Life histories across time and space : methods for including geographic factors on the micro-level in longitudinal demographic research

Hedefalk, Finn

2014

Life histories across space and time

Methods for including geographic factors on the micro-level in longitudinal demographic research

LUNDS UNIVERSITET

Finn Hedefalk

AKADEMISK AVHANDLING

som för avläggande av filosofie licentiatexamen

vid Naturvetenskapliga fakulteten, Lunds universitet,

kommer att offentligen försvaras i Biosfären, Geocentrum I,

Sölvegatan 10, Lund, fredagen den 3:e oktober, kl 09.15

Organization	Document name
LUND UNIVERSITY	LICENTIATE DISSERTATION
	Date of issue: 2014-10-03
Author(s): Finn Hedefalk	Sponsoring organization: eSSENCE

Title and subtitle

Life histories across space and time: Methods for including geographic factors on the micro-level in longitudinal demographic research

Abstract

Historical demography, which is the study of human population dynamics in the past, is central for understanding human behaviours and traits, such as fertility, mortality and migration. An important factor in demographic research is the geographic context. Where people lived often determined their social ties, exposure to diseases and economic development. Such information is essential not only for historical demographic research but also for a wide range of disciplines. While the geographic context on an aggregated level has an important role in longitudinal historical studies, geographic contexts on a micro-level have only played a minor role.

This licentiate contributes to historical demographic research by studying how geographic factors on the micro-level can be included in longitudinal historical analyses. A primary focus is the methodological development for creating longitudinally detailed locations that can be linked to individuals in demographic databases. This research should offer a variety of possibilities for studying how geographic factors on the micro-level affected human living conditions throughout history.

The thesis has four research objectives. The first objective is to extend a standardised data model for longitudinal demographic data to include geographic data. This is achieved by introducing IDS-Geo, which is a geographically extended version of the standardised data model IDS. The second objective is to develop and evaluate harmonisation methods to ensure that source data comply with standardised data models. This is achieved by testing and developing a method for first harmonising Swedish environmental data and metadata and then testing the data for compliance against standardised data models and specifications. The third objective is to develop a methodology for creating integrated longitudinal demographic and geographic databases that include geographic factors on the micro-level in demographic research. The core of the methodology is to transform geographic objects in snapshot time representations (digitised from historical maps) into longitudinal object lifeline time representations, and to link individuals to these geographic objects using standardised locations. The methodology is implemented in a case study in which we integrate information from approximately 60 digitised historical maps with longitudinal individual-level data from the Scanian Economic Demographic Database (SEDD). We link 80,431 individuals in five rural parishes in Sweden during 1813-1914 to the property units where they lived. The resulting database is tested using fundamental queries for spatio-temporal data. Additional historical geographic data used for computing context variables are constructed. The results are a unique contribution in terms of linking individuals over such long time periods to longitudinal geographic data on the micro-level. Lastly, the fourth objective of the thesis is to perform longitudinal demographic analyses where geographic factors can subsequently be included. This is performed by analysing the intergenerational effects of child bearing by relatively older women on the longevity of adult offspring in pre-transitional Utah, USA.

Key words: longitudinal historical data, geographic factors, geodemographic databases, micro-level

Classification system and/or index terms (if any)

Supplementary bibliographical information: Avhandlingar från Institutionen för naturgeografi och ekosystemanalys. Avhandlingsnummer 14.	Language: English
ISSN and key title	ISBN: 978-91-85793-41-9
Recipient's notes	Number of pages: 133 / Security classification	Price

Signature _____ Date 2014-08-22

Life histories across space and time

Methods for including geographic factors on the micro-level in longitudinal demographic research

LUNDS UNIVERSITET

Finn Hedefalk

Centre for Geographical Information Systems (GIS Centre)

Department of Physical geography and Ecosystem Science

Faculty of Science

Lund University

A licentiate thesis at a university in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summarizes the accompanying papers already published or manuscripts at various stages (in press, submitted or in preparation)

Cover maps from the Lantmäteriet historical archive. Front page: Kågeröd parish cadastral map from 1833 (*Karta öfver inägorna till sätesgården Knutstorp*); back page: an economic map (*Häradsekonomiska kartan*) from 1910-1915 over the same area.

Drawings by Finn Hedefalk

Centre for Geographical Information Systems (GIS Centre)

Department of Physical Geography and Ecosystem Science

Faculty of Science, Lund University

Sölvegatan 12, SE-223 62 Lund, Sweden

ISBN 978-91-85793-41-9

Printed in Sweden by Media-Tryck, Lund University
Lund 2014

—You remember various details. But not even all of them together shew your intention. It is as if a snapshot of a scene had been taken, but only a few scattered details of it were to be seen: here a hand, there a bit of a face, or a hat—the rest is dark. And now it is as if we knew quite certainly what the whole picture represented. As if I could read the darkness.

Ludwig Wittgenstein

Abstract

Historical demography, which is the study of human population dynamics in the past, is central for understanding human behaviours and traits, such as fertility, mortality and migration. An important factor in demographic research is the geographic context. Where people lived often determined their social ties, exposure to diseases and economic development. Such information is essential not only for historical demographic research but also for a wide range of disciplines. While the geographic context on an aggregated level has an important role in longitudinal historical studies, geographic contexts on a micro-level have only played a minor role.

This licentiate contributes to historical demographic research by studying how geographic factors on the micro-level can be included in longitudinal historical analyses. A primary focus is the methodological development for creating longitudinally detailed locations that can be linked to individuals in demographic databases. This research should offer a variety of possibilities for studying how geographic factors on the micro-level affected human living conditions throughout history.

The thesis has four research objectives. The first objective is to extend a standardised data model for longitudinal demographic data to include geographic data. This is achieved by introducing IDS-Geo, which is a geographically extended version of the standardised data model IDS. The second objective is to develop and evaluate harmonisation methods to ensure that source data comply with standardised data models. This is achieved by testing and developing a method for first harmonising Swedish environmental data and metadata and then testing the data for compliance against standardised data models and specifications. The third objective is to develop a methodology for creating integrated longitudinal demographic and geographic databases that include geographic factors on the micro-level in demographic research. The core of the methodology is to transform geographic objects in snapshot time representations (digitised from historical maps) into longitudinal object lifeline time representations, and to link individuals to these geographic objects using standardised locations. The methodology is implemented in a case study in which we integrate information from approximately 60 digitised historical maps with longitudinal individual-level data from the Scanian Economic Demographic Database (SEDD). We link 80,431 individuals in five rural parishes in Sweden during 1813-1914 to the property units where they lived. The resulting database is tested using fundamental queries for spatio-temporal data. Additional historical geographic data used for computing context variables are constructed. The results are a unique contribution in terms of linking individuals over such long time periods to longitudinal geographic data on the micro-level. Lastly, the fourth objective of the thesis is to perform longitudinal demographic analyses where geographic factors can subsequently be included. This is performed by analysing the intergenerational effects of child bearing by

relatively older women on the longevity of adult offspring in pre-transitional Utah, USA.

Acknowledgements

I am now halfway in my PhD project and there are many people I want to thank so far. I find myself very lucky with my three supervisors. First and foremost I want to thank my main supervisor Lars Harrie; always razor-sharp, always there and with a never-empty store of good ideas. I deeply appreciate all your support. My second supervisor Patrick Svensson has not only been a great advisor, but his dedication to his research in agricultural history has also spurred my own interest in this subject. I also want to thank my third supervisor, Ali Mansourian, for his valuable discussions and support.

Furthermore, I want to thank Tommy Bengtsson, Clas Andersson, Daniel Persson, Irene Rangel Öhrn, Lena Arvidsson and Luciana Quaranta for the cooperation and involvement in the eSSENCE project. Without your help, I would not have been in the situation I am in today. In particular I want to thank Tommy for welcoming me to the Centre for Economic Demography.

I also want to thank co-authors and colleagues, current and past; many thanks to my co-authors Saskia Hin and Bartosz Ogórek, and Siddartha Aradhya and Jonas Helgertz. It has been, and it still is, a pleasure doing research with you. When looking back at the past, I have very much to thank Anders Östman, my previous supervisor and professor at the University of Gävle. He always inspired me of pursuing my career in academia. Big thanks also to Solgerd Tanzilli, Xin He, Helen Eriksson and Johanna Fröjdenlund Runarson for the times in the GeoTest project. Moreover, I want to address a special thanks to Junjun Yin and Alexey Tereshenkov for all their valuable help in GIS related issues. Finally, I thank Isak Willebrand and Rolf-Erik Keck for giving invaluable comments to my PhD application letter.

My colleagues at the GIS centre are lightening up many of my days as well as several evenings: thanks Abdulghani, Alex, Andreas, Ehsan, Karin, Lars E, Lina, Martin, Micael, Mitch, Mohammadreza, Mojgan, Petter, Roger, Sam, Stefan and Ulrik. Special thanks go to Petter Pilesjö for making the GIS Centre to the excellent place it is.

I thank the Department of Physical Geography and Ecosystem Science, the Centre for Economic Demography and the Department of Economic History as well as all their employees for making my time there very pleasant. I thank my fellow PhD-students at the department, and particularly those that have become my good friends (I assume you know who you are). I also appreciate the people in the football group, the badminton group and the ping-pong group for making the time so much more fun. Moreover, I want to thank my colleagues at the Centre for Economic Demography for making my second workplace enjoyable: Andy, Anna, Annika, Björn, Fredrik, Joe, Kirk, Madeleine, Maria, Martin, Mats, Volha and Zeyuan.

Almost finally, I want to thank all my other friends for the life outside work and many thanks go to my family: Anna, Carl-Magnus, Dag and Nea for all their help and love throughout my life.

Finally, I am deeply grateful to my most lovely Jing for her endless support and love in all kinds of times.

Table of Contents

1 Introduction

1.1 Motivation for the licentiate thesis

Historical demography is the research of past human population dynamics. It studies aspects of fertility, mortality, nuptiality and migration, as well as the relationship between populations and the larger society. Essential resources in demographic research are data on the individual-level that cover long time periods, i.e., longitudinal micro-data. Studies can also be performed on the aggregated level, i.e., population groups. Micro-data and aggregated data that cover long time periods make it possible to construct robust demographic models.

A key factor in demographic research is the geographic context. The places people lived often determined their social ties, exposure to diseases and economic development. Such information is important not only for historical demographic research but also for a wide range of applications in other fields, such as epidemiology, medicine and geography.

While geographic contexts on the aggregated level has been an important component of longitudinal historical studies, geographic contexts on the micro-level have only played a minor role because large historical datasets in which individuals are linked to detailed physical locations are sparse. Therefore, we cannot account for the spatial variation that occurs within the aggregated regions. Modern demographic data can easily link individuals to standardised addresses, but the time periods for which these addresses remain constant are often short. Historical demographic data, however, can cover much longer time periods and permit the study of several generations. In addition, because these data are usually less constrained by integrity laws than modern data, they can be used more freely. However, standardised addresses are seldom available in historical data; therefore, individuals need to be linked to geographic features (e.g., buildings or property units).

The ability to track individuals on the micro-level across space and time would provide many new insights about how geographic factors have affected human living conditions throughout history. This is especially true when using longitudinal data; because of the long time periods these datasets cover, it is possible to track several generations and accurately analyse their social and

biological traits through time. Therefore, it would be possible to study how population densities, social networks and land use affect mortality, fertility rates and migration in both the short term and long term.

Moreover, most historical longitudinal micro-data are available from approximately the end of the 17th century to the beginning of the 20th century. This time period is especially interesting to study because it encompasses some of the most extensive changes to human populations. Particularly, the period includes the demographic transition, which began in the early 19th century in Europe and North America, when mortality and fertility began to decline. Before this period, mortality was mainly determined by widespread epidemic diseases, and the population used to fluctuate. The agricultural and industrial revolutions, which substantially changed our society, are also linked to the demographic transition. One of the main reasons for the mortality decline is thought to be the improvements in nutrition following the agricultural revolution. The health of people improved, which boosted the economic development and positively affected agricultural and industrial advances. Because the demographic transition continues in several developing countries, research on this subject is also important for modern societies. By including geographical factors at the micro-level, further research may provide new and important insights into the demographic transition.

