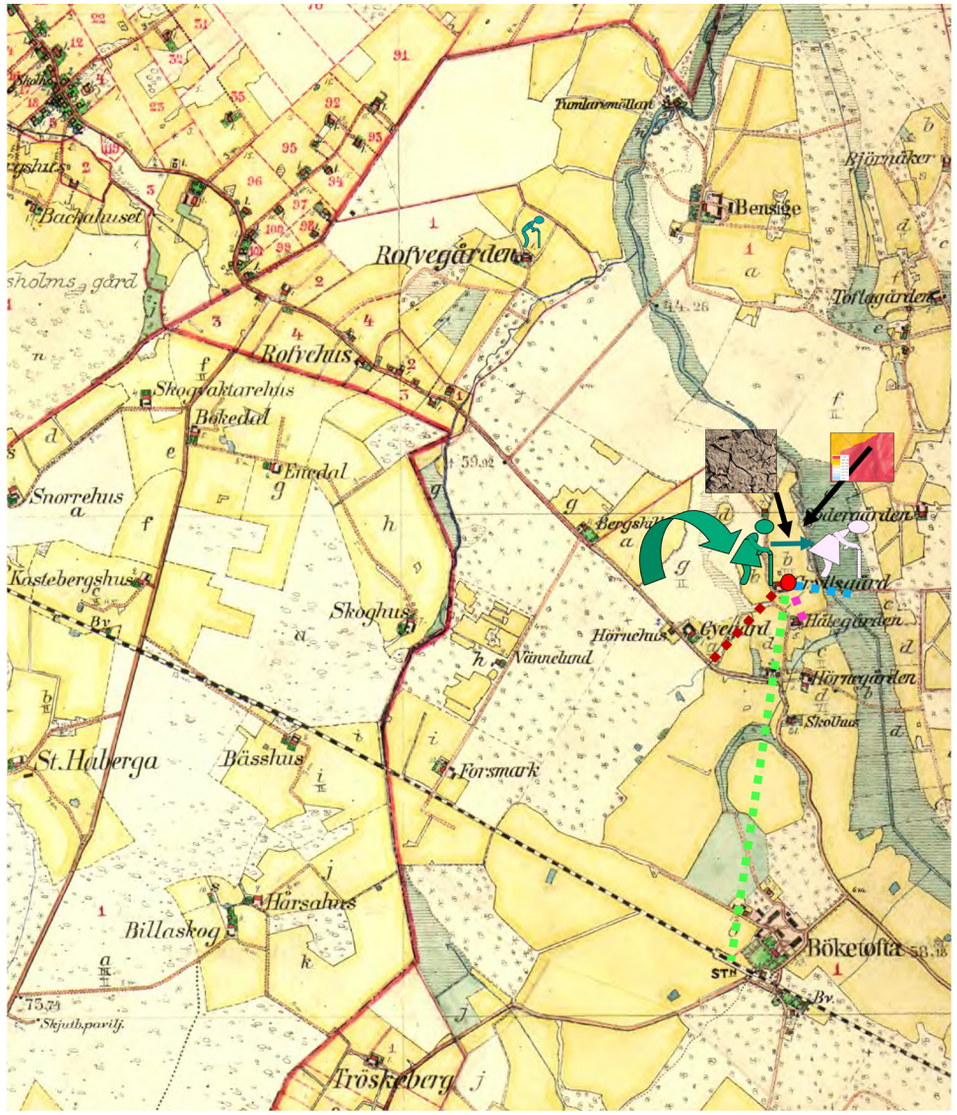
Life paths through space and time: Adding the micro-level geographic context to longitudinal historical demographic research

Hedefalk, Finn

2016

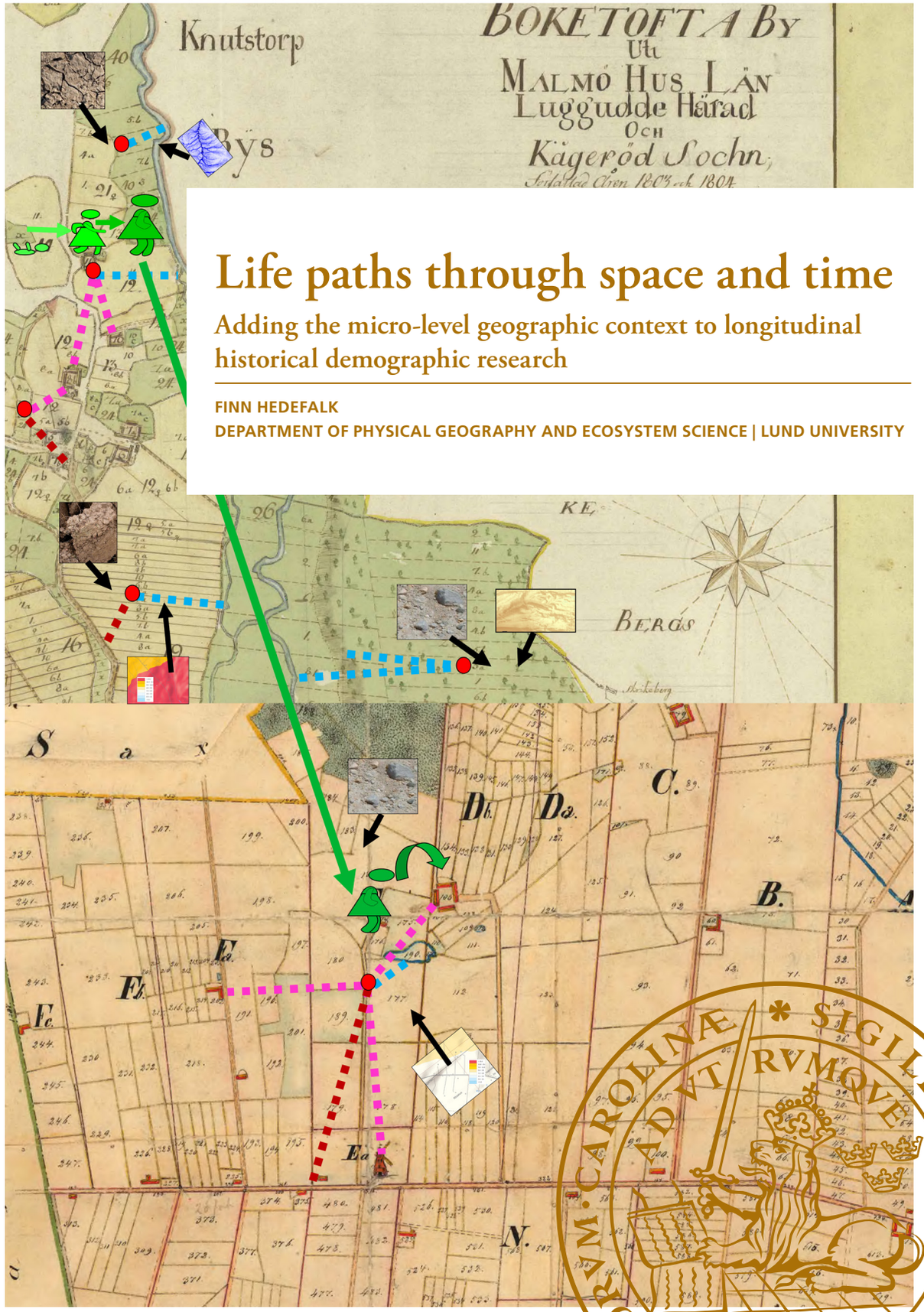
Document Version:
Publisher's PDF, also known as Version of record



Printed by Media-Tryck, Lund University 2016
Nordic Ecolabel 3041 0903

FINN HEDEFALK

Life paths through space and time



Life paths through space and time

Adding the micro-level geographic context to longitudinal historical demographic research

FINN HEDEFALK
DEPARTMENT OF PHYSICAL GEOGRAPHY AND ECOSYSTEM SCIENCE | LUND UNIVERSITY

LUND UNIVERSITY

Lund University
Faculty of science
Department of Physical Geography and Ecosystem Science
ISBN 978-91-85793-63-1
E-ISBN 978-91-85793-64-8

2016

Life paths through space and time

Adding the micro-level geographic context to longitudinal historical demographic research

Finn Hedefalk

LUND
UNIVERSITY

DOCTORAL THESIS
by due permission of the Faculty of science, Lund University, Sweden.
To be defended at Världen auditorium, Geocentrum I, Sölvegatan 10, Lund.

Friday, November 4, 2016, at 10:15.

