2009b.

D. D. McCarthy and G. Petit, editors. *IERS Conventions (2003)*, chapter 4. Verlag des Bundesamtes für Kartographie und Geodäsie, 2004. URL `http://www.iers.org/nn_11216/IERS/EN/Publications/TechnicalNotes/tn32.html`.

N. Meratnia and R. A. de By. Spatiotemporal compression techniques for moving point objects. In *9th International Conference on Extending Database Technology (EDBT'04)*, pages 765–782. Springer, 2004.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London*, 209:415–446, 1909.

H. J. Miller. A measurement theory for time geography. *Geographical Analysis*, 37:17–45, 2005.

A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. WhereNext: a location predictor on trajectory pattern mining. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'09)*, pages 637–646. ACM, 2009.

A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement data anonymity through generalization. *Transactions on Data Privacy*, 3(2):91–121, 2010.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrica*, 37:17–23, 1950.

J. Muckell, J.-H. Hwang, C. T. Lawson, and S. S. Ravi. Algorithms for compressing gps trajectory data: an empirical evaluation. In *Proc. of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACMGIS '10)*, pages 402–405. ACM, 2010.

G. D. Murray and J. G. Findlay. Correcting for bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, 7:941–946, 1988.

M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems (JIIS)*, 27(3):267–289, 2006.

M. Nanni, B. Kuijpers, C. Körner, M. May, and D. Pedreschi. Spatiotemporal data mining. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 10. Springer, Berlin Heidelberg, 2008.

P. Newson and J. Krumm. Hidden markov map matching through noise and sparseness. In *Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS'09)*, pages 336–343. ACM, 2009.

A. Okabe, B. Boots, and K. Sugihara. *Spatial Tessellations - Concepts and Applications of Voronoi Diagrams.* Wiley & Sons, 1992.

I. Olkin and R. F. Tate. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32:448–465, 1961.

Out-of-Home Research & Services GmbH, Fachverband Aussenwerbung e.V. Plakat & Media Grand Prix (PlakaDiva), 2009. URL `http://www.plakadiva.com`.

G. Paaß and J. Kindermann. Current approaches to spatial statistics and bayesian extensions. Technical report, GMD - Forschungszentrum Informationstechnik, 2000.

A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proc. of the 2008 ACM symposium on Applied Computing (SAC'08)*, pages 863–868. ACM, 2008.

C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, and Y. Theodoridis. Segmentation and sampling of moving object trajectories based on representativeness. *IEEE Transactions on Knowledge and Data Engineering*, PrePrints(99), 2011.

D. Papadias and Y. Theodoridis. Spatial relations, minimum bounding rectangles, and spatial data structures. *International Journal of Geographical Information Science (IJGIS)*, 11(2): 111–138, 1997.

T. Park and C. S. Davis. A test of the missing data mechanism for repeated categorical data. *Biometrics*, 49:631–638, 1993.

M. Pasquier, U. Hofmann, F. H. Mende, M. May, D. Hecker, and C. Körner. Modelling and prospects of the audience measurement for outdoor advertising based on data collection using GPS devices (electronic passive measurement system). In *Proceedings of the 8th International Conference on Survey Methods in Transport*, 2008.

K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, June 2005.

N. Pelekis, B. Theodoulidis, I. Kopanakis, and Y. Theodoridis. Literature review of spatio-temporal database models. *Knowledge Engineering Review*, 19(3):235–274, 2004.

N. Pelekis, I. Kopanakis, G. Marketos, I. Ntoutsi, G. Andrienko, and Y. Theodoridis. Similarity search in trajectory databases. In *Proc. of the 14th International Symposium on Temporal Representation and Reasoning (TIME'07)*, pages 129–140. IEEE Computer Society, 2007.

N. Pelekis, A. Raffaetà, M. l: Damiani, V. Vagenot, G. Marketos, E. Fretsos, I. Ntoutsi, and Y. Theodoridis. Towards trajectory data warehouses. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 7. Springer, Berlin Heidelberg, 2008.

N. Pelekis, G. Andrienko, N. Andrienko, I. Kopanakis, G. Marketos, and Y. Theodoridis. Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems*, pages 1–49, 2011a.

N. Pelekis, E. Frentzos, N. Giatrakos, and Y. Theodoridis. HERMES: A trajectory db engine for mobility-centric applications. *International Journal of Knowledge-based Organizations (IJKBO)*, 2011b. in press.

N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos, and Y. Theodoridis. Clustering uncertain trajectories. *Knowledge and Information Systems*, 28(1):117–147, 2011c.