This licentiate thesis aims to improve historical demographic analysis by studying how geographic factors on the micro-level can be included in longitudinal historical research. To include this information, methodological developments for creating detailed longitudinal geographic data that can be integrated with longitudinal demographic micro-data are required. Such development should pave the way for future studies that exploit the integrated data, for example, demographic studies on factors that affected the mortality and fertility declines.

Another essential aspect of historical demographic research is the possibility of comparing patterns among population groups and regions. Then, it becomes possible to answer fundamental questions regarding which demographic outcomes are determined by society, biology, or both. To conduct such comparisons, we need standardised data from several regions. Hence, the secondary purpose of this thesis is to determine how current standards for demographic data can be used and extended for integrated longitudinal demographic and geographic data.

1.2 Research questions and objectives

The overall research question is how to include geographic factors on the micro-level in longitudinal demographic research. To answer this question, we foremost need to link individuals in demographic databases to longitudinal detailed locations. Additional historical geographic data used for computing context variables also need to be constructed. The primary focus of this thesis is the methodology for creating such data.

Moreover, to facilitate comparative studies, an additional aim is to determine how current standardised data models for demographic data can be extended for integrated longitudinal demographic and geographic data and to develop a process for transforming source data into such data models. Lastly, we aim to conduct longitudinal demographic analyses and exemplify how these analyses can be improved by adding geographic factors on the micro level.

To answer the research questions, this thesis focuses on four research objectives.

- The first objective is to extend a standardised data model for longitudinal demographic data to include geographic data (Paper I).
- The first objective requires that source data be transformed into a target data model. Thus, the second objective is to develop and evaluate such transformation methods (Paper II).
- The third objective is to develop methods for creating integrated longitudinal demographic and geographic databases (called geodemographic databases hereafter) that can include geographic factors on the micro-level in demographic research (Paper III).
- The fourth objective is to perform longitudinal demographic analyses where geographic factors could later be included (Paper IV).

This licentiate thesis is the first part of a PhD project. In the second part, the primary aim is to study how the integrated data constructed in Paper III can be used to construct context variables and thus improve longitudinal demographic analysis.

1.3 Thesis organisation

Chapter 2 presents the overall methodology applied in the thesis. In Chapter 2, the literature is reviewed. Section 2.1 describes the nature of geographic data that endure over time (spatio-temporal data) and how these data are represented conceptually. One aim of this chapter is to specify the terminology used throughout the thesis. Section 2.2 further describes the representation of the historical spatio-temporal data on a logical level and reviews common data models used by historical geographic databases. In section 2.3 the data transformation process is described. The aim of this section is to provide a background for the methods carried out in Paper II and III. Section 2.4 describes some of the main methods for analysing longitudinal historical demographic data, whereas section 2.5 describes the requirements and sources for creating longitudinal historical geodemographic data. Lastly, section 2.6 reviews related studies of integrating longitudinal demographic and geographic data on the micro-level. Chapter 3 describes the study area of Paper I and III, Chapter 4 summarises the papers that are the basis for the thesis, and Chapter 5 presents some concluding remarks.

The second part of the thesis contains the four papers of which the overall methodology is based on. These are presented below.

List of papers

I. Hedefalk, F., Harrie, L., and Svensson, P. 2014. Extending the Intermediate Data Structure (IDS) for longitudinal historical databases to include geographic data. *Historical Life Course Studies* 1:27-46.

II. Hedefalk, F., and Östman, A. 2011. Making Swedish Environmental Geodata INSPIRE Conformant: A Harmonization Case Study. *Mapping and Image Science* 3:30-39.

III. Hedefalk, F., Harrie, L., and Svensson, P. 2014. Methods to create a longitudinal integrated demographic and geographic database on the micro-level: a case study of five Swedish rural parishes, 1813-1914. *Submitted*.

IV. Hin, Saskia., Ogórek, B., and Hedefalk, F. 2014. An old mom keeps you young: Mother's age at last birth and offspring longevity in 19th century Utah. *Manuscript*.

List of contribution

I. FH carried out the practical part of the study. All authors contributed to the manuscript and study design. FH was the lead author.

II. FH carried out the practical parts of the study. Both authors contributed to the manuscript and study design. FH was the lead author.

III. FH carried out the practical parts of the study. All authors contributed to the manuscript and study design. FH was the lead author.

IV. FH carried out the data processing and contributed to the data analysis. All authors participated equally in the study design and interpretation of the results. SH was the lead author. FH and BO contributed equally to the writing of the paper.

1.4 Methodology

This section presents the overall methodology used in the thesis. The four papers of this thesis are connected in the following way (Figure 1). Papers I-III address methods that enable the inclusion of geographic factors on the micro-level in historical demographic research. Paper I extends a standardised data model for longitudinal historical data to include geographic data. Paper II tests methods for transforming source data to comply with such standardised data models. In Paper III, we first create and transform geographic snapshot data into an object-lifeline data model. Here, the data model is based on the principles of the model developed in Paper I, and the transformation processes are influenced by the models used in Paper II. Then, we integrate the geographic data and the longitudinal demographic data (i.e., link individuals to physical locations). The result is a longitudinal geodemographic database on the micro-level. In Paper IV, we perform a longitudinal demographic analysis where geographic factors can be included future studies. In these future studies, the database from Paper III can be utilised for longitudinal demographic analysis with geographic factors.

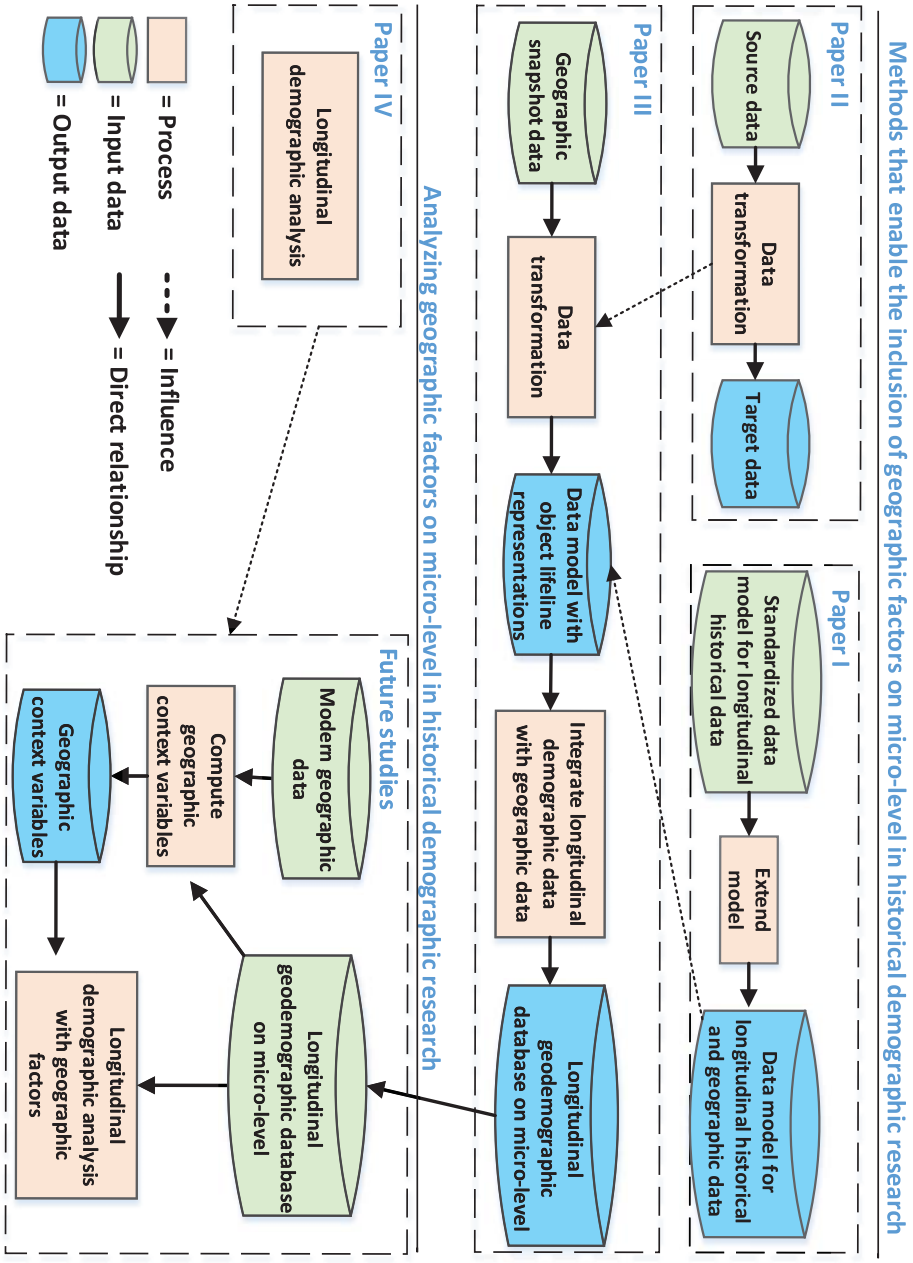

Figure 1: Overall methodology used in the licentiate thesis.

2 Literature review

2.1 Representation of spatio-temporal data

Because a main component of the thesis is to create longitudinal geographic data, this section discusses the nature of spatial data that endure over time (called spatio-temporal data) and how these data are represented.

2.1.1 The nature of spatio-temporal data

First, we need to create, represent and analyse things that occur or exist in space and time. Such things are commonly called *entities*. If the entities relate to Earth, then they are called *geographic entities* (Grenon and Smith, 2004). Throughout this thesis, the terminology in Figure 2 is used, which is based on the general philosophical literature (see, e.g., Casati and Varzi, 2010) and on definitions for spatio-temporal data (cf. Grenon and Smith, 2004; Worboys, 2005; Yuan and Hornsby, 2010). In Figure 2, the entities are either *objects* or *events*.

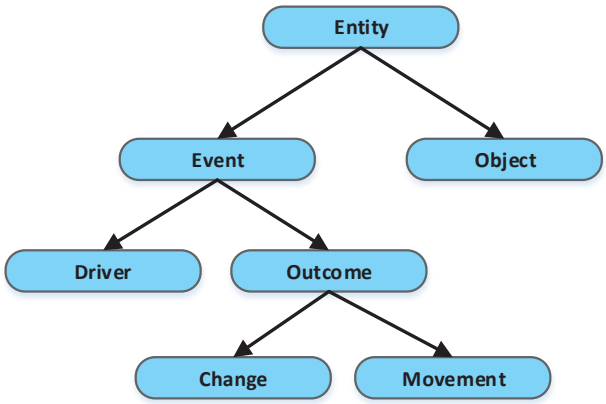

Figure 2: Terminology of the spatio-temporal data

Objects are entities that endure over time and survive changes (Grenon and Smith, 2004). They are often concrete physical objects that occupy space, for example, people, buildings, cadastral parcels, roads, and wetlands; they may also be sites (e.g., the place name "Lund") and attributes of other objects (e.g., the height of a building). The geographical spaces where the entities exist are also objects (e.g., a whole space or a sub-space that is man-made or natural).

Events are things that occur instantaneously or over a period of time and then disappear, for example, the birth of a child, the construction of a building, rainfall, or an enclosure rearranging property units. These events involve objects, and they exist in temporal and spatio-temporal regions (Grenon and Smith, 2004). Events are further divided into *drivers* and *outcomes* (Yuan and Hornsby, 2010). Drivers are the events that cause the outcomes for the objects. For example, a construction event (driver) changes the geometry (outcome) of a building (object). Outcomes can be divided into *change* and *movement*. Change refers to changes to both geometric and non-geometric properties of objects, in which the geometric changes may be both external and internal. Movement refers to changes in the physical location of the object, for example, an individual migrates and changes locations (movement). A central aspect of an outcome is whether an object's identity is retained when it changes or moves (Yuan and Hornsby, 2010). Specifically, when an object changes or moves, does it keep its identity or does it cease to exist?

2.1.2 Representing spatio-temporal data (conceptually)

In their general theory for geographic representation, Goodchild, Yuan and Cova (2007) define the geo-atom as the smallest building block of geographical entities. Geo-atoms are points located in space and time that have a descriptive property. For example, at the spatio-temporal location *x*, the altitude (a property) is 39 metres (the value of the property). To represent geographic data using these atomic elements, there are two fundamental views we can apply (Worboys and Duckham, 2004): *discrete objects*[1] and *continuous fields*. For discrete objects, countable entities with well-defined boundaries (e.g., mountains, lakes, or people) occupy space-time in an otherwise empty world. For continuous fields, the world is continuously represented by a number of variables whose values vary throughout space and time (e.g., a field with varying temperature values) (Longley et al., 2010). These two conceptual world views can be represented in computer systems

[1] Despite its name, *discrete objects* involves both objects and events

as either raster data (space is composed of cells/pixels that each contain a value of a property) or vector data (geographic objects are represented by points, lines or polygons).