Faculty opponent
Professor Chris Dibben

University of Edinburgh

Organization	Document name
LUND UNIVERSITY Department of Physical Geography and Ecosystem Sciene, Sölvegatan 12, SE-223 Lund, Sweden	DOCTORAL DISSERTATION
	Date of issue: 2016-10-03
Author(s): Finn Hedefalk	Sponsoring organization: eSSENCE

Title and subtitle: Life paths through space and time: Adding the micro-level geographic context to longitudinal historical demographic research

Abstract: Historical demographic research is central to understanding past human behaviours and traits, such as fertility, mortality and migration. An essential part of historical demography is conducting longitudinal analyses at the micro-level, which involves the detailed follow-up of individuals over long time periods throughout their lives. By including the geographic context in such analyses, we can study how the environment has affected human living conditions over long time periods. However, the use of micro-level geographic factors in historical longitudinal analyses is seldom feasible because of the absence of data. Thus, studies have been primarily limited to examining the geographic context on an aggregated level.

In five papers, this thesis contributes to historical demographic research by adding and utilising micro-level geographic factors in longitudinal historical analyses. First, we develop and implement methods for creating detailed longitudinal geographic data that are integrated with longitudinal demographic micro-level data. We then perform novel studies of the effect of the environment on demographic outcomes at the micro-level.

Papers I-III include micro-level geographic factors with longitudinal historical analyses. Paper I contributes to the standardisation of longitudinal demographic data by geographically extending the Intermediate Data Model (IDS) using standardised exchange formats. Paper II presents methods for geocoding longitudinal demographic databases. The core part of the process is to transform geographic objects as snapshots (digitised from historical maps) into longitudinal object-lifeline time representations (with information about the creation, changes and ends of each object). Individuals are subsequently linked to these geographic objects. We geocoded the Scanian Economic Demographic Database (SEDD) from 1813 to 1914. Approximately 53,000 individuals who lived in five rural parishes in southern Sweden are linked to the property units where they lived. Geographic snapshot data (e.g., roads and buildings) are also created. Paper III improves and evaluates the geocoded database, and wetlands in object-lifelines are added.

Paper IV investigates how longitudinal demographic analyses are affected by different geocoding levels and presents methods for quantifying geographic factors. In a novel case study, we use a geocoded database to analyse the effect of population density and proximity to wetlands on the risk of dying for the period 1850-1914. We show that even small differences between the property units and coarser geographic levels and the choice of method for quantifying the geographic factors substantially affected the results of the demographic analyses. Therefore, geocoding to property units is likely needed for fine-scale analyses at distances within a few hundred metres. In addition, proximity to wetlands affected the mortality of women, which may indicate exposure to malaria-transmitting mosquitoes.

Paper V focuses on the role of nutrition in historical societies by analysing the effect of soil type on child mortality in the five parishes between 1850 and 1914. Certain soil types seem to have influenced agricultural productivity, which in turn affected the nutrition of farmers' children and their risk of dying. This study adds new findings about the importance of nutrition and agricultural productivity regarding child mortality in preindustrial Sweden.

To conclude, this thesis enables the novel inclusion of geographic micro-level factors into historical longitudinal studies. The results increase our understanding about how the micro-level geographic context affected individual living conditions throughout history. The geocoding of the demographic database has also proved to be a unique and important resource for historical and geographic research and a starting point for additional research that includes the micro-level geographic context.

Key words: Geographic micro-level factors, longitudinal historical data, geocoding, historical demography, 19th century, individual level, spatio-temporal analysis, detailed geographic data, property units, geographic context.

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language: English
ISSN and key title		ISBN: 978-91-85793-63-1
Recipient's notes	Number of pages 241	Price
	Security classification	

Life paths through space and time

Adding the micro-level geographic context to longitudinal historical demographic research

Finn Hedefalk

LUND
UNIVERSITY

Cover maps from the Lantmäteriet historical archive.

Front page: Kågeröd parish land survey map from 1804 (*Karta öfver inägorne till Böketofta by*) (top), and Halmstad parish enclosure map from 1863 (*Karta övfer ägorne till Oregården mfl*.) (bottom).

Back page: an economic map from 1910-1915 (*Häradsekonomiska kartan*) over the previous village of Böketofta.

Drawings by Finn Hedefalk

Faculty of science
Department of Physical Geography and Ecosystem Science

ISBN 978-91-85793-63-1
E-ISBN 978-91-85793-64-8

Printed in Sweden by Media-Tryck, Lund University
Lund 2016

— For a while, Criticism travels side by side with the Work, then Criticism vanishes and it's the Readers who keep pace. The journey may be long or short. Then the Readers die one by one and the Work continues on alone, although a new Criticism and new Readers gradually fall into step with it along its path. Then Criticism dies again and the Readers die again and the Work passes over a trail of bones on its journey toward solitude. To come near the work, to sail in her wake, is a sign of certain death, but new Criticism and new Readers approach her tirelessly and relentlessly and are devoured by time and speed. Finally the Work journeys irremediably alone in the Great Vastness. And one day the Work dies, as all things must die and come to an end: the Sun and the Earth and the Solar System and the Galaxy and the farthest reaches of man's memory. Everything that begins as comedy ends in tragedy.

Roberto Bolaño, The Savage Detectives

Abstract

Historical demographic research is central to understanding past human behaviours and traits, such as fertility, mortality and migration. An essential part of historical demography is conducting longitudinal analyses at the micro-level, which involves the detailed follow-up of individuals over long time periods throughout their lives. By including the geographic context in such analyses, we can study how the environment has affected human living conditions over long time periods. However, the use of micro-level geographic factors in historical longitudinal analyses is seldom feasible because of the absence of data. Thus, studies have been primarily limited to examining the geographic context on an aggregated level.

In five papers, this thesis contributes to historical demographic research by adding and utilising micro-level geographic factors in longitudinal historical analyses. First, we develop and implement methods for creating detailed longitudinal geographic data that are integrated with longitudinal demographic micro-level data. We then perform novel studies of the effect of the environment on demographic outcomes at the micro-level.

Papers I-III include micro-level geographic factors with longitudinal historical analyses. Paper I contributes to the standardisation of longitudinal demographic data by geographically extending the Intermediate Data Model (IDS) using standardised exchange formats. Paper II presents methods for geocoding longitudinal demographic databases. The core part of the process is to transform geographic objects as snapshots (digitised from historical maps) into longitudinal object-lifeline time representations (with information about the creation, changes and ends of each object). Individuals are subsequently linked to these geographic objects. We geocoded the Scanian Economic Demographic Database (SEDD) from 1813 to 1914. Approximately 53,000 individuals who lived in five rural parishes in southern Sweden are linked to the property units where they lived. Geographic snapshot data (e.g., roads and buildings) are also created. Paper III improves and evaluates the geocoded database, and wetlands in object-lifelines are added.

Paper IV investigates how longitudinal demographic analyses are affected by different geocoding levels and presents methods for quantifying geographic factors. In a novel case study, we use a geocoded database to analyse the effect of population density and proximity to wetlands on the risk of dying for the period

1850-1914. We show that even small differences between the property units and coarser geographic levels and the choice of method for quantifying the geographic factors substantially affected the results of the demographic analyses. Therefore, geocoding to property units is likely needed for fine-scale analyses at distances within a few hundred metres. In addition, proximity to wetlands affected the mortality of women, which may indicate exposure to malaria-transmitting mosquitoes.

Paper V focuses on the role of nutrition in historical societies by analysing the effect of soil type on child mortality in the five parishes between 1850 and 1914. Certain soil types seem to have influenced agricultural productivity, which in turn affected the nutrition of farmers' children and their risk of dying. This study adds new findings about the importance of nutrition and agricultural productivity regarding child mortality in preindustrial Sweden.

To conclude, this thesis enables the novel inclusion of geographic micro-level factors into historical longitudinal studies. The results increase our understanding about how the micro-level geographic context affected individual living conditions throughout history. The geocoding of the demographic database has also proved to be a unique and important resource for historical and geographic research and a starting point for additional research that includes the micro-level geographic context.

Sammanfattning

Historisk-demografisk forskning är central för att förstå mänskliga beteenden och egenskaper genom historien, såsom fertilitet, dödlighet och migration. En väsentlig del inom historisk demografi är att utföra longitudinella analyser på mikronivå. Detta innebär att individer följs kontinuerligt över livscykeln utifrån demografiska händelser (födelse, giftermål, flyttningar och död) under flera generationer. Genom att inkludera geografiska sammanhang i sådana analyser kan vi studera hur geografiska faktorer har påverkat livsförhållanden genom historien. Användningen av geografiska faktorer på mikronivå i longitudinella analyser av historiska befolkningar är emellertid sällan möjligt eftersom data saknas. Därför har analyserna i huvudsak varit begränsade till att utföras på större geografiska regioner.