M. A. Quddus, W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312 – 328, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`.

J. Renz and D. Mitra. Qualitative direction calculi with arbitrary granularity. In *Proc. of the 8th Pacific Rim International Conference on Artificial Intelligence(PRICAI)*, volume 3157 of *Lecture Notes in Computer Science*, pages 65–74. Springer, 2004.

B. Richmond. *Time Measurement and Calendar Construction.* Leiden, 1956.

P. Rigaux, M. Scholl, and A. Voisard. *Spatial Databases. With Application to GIS.* Morgan Kaufmann, 2001.

S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7 (3):225–239, 2008a.

S. Rinzivillo, F. Turini, V. Bogorny, C. Körner, B. Kuijpers, and M. May. Knowledge discovery from geographical data. In F. Giannotti and D. Pedreschi, editors, *Mobility, Data Mining and Privacy*, chapter 9. Springer, Berlin Heidelberg, 2008b.

A. H. Robinson, J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. *Elements of Cartography*, chapter 4. Wiley & Sons, 1995a.

A. H. Robinson, J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. *Elements of Cartography*, chapter 5. Wiley & Sons, 1995b.

C. Roever, N. Raabe, K. Luebke, U. Ligges, G. Szepannek, and M. Zentgraf. *Reference Manual: Package klaR*, 2011. URL `http://cran.r-project.org/web/packages/klaR/klaR.pdf`.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, 1987.

S. Saltenis, C. S. Jensen, S. T. Leutenegger, and M. A. Lopez. Indexing the positions of continuously moving objects. In *Proc. of the 2000 ACM SIGMOD International Conference on Management of Data (SIGMOD'00)*, pages 331–342. ACM, 2000.

J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Journal of Data Mining and Knowledge Discovery*, 2:169–196, 1998.

J. L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, 1997.

J. L. Schafer. *mix: Estimation/multiple Imputation for Mixed Categorical and Continuous Data*, 2010. URL `http://CRAN.R-project.org/package=mix`. Original by J. L. Schafer, R package version 1.0-8.

J. L. Schafer and J. W. Graham. Missing data: Our view on the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

R. Schlich and K. W. Axhausen. Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30:13–36, 2003.

S. Schönfelder and K.W. Axhausen. Analysing the rhythms of travel using survival analysis. In C. Kaspar, C. Laesser, and T. Bieger, editors, *Jahrbuch 2000/2001 Schweizerische Verkehrswirtschaft*, pages 137–162. Universität St. Gallen, 2001.

N. Schuessler and K. W. Axhausen. Processing raw data from global positioning systems without additional information. *Transportation Research Record*, 2105:28–36, 2009.

H. Seeger. Spatial referencing and coordinate systems. In *Geographical information systems*, volume 1, chapter 30. John Wiley & Sons, 1999.

S. Shekhar and S. Chawla. *Spatial Databases: A Tour*, chapter 7. Prentice Hall, 2003.

S. Shirali and H. L. Vasudeva. *Metric Spaces*, pages 103–104. Springer, 2006.

J. Z. Sissors and R. B. Baron. *Advertising Media Planning*, chapter 4-5. McGraw-Hill, 2002.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. NeuroCOLT Technical Report NC2-TR-1998-030, 1998.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human motion. *Science*, 327(5968):1018–1021, 2010.

L. Song and X. He. Evaluating next-cell predictors with extensive wi-fi mobility data. *IEEE Transactions on Mobile Computing*, 5(12):1633–1649, 2006.

S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data and Knowledge Engineering*, 65(1):126–146, 2008.

Stiftung Werbestatistik Schweiz. Werbeaufwand Schweiz 2011, Datenerhebung 2010 (Advertising expenditure Switzerland 2011, survey year 2010), 2011. URL `http://www.agma-mmc.de/03_forschung/plakat.asp?topnav=10&subnav=199`.

P. R. Stopher. Collecting and processing data from mobile technologies. In *Transport Survey Methods - Keeping Up With a Changed World*, chapter 21. Emerald Group Publishing Limited, 2009.

N. Strobel. *Astronomy Notes.* McGraw-Hill, 2007. URL `http://www.astronomynotes.com`.

Swiss Poster Research Plus (SPR+). SPR+ Strassenstudie, 2011a. URL `http://www.mobitrack.ch/main.aspx?TabID=289`.

Swiss Poster Research Plus (SPR+). SPR+ Straßenstudie, 2011b. URL `http://www.spr-plus.ch/Userfiles/PDF/Brochuren/SPRPlus_Strassenstudie.pdf`.