Although continuous fields can be successfully applied in many models that represent dynamic geographic data, the main focus in this thesis is the discrete object conceptualisation using the vector representation. This is because individuals in demographic databases are modelled as discrete objects that endure over time. Thus, when geographic objects are created for locating individuals in space, they need to be represented as discrete objects as well. Nevertheless, when creating geographic data for calculating context variables, both continuous fields and discrete object views may be applied.

Discrete objects can be represented as *geo-objects* (Goodchild et al., 2007), which are aggregated geo-atoms that share some particular property values. When these geo-objects are represented in both space and time, Goodchild et al. (2007) describe them as either static or dynamic in terms of their external geometry, internal geometry and movement. In this thesis, the changes in external and internal geometries are considered to be the same type, and we add non-spatial properties as a third type of change to geographic objects (Figure 3). Thus, the states of these objects are related to the outcomes of change and movement (Figure 2).

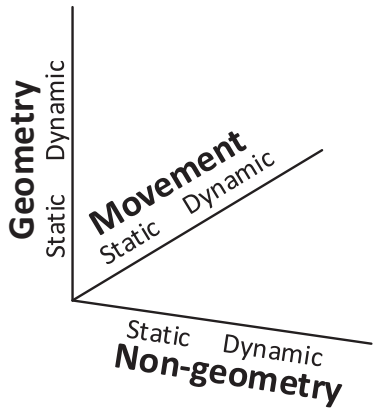

Figure 3: Three types of conditions and the combinations that describe the state of an object.

When applying the concept of dynamics in Figure 3 to the historical geographic data in this thesis, individuals that have point locations are modelled as static in their geometry but dynamic in their movement and non-spatial properties. Property units (physical locations that individuals can be linked to), however, are

dynamic in their geometry and non-spatial properties but static in their movement in most cases.

2.2 Data models for historical geographic databases

Section 2.1 describes the nature of spatio-temporal data and how to represent their associated changes conceptually. This section reviews the specific database models used to store the changes that occur to objects and the events that affect the objects. First, the most fundamental models for spatio-temporal data are described; then, the models applied for modern and historical geographic databases are reviewed.

2.2.1 From snapshot models to event chronicles

The sources of historical geographic data are often scanned historical maps, which can be regarded as snapshots of the conditions at a certain time. From these historical maps, objects, such as property units and buildings, can be digitised. Thus, one of the simplest models for storing spatio-temporal data is to assign each digitised object a time-stamp that corresponds to the date of the historical map (Table 1). Models for storing such time-stamped objects are usually called *snapshot models* (Armstrong, 1988) or *temporal snapshots* (Worboys, 2005). Temporal snapshots are simple to create, but they are not suited for tracing the changes of objects through time. In Table 1, a property unit named "Hög 5" is digitised (from three historical maps) and stored as three different objects in a relational database. Each object is assigned a time-stamp that represents the creation date of the historical map. Note that the geometry changes in the last two rows (i.e., *polygon 2a* and *polygon 2b*). The reason for this change may be that an area of the property unit is subdivided. Also note that the geometry is not exactly the same in these two rows (indicated by indexes *a* and *b*) because the rows are based on different maps with non-perfect geometries.

Table 1: Temporal snapshots of the property unit "Hög 5" (or "Höj 5" in some historical maps). The *timeStamp* attribute represents the creation date of each historical map.

Id	name	timeStamp	geometry
15	Höj 5	1804-01-01	(polygon 1)
22	Hög 5	1820-01-01	(polygon 2a)
89	Hög 5	1865-01-01	(polygon 2b)

From Table 1, we can determine that the property unit experienced a change in geometry. However, it does not store the time of the change or whether other changes occurred between the snapshots. Moreover, the three rows are snapshot observations of, presumably, one object. However, because there is no common identifier that binds the observations, each observation is static.

To enable the identification of changes and to trace the objects through time, *object lifelines* models (Table 2) can be used (Worboys and Duckham, 2004). In this time-representation model, each state of the object is assigned a time period. In Table 2, the objects in Table 1 are linked to a common identifier (*propertyUnitId*). Moreover, additional textual sources are used to attain a more precise estimation of the period during which the property unit and its geometries existed in the real world. In Table 2, the geometry *polygon 2b* is used instead of *polygon 2a* because the source data is higher quality than that for polygon 2a.

Table 2: The property unit 'Hög 5' stored as object lifelines. *startDate* and *endDate* represent the valid time period of the object.

id	propertyUnitId	name	startDate	endDate	geometry
4	pu_hog_52	Hög 5	1790-08-01	1815-08-01	(polygon 1)
5	pu_hog_52	Hög 5	1815-08-01	1890-08-01	(polygon 2b)

Various implementations of object lifeline models are widely used in the GIS domain, but they do not describe the events or the drivers of the outcomes. For example, in Table 2, we cannot identify what event that caused the property unit to change its geometry. To represent the relationships between the drivers and outcomes, the *event chronicles* model (Table 3) is proposed (Worboys, 2005). Here, the focus shifts from the objects to the events. Specifically, instead of describing the states of objects, the drivers and the outcomes are described. Table 3 illustrates how drivers that affected the property unit Hög 5 can be represented. In this example, the property unit was created on 1790-08-01. Then, a part of Hög 5 was subdivided on 1815-08-01 into the new property unit Hög 5a. Finally, it was partitioned on 1890-08-01 into two new property units: Hög 5b and Hög 5c. If such information about a property unit is available, then storing it as event chronicles could permit a more detailed description of the events and objects. Note that Table 3 shows a simplified example using a relational database. To model a large process (in which many events are involved) as a whole, we can describe the events in more detail and better model the relationships between the events and how they are involved in the process (cf. Yuan and Hornsby, 2010).

Table 3: Events linked to the property unit Hög 5 and stored as event chronicles. For readability, the names in the propertyUnit attribute are used as the identifiers.

eventId	eventName	date	propertyUnit
43	Created	1790-08-01	Hög 5
25	Subdivided	1815-08-01	Hög 5; Hög 5a
69	Partitioned	1890-08-01	Hög 5 -> Hög 5b; Hög 5c

2.2.2 Common models for historical geographic databases

Historical GIS databases must often handle two types of changes. These include changes to the geometry or movement of individual objects and changes to the geometry of line networks and polygon partitions, such as communication networks and administrative boundaries (i.e., a change in one object affects its adjacent objects) (Gregory and Ell, 2007). To tackle such changes, most current national historical GIS databases use data models that implement some form of object lifeline representations, e.g., the Great Britain Historical GIS (Gregory and Southall, 2005), the China Historical GIS (Berman, 2003), the Belgian Historical GIS (Vanhaute, 2003) and the Danish DigDag project (Dam, 2013). Moreover, models based on temporal snapshots in combination with object lifelines are also used in several datasets, e.g., the US National Historical GIS (Fitch and Ruggles, 2003) and the HUE dataset (Villarreal, 2014). However, longitudinal demographic databases focus more on the modelling of events, which facilitates longitudinal analysis (Alter, Mandemakers and Gutmann, 2009). For example, the SEDD uses variations of event chronicles and object lifelines (Bengtsson et al., 2012). Moreover, within the standardisation work for geographic data, object lifeline representations are used in several data models for exchanging geographic data. For example, the 34 data specifications developed by the Infrastructure for Spatial Information in the European Community (INSPIRE) Directive use object-lifelines (INSPIRE, 2014).

Within the research of spatiotemporal data models, several models that focus on events have been developed, e.g., the event-based Spatio-Temporal Data Model (ESTDM) for raster data (Peuquet and Duan, 1995) and the three domain model (Yuan, 2000). Other more recent developments include the conceptual Continuous Spatio-Temporal Model (CSTM) (Van de Weghe, et al., 2014) and the Extended dynamic GIS model (EDGIS) (Pultar et al., 2010); the latter model based on the proposed general GIScience theory from Goodchild et al. (2007), which includes geo-atoms and geo-objects. Moreover, Yi et al. (2014) created a framework for representing changes and interactions between spatiotemporal entities. This framework was applied to model the dynamics of spatiotemporal entities in the

ocean, which were observed by snapshots from multiple remote sensing images. Models have also been developed to specifically represent the spatiotemporal paths of individuals, e.g., Shaw, Yu and Bombom (2008).

2.3 Data transformation concepts

When data are stored in different systems and models, heterogeneities and conflicts often occur. The overall process for resolving these conflicts is called *data harmonisation*. This process aims to solve heterogeneities that exist between the data to combine them in a meaningful way. A common step in a data harmonisation process is to create standardised data models (one example in this thesis is the IDS data model (Alter and Mandemakers, 2014)). Then, data from several sources and varying data models can be adapted to these standardised models and thus be compared and analysed together. However, this implies that the original data, usually called *source data*, need to be transformed into the new *target* data model. Formally, this transformation f can be defined as:

$$f: \Omega_{source} \rightarrow \Omega_{target}$$

where Ω represents all entities.

The heterogeneities among the data that a harmonisation attempts to address can be categorised into three types (Sheth and Larson, 1990):

- syntactical: differences in formats and data types
- structural: differences in the data models and structures
- semantic: differences in the meaning of concepts.

For example, two datasets that describe the same type of buildings may use two different data formats (Figure 4a), use different data models (Figure 4b), or have different definitions of what constitutes a river (Figure 4c). Of these heterogeneities, the semantic type is often the most problematic (Lutz et al., 2009, Paper II). Syntactical and structural differences can often be solved automatically, although there is sometimes a risk of quality loss during the transformation.

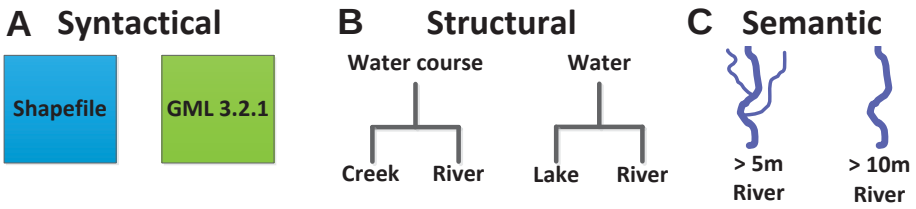

Figure 4: Three types of heterogeneities. A) Two datasets are stored in different GIS formats: Shapefile and GML version 3.2.1. B) Two database models differ in their representation of a river. C) In the first dataset, a river is defined as a watercourse that is wider than 5 metres; in the second dataset, the width is >10 metres.

A transformation process in a database is often called a schema translation (Dumpala and Arora, 1981). This procedure consists of three main parts.

- Schema matching: Finding concepts[2] in the source data model that share semantics with concepts in the target data model.
- Schema mapping: Specifying how concepts in the source data model relate to concepts in the target data model, specifically, defining the operations needed to transform the source data to the target data model.
- Schema transformation: Executing the specified operations in the schema mapping.

The aim of schema matching is to identify which concepts in the source data model can be transformed to the target data model. During the schema matching process, concepts in the source data that share the same or similar meaning with concepts in the target data model are identified, for example, if the table "Building" in the source data corresponds to a table in the target data model.

During the schema mapping, operations needed to transform the source data to the target data model are specified. These operations differ depending on what type of heterogeneities needs to be solved (i.e., syntactic, structural or semantic) and whether geometric and/or non-geometric properties are to be transformed. In addition, the mapping operations can also be defined on different levels of granularity (e.g., database table, attribute, or cell value). Common operations for

[2] The term *concept* includes all types of entities that are modelled and represented at various levels of granularities. For example, a concept may be a database table that describes buildings, an attribute specifying building addresses, or an attribute value containing address names.

both geometric and non-geometric properties are filtering of information, reclassification, aggregation, merging, splitting, value conversions, morphing and augmentation (for non-geometric, cf. Lehto, 2007, Paper II; for geometric, cf. Regnauld and McMaster, 2007). For example, non-geometric mapping may be used to specify how a date type should be converted into text or how a reclassification should be conducted. When transforming the geometric properties of data, the procedure is usually called Spatial Extract Transform Load (ETL) (Bédard, Merrett, and Han, 2001), and specialised tools are often needed to conduct such transformations. A geometric operation may be designed to collapse property unit polygons to points that represent their centroids or to transform road segments into a road network.