I denna avhandling ingår fem artiklar som syftar till att stärka och utveckla historisk-demografisk forskning genom att inkludera och studera geografiska faktorer på mikronivå i longitudinella historiska analyser.

Den första artikeln bidrar till standardiseringen av longitudinella demografiska data genom att utvidga en standardiserad datamodell så att geografiska data kan inkluderas. Den andra artikeln integrerar historiska kartor med demografiska data på individnivå. En central del är utveckling av konceptuella metoder för integrering av data och modeller för att hantera olika representationer av tid. Baserat på denna metodologi har vi integrerat historisk och modern detaljerad geografisk information med historisk-demografisk data på mikronivå från Skånes ekonomisk-demografiska databas (SEDD). Databasen omfattar alla individer som levt i fem skånska församlingar under perioden 1646 till nutid. Individerna följs kontinuerligt över livscykeln utifrån demografiska händelser och till dessa data har ekonomiska data (yrke, inkomst, jordinnehav) tillfogats. Vi geokodade c:a 53 000 individer i dessa församlingar till de fastigheter där de levde för perioden 1813-1914. Vidare har vi skapat årlig information om när varje fastighet skapades och upphörde att existera, och om eventuella förändringar i fastighetens geometri. Artikel nummer tre förbättrar och utvärderar den geokodade databasen och inkluderar årlig information om våtmarker.

I den fjärde artikeln studerar vi hur olika geografiska skalor i geokodningen och olika definitioner av geografiska faktorer påverkar kvaliteten på historisk-demografiska longitudinella analyser. Genom att använda den geokodade databasen analyserar vi hur befolkningstäthet och närhet till våtmarker påverkar

risken att dö att under perioden 1850-1914. Studien visar att även små skillnader mellan den mest detaljerade skalan (dvs., fastighetsnivå) och de något grövre geografiska skalorna, och valet av metod för att definiera geografiska faktorer, väsentligt påverkade resultaten av de historisk-demografiska analyserna. Därför behövs sannolikt en geokodning till fastigheter för att möjliggöra finskaliga analyser där geografiska faktorer studeras inom ett par hundra meter. Dessutom visar studien att närhet till våtmarker påverkade dödligheten hos kvinnor, vilket indikerar att våtmarker under denna period ökade exponering för malaria-överförande myggor.

Slutligen undersöker vi i artikel nummer fem hur geografin har påverkat levnadsförhållanden i de fem församlingarna under perioden 1850-1914. Genom att fokusera på kost undersöker vi hur jordtyp har påverkat barnadödligheten på gårdsnivå (via näringsstatus). Studien visar att vissa jordtyper verkar ha påverkat produktiviteten inom jordbruket, vilket i sin tur påverkade näringen för barn i åldrarna 1-15 (vars föräldrar var beroende av jordbruk) och därmed risken för att dö. Denna studie kommer med nya rön om vikten av kost och produktivitet inom jordbruket med avseende på barnadödligheten i det förindustriella Sverige.

Avslutningsvis bidrar denna avhandling med att införa geografiska faktorer på mikronivå i longitudinella historiska studier. Resultaten från avhandlingen ökar vår förståelse för hur geografiska faktorer har påverkat livsförhållanden genom historien. Geokodningen av den demografiska databasen har också visat sig vara en unik och viktig resurs för historisk och geografisk forskning, särskilt eftersom dessa detaljerade data täcker en sådan lång tidsperiod (1813-1914). Förhoppningsvis blir resultatet från denna avhandling en utgångspunkt för ytterligare studier kring spatiala mönster och exponering på mikronivå.

Content

Acknowledgement

When thinking back at my work, I remember one late evening two years ago when I was digitising some property units in Kågeröd parish. I then suddenly got a strong feeling of being a part of something that produces new knowledge. That I was thickening a branch of knowledge, or, perhaps, creating a new small branch, on which sprouts of speculations, or leaves of hypotheses, could grow. That feeling has followed me throughout the rest of the work and given me a sense of purpose that has motivated me to complete this thesis. However, what has motivated me much more, and also made this thesis possible, is all the support of my colleagues, friends and family.

This thesis is the result of collaboration between the Centre for Economic Demography (CED) and the Department of Physical Geography and Ecosystem Science (INES), which has been funded by the Swedish Research Council through the project eSSENCE. I am grateful for the financial support that this project has given me, which has enabled my research project.

First and foremost I want to thank my main supervisor Lars Harrie. I have often been amazed by his efficiency, sharpness and creativity. But what I have appreciated most is that he truly cares for his PhD students in a very unselfish way. He also has a great pedagogical ability and open mind; he has carefully listened to all of my ideas, good or bad, which I think has helped me to dare to think more freely and to becoming an independent scientist. For this, 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. Also Patrick has been a pedagogical supervisor. Specifically I have enjoyed our motivating and relaxed meetings and discussions, where we have bounced ideas back and forth with each other. In some way or another I have always walked away from these meetings with a good feeling.

I also want to thank my third supervisor, Ali Mansourian. My research went in another direction than Ali's; however, we ended up teaching together for several years. During this time, Ali has been a great and patient supervisor, and I have learned a lot on how to hold and plan a course.

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

want to thank Tommy for welcoming me to CED. And I have appreciated all the help from Clas with regards to the SEDD database.

I have enjoyed working together with Luciana and Tommy on Paper V. I believe I have learned much from Luciana's critical thinking and outstanding ability in problem solving, and from Tommy's wide and deep knowledge and expertise in how to pitch a paper. For Paper IV, I have also enjoyed collaborating with Karolina Pantazatou (in addition to Luciana and Lars in this paper), who worked in a very structured and organized way, which eased the work for everyone, and who had many good ideas and an excellent eye for details.

I also want to acknowledge my co-authors on other papers. I enjoyed the cooperation with Saskia Hin and Bartosz Ogórek; especially the process of first working together as group members in the LAHDD course in Michigan (here I take the opportunity to thank the teachers in this excellent course), and then being able to produce something real from it. It has also been a pleasure, as well as a very creative process, to write an article together with Siddartha Aradhya, Jonas Helgertz and Kirk Scott.

Throughout my PhD study, I have had two offices simultaneously (at three places in total) and thus shared space and experiences with many great people. The first two years of my PhD studies I spent at the GIS centre. I truly appreciated the relaxed and familiar atmosphere, the joint travels and the many evenings spent together. Thanks to: Abdulghani, Alex, Andreas, Ehsan, Karin, Lars E, Lina, Micael, Mitch, Mohammadreza, Petter, Roger, Stefan and Ulrik. Special thanks go to Petter Pilesjö for making the GIS Centre to the excellent place it is. Moreover, my appreciation goes to my former and present colleagues (including the previously mentioned co-authors) at the CED for constituting a very stimulating, active and bright group: Andy, Anna, Annika, Björn, Clas, Jeff, Joe, Kriss, Madeleine, Maria, Martin, Martin, Mats, Volha and Zeyuan. I also want to thank my colleagues at my office at INES (too many names to list here). With you, I have enjoyed the sunny lunches at the roof top or in the botanical garden, and all the fika and talks. Finally, I want to thank the colleagues and the staff at the Department of Economic History for making me feel welcome there as well.

I have had some office mates that I want to thank: Christian and Björn. I think I have been lucky with both of you. We have had mutual respect for each other's time to focus on our work, and we have also been able to get into many stimulating discussions. Lastly: Christian for keeping his office clean and organised, and Björn for sharing my own views on keeping it really messy.

A special appreciation goes to those that have been, and still are, my fellow PhD-students (at both departments): for all the good and stimulating time, for the shared experience of travelling towards the same goal; for the PhD fikas and lunches, for the travels and after-works, for the ping-pong games, for the feedback and support on presentations and drafts. In particular I want to acknowledge those of you that have become my good friends (because it is difficult to define sharp borders in

friendship, I leave out the names; however, I assume you know who you are). We have shared thoughts and feelings, days and evenings, joy and stress, and failure and success. Although you deserve a longer paragraph of dedication than this, I hope you realize that these sentences are written with a very high concentration of appreciation.

When looking back at the past, I am grateful to 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 who motivated me to test something new. Moreover, I want to address a special thanks to Junjun Yin and Alexey Tereshenkov for all their valuable help in GIS related issues and for being some of the core motivators to do a PhD. Finally, I thank Isak Willebrand and Rolf-Erik Keck for giving invaluable comments to my PhD application.