Y. Tao and D. Papadias. Time-parameterized queries in spatio-temporal databases. In *Proc. of the 2002 ACM SIGMOD International Conference on Management of Data (SIGMOD'02)*, pages 334–345. ACM, 2002.

Y. Tao, D. Papadias, and J. Sun. The TPR*-tree: An optimized spatio-temporal access method for predictive queries. In *Proc. of 29th International Conference on Very Large Data Bases (VLDB'03)*, pages 790–801. Morgan Kaufmann, 2003a.

Y. Tao, J. Sun, and D. Papadias. Analysis of predictive spatio-temporal queries. *ACM Transactions on Database Systems (TODS)*, 28(4):295–336, 2003b.

T. Therneau. *survival: Survival analysis, including penalised likelihood.*, 2009. URL `http://CRAN.R-project.org/package=survival`. Original R port by T. Lumley, R package version 2.35-8.

W. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.

G. Upton and B. Fingleton. *Spatial Data Analysis by Example*, volume 1. John Wiley & Sons, 1994.

M. Wachowicz, R. Ong, C. Renso, and M. Nanni. Finding moving flock patterns among pedestrians through collective coherence. *International Journal of Geographical Information Science (IJGIS)*, 25(11):1849–1864, 2011.

H. Wackernagel. *Multivariate Geostatistics.* Springer, Berlin, 1998.

Y. Wang and I. Witten. Inducing model trees for continuous classes. In *Proceedings of ECML 1997*, pages 128–137, 1997.

Y. Wang, E.-P. Lim, and S.-Y. Hwang. On mining group patterns of mobile users. In *Proc. of the 14th International Conference on Database and Expert Systems Applications (DEXA'03)*, pages 287–296. Springer, 2003.

WebFinance Inc. Gross rating point (GRP) — BusinessDictionary.com, 2010. URL `http://www.businessdictionary.com/definition/gross-rating-point-GRP.html`. [Online; accessed May 2010].

Westburn Publishers Ltd. Opportunities-to-see (OTS) — The Westburn Dictionary of Marketing, 2010. URL `http://www.westburnpublishers.com/marketing-dictionary/o/opportunities-to-see-%28ots%29.aspx`. [Online; accessed May 2010].

Wikipedia. Out-of-home advertising — Wikipedia, the free encyclopedia, 2010a. URL `http://en.wikipedia.org/wiki/Out-of-home_advertising`. [Online; accessed September 2010].

Wikipedia. Reach (advertising) — Wikipedia, the free encyclopedia, 2010b. URL `http://en.wikipedia.org/wiki/Reach_%28advertising%29`. [Online; accessed May 2010].

Wikipedia. Coordinated universal time — Wikipedia, the free encyclopedia, 2010c. URL `http://en.wikipedia.org/wiki/Utc`. [Online; accessed March 2010].

Wikipedia. Zeit — Wikipedia, the free encyclopedia, 2010d. URL `http://de.wikipedia.org/wiki/Zeit`. [Online; accessed Feb. 2010].

Wikipedia. Great-circle distance — Wikipedia, the free encyclopedia, 2011. URL `http://en.wikipedia.org/wiki/Great-circle_distance`. [Online; accessed September 2011].

S. Willard. *General Topology*. Courier Dover Publications, 2004.

J. Wolf, R. Guensler, and W. Bachman. Elimination of the travel diary: An experiment to derive trip purpose from gps travel data. *Transportation Research Record*, (1768):125–134, 2001.

Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology (EDBT'11)*, pages 259–270. ACM, 2011a.

Z. Yan, N. Giatrakos, V. Katsikaros, N. Pelekis, and Y. Theodoridis. SeTraStream: Semantic-aware trajectory construction over streaming movement data. In *Proc. of the 12th International Symposium on Advances in Spatial and Temporal Databases (SSTD'11)*, pages 367–385, 2011b.

J. Yang and M. Hu. Trajpattern: Mining sequential patterns from imprecise trajectories of mobile objects. In *Proc. of the 10th International Conference on Extending Database Technology (EDBT'06)*, pages 664–681. Springer, 2006.

G. Yavas, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data and Knowledge Engineering*, 54(2):121–146, 2005.

A. Zanda, C. Körner, F. Giannotti, D. Schulz, and M. May. Clustering of german municipalities based on mobility characteristics: An overview of results. In *Proc. of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS'08)*, pages 1–4. ACM, 2008.

Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from GPS trajectories. In *Proc. of the 18th International Conference on World Wide Web (WWW'09)*, pages 791–800. ACM, 2009.

C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, 25(3), 2007.

Index