2.4 Methods for longitudinal analysis of historical data

Longitudinal individual-level data contain continuous information about each individual in the sample. Therefore, they require specific analytical methods that can handle these longitudinal data (Alter et al., 2012). The aim of this section is to describe some of the methods that are included in the term "survival analysis".

2.4.1 Survival analysis – General concepts

A common approach when analysing longitudinal historical demographic data is to perform survival analysis (often called "event history analysis" in historical demography). Survival analysis is a collection of statistical methods, mainly regression models, that are adapted for longitudinal data and used in a variety of fields, such as epidemiology, medicine and engineering. These models examine the time up to a particular event occurrence (Mills, 2011). The dependent variable in these models is the time it takes a particular event to occur, for example, migration, death or birth.

When an event takes place, the term *failure* (which can refer to a positive event) is used, and the term *survival time* describes the time it takes for a failure to occur. The survival time is analysed in terms of how it is affected by one or more independent variables. There are three core concepts in survival analysis: the probability density function, the survival function and the hazard function.

If the dependent variable is the survival time T it takes for a failure to occur, then its probability density function (also referred to as the instantaneous failure rate (Stevenson, 2009)) $f(t)$ describes the instantaneous probability that a failure will occur at time t. Mathematically, it can be expressed as (Mills, 2011):

$$f(t) = \frac{dF(t)}{d(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (1)$$

where *F(t)* is the cumulative distribution function, and *P* is the probability of occurrence between the time interval (t, t + Δt).

The survival function *S(t)* describes the probability that *T* is equal to or greater than a specific time *t*, for example, the probability that an individual survives beyond 80 years. Mathematically, *S(t)* can be defined as (Cleves et al., 2010):

$$S(t) = 1 - F(t) = P(T \geq t) \quad (2)$$

Thus, *S(t)* decreases over time. When no failures occurred, *S(t) = 1*, and when all failures occurred, *S(t) = 0*.

Lastly, the hazard function (or hazard rate) *h(t)* expresses the probability that a failure will occur at time *t* given that a failure has not yet occurred. Mathematically, it is defined as (Mills, 2011):

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \quad (3)$$

Notably, the difference between the density function and the hazard function is that the former is an unconditional probability unaffected by any independent variables, whereas the latter is a conditional probability that changes over time and by independent variables (Sainani, 2008).

2.4.2 Censoring and truncation

In longitudinal demographic studies, it is seldom possible to cover the survival time for all the individuals completely. This is an important aspect to consider; otherwise, serious biases may be introduced in the results. Censoring means that failure most likely occurred, but it has not been observed; truncation means that no information exists about the occurrence of the failure (Mills, 2011). However, in survival analysis, when dealing with failures that we know remove the entity from the observation, such as death, truncation often means that failure cannot possible have occurred before the observation (Cleves et al., 2010). There are various types of censoring and truncation, but the two most common in longitudinal demographic studies are “right censoring” and “left truncation” (Cleves et al., 2010):

Right censoring: Here, the observation of a subject is lost before the failure has occurred. This is common if a demographic sample is taken of a specific geographic area and an individual migrates from this area before the event (e.g., death) has occurred. Right censoring is also common when a specific time period is studied and an individual does not experience the event during the period. This censoring is usually assumed to occur randomly; therefore, the observations are

included in the analysis until they are censored (if right censoring occurs in a non-random way, however, it is more problematic).

Left truncation: This is common when individuals enter an ongoing observation such that there is no knowledge about their survival time prior to their observation, for example, when a person migrates into the geographic study area. If mortality is the event, then it cannot possibly have occurred before the person migrated (if it did, it would not have been observed). Left truncation may introduce biases depending on what is studied and if the biases are non-random; however, they are usually handled the same as random right-censored observations.

Most survival models for longitudinal data are able to account for these types of censoring and truncation. Thus, individuals are analysed during the time they are observed until the event occurs or until they are censored. Therefore, we avoid biases that could be introduced if only subjects that experienced the event are studied (Alter et al., 2012).

2.4.3 Survival models

The main objective of survival models is to study how one or more independent variables affect the hazard rate and to compare the relative risks of these variables (Alter et al., 2012). For data that contain continuous observations[3] in which we know the exact date of the event, the basic types of survival models are non-parametric, semi-parametric and parametric. Non-parametric models do not include assumptions about the shape of the hazard function nor do they model the possible effect that the independent variables have on the hazard function. Two common non-parametric models are the life table and the Kaplan-Meier method. These models are useful as a first step for describing the data, often visually. They also aid in the identification of the shape of the hazard function (for a specific population) to determine whether it follows a particular distribution.

Semi-parametric and parametric models are able to model how multiple independent variables affect the hazard function. These variables can be both fixed and time varying (i.e., the value of a variable changes over time). The main difference between these two models is that the semi-parametric model does not make any assumptions about the shape of the hazard, whereas the parametric

[3] For data based on discrete observations in which it is only known that an event occurred between two observations, discrete time methods must be applied (not addressed in this thesis).

model does. Thus, semi-parametric models are best for hazard functions that do not follow a particular form.

The Cox proportional hazard model (or Cox regression) is a common semi-parametric models used in longitudinal survival analysis. The cox model is a product of two functions: the baseline hazard h_0 and a linear function of the independent variables describing the relative risk. These functions are fitted separately to the data using a partial likelihood function. The Cox model is defined as (Sainani, 2008):

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}} \quad (4)$$

where x is an individual-specific independent variables and β is an unknown parameter. If all the x values are 0, then $h_i(t) = h_0$; therefore, h_0 is called the baseline hazard (Stevenson, 2009). The baseline hazard can take any form, but the Cox model assumes that the effects of the independent variables are constant in proportion to the baseline hazard. In other words, the ratio between two hazard functions should be constant over time (this is called the proportional hazards assumption) (Mills, 2011). For example, the difference of the hazards between a smoker and a non-smoker should not change over time. To test this proportional hazard assumption, a test of the correlation between Schoenfield residuals and the survival time is conducted. If the proportional hazard assumption is violated, then a common approach is to stratify the model, either in time or by the independent variable that violates the assumption, or create a time-dependent version of the violating variable (Sainani, 2008). There are several other diagnostics that can be applied to the survival models, such as Martingale residuals that check for non-linearity of the independent variables and a Goodness-of-fit that tests the overall fit of the model (cf. Cleves et al., 2010).

Overall, these survival models can be successfully applied to longitudinal historical demographic data. Geographic context variables can be constructed on the micro-level and included as either fixed or time-varying independent variables.

2.5 Sources for longitudinal historical geographic data

This section describes the requirements and sources for creating longitudinal historical geographic data. Note that the focus is rural areas rather than urban areas because the study area and available source data (see Chapter 3) cover rural parishes. Thus, some sources specific to urban areas are not covered in this section.

2.5.1 Requirements

Geographic data that are created for integration with longitudinal demographic data need to fulfil particular requirements. Primarily, they need to be longitudinal, i.e., have a sufficient resolution and complete coverage with respect to the demographic data that are to be studied (Campbell et al., 2004). Specifically:

First, all geographic data need to cover the same space and times as the demographic data. Demographic data that spread over an area that is too large may reduce the chances of finding available sources. For instance, creating geographic data for demographic data that are sampled from one or several specific regions may be less problematic than creating data for a demographic dataset that follows individuals throughout their lifecycles, regardless of where they live and move.

Second, the data need to be longitudinal, specifically, the geographic objects need to be traced through time and their changes (if they are not static) need to be recorded. Only then will we be able to account for those variables that are time-varying. The events associated with the objects may be recorded as well, but this may be more important if the geographic objects themselves are analysed.

Third, the geographic data need to have sufficient quality. Geographic quality is described according to the terms used by the ISO 19157:2013 standard *Geographic information -- Data quality* (ISO, 2013). Relevant quality elements from this standard are *completeness* (presence or absence of objects), *thematic accuracy* (e.g., the classification correctness of objects), *temporal quality* (e.g., accuracy of time measurements) and *positional accuracy* (e.g., geometrical accuracy of an object). The quality of geographic data must be defined in relation to an application. For example, Zandbergen (2007) studied the impact of geocoded streets' (used as residence locations) positional accuracies when analysing individual-level exposure to traffic-related air pollution. He found that the locations, which had a median positional accuracy of approximately 40 metres or higher, introduced major biases to the exposure analysis. Generally, the better the data quality is the more accurate results can be obtained; however, creating high quality datasets may be very costly which has to be considered.

Fourth, the objects used for assigning a location to individuals must represent an area where the specific persons lived or spent most of their time. In a rural region, the location is the building a person lived in or the field they worked in; in an urban region, the location is the building or city block they lived in or the place they worked.

2.5.2 Sources for geographic data: Historical maps

The most common sources of historical geographic data are maps. Various types of cadastral and economic maps link individuals to locations and create context variables. These maps contain detailed information about property unit boundaries and buildings, which can be used to link individuals to their place of living. Topographic maps, such as military maps, can also be used for creating context variables. Additionally, modern geographic data can be used for estimating geographic context variables that are static in time, for example, soil data and elevation.

Large-scale cadastral maps were first created in 16th century Europe. These maps contain information about properties and their owners. A textual document with details about the owner(s), the area of the property unit and its taxation value is often linked to the map (Kain and Baigent, 1992). Cadastral maps were created either by individuals who wanted an inventory of their lands or by the state to keep track of taxable property units or to plan land reforms. The maps generally covered the properties within a parish, a city, a town, or one or a few property units (Beech and Mitchell, 2004). The information we can acquire from cadastral maps is mainly boundaries of property units and settlements. However, they also often contain information about land use, vegetation and communications, which can be used for creating context variables.

Medium-scale military maps can be a resource for creating geographic context variables. Military maps were mostly topographic maps with the purpose of mapping terrain, communications (e.g., roads and railways), and physical objects, such as buildings, rivers, wetlands, and forests. The quality of roads and different types of forests were sometimes described. These maps did not usually include economic boundaries or documented information about the individuals living in the areas. Hence, military maps can mainly be used for obtaining information about the physical objects, but they are seldom suited for linking individuals to locations. However, because objects such as buildings were mapped, they may be used in combination with cadastral maps to determine if a building that existed at one point in time on a cadastral map also existed later or earlier on a military map.

Maps are snapshots of geography at specific times; however, they can be merged into sequential snapshots and combined with textual sources to fulfil the longitudinal requirement (cf. Paper III). In terms of the data quality requirements, cadastral maps generally had a higher resolution than military maps and thus most likely a better positional accuracy. For example, in Sweden, the scales of 19th century land cadastral and economic maps were approximately 1:1,000 – 1:2,000 when covering single property units, 1:4,000 – 1:8,000 when covering a parish or a town, and approximately 1:20,000 when covering several parishes or towns. The

military maps, however, vary between scales of 1:20,000 and 1:200,000. A common rule of thumb is that the positional accuracy of objects on (modern) maps is approximately 0.5 mm multiplied with the scale of the map (Longley, et al., 2010). Thus, the positional accuracies of a cadastral map with a scale of 1:4,000 and a military map with a scale of 1:50,000 are approximately 2 metres and 25 metres, respectively. However, the final positional accuracy of objects digitised from such maps also depend on the methods used to create the historical map, the georeference process, the reference maps used during the georeferencing, and the quality of the digitisation of the objects (see, e.g., Podobnikar, 2009). Thus, the positional accuracy of the digitised objects is expected to be lower than the original accuracy of the historical map. Nevertheless, the more large-scale maps we can use, the better the final positional accuracy is expected to be.

Furthermore, the temporal accuracy is usually accurate for military maps because they documented how the area looked at the specific time. However, the reported date of the map creation is more uncertain. Cadastral maps that documented land reforms, however, were often maps of planned areas. Therefore, there is uncertainty whether some of the planned areas on these maps were actually implemented, and if so, at what time (Olsson, 2012). Lastly, completeness may be an issue for the maps. Here, completeness means the under- or over-representation of objects on the maps. Such issues may depend on the purpose of each map (i.e., what objects were considered important to the document), the skills of the surveyors', or whether there were any time constraints during the mapping process that resulted in missing objects (Olsson, 2012).