One core part of my time as a PhD candidate has been all the ping-pong, badminton, squash, paddle and football, often with colleagues and friends at the work place. These activities have been refreshing sparks which have lightened up my weeks. Especially I want to acknowledge the Monday and Friday football groups at the departments. There is something special with football and I have a hard time finding any other sport that is as close to perfection. I truly enjoy the endless possibilities of various passing combinations, the chess-type kind of games, and the connection you get with the team mates when you together find patterns and create opportunities. The latter is also related to when you have played with the same people for several years; therefore, a special thanks to you within the football groups that have persistently joined the football activities with me for such a long time. And, of course, I want to highlight the past and present members in our football team Poetry in motion: Wenxin, Willy, Unn, Payam, Lin, Jing, Ehsan, Cecilia, Bakhtiyor, Andrew and Ali.

Before I forget it, I am grateful for the existence of the botanical garden in Lund. Without it, I would not have been able to have those almost daily walks in nature, which has been important for reflection and relaxation.

I sincerely appreciate all my other friends for the life outside work. In particular the core part of old friends; jewels that I have the fortune to still keep regular contact with. You are an important foundation in my life, and you have been central motivators to me throughout this process.

Almost finally, many thanks goes to my family: Anna, Carl-Magnus, Dag and Nea for all their support throughout my life, and for shaping me to become the person I am today. Besides that you have been open-minded, reasonable and loving, I have always being able to count on you, which has meant a lot to me.

Finally, I am deeply grateful to my most lovely fiancée Jing for her endless support and love in all kinds of times. I promised her a whole page of acknowledgement, so here it is. Somehow you always manage to laugh, smile and see the positive side of things, even when life has been very stressful. With this, you have made my life happy and colourful as well. You are a tough doer who makes things happen. You are empathic and observant, extremely talented but humble and unselfish. With regards to my PhD research, you have also been an important peer in which I have been able to discuss with you on how to find solutions to difficult problems. Lastly, I owe you a lot of work after my defence for the burden you have taken by fixing our new house.

1 Introduction

1.1 Motivation

Historical demography is the study of past human population dynamics, including fertility, mortality, nuptiality and migration, as well as the relationship between populations and the larger society. Essential resources in demographic research include individual-level data that encompass long time periods, i.e., longitudinal micro-data. Studies can also be performed on aggregated levels, i.e., population groups. Micro-data and aggregated data that encompass long time periods enable the construction of robust demographic models.

A key factor in demographic research is the geographic context. The places where 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 an extensive range of applications in other fields, such as epidemiology, medicine and geography.

Although geographic contexts on the aggregated level have been important components of longitudinal historical studies, geographic contexts on the micro-level have only had 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 are short. Historical demographic data, however, can cover much longer time periods. As these data are usually less constrained by integrity laws than modern data, they can be more freely used. However, standardised addresses are seldom available in historical data; therefore, individuals need to be linked to geographic objects (e.g., buildings or property units).

The ability to track individuals at the micro-level across space and time would provide many new insights into how geographic factors have affected human living conditions throughout history, especially when longitudinal data are used. Longitudinal data enable us to track several generations and accurately analyse their social and biological traits through time, which would facilitate the study of the effect of population densities, social networks and land use on mortality, fertility rates and migration in both the short term and long term.

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. In particular, the period includes the demographic transition, which encompasses the time when mortality, followed by fertility, began to decline in Europe and North America. Before this period, mortality was primarily determined by numerous epidemic diseases, and the population fluctuated. The agricultural and industrial revolutions, which substantially changed our society, are also linked to the demographic transition. People's health improved, which boosted economic development and positively affected agricultural and industrial advances. Before and during the period of mortality decline, mortality rates also substantially varied within and between regions. A discussion remains about the factors that caused these differences; for example, to what extent were the differences caused by variations in nutritional levels or by exposure to diseases (e.g., McKeown, 1976; Smith, 1983; Fogel 1994; 2004; Puleston and Tuljapurkar, 2008; Floud et al., 2011; Bengtsson and Dribe, 2010; 2011). By analysing geographic micro-level factors within these regions, the contribution of new and important knowledge about the causes of these differences and how the geographic context affected demographic outcomes for the study period is possible. Research on this subject is also important for modern societies because the demographic transition in ongoing in several developing countries.

To utilise the micro-level geographic context, we need to geocode individuals in longitudinal demographic databases; i.e., link individuals to physical places where they have lived. Such geocoding requires historical maps, from which geographic objects used for locating people can be digitised (e.g., property units and buildings). These maps also often contain other geographic information that can be used to analyse how various geographic factors affect demographic outcomes. When studying long time periods, geographic changes should be considered. To account for these changes, information should be digitised from multiple maps that cover the same area at several points in time. However, historical maps only show snapshots in time; thus, the events occurring during the time before the maps are produced are unknown. Therefore, the development of methods that transform these snapshots into continuous information about the geography is necessary. To achieve this goal, we can use supplementary textual sources that contain such information about changes in time. Thereafter, we can geocode the demographic databases to account for the changing environment in the analyses.

This thesis aims to improve historical demographic analysis by adding and utilising geographic factors at the micro-level in longitudinal historical research. To include this information, the abovementioned methodological developments for creating detailed longitudinal geographic data, which can be integrated with longitudinal demographic micro-data, are required. Subsequently, we can exploit

this geocoded longitudinal demographic data to contribute to the literature on the effect of the environment on demographic outcomes at the micro-level.

A detailed geocoding of longitudinal demographic databases often involves high costs in terms of time and money. Therefore, this thesis also aims to study methodological issues involved in performing demographic analyses on different levels of aggregation and using different definitions of geographic context variables. These analyses provide insights that will be valuable for other projects that aim to geocode demographic databases on the micro-level.

Another essential aspect of historical demographic research is the ability to compare patterns among population groups and regions, which may provide answers to fundamental questions regarding the demographic outcomes that are determined by society, biology, or both. To conduct such comparisons, we need to obtain standardised data from several regions. Thus, 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 overarching research question of this thesis is how to add and use the micro-level geographic context in longitudinal demographic research. To answer this question, we must link individuals in demographic databases to longitudinal detailed locations. Additional historical geographic data used for computing context variables also need to be constructed. To facilitate comparative studies, we also need to determine how current standardised data models for demographic data can be applied to integrated longitudinal demographic and geographic data. Tests of the geocoded data and demographic studies that use micro-level geographic factors can then be conducted to exemplify how the geographic context can be utilised.

Consequently, the aim of this thesis is to improve historical demographic research by adding the micro-level geographic context to longitudinal historical analysis. This aim consists of three research objectives: (1) to develop methods for including micro-level geographic factors in longitudinal historical analyses; (2) to study how the geographic geocoding level and variable definitions affect the results of demographic analyses; and (3) to investigate the use of micro-level geographic information to improve and extend knowledge of demographic change. The first research objective focuses on the methodology for creating a geocoded database; the second objective focuses on method development with applied analysis; and the third object focuses on applied analysis. The research objectives and related papers are constructed as follows:

- Research objective 1
    - To extend a standardised data model for longitudinal demographic data to include geographic data (Paper I).
    - To develop and implement methods for creating geocoded longitudinal demographic databases that include micro-level geographic factors in demographic research (Papers II, III).
- Research objective 2
    - To study how the demographic longitudinal analyses are affected by different temporal models and geographic levels in the geocoding, as well as the methods used to quantify the geographic context variables (Paper IV)
- Research objective 3
    - To perform longitudinal demographic analyses with geographic factors. We aim to obtain new insights into how geographic factors affect human living conditions at the micro-level. Specifically, the aim is to study the role of nutrition in preindustrial societies by analysing the effect of soil type on the mortality rates of children who lived in five rural parishes in southern Sweden during the period 1850-1914. (Paper V).

1.3 Thesis organisation

Section 1.4 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 section 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. Thereafter, section 2.3 reviews briefly the standardisation work for historical and geographic data. Section 2.4 describes some of the main methods for analysing longitudinal historical demographic data, whereas section 2.5 focuses on methods for analysing historical demographic data with geographic factors. Section 2.6 describes the requirements and sources for creating longitudinal historical geodemographic data. Lastly, section 2.7 reviews related studies of integrating longitudinal demographic and geographic data on the micro-level. Chapter 3 describes the study area of the papers included in the thesis, Chapter 4 summarises the papers that are the basis for the thesis, and Chapter 5 presents the main conclusions and future studies.

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

1.3.1 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., 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. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48(3):153-17.

III. Hedefalk, F., Harrie, L., and Svensson, P. 2016. Spatiotemporal historical datasets at micro-level for geocoded individuals in five Swedish parishes, 1813-1914. *Submitted*.

IV. Hedefalk, F., Pantazatou, K., Quaranta, L., and Harrie, L. 2016. Importance of the geocoding level for historical demographic analyses: A case study of rural parishes in Sweden, 1850-1914. *Submitted*.

V. Hedefalk, F., Quaranta, L., and Bengtsson, T. 2016. Unequal lands: Soil type, nutrition and child mortality in southern Sweden, 1850-1914. *Under review*.

1.3.2 List of contribution

I. FH led the study design, carried out the practical part of the study, interpreted the results together with the co-authors and led the writing.

II. FH led the study design, carried out the practical part of the study, interpreted the results together with the co-authors and led the writing.

III. FH led the study design and carried out the practical part of the study, interpreted the results together with the co-authors and led the writing.

IV. FH contributed to the study design. FH carried out most of the practical parts of the study except for the computations of the distance to wetland variables which was performed by KP. FH interpreted the results together with the co-authors and led the writing.

V. FH contributed to the study design, carried out the practical part of the study, in which LQ contributed to the set-up of the template in Stata. FH interpreted the results together with the co-authors and led the writing.

1.3.3 Related papers

The author has also been involved in the following related papers.

Hedefalk, F., and Östman, A. 2011. Making Swedish Environmental Geodata INSPIRE Conformant: A Harmonization Case Study. *Mapping and Image Science*, 3:30-37. (Manuscript included in the licentiate thesis).

Hin, Saskia., Ogórek, B., and Hedefalk, F. 2016. An old mom keeps you young: Mother's age at last birth and offspring longevity in 19th century Utah. *Biodemography and Social Biology*, 62(2):164-181. (Manuscript included in the licentiate thesis).

Aradhya, S., Hedefalk, F., Helgertz, J., and Scott, K. 2015. Region of Origin: Settlement Decisions of Turkish and Iranian Immigrants in Sweden, 1968-2001. *Population, Space and Place*. In press.

1.4 Methodology

This section presents the overall methodology used in the thesis. The five papers are connected in the following way (Figure 1.1). Papers I-III addresses 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. In Paper II and 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. Then, we geocode the longitudinal demographic data (i.e., link individuals to physical locations). The result is a geocoded longitudinal historical database on the micro-level. This database is thereafter used in Paper IV and V. In Paper IV, we compute geographic context variables on different geographical levels and by using different quantification methods. Thereafter we compare the results of the variables and the results from the demographic analyses with geographic factors. In Paper V, we perform a longitudinal demographic analysis with geographic factors.

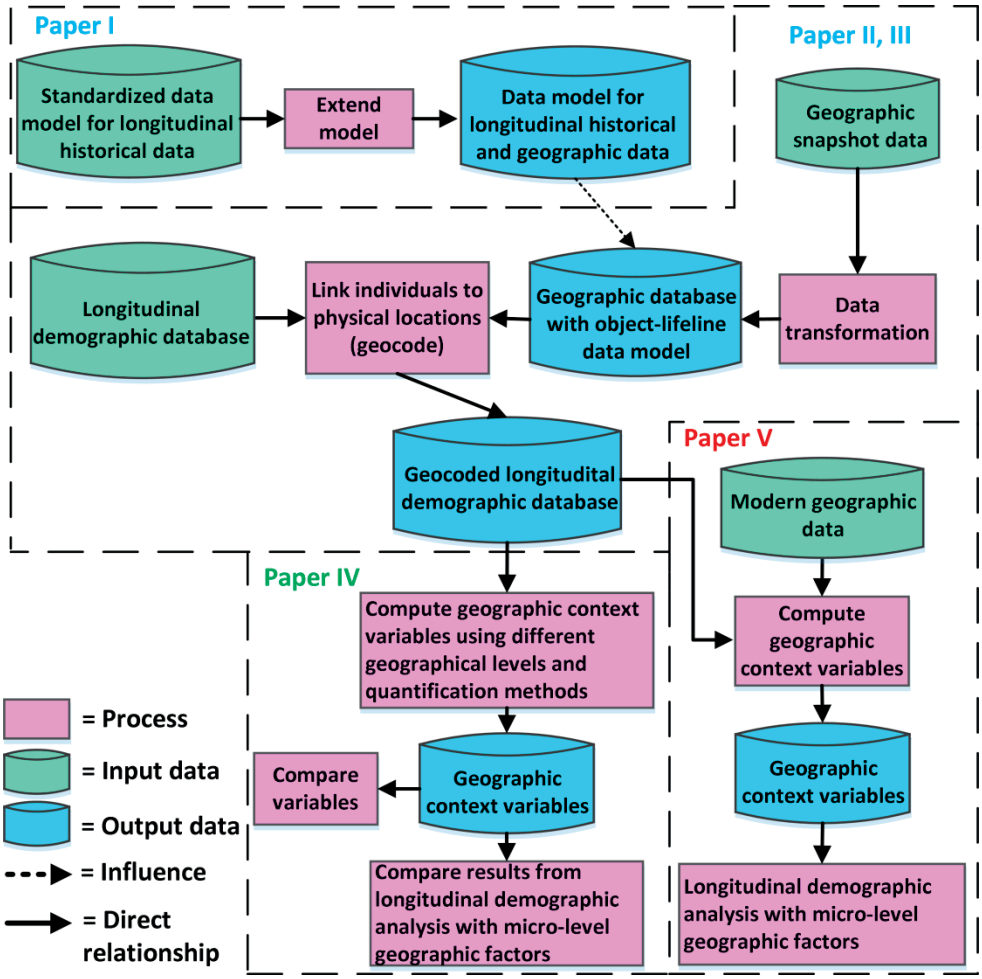


Figure 1.1: Overall methodology used in the licentiate thesis.

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

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 2.1 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 2.1, the entities are either objects or events.

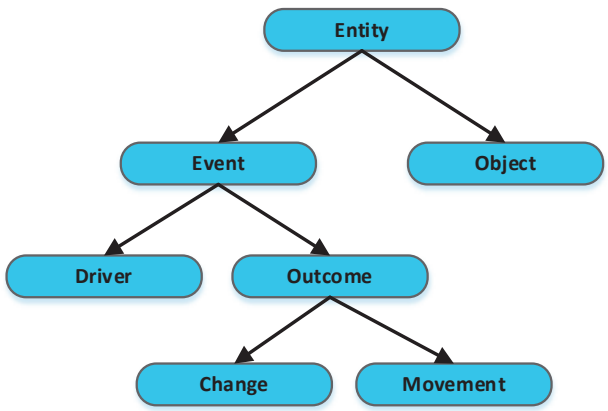


Figure 2.1: Terminology of the spatio-temporal data

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?

It is important to consider the abovementioned relationships between the spatio-temporal entities when modelling and creating geographic longitudinal data. Modelling such relationships in an appropriate way is the base for analysis of the data.

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.,

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

2010). These two conceptual world views can be represented in computer systems 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 Papers I-III 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 2.2). Thus, the states of these objects are related to the outcomes of change and movement (Figure 2.1).

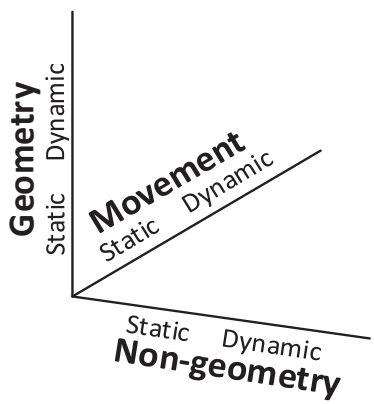


Figure 2.2: Three types of conditions and the combinations that describe the state of an object.

When applying the concept of dynamics in Figure 2.2 to the historical geographic data in this thesis (Papers I-III), individuals that have point locations are modelled as static in their geometry but dynamic in their movement and non-spatial properties. Property units (e.g., used as 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.

2.2.1 Temporal snapshots, object-lifelines and event-chronicles.

This section briefly describes some of the most fundamental models for spatio-temporal data (cf. Paper I for a more detailed description).