2.5.3 Sources for geographic data: Textual sources

Textual sources for historical geographic data are those sources that can be linked and combined with historical maps. These may be textual sources that document and plan changes to the geographic objects or demographic data, such as household registers, parish registers, vital statistics and censuses. The key is that the sources contain a locator of sufficient resolution that can be linked to an object digitised from the historical maps. Then, they can be used as an observation that helps estimate the lifeline of the objects. However, to determine whether changes have occurred to a geographic object, the textual sources need to provide indications of a geometric change. For property units, the sources are commonly periodical tax registers, which often provide information about the productivity and size of a farm.

2.5.4 Examples of Swedish sources

The following paragraphs provide an overview of Swedish sources of historical geographic data. Regarding historical maps, the most important historical maps available are geometrical maps, enclosure maps, military topographic maps, and economic maps.

The first large-scale (1:5,000) historical maps in Sweden were geometrical maps (*Geometriska jordeböcker*) created during 1630-1650 by the newly established Swedish Land Survey (Lantmäteriet, 2014a). Taxation was the main purpose of these maps, which mapped villages and property units, as well as relevant meadows and forests (Kain and Baigent, 1992). Buildings can also be identified on the maps. However, these maps are unevenly scattered across the country; thus, they are not available for all areas (Lantmäteriet, 2014a). For example, the study area described in Chapter 3 is not covered by these maps.

Enclosure maps/land survey maps/ (1750 – 1927) are another important resource for geographic data (Figure 5a). The main purpose of these maps was to map the three enclosure movements/land reforms (*storskifte*, *enskifte*, and *laga skifte*) (Lantmäteriet, 2014b). The scale of these maps is usually 1:4,000 for the infield areas (croplands and meadows) and 1:8,000 for the outfields (woodlands). Buildings and property units are the main objects that can be identified on these maps. The names of objects are often included on the maps, which makes it possible to link them with textual sources. Additional textual documents that describe the property units and their owners are commonly available.

The military topographic survey maps of Scania (1815-1820) (Swedish, *Skånska rekognosceringskartan*) contain topographical descriptions and textual information about parishes and villages. The maps have a scale of 1:20,000, and they were created for military interests. Hence, the topography and physical objects were well documented. Land cover and land use, such as buildings, roads, water bodies, wetlands and forest were mapped; however, juridical and economic borders, such as property units, were not mapped (except for fences, which may indicate the area of a property unit). The maps only contain village-level names; therefore, it is not possible to identify houses or property units by name (Fältmäteribrigaden, 1986).

Another military topographic map is the *Generalstabskartan* (1827-1971) (Lantmäteriet, 2006). Similarly to the military topographical survey maps, this topographic map describes the landscape, including elevation, land use, communications and buildings. At a scale of 1:50,000 - 1:200,000, these maps have a low resolution compared to the other historical maps. On these maps, property units are not possible to identify, and buildings are only point objects on the map, which makes them difficult to identify.

Finally, *Häradsekonomiska kartan*, an economic map, (1859-1934) was produced in several map series between 1859 and 1934 and is usually at a scale of 1:20,000. This map was partly based on the land survey maps, and it describes land use, vegetation, settlements, communications and economic boundaries (Lantmäteriet, 2014c). Each property unit on the map has linked textual information about the address, owner, taxation value, etc.

Additionally, cadastral dossiers often describe the geometry of a specific property unit. Lantmäteriet (the Swedish mapping, cadastral and land registration authority) has archived these dossiers from the mid-1700s to the present; thus, all cadastral procedures that were conducted during this time were saved. The cadastral dossiers reported cadastral procedures, such as subdivisions and partitions for property units. They often contain a map describing the borders of the property units, both before and after the cadastral procedure. Lastly, they contain protocols that describe how different ordinances were implemented, as well as the rights for each of the property units and information about the owners (Lantmäteriet, 2014d).

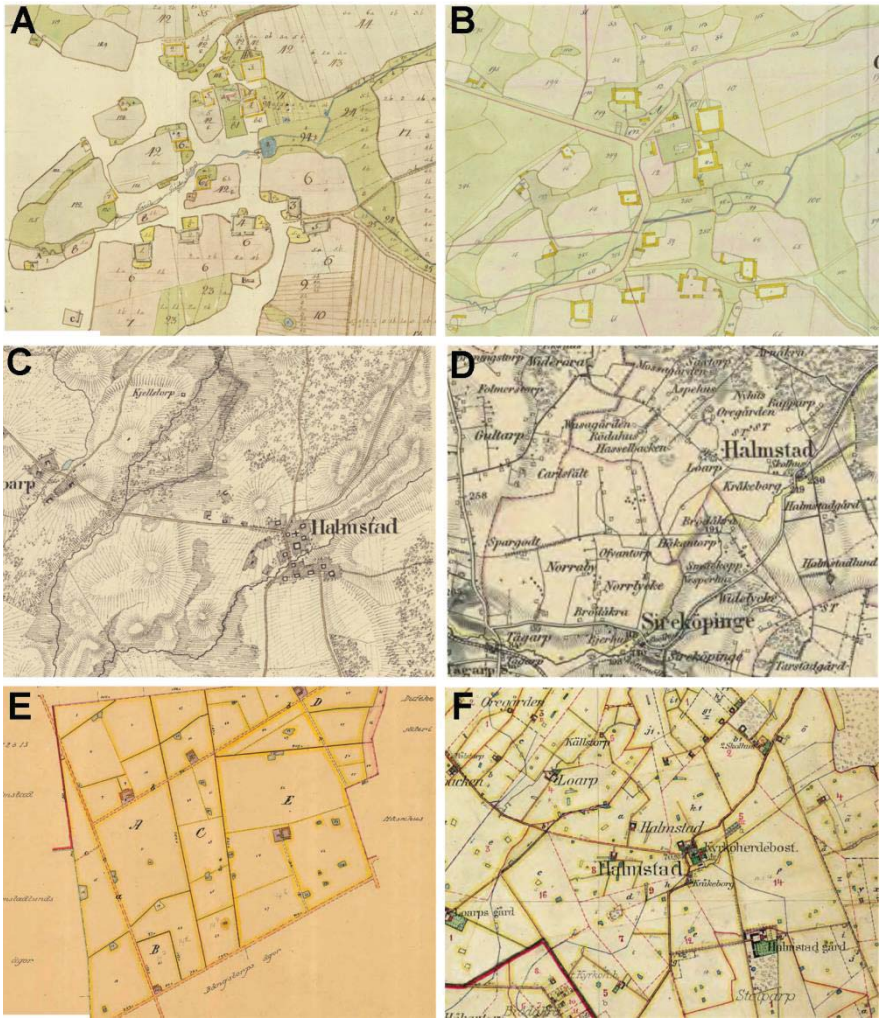

Figure 5: Examples of Swedish historical maps covering Halmstad Village. Maps A-C show the village before the land reform, whereas maps D-F show the village after the land reform. A) Land survey map (*Inägodelning*) 1796; B) Land survey map (*Enskifte*) 1827; C) Military topographic survey (*Skånska rekognosceringskartan*) 1815-1820; D) Topographic map (*Generalstabskartan*) 1860; E) Cadastral dossier in 1913 registering a subdivision; and F) Economic map (*Häradsekonomiska kartan*) 1910-1915 (Source: Lantmäteriet 2014e).

The annual Swedish poll-tax registers (Swedish *Mantalslängder*) are an important information source for geographic data. The poll-tax registers were established for per head taxes of all household members and existed from 1635 to 1938. Foremost they include the addresses of each household and its members as well as the taxation value of their property units (Swedish *Mantal*). Although the taxation value was a measure of the productivity of the farm and not the area, they

remained fixed over time unless a change such as a subdivision or partition in the property unit occurred (Wannerdt, 1982; Svensson, 2001). This means that a change in the taxation value is indicative of a geometrical change in the property unit's boundary through some cadastral procedure. Therefore the poll-tax registers can be a source for estimating the object lifelines of property units.

2.6 Studies of integrating geographic and demographic data

Many historical demographic studies have utilised geographic data and made important contributions to our understanding of how the geographic environment shapes human lives (e.g., Ekamper, Poppel and Mandemakers, 2011; Gregory, 2008; Gutmann et al., 2005; Haines and Hacker, 2011; Schmertmann et al., 2011; DeBats, 2008, 2011; Ekamper, 2010; Gilliland, Olson and Gauvreau, 2011; Villarreal et al., 2014). Important sources for such studies are the national historical GIS databases that were created in the last decade. However, as mentioned, most databases on the national level contain geographic data in low resolution (e.g., parish boundaries).

An example of a historical GIS database that contains longitudinal medium- or high-resolution geographic data is the Danish project DigDag (Dam, 2013). In this project, approximately 25,000 small administrative units called *ejerlav* were digitised from 1660 to the present in Denmark. These units represented the total area used for agriculture in a village or for a noble farm (after the land reforms, they remained as administrative areas) (Dam, 2013). Because ejerlavs often represent small areas, they are an important resource if they can be integrated with historical demographic micro-data.

Studies integrating longitudinal geographic and historical demographic micro-data in rural areas are rare. Using large-scale historical maps for studying landscape changes is more common. For example, Bender et al. (2005a; 2005b) analyse the long-term change in the land use in Germany by digitising historical cadastral maps (scales 1:5,000) for the period 1850-2000. In combination with cadastral maps, they add information from land registers and aerial photographs. From the maps and the registers, they obtain information about ownership of property units, soil quality, age of farm owners and other socio-economic information. They also use a digital elevation model (DEM) to obtain altitude and slope data. The resulting data model uses a temporal snapshot for geographic and textual information. Similar studies that analyse landscape changes with help of historical maps include Skaloš et al. (2011), who use historical military maps and

orthophotos for several periods, along with land registers, to study long-term changes. Moreover, Hamre et al. (2007) digitise a large-scale (1:2,000) cadastral map from 1865, combined with a field survey using the total number of stations, to detect land cover and structural changes in detail.

Regarding geographic data at a larger scale in urban areas, Ekamper (2010) digitised a cadastral map of the Dutch city of Leeuwarden from 1832 and linked its cadastral units with population censuses and registers from 1839. By doing so, it was possible to explore and reveal patterns of population density, infant mortality and socio-economic differences at detailed geographic levels (at a single moment in time).

Another noteworthy study in urban areas was conducted by Villarreal et al. (2014), who created the Historical Urban Ecological (HUE) dataset. The authors reconstructed historical ward boundaries (administrative units within cities) for the period 1830-1930 in seven large cities in the USA (Baltimore, Boston, Brooklyn, Chicago, Cincinnati, New York and Philadelphia) and linked them to ward-level demographic data. The ward boundaries were reconstructed using digitised historical street centre lines from 1930. In addition, they digitised sewer and water pipelines. A temporal snapshot model was used to represent these data. The street centrelines were then used to geocode longitudinal individual-level data of the Union Army veterans and U.S. Colored Troops from 1816 to 1949. To geocode the veterans' places of residence, they used addresses reported in censuses, pensions and medical documents; however, these addresses were not standardised. In addition, the address systems within the cities, as well as the street names, changed frequently throughout the study period as the cities developed. Thus, many addresses were only valid for short time periods. Therefore, the authors manually geocoded the addresses and used complementary historical sources, such as address change documents and historical maps, to identify the correct spatial location of each address. By doing so, they were able to identify the streets on which most of the individuals lived (Villareal et al., 2014). According to CPE (2013), 28,538 address records of 7,302 veterans were geocoded (information about the data quality of the geocoded individuals, however, does not appear to be available). This extensive work presents a great opportunity to analyse geographic context factors on the individual level. However, the changing geography is not fully modelled using object lifelines or similar techniques, which may be required when creating dynamic context variables.

There are several studies that use modern detailed geographic data and longitudinal micro-data, although many of them have studied short time periods, used aggregated data or have not accounted for human mobility (Meliker and Sloan, 2011). One reason for these shortcomings is the integrity concerns of attaching detailed geographic identifiers to individuals (Meliker and Sloan, 2011).

Accounting for human mobility refers to the ability to trace individuals' residences throughout their life, not simply their place of birth and death. For large datasets, this is commonly performed by geocoding addresses in civic registration registers that contain residential histories of individuals.