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 (Figure 2.3). Models for storing such time-stamped objects are usually called *temporal snapshots* (Worboys, 2005; Armstrong, 1988). These models are simple to create, but less suited for tracing the changes of objects through time. In Figure 2.3, property unit "A" has been digitised from a map in 1830, and property units "B" and "C" have been digitised from a map in 1862. Property unit "B" and "C" cover the same area as "A", which indicate that a geometric change has occurred (e.g., that "A" has been partitioned or subdivided into "B" and "C"). When storing these objects as temporal snapshots, the geometry, the observation date (i.e., the date of the historical map), as well as additional information such as the name and address, can be stored in a database. However, we have no information about when the objects were created or changed, or whether other changes occurred between the two snapshots.

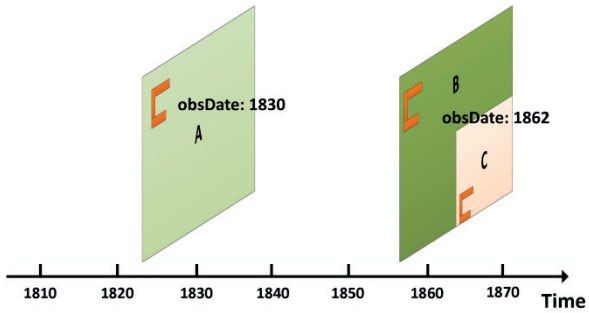


Table 2.3: Temporal snapshots of the property units ”A”, ”B” and ”C”. The ”obsDate” attribute represents the creation date of the historical map from which the property units were digitised.

To enable the identification of changes and to better trace the objects through time, *object-lifelines* models (Figure 2.4) can be used (Worboys and Duckham, 2004). In this time-representation model, each state of the object is assigned a time period. In Figure 2.4, supplementary historical sources (e.g., poll-tax registers) have been used to attain a more precise estimation of the period during which the property units existed in the real world.

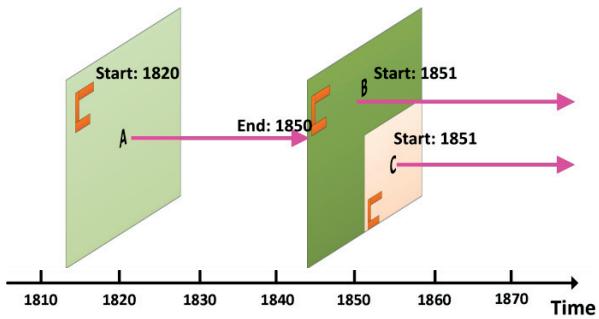


Figure 2.4: Object-lifeline model of the property units ”A”, ”B” and ”C”. The “Start” and “End” attributes define the period for which the objects exist (i.e., their lifelines).

Various implementations of temporal snapshots and object-lifeline models are widely used in historical GIS databases. However, they do not fully describe the events or drivers of the outcomes. For example, in Figure 2.4, we cannot identify what event that caused the property unit to change its geometry. To represent the relationships between the drivers and outcomes, an *event chronicles* model (Table 3) can be implemented (Worboys, 2005). These models focus more on describing the events, such as drivers and outcomes, instead of describing the state of the objects. Figure 2.5 illustrates how drivers that affected the property units can be represented. In this simplified example, the property unit “A” was created in 1820 by an event with id “21”. Then, in 1851 it was partitioned into the two new property units “B” and “C” (event “47”). When storing such information in a

database, the events can be stored separately with links to each object that they affect, as well as events they are related to. Thus, in addition to being able to obtain the lifelines of the objects by using the event information, it is also relatively simple to establish links between each successor and predecessor (e.g., that object "A" is the predecessor to "B" and "C"). 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.

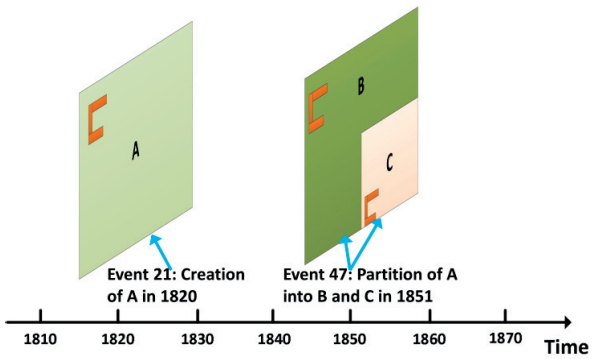


Figure 2.5: Simplified example of event chronicles. Two events ("21" and "47") are linked to the property units "A", "B" and "C".

As mentioned in Section 2.1.1, an important aspect related to the three models described in this section is how to define an object's identity. In the above examples, object "A" ceases to exist when the partitioning occur. However, it may also retain its identity if we, e.g., assume that that "A" and "B" represent the same object/property unit which undergo a geometrical change. Such identifiers binding the observations can also be applied to the temporal snapshots.

Finally, note that the actual database implementations of the above described models may look very different. In addition, 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).

2.2.2 Common models for historical and modern spatio-temporal databases

This section summarises models applied for modern and historical geographic databases. For a more comprehensive review, see Paper II.

Historical GIS databases that contain spatio-temporal data must often handle: 1) changes to the geometry or movement of individual objects; and 2) 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). Briefly, most historical GIS database use various forms of object-lifeline representations and temporal snapshots to handle such changes (e.g., Gregory and Southall, 2005; Berman, 2003; Vanhaute, 2003; Dam, 20013; Fitch and Ruggles, 2003; Villarreal, 2014). Object-lifeline representations are also used in those standardised data models, developed by the INSPIRE directive, that must handle spatio-temporal data (INSPIRE, 2014). Longitudinal demographic databases, on the other hand, have instead a more event-oriented approach (Alter, Mandemakers and Gutmann; Bengtsson et al., 2014). Therefore, it is important to consider such models when the aim is to integrate historical geographic data with longitudinal demographic data.

Within GIScience there has also been much research regarding spatio-temporal data models. Several studies have focused more on modelling events than objects (e.g., Peuquet and Duan, 1995; Yuan 2000), and some studies have developed data models based on the general GIScience theory from Goodchild et al. (2007) (Pultar et al., 2010) (cf. Paper II).

This thesis does not aim to contribute to the research and developments of spatio-temporal data models. Instead, it aims to include the geographic context in historical demographic research. Therefore, we aim to use basic and generic models that facilitate longitudinal demographic analyses with geographic factors. The IDS-Geo model developed in Paper I uses a generic structure that allows for basic temporal representations such as temporal snapshots, object-lifelines and event chronicles. In Paper II and III we use an object-lifeline model to store the geographic longitudinal data and to enable the geocoding of individuals.

2.3 Standardised data models for historical and geographic data

This section describes standardisation of longitudinal historical data and geographic data, which relates to the IDS-Geo data model developed in Paper I.

As section 2.2 shows, there are many ways to represent historical geographic data. When data are stored in such different systems and models, heterogeneities and conflicts often occur. Such heterogeneities can be of different types: syntactical (differences in formats and data types, Figure 2.6a); structural (differences in the data models and structures, Figure 2.6b); and semantic (differences in the meaning of concepts, Figure 2.6c) (Sheth and Larson, 1990). The semantic differences are often the most problematic (Lutz et al., 2009), whereas syntactical and structural

differences can often be solved automatically, although there is sometimes a risk of quality loss during the transformation.

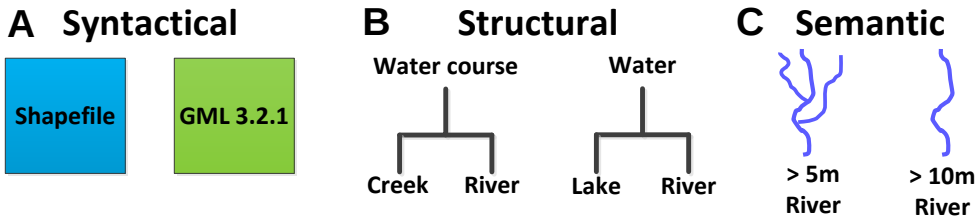


Figure 2.6: Three types of heterogeneities. A) Two datasets are stored in different GIS formats: Shapefile and GML version 3.2.1. B) Two data schemas differ in their representation of a river. C) In the first dataset, a river is defined as a watercourse that is wider than 5 metres; in the second dataset, the width is >10 metres.

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. Then, data from several sources and varying data models can be adapted to these standardised models and thus be compared and analysed together. A further step is to create an infrastructure for facilitating the access to harmonised data and to hence the data sharing among organisations and countries. Such infrastructures for spatial data, Spatial Data Infrastructures (SDIs), are being implemented across the world; e.g., within the European Union (EU) the SDI development is harmonised by the *Infrastructure for Spatial Information in the European Community* (INSPIRE) Directive. The following sections describe briefly such standardisation work for longitudinal historical data and geographic data.