For example, Nordsborg et al. (2014) performed a space-time cluster analysis of breast cancer occurrences in Copenhagen, Denmark, for the period 1971-2003. They adjusted the space-time cluster analysis by controlling for (aggregated) socioeconomic factors and individual-level reproductive factors, both which are known to increase the risk of breast cancer (using parametric logistic regression analysis). Thus, they identified geographic areas that correlate with breast cancer in time and space that could not be explained by non-spatial factors. They geocoded the residences of approximately 9,000 individuals (consisting of one group diagnosed with cancer and two independent control groups). The geocoding process was straightforward; they used the unique personal identification numbers of the individuals to trace their place of living by matching these numbers to the Danish Civil Registration System. This civil registration contained the residential addresses (on the building level) of the individuals and the dates of moves. Then, the addresses in the registration system were matched to Danish standardised and official addresses, which contained geographic coordinates. The study does not reveal whether the addresses remained constant for the entire period or whether they were time-dependent for specific years.

Moreover, in a study analysing environmental effects on the disease Amyotrophic Lateral Sclerosis (ALS), Sabel et al. (2009) geocoded addresses from the Finnish Central Population Register. They tracked the residential histories down to the building level of 1,000 individuals diagnosed with ALS and 1,000 control persons from 1964 to 1985-1995 (years of their deaths). However, because most of the individuals were born before 1964, which is when the Finnish population register began collecting digital addresses, truncation occurred for most of the subjects.

When using smaller (modern) datasets, structural interviews are an accurate technique for obtaining detailed information about residential histories. For example, Meliker et al. (2010) conducted a population-based case-control study in Michigan, USA, in which they analysed moderate arsenic intake in drinking water. They studied 411 individuals diagnosed with bladder cancer between 2000 and 2004 and 566 individuals from a control group. They traced both the residential histories and the places of work of the individuals through interviews (as well as other characteristics, such as health habits) by asking the subjects where they lived throughout their lives. Information was also gathered about where the fluids they drank came from. Thus, they were able to obtain accurate estimates of exposure not only at their homes but also from their work locations and other sources of exposure. By estimating the arsenic concentrations of the nearby wells, they were

able to analyse their lifetime exposure to arsenic. Studies using similar data collection techniques have been conducted (Gallagher et al., 2010; De Roos et al., 2010; Pronk et al., 2013; James et al, 2013). For example, James et al. (2013) studied lifetime exposures to arsenic in drinking water and their effects on diabetes. They conducted similar structural interviews for 141 cases and 488 control individuals to trace the residential histories and other variables of the individuals. They also collected information about the wells over the time period to create longitudinal geographic context variables. Furthermore, De Roos et al. (2010) analysed the residential proximity to industrial facilities and the risk of the disease non-Hodgkin lymphoma (NHL). They first measured the coordinates using GPS receivers to locate the current home of each of the 864 cases and 684 controls that were recruited during 1998-2000. Then, they combined the coordinates with the residential histories for the last 10 years, as obtained by interviews. The authors obtained the locations and information of industrial facilities for this period. Thus, it was possible to construct geographic context variables for a 10-year period by calculating the proximity of the individuals to the industries (anywhere in the US). Using interviews to collect historical data is not possible, but a qualitative approach using a smaller sample could allow us to trace a few individuals in more detail.

Overall, there are few historical studies that have constructed or used longitudinal geodemographic databases on the micro-level. This is more common in studies using modern data because of available digital civic registers that contain standardised addresses and the possibility to directly communicate with the individuals in the study sample. However, the above studies seldom model the long-term changes in geography (sometimes because there is no need for it); when long-term changes are considered, temporal snapshots are most often used instead of object lifelines. Using the latter would enable the creation of a "true" spatio-temporal database that allows us to create time-varying geographic context variables.

3 Data and study area

A central outcome of Paper III is the longitudinal geodemographic database on micro-level. In addition, this database will be used in future studies for longitudinal demographic analysis with geographic factors. Hence, it is important to describe the source data that the database is based on.

3.1 Study area

Papers I and III use longitudinal demographic data from the Scanian Economic Demographic Database (SEDD), which was created by the Centre for Economic Demography (CED) at Lund University (Bengtsson, Dribe and Svensson, 2012) in collaboration with the Regional Archives in Lund. SEDD contains longitudinal and individual-level demographic and economic information about all persons that have lived in nine parishes in southern Sweden (Scania) from the 17th century onwards (cf. Bengtsson and Dribe, 1997; CED, 2013). Out of these nine parishes, 60 historical maps for five rural parishes have been georeferenced and digitised, namely, Hög, Kävlinge, Kågeröd, Sireköpinge and Halmstad. Hence, these five parishes constitute the study area (a total of approximately 130 km^2) in Papers I and III (Figure 6). Although several of the historical maps originated in the 18th century, the study period is 1813-1914 due to the demographic data availability; catechetical examination registers documenting migration and household compositions are only available for these parishes since 1813.

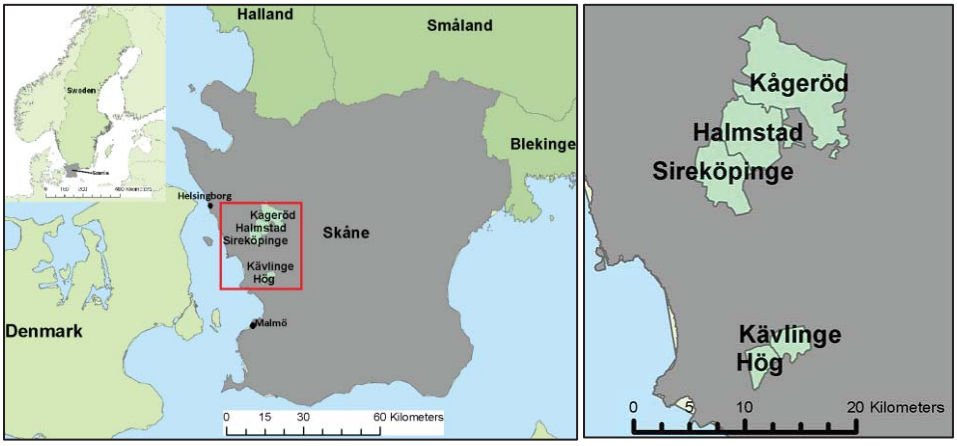

Figure 6: The five parishes of Hög, Kävlinge, Sireköpinge, Halmstad and Kågeröd that constitute the study area for Papers I and III, as well as for future studies (Source: Paper III).

3.2 Demographic data

There are two types of sources in SEDD that have a reference to places and which can be linked with geographic information. The first source contains information about individuals' locations of births, marriages and in- and out-migrations (the individuals are traced from when they are born or in-migrate, to when they die or out-migrate (CED, 2013)). The information comes from vital registers (i.e. birth-, death- and marriage-registers) and catechetical examination registers/parish registers. The sources for the birth locations are the birth and baptism registers where the residence of the child's mother and father was registered. However, only the parish or a vague address is often given and therefore the birth location is known with a low resolution. The marriage and migration locations, which are obtained from other vital registers, have usually a more specific location, e.g., the property unit. The catechetical examination/parish registers, which are available from 1813 and onwards for the study area, observe migrations both within and between parishes. Therefore it is possible to determine when individuals moved within the parish (Dribe and Lundh, 2005).

However, these addresses are not standardised and they do not reflect the changes occurring in the property units. That is, when property units were subdivided or partitioned into smaller units, they did not receive new designations. Because the number of such cadastral procedures became gradually more common in the latter half of the 19th century, property units that share addresses were especially common in the end of the 19th century and beginning of the 20th century.

Therefore, it is often not possible to distinguish which property unit the location names in the birth-, -marriage and migration-registers point to.

The second information source of location names is the Swedish poll-tax registers. Except for the address of each person who had to pay taxes, they also contain information about taxation values, owners and other information related to the property units. Thus, property units that share addresses can be separated by their different tax values and owner names. Moreover, the individuals in SEDD are linked by their households and families to the individuals in the poll-tax registers. Therefore, the addresses in the poll-tax registers can be geocoded and then be used to link individuals to the property units in which they lived. This process is described in more detail in Paper III.

3.3 Geographic data

Approximately 60 historical maps from four map series have been used in this thesis: various land surveyor maps, the military topographical survey map of Scania, topographic maps and economic maps. These maps, which were obtained in digital format from the Lantmäteriet, have been geocoded and digitised within an ongoing project (Table 4). So far, approximately 900 property units, 3000 buildings as well as a substantial amount of roads, railways, streams and wetlands have been digitised.

Table 4: Summary of the digitised historical maps

Map series	Years	No. of map documents	Scale
Land Survey Maps	1757-1863	39	1:4,000-1:8,000
Military Topographical survey	1812-1820	11	1:20,000
Topographic maps	1860-1865	2	1:100,000
Economic maps	1910-1915	7	1:20,000

4 Summary of papers

Chapter 4 summarises the papers that are the basis for the thesis. How the papers are related is described in Section 1.4 and Figure 1.

4.1 Paper I: Extending the Intermediate Data Structure (IDS) for longitudinal historical databases to include geographic data

The aim of this study is to create a data model for standardised the storage and export of longitudinal geodemographic micro-data (first objective in Section 1.2). Such a model facilitates comparative studies in historical demography using geographic data. We accomplish this by extending the Intermediate Data Structure (IDS) version 4 (Alter and Mandemakers, 2014) to include geographic data. IDS is a de-facto standard data model for sharing longitudinal historical data. For example, within the European Historical Population Samples Network (EHPS-Net), at least 15 longitudinal historical databases worldwide aim to transform their data to comply with IDS (Brändström, Mandemakers and Matthijs, 2009).

The IDS data model includes individuals, contexts and the relationships between them. It also includes the possibility to link external geographic data to the contexts and to store point data. However, it cannot store and export common geometric data types in a standardised way. When individuals in historical demographic data are to be linked to detailed physical locations, such as buildings and property units, geometric data types other than points need to be stored in IDS. Thus, we offer the possibility of integrating detailed geographic data within IDS in a new model coined IDS-Geo.

IDS-Geo is a slightly modified IDS model in which we add standardised geometric data types to permit the storage of geometric representations of objects. The IDS-Geo model is designed conceptually, and an eXtensible Markup Language (XML) Schema (with GML elements that specify the geographic data)

is created for the data export. Both of these models allow only geometries based on the OGC/ISO Simple Feature specification.

The conceptual IDS-Geo model is implemented in a case study using historical property units (see Section 3.3 for description of the source data) for the period 1804 to 1914. Thereafter, the data are exported from the database into XML files that are compliant with the specified IDS-Geo XML Schema. To enable integration of longitudinal demographic data and geographic data in IDS-Geo, we included an object lifeline representation for storing the geographic objects. The case study verifies that the IDS-Geo model is capable of handling geographic data that can be linked to demographic data. However, more research is required to test the usability of the model by using fully integrated individual-level demographic and geographic data.

Including geographic data in IDS will improve longitudinal analyses by enabling individual-level spatial analysis, and IDS-Geo facilitates the linkages between individuals and (geocoded) geographic objects. Because the main aim of the IDS structure is to simplify the exchange of historical demographic data, we believe that only geographic data that can link individuals in IDS to spatial locations should be stored in IDS-Geo. Nevertheless, when standard addresses are available, they should be used instead of geographic data. Finally, using standardised exchange formats, such as the specified XML Schema, should aid in data sharing and in the development of extraction and transformation programs.

The scientific contribution of this study is (1) the demonstration of how detailed geographic data can be described and distributed in a standardised way in combination with longitudinal demographic data and (2) the potential future development of the IDS structure.

4.2 Paper II: Making Swedish environmental geodata INSPIRE compliant: A harmonization case study

The purpose of standardised data models, such as IDS, is to allow data from different sources to be combined in a meaningful way. However, as mentioned in Section 2.3, this means that the source data need to be transformed for the standardised (target) data model. If there are large differences between the source data model and the target data model, then the transformation process can be costly (in terms of time or data quality). Thus, the aim of this study is to test and evaluate methods for data harmonisation (second objective in Section 1.2).

We test a data harmonisation approach to make Swedish environmental geographic data and metadata compliant with Nature-SDIplus test versions[4] of the INSPIRE (The Infrastructure for Spatial Information in the European Community) data specifications. INSPIRE is a EU Directive that aims to build a European Spatial Data Infrastructure (SDI) that could provide easy access to harmonised data and metadata (EC, 2009).

The main aim of the data harmonisation is to solve the heterogeneities that exist between the source and the target data model. We split the transformations into geographic and non-geographic components and use standardised formats to permit vendor neutrality. Then, we conduct compliance tests on the data and metadata specifications by validating them against both XML Schema and Schematron. Finally, we identify transformation processes that may be costly or have negative impacts on the data quality. The harmonised data and metadata are published as network services compliant with OGC Web Service specifications. The output from our method is data and metadata that are valid according to the Nature-SDIplus data specifications and metadata profiles.

By splitting the harmonisation into two manageable components, we avoided some limitations of XML Schema translations in the current (2011) spatial transformation tools. We successfully harmonise the Swedish environmental geodata and metadata such that they are compliant with the Nature-SDIplus data specifications and metadata profiles; thus, the method is feasible. When testing against the Schematron rules, non-compliances that had been missed during the XML Schema compliance tests are found. This indicates the importance of supplementary semantic rules to ensure compliance for a data specification. Finally, costly processes are identified, which are caused by missing elements and by unstructured information in the source data; degradation of the positional and thematic accuracies occur during the harmonisation.

The main scientific contribution of this study is the development of a feasible method for harmonising and testing the compliance of Swedish environmental data against standardised data specifications. Such harmonisation methods are important for the authorities that need data that complies to INSPIRE and other standardised data specifications.

---

[4] The European project Nature-SDIplus developed data and metadata specifications for three INSPIRE Annex III themes: habitats and biotopes, bio-geographical regions and species distributions. These served as a foundation for the thematic groups that corresponded to the INSPIRE specifications that were later released.

4.3 Paper III: Methods to create a longitudinal integrated demographic and geographic database on the micro-level: a case study of five Swedish rural parishes, 1813-1914

The aim of this paper is to develop a general methodology for creating databases that can be used for adding geographic factors on the micro-level to longitudinal historical demographic analysis (third objective in Section 1.2). The first steps of this methodology are to scan, geocode and digitise historical maps. These steps result in a set of temporal snapshots of geographic objects, such as property units, buildings, roads and land cover. However, these snapshots only show information about the state of an object for single points in time. Longitudinal demographic data, however, have continuous object lifelines; individuals are traced from birth to death and related events are recorded. To link an individual to a geographic object, we need to model when the geographic object was created, changed and ceased to exist (including its geometry through the different stages). Thus, in the second step, objects in temporal snapshots are transformed into an object-lifeline data model. This transformation is performed by combining the historical maps with supplementary longitudinal register data, such as poll-tax registers, and snapshot data, such as land survey acts. In the last phase, the individuals in the demographic data are linked to one or several physical locations (depending on how they moved) using standardised locations. By utilising historical textual sources (poll-tax registers, church books, cadastral dossiers, etc.), it is possible identify the property units where the individuals lived. By identifying these names in the historical map information, we can link the individuals to a location.

The methodology is evaluated in a case study using longitudinal individual-level data from the SEDD database (Bengtsson et al., 2012). First, we geocode and digitise approximately 60 Swedish historical maps. In total, approximately 900 property units, 3,000 buildings and a substantial number of roads, railways, streams and lakes are digitised. Of these objects, we transform the snapshots of the property units into longitudinal object lifelines. Then, we link 80,431 individuals in the five parishes (described in Chapter 3) for the period 1813-1914 to the property units they lived in. The resulting database is tested using fundamental queries for spatio-temporal data. These tests imply that the methodology is feasible and that the resulting longitudinal geodemographic data can be used in demographic studies and with geographic factors.

The scientific contribution lies in both the methods for creating longitudinal geodemographic databases on the micro-level and in the database resulting from the case study. To our knowledge, the result is a unique contribution in terms of

linking individuals over such long time periods to longitudinal geographic data on the micro-level. Using the database will present a variety of opportunities to study how geographic factors on the micro-level affected human living conditions throughout history.

4.4 Paper IV: An old mom keeps you young: Mother's age at last birth and offspring longevity in 19th century Utah

The aim of this paper is to analyse the intergenerational effects of late child bearing on offspring's adult longevity in pre-transitional Utah, USA. Hence, this paper performs longitudinal demographic analysis where geographic factors could be included on the micro-level (fourth objective in Section 1.2).

Studies have found that the life expectancy of women who experience late menopause and prolonged reproduction increases. The positive correlation is believed to be caused by biological or genetic factors. If this is the case, then we would also expect to find evidence for the intergenerational transmission of life longevity benefits. Thus, we used longitudinal data for the 19th and 20th centuries from the Utah Population Database, which reflects a natural fertility population, to investigate the relationship between prolonged natural fertility of mothers and their offspring's survival rates in adulthood.

We used five parametric Gompertz proportional hazard models with gamma-distributed frailty to analyse the effect of a mother's age at her last childbirth on her offspring's survival beyond age fifty. In all models, the dependent variable is an individual's age at death. The mother's age at her last childbirth is chosen as the independent variable. These models reveal that the offspring of mothers who were naturally fertile until a relatively old age lived significantly longer. These results, which control for various childhood, adult and senior health conditions, as well as shared frailty, indicate that there is a positive relationship between a mother's age at her last childbirth and her offspring's longevity; this strengthens the notion that menopause age is a good predictor of this relationship.

Notably, in all five models, the shared frailty factor for siblings was approximately 30%. A high value of shared frailty indicates that there are unobserved characteristics that are shared between siblings. Such factors may be geographic; therefore, these models can be better fine-tuned when geographic factors are included at the micro-level.

To our knowledge, this study is the first to apply an intergenerational demographic analysis on this specific topic. Specifically, there is a positive correlation between a mother's age at her last childbirth and her offspring's longevity, and menopause age is a good predictor for this relationship. Hence, the scientific contribution of this research is a broader understanding of genetic factors that affect mortality patterns beyond age fifty.

5 Concluding remarks

5.1 Conclusions

The overall aim of this licentiate thesis was to improve historical demographic research by including geographic factors on the micro-level in longitudinal historical analysis. This was foremost achieved by developing methods for creating integrated longitudinal demographic and geographic databases (called longitudinal geodemographic databases) on the micro-level (Paper III) and by contributing to the standardisation of longitudinal historical databases that include geographic data (Paper I and Paper II).

Four objectives were outlined for this licentiate thesis. The first objective was to extend a standardised data model for longitudinal demographic data to include geographic data. This was achieved in Paper I by introducing IDS-Geo, which is an extended version of the standardised data model IDS. In IDS-Geo, we offered the possibility of integrating detailed geographic data within IDS by adding standardised geometric data types to permit the storage of geometric representations. The option to include geographic data in IDS will improve longitudinal analyses by facilitating individual-level spatial analysis. Such inclusion is a good option when address data are not available. To address data heterogeneities, it is also important to use standardised exchange formats when distributing the data. In this study, we designed an XML Schema that should aid data sharing and the development of extraction and transformation programs.

The second objective was to develop and evaluate harmonisation methods for making source data compliant with standardised data models (such as IDS-Geo). This was performed in Paper II by testing and developing a method for harmonising Swedish environmental data and metadata and compliance testing the data against standardised data models and specifications. Although Paper II did not harmonise historical demographic and geographic data, the methods therein can be applied for such data.

The third objective was to develop methods for creating longitudinal geodemographic databases that can include geographic factors on the micro-level in demographic research. The objective was achieved in two ways (Paper III). First, we developed a methodology for creating integrated databases. The central

objective was to transform snapshots of geographic objects (digitised from historical maps) into longitudinal object lifeline time representations, and to link individuals to these geographic objects. Second, the longitudinal geodemographic database, which links 80,431 individuals over long time periods (1813-1914) to longitudinal geographic data on the micro-level, will be an important and unique resource for historical demographic research.

The fourth objective was to perform longitudinal demographic analysis where geographic factors can later be included. In Paper IV, we conducted a longitudinal study in which we analysed the intergenerational effects of late child bearing on the offspring's adult longevity in pre-transitional Utah, USA. We found a positive correlation between a mother's age at her last childbirth and her offspring's longevity; menopause age is a good predictor for this relationship. We also found high values of shared frailty among siblings, which indicates that unobserved characteristics are shared between them. Such factors may be geographic; therefore, these models can be better fine-tuned by including geographic factors at the micro-level.

5.2 Future studies

In future studies, we aim to both improve and use the longitudinal geodemographic database created in Paper III.

The improvement of the database includes, for example, the study of how additional longitudinal geographic data can be constructed for computing context variables. It also includes more accurate estimations of the residential areas of individuals, specifically, determining residences within property units (using building data). Additionally, the issue of the database's quality will be addressed.

We aim to obtain new insights regarding how geographic factors affect human living conditions on the micro-level. With the help of the geodemographic database, we have information about the residential areas of individuals throughout their lives. Additionally geographic data can be used to estimate relevant context variables. Our principal aim is to study exposure to diseases in terms of population density, accessibility to road networks and distance to wetlands (i.e., habitats for mosquitos carrying malaria). We also aim to extend the study of population density by performing social network analysis, for example, by computing network centrality measures to estimate central individuals in parishes and how such centrality affected their exposure to diseases. Such analysis, however, requires longitudinal and accurate road data; thus, improvements to the database are needed. Figure 1 (bottom right) shows the overall procedure for using

the geodemographic database. The database is first used in combination with modern geographic data to compute relevant geographic context factors. These context variables are then used in conjunction with the physical location of the individuals to perform longitudinal survival analysis with geographic factors.

References

Alter, G. C., Gutmann, M. P., Leonard, S. H., and Merchant, E. R. 2012. Introduction: Longitudinal analysis of historical-demographic data. *Journal of Interdisciplinary History* 42(4):503-517.

Alter, G., and Mandemakers, K. 2014. The Intermediate Data Structure (IDS) for Longitudinal Historical Microdata, version 4. *Historical Life Course Studies* 1:1-26.

Alter, G., Mandemakers, K., and Gutmann, M. P. 2009. Defining and Distributing Longitudinal Historical Data in a General Way Through an Intermediate Structure. *Historical Social Research-Historische Sozialforschung* 34(3):78-114.

Armstrong, M. P. 1988. Temporality in spatial databases. *Proceedings from GIS/LIS* 88(2):880-889. San Antonio, TX.

Bédard, Y., Merrett, T., and Han, J. 2001. Fundamentals of spatial data warehousing for geographic knowledge discovery. *Geographic data mining and knowledge discovery* 2:53-73.

Beech, G., and Mitchell, R. 2004. *Maps for family and local history*. Toronto: Dundurn.

Bender, O., Boehmer, H. J., Jens, D., and Schumacher, K. P. 2005a. Using GIS to analyse long-term cultural landscape change in Southern Germany. *Landscape and Urban Planning* 70(1):111-125.brink.

Bender, O., Boehmer, H. J., Jens, D., and Schumacher, K. P. (2005b). Analysis of land-use change in a sector of Upper Franconia (Bavaria, Germany) since 1850 using land register records. *Landscape Ecology* 20(2):149-163.

Bengtsson, T. and Dribe, M. 1997. Economy and Demography in Western Scania, Sweden, 1650-1900. *EAP Working Series Paper No.10*. Kyoto: International Research Center for Japanese Studies.

Bengtsson, T., Dribe, M. and Svensson, P. 2012. *The Scanian Economic Demographic Database, version 2.0*. Lund: Lund University, Centre for Economic Demography.

Berman, L. M. 2003. *A Data Model for Historical GIS: The CHGIS Time Series*. Technical Report. Cambridge, MA: Harvard Yenching Institute.

Brändström, A., Mandemakers, K., and Matthijs, K. 2009. *Proposal for an ESF Research Networking Programme – Call 2009*. Retrieved from http://www.iisg.nl/hsn/documents/ehps-net.pdf.

Campbell, C., Kurosu S., Manfredini, M., Neven, M., and Bengtsson, T. 2004. Appendix: Sources and Measures. In *Life Under Pressure: Mortality and Living Standards in Europe and Asia, 1700-1900*, edited by T. Bengtsson, C. Campbell, and J. Z. Lee. Cambridge, MA: MIT Press Books.

Casati, R., and Varzi, A. 2010. Events. In *The Stanford Encyclopedia of Philosophy*, edited by Z. N. Edward. Retrieved from http://plato.stanford.edu/archives/spr2010/entries/events/.