For historical data, several projects and countries are creating longitudinal demographic databases according to the Intermediate Data Structure (IDS) data schema. The IDS schema has been developed within the European Historical Population Samples Network (EHPS-Net) (Alter and Mandemakers 2014) and aims, through standardisation, to facilitate the storage and sharing of longitudinal historical demographic data on the micro-level. Moreover, to enable the interpretation of data from different countries, a metadata registry has been created that contains code lists explaining the variables and values used. The core parts of the IDS schema are individuals and contexts (both geographic and social), and the relationships between them. A main principle adapted in the schema is the Entity Attribute Value (EAV) model (Stead, Hammond and Straube, 1982). Here, each tuple (e.g., row in a table) contain only one attribute value (e.g., “Kågeröd”), which is explained by one type attribute (e.g., “Parish”). Thus, the IDS schema has a very generic structure to facilitate the storage of heterogeneous data. Moreover, to use the data in IDS for longitudinal analyses, extraction programs have been developed that transforms the data into usable datasets (Quaranta, 2015; 2016). In

terms of data heterogeneities, the IDS solve foremost structural and semantic heterogeneities, but syntactical may occur because there is no specification of a low-level and standardised exchange file-format.

For geographic data, SDIs are being implemented across the world. In Europe, the INSPIRE Directive is a core part which aims to establish a SDI within the EU. The main goal of this directive, which has currently been implemented in the member states, is to provide better access to standardised geographic data and metadata needed for environmental applications (EC, 2009). A central part of the INSPIRE Directive is that the member states should provide standardised network services which can be used to discover, view and download harmonised geographic data. These geographic data should comply with one of the 34 data specifications that have been created for certain spatial data themes (e.g., cadastral parcels, buildings, hydrography, land use, etc.) (EC, 2009). These specifications are based on international standards for describing geographic data, developed by the Open Geospatial Consortium (OGC) and the ISO Technical Committee 211 (TC/211).

An important part of an SDI is the use of standardised transfer formats. By using such formats, standardised languages for transforming the datasets; e.g., into a dataset usable for longitudinal analyses, can be used. Within the INSPIRE Directive, each data specification specifies the schema and export format that the data should comply with. As export format, the eXtensible Markup Language (XML) in which Geography Markup Language (GML) is used to describe the geographic data. GML is an XML grammar and an ISO standard for encoding geographic information, developed by OGC (Portele, 2012). These datasets, henceforth called GML-files, should also comply with a GML Application schema (i.e., an XML Schema that uses GML to define the geometric data types) (EC, 2009). Such schema specifies the allowed structure and syntax of the data. In short, a GML Application schema is the XML/GML correspondence to a database schema such as the IDS, but on a lower level.

A benefit with using GML-files is that they enable the use of Linked Data (Schade and Cox, 2010). Shortly, Linked Data refers to connecting data and information one the web. A central part of Linked Data is to use HTTP Uniform Resource Identifiers (URIs) to identify entities and which have links to places on the web that contain information about these entities. This information may then, in turn, be linked to other entities, creating a web of linked information. Such information is obtained using standard web technologies such as HTTP, Resource Description Framework (RDF) and URI (Berners-Lee, 2006). Linked Data also facilitates the interpretation of relationships among entities, and can be used to solve semantic heterogeneous by applications automatically (W3C, 2016). Within the INSPIRE framework, a central online code register has been set up, which contain information and definitions of entities used in the data specifications (EC, 2014). Such codes can then be directly linked to within the GML-files through the use of URIs (Schade and Cox, 2010). This will ensure that only terms and codes that

have been agreed upon are used, and that information defining such terminology can be easily accessed through standardised ways. In addition, establishing such links will in general facilitate data integration and knowledge sharing in general. To transform data in IDS to Linked Data is straightforward on the conceptual level, because the generic data models of IDS and RDF share some core similarities. Such transformation presents new possibilities for reasoning of the data. It remains to be seen if this will be utilised in the future.

To permit the storage and exchange of geographic data in IDS in a standardised way, we introduce a slightly modified data model named IDS-Geo. Here, standardised geometric data types are added based on the OGC/ISO Simple Feature specification. As exchange format we specify GML Application/XML Schemas which the GML-files should comply to. Thus, the abovementioned benefits of using standardised and open formats applies also to IDS-Geo. However, in Paper III we do not use the IDS-Geo data schema and exchange format when distributing the geographic data (we use, however, common principles for the two data schemas). Instead we use the ESRI Shapefile as exchange format, and do not follow the EAV model used in IDS and IDS-Geo. The reason for this is that shapefiles are still one of the most common exchange GIS formats, and therefore we adapt to this situation. In addition, storing geographic data in shapefiles using the EAV principles is not feasible.

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, which 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):

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

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):

$$S(t) = 1 - F(t) = P(T \ge t) \tag{2}$$

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):

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} \tag{3}$$

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 commonly means that failure likely occurred, but it has not been observed; truncation means that no information exists about the occurrence of the failure (Cleves et al., 2010; 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[2] 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 (Cleves et al., 2010). 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

[2] 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).

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 model does. Thus, semi-parametric models are best for hazard functions that do not follow a particular form (Cleves et al., 2010).

The Cox proportional hazard model (or Cox regression) is a common semi-parametric model 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):

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

where x is an individual-specific independent variable 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 applied to longitudinal historical demographic data. In this thesis, geographic context variables have been constructed on the micro-level and included as either fixed or time-varying/dynamic independent variables. However, as Section 2.5 reveals, it is important to consider issues such as spatial autocorrelation.

2.5 Longitudinal analyses with geographic micro-level factors

2.5.1 Geographic context factors

Geographic context is a broad term; in this thesis, this term includes all geographic factors that can affect demographic outcomes such as mortality, migration and fertility. To analyse the effect of geographic factors, we must quantify them using certain methods. Thus, we denote these quantified context factors as *geographic context variables*. In historical and modern societies, many context variables have affected demographic outcomes, such as population density, land conditions, and proximity to communication networks, health centres and wetlands. This thesis focuses on micro-level geographic context variables. These variables can be estimated at the individual level; e.g., by measuring the distance to certain geographic objects[3] from an individual's home or work (e.g., building or property unit). When constructed, these variables can be included in survival analyses or various spatial analyses.

Geographic context variables can be both static (unchanged with time) and dynamic (changing with time) (cf. Section 2.1.2), depending on the geographic objects (e.g., soil types, wetlands and elevation) and on the residential histories of the individuals. If both the geographic objects and the residential histories are static, the context variable for an individual will not change over time. If either the geographic objects or the residential histories are dynamic, they may produce dynamic variables for the individuals. That is, if the geographic objects are static, the context variables can only be dynamic for individuals who change their location. For individuals who remain at their location, the context variables can only be dynamic if the geographic objects change. If both the geographic objects and the residential histories are dynamic, the context variables will be dynamic. Note that some variables may change more frequently than other variables. For example, the population density may vary on a daily, monthly or yearly basis, whereas road networks change less often.

Figure 2.7 shows an area in Halmstad parish. In this figure, five changes to geographic objects have occurred. Between the years 1880 and 1890, two merges of property units and one wetland drainage occurred. Between the years 1890 and 1910, one property unit was subdivided into several smaller units, and one wetland was drained. Because of these changes, context variables such as distance to wetlands, population density, and soil type coverage will be dynamic. The distances from one property unit to, e.g., the closest wetland border, will

[3] The term 'object' also includes entities with less defined borders, e.g., air pollution.

inherently change if this wetland is drained. The two property units *Saxtorp 01*, which are observed in 1880, merged in 1890. This change will also influence the distance to wetland variable for individuals who reside within the merged unit. In addition, this merge may change variables such as population density (especially if it is geographically weighted) and variables that are computed from static data, such as soil type and elevation, because of a geometric change in the property unit. Thus, changes in the objects that were used for geocoding may also produce dynamic variables.