Center for Population Economics (CPE). 2013. *Historical Urban Ecological (HUE) Data Set Historical Geocoding Guide*. Technical Report. Retrieved from http://hue.uadata.org/assets/documentation/Geocoding_the_UA_and_USCT.pdf.

Cleves, M., Gould, W., Gutierrez, R., and Marchenko, V. Y. 2010. *An introduction to survival analysis using Stata, Third edition*. College Station, TX: Stata Press.

Dam, P. 2013. *Integrating time and space in a digital-historical administrative atlas*. Unpublished manuscript.

De Roos, A. J., Davis, S., Colt, J. S., Blair, A., Airola, M., Severson, R. K., ... and Ward, M. H. 2010. Residential proximity to industrial facilities and risk of non-Hodgkin lymphoma. *Environmental research* 110(1):70-78.

DeBats, D. A. 2008. A tale of two cities: Using tax records to develop GIS files for mapping and understanding nineteenth-century US cities. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 41(1):17-38.

DeBats, D. A. 2011. Political Consequences of Spatial Organization Contrasting Patterns in Two Nineteenth-Century Small Cities. *Social Science History* 35(4):505-541.

Dumpala, S. R., and Arora, S. K. 1981. Schema Translation Using the Entity-Relationship Approach. *Proceedings of the Second International Conference on the Entity-Relationship Approach to Information Modeling and Analysis* 81:337-356.

Ekamper, P. 2010. Using cadastral maps in historical demographic research: Some examples from the Netherlands. *History of the Family* 15(1):1-12.

Ekamper, P., Poppel, F., and Mandemakers, K. 2011. Widening Horizons? The Geography of the Marriage Market in Nineteenth and Early-Twentieth Century Netherlands. In *Navigating time and space in population studies*, edited by M. P. Gutmann, G. D. Deane, E. R. Merchant, and K. M. Sylvester. Dordrecht: Springer.

European Commission (EC). 2009. Commission Regulation (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards the Network Services, *Official Journal of the European Union* 274:9-18.

Fitch, C. A., and Ruggles, S. 2003. Building the national historical geographic information system. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 36(1):41-51.

Fältmäteribrigaden. 1986. *Skånska rekognosceringskartan*. Gävle: Lantmäteriet.

Gallagher, L. G., Webster, T. F., Aschengrau, A., and Vieira, V. M. 2010. Using residential history and groundwater modeling to examine drinking water exposure and breast cancer. *Environmental health perspectives* 118(6):749-755.

Gilliland, J. A., Olson, S. H., and Gauvreau, D. 2011. Did Segregation Increase as the City Expanded? The Case of Montreal, 1881–1901. *Social Science History* 35(4):465-503.

Goodchild, M. F., Yuan, M., and Cova, T. J. 2007. Towards a general theory of geographic representation in GIS. *International journal of geographical information science* 21(3):239-260.

Gregory, I. N. 2008. Different Places, Different Stories: Infant Mortality Decline in England and Wales, 1851-1911. *Annals of the Association of American Geographers* 98(4):773-794.

Gregory, I. N., and Ell, P. S. 2007. *Historical GIS: technologies, methodologies, and scholarship*. Cambridge: Cambridge University Press.

Gregory, I., and Southall, H. 2005. The Great Britain Historical GIS. *Historical Geography* 33:132-34.

Grenon, P., and Smith, B. 2004. SNAP and SPAN: Towards dynamic spatial ontology. *Spatial cognition and computation* 4(1):69-104.

Gutmann, M. P., Deane, G. D., Lauster, N., and Peri, A. 2005. Two population-environment regimes in the Great Plains of the United States, 1930–1990. *Population and Environment* 27(2):191-225.

Haines, M. R., and Hacker, J. D. 2011. Spatial aspects of the American fertility transition in the nineteenth century. In *Navigating time and space in population studies*, edited by M. P. Gutmann, G. D. Deane, E. R. Merchant, and K. M. Sylvester. Dordrecht: Springer.

Hamre, L. N., Domaas, S. T., Austad, I., and Rydgren, K. 2007. Land-cover and structural changes in a western Norwegian cultural landscape since 1865, based on an old cadastral map and a field survey. *Landscape ecology* 22(10):1563-1574.

INSPIRE. 2014. *D2.5: Generic Conceptual Model, Version 3.4*. Framework Document. Retrieved from http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4rc3.pdf.

International Organization for Standardization (ISO). (2013). *ISO 19157: Geographic information -- Data quality*. Geneva, Switzerland: ISO/TC 211.

James, K. A., Marshall, J. A., Hokanson, J. E., Meliker, J. R., Zerbe, G. O., and Byers, T. E. 2013. A case-cohort study examining lifetime exposure to inorganic arsenic in drinking water and diabetes mellitus. *Environmental research* 123:33-38.

Kain, R. J., and Baigent, E. 1992. *The cadastral map in the service of the state: A history of property mapping*. Chicago: University of Chicago Press.

Lantmäteriet. 2014a. *Geometriska jordeböcker*. Retrieved May 21, 2014, from http://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Lantmateristyrelsens-arkiv---LMS/Geometriska-jordebocker/.

Lantmäteriet. 2014b. *Skifteskartor*. Retrieved May 21, 2014, from http://www.lantmateriet.se/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Lantmateristyrelsens-arkiv---LMS/Skifteskartor/.

Lantmäteriet. 2014c. *Häradsekonomiska kartan*. Retrieved May 21, 2014, from http://www.lantmateriet.se/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Rikets-allmanna-kartverks-arkiv---RAK/Haradsekonomiska-kartan/.

Lantmäteriet. 2014d. *Förrättningsakter i Arken*. Retrieved May 21, 2014, from http://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Lantmateriets-arkiv/Forrattningsakter-i-Arken/.

Lantmäteriet. 2014e. *Historical Maps*. Retrieved from http://historiskakartor.lantmateriet.se/arken/s/search.html.

Lantmäteriet Geodesienheten. 2006. *Generalstabskartan*. Retrieved from http://www.lantmateriet.se/Global/Kartor%20och%20geografisk%20information/GPS%20och%20m%C3%A4tning/Geodesi/Ordlista/Generalstabskartan.pdf.

Lehto, L. 2007. *Schema translations in a web service based SDI*. Paper presented at the 10th AGILE International Conference on Geographic Information Science, Aalborg University, Denmark. Retrieved from http://www.agile-online.org/Conference_Paper/CDs/agile_2007/PROC/PDF/29_PDF.pdf.

Longley, P., Goodchild, M. F., Maquire, J. D., and Rhind, W. D. 2010. *Geographic information systems and science*. Hoboken, NJ: John Wiley & Sons.

Lutz, M., Sprado, J., Klien, E., Schubert, C., and Christ, I. 2009. Overcoming semantic heterogeneity in spatial data infrastructures. *Computers & Geosciences* 35(4):739-752.

Meliker, J. R., and Sloan, C. D. 2011. Spatio-temporal epidemiology: principles and opportunities. *Spatial and Spatio-temporal Epidemiology* 2(1):1-9.

Meliker, J. R., Slotnick, M. J., AvRuskin, G. A., Schottenfeld, D., Jacquez, G. M., Wilson, M. L., ... and Nriagu, J. O. 2010. Lifetime exposure to arsenic in drinking water and bladder cancer: a population-based case–control study in Michigan, USA. *Cancer causes & control* 21(5):745-757.

Mills, M. (2011). *Introducing survival and event history analysis*. Newbury Park, CA: SAGE Publications.

Nordsborg, R. B., Meliker, J. R., Ersbøll, A. K., Jacquez, G. M., Poulsen, A. H., and Raaschou-Nielsen, O. 2014. Space-time clusters of breast cancer using residential histories: A Danish case-control study. *BMC cancer* 14:255.

Olsson, P. 2012. *Ömse sidor om vägen; Allén och landskapet i Skåne 1700-1900*. PhD diss., Lund University.

Peuquet, D. J. 1994. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers* 84(3):441-461.

Podobnikar, T. 2009. Georeferencing and quality assessment of Josephine survey maps for the mountainous region in the Triglav National Park. *Acta Geodaetica et Geophysica Hungarica* 44(1):49-66.

Pronk, A., Nuckols, J. R., De Roos, A. J., Airola, M., Colt, J. S., Cerhan, J. R., ... and Ward, M. H. (2013). Residential proximity to industrial combustion facilities and risk of non-Hodgkin lymphoma: a case–control study. *Environmental Health*, 12(1):20.

Pultar, E., Cova, T. J., Yuan, M., and Goodchild, M. F. 2010. EDGIS: a dynamic GIS based on space time points. *International Journal of Geographical Information Science* 24(3):329-346.

Regnauld, N., and McMaster, R. B. 2007. A synoptic view of generalisation operators in *Generalisation of geographic information: Cartographic modelling and applications*, edited by W. A. Mackaness, A. Ruas, and L. T. Sarjakoski. Amsterdam: Elsevier.

Sabel, C. E., Boyle, P., Raab, G., Löytönen, M., and Maasilta, P. 2009. Modelling individual space–time exposure opportunities: A novel approach to unravelling the genetic or environment disease causation debate. *Spatial and spatio-temporal epidemiology* 1(1):85-94.

Sainani, K. 2008. *Introduction to Survival Analysis*. PowerPoint slides. Retrieved from http://www.pitt.edu/~super1/lecture/lec33051/index.htm.

Schmertmann, C. P., Potter, J. E., and Assunção, R. M. 2011. An Innovative Methodology for Space-Time Analysis with an Application to the 1960–2000 Brazilian Mortality Transition. In *Navigating time and space in population studies*, edited by M. P. Gutmann, G. D. Deane, E. R. Merchant, and K. M. Sylvester. Dordrecht: Springer.

Shaw, S. L., Yu, H., and Bombom, L. S. 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12(4):425-441.

Sheth A. P., and Larson J. A. 1990. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Comput Surveys* 22(3):183–236.

Skaloš, J., Weber, M., Lipský, Z., Trpáková, I., Šantrůčková, M., Uhlířová, L., and Kukla, P. 2011. Using old military survey maps and orthophotograph maps to analyse long-term land cover changes–Case study (Czech Republic). *Applied geography* 31(2):426-438.

Stevenson, M. 2009. *An introduction to survival analysis*. Retrieved from http://www.ngatangata.ac.nz/massey/fms/Colleges/College%20of%20Sciences/Epicenter/docs/ASVCS/Stevenson_survival_analysis_195_721.pdf.

Svensson, P. (2001). *Agrara entreprenörer. Böndernas roll i omvandlingen av jordbruket i Skåne ca 1800-1870*. PhD diss., Lund University.

Van de Weghe, N., De Roo, B., Qiang, Y., Versichele, M., Neutens, T., and De Maeyer, P. 2014. The continuous spatio-temporal model (CSTM) as an exhaustive framework for multi-scale spatio-temporal analysis. *International Journal of Geographical Information Science* 28(5):1047-1060.

Vanhaute, E. 2003. *Construction of a GIS for the territorial structure of Belgium.* Technical Report. Retrieved from http://www.hisgis.be/start_en.htm.

Villarreal, C., Bettenhausen, B., Hanss, E., and Hersh, J. 2014. Historical Health Conditions in Major US Cities: The HUE Data Set. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 47(2):67-80.

Wannerdt, A. 1982. *Den svenska folkbokföringens historia under tre sekler*. Retrieved June 15 from https://www.skatteverket.se/privat/folkbokforing/omfolkbokforing/folkbokforingigaridag/densvenskafolkbokforingenshistoriaundertresekler.4.18e1b10334ebe8bc80004141.html.

Worboys, M. 2005. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science* 19(1):1-28.

Worboys, M., and Duckham, M. 2004. *GIS: A computing perspective*. Boca Raton, FL: CRC Press.

Yi, J., Du, Y., Liang, F., Zhou, C., Wu, D., and Mo, Y. 2014. A representation framework for studying spatiotemporal changes and interactions of dynamic geographic

phenomena. *International Journal of Geographical Information Science* 28(5):1010-1027.

Yuan, M. 2000. Modeling geographic information to support spatiotemporal queries. In *Life and Motion of Socio-Economic Units*, edited by A. U. Frank, J. Raper, and J. P. Cheyland. London: Taylor and Francis.

Yuan, M., and Hornsby, K. S. 2010. *Computation and visualization for understanding dynamics in geographic domains: a research agenda*. Boca Raton, FL: CRC Press.

Zandbergen, P. A. 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC public health* 7(1):37.