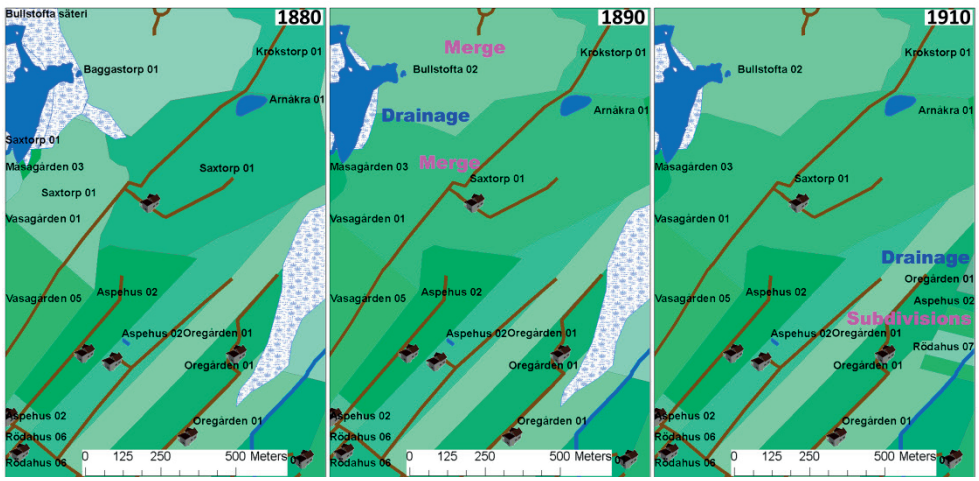


Figure 2.7: Three maps of an area in Halmstad parish for the years 1880, 1890 and 1910. White-blue areas represent wetlands, and blue areas represent lakes/water. The buildings and roads are digitised from the EM map. During the study period, geographic changes such as drainage and cadastral procedures (i.e., changes in the property units) can be observed.

2.5.2 Methods of longitudinal analyses with geographic micro-level context factors

The Cox model described in Section 2.4.3 and other survival models assume that all independent variables that affect a hazard are included in the model. However, other variables that affect the hazard are often not included (e.g., due to a lack of information). Excluding these variables will affect the effect of the included variables. To account for the excluded variables, a shared frailty is commonly included in the Cox model. A shared frailty accounts for the fact that some groups have a higher hazard or are frailer due to excluded variables. The shared frailty mirrors a within-group correlation; that is, observations within a group are correlated because they share the same frailty (Cleves et al., 2010). For example, individuals within a family or household may be correlated, in which some families are frailer than others. The Cox model with shared frailty is defined as

$$h_{ij}(t) = h_0(t)e^{\beta x_{ji} + V_j} \quad (5)$$

where i is nested in the group j, and V_j is the logged shared frailty for the individuals in group j.

With regard to the geographic context, a problem with the shared frailty model is that it assumes that the frailties are independent from each other (Darmofal, 2009). For example, it assumes that individuals within the same family share a frailty but these frailties are independent from neighbouring families. However, many variables in historical demography are spatially autocorrelated; that is, similar values are clustered in space. For example, certain families may be clustered within an area that has known and unknown characteristics that affect their hazards. In this case, the assumption that the frailties of these families are independent is unrealistic. This spatial autocorrelation violates the model's assumption that groups are independent and will affect the estimation of the variables' effects (Darmofal, 2009).

In health geographic information systems (GISs) and epidemiology, multi-level models and spatial regression models are widely used (e.g., Cromley and McLafferty, 2012). Although multi-level models are similar to shared-frailty models, the hazard of individuals is affected by factors that may operate on multiple hierarchical levels. For example, they may include individual-level factors and factors that are common to the neighbourhood in which individuals reside. These models are often used to measure area effects. For example, Dibben, Sigala and Macfarlane (2006) use detailed neighbourhood indices of deprivation in England to measure the area effect of low birth weight for the period 1996-2000. They discovered that mothers who lived in areas with income deprivation generally had a higher risk of giving birth to children with low birth weight regardless of their individual characteristics. Thus, space had a strong effect on their outcomes.

Multi-level models have, however, been criticised for not explicitly considering the spatial autocorrelation (Cromley and McLafferty, 2012). Spatial regression models, on the other hand, consider spatial autocorrelation and incorporates information from nearby areas. Commonly, a spatial weights matrix that describes how the estimation of a variable is influenced by observations from nearby areas is used. A simple spatial regression model is an extended ordinary least square (OLS) regression that is defined as

$$y_i = \alpha + \sum_k \beta_k x_{ik} + \lambda \sum_j w_{ij} e_j + \mu_i \quad (6)$$

where y_i is the observed outcome for i, α is the intercept, x represents the independent variable, β_k is the parameter estimated for the variable k, λ is a spatial autoregressive coefficient, $\sum_j w_{ij} e_j$ is the sum of all the spatial weights which has been multiplied by a spatially dependent error, and μ_i is the random error (Cromley and McLafferty, 2012).

For example, Lee, Ferguson and Mitchell (2009) applied spatial regression to consider the spatial autocorrelation when analysing the effect of air pollution on health outcomes in Scotland.

For survival analyses, Bayesian spatial survival models that utilise spatial weight matrices to consider spatial autocorrelation are commonly used in fields such as biostatistics (but less common in historical demography). The Cox model can be extended to a Bayesian Cox model with spatial frailties. This model is described in Bernardinelli and Montomoli (1992), Banerjee, Wall and Carlin (2003) and Darmofal (2009). The Cox model uses a *conditionally autoregressive* (CAR) prior that model the influence of neighbouring frailties by using spatial weight matrix. The main idea for this CAR prior is that at any location certain values are estimated, their probability is conditional on the levels of neighbouring values (Darmofal, 2009). Thus, the model assumes that individuals who live near each other share frailties (note that the use of Cox models that include both spatial and non-spatial frailties is possible).

An important part of estimating spatial frailty models is defining neighbours and determining the effect of their influence on distance. For example, how near should two objects be located to each other to be considered neighbours? Do neighbours in close proximity to a given object have a greater influence than more distance neighbours?

With regard to the definition of a neighbourhood, the method for defining neighbours differs depending on the data used; e.g., raster versus vector. For polygon vector data, which are used for the property units in this thesis, neighbours of a polygon can be defined as polygons that are adjacent to the given polygon or as polygons that lie within a specified distance of the given polygon. For the latter, the distance is commonly calculated between the centroids of polygons, and a bandwidth is established to define the neighbourhood. In addition to specifying the size of the bandwidth, the selection of the type of distance measure is critical. A simple and commonly used measure is the Euclidean distance. However, a more realistic distance may, for example, be a network distance that is measured along a street network. For the rural parishes in this thesis, topographic information such as forests, water and elevation, and road networks, should be included. That is, two areas may be neighbours in space but separated by impassable terrain. In Figure 2.8, a Euclidean bandwidth is specified for the property unit *Jordkull 01* to identify its neighbours for the year 1870. A wetland is separating Jordkull 01 and some of the property units that fall within the specified bandwidth. In this case, another distance measure that also considers wetlands may produce a more realistic neighbourhood. Thus, this type of neighbourhood would not be in the form of a circle, as shown in the figure, but adapted depending on the terrain. Note that when using longitudinal geographic data, it is important to account for the changing geography, which may affect the shape and size of the neighbourhood.

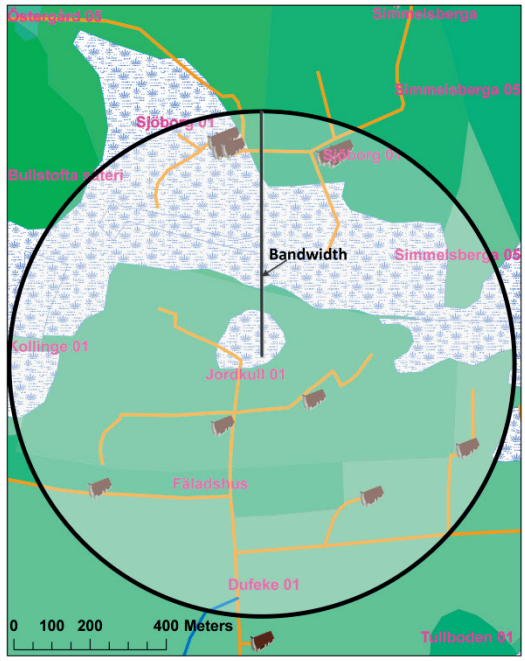


Figure 2.8: A bandwidth using Euclidean distance to define the neighbours of the property unit Jordkull 01 (using the centroid as starting point). Property unit centroids that fall within the bandwidth are considered neighbours of Jordkull 01.

In addition to defining how neighbourhoods should be computed, we must define how neighbours' effects are influenced by distance. The simplest approach is to give each neighbour a weight of 1 and each non-neighbour a weight of 0. That is, all neighbours equally influence the parameter estimations for a given property unit. However, this method is seldom realistic because it makes a sharp distinction between neighbours and non-neighbours. If we assume that nearby units have a greater influence than farther units but remain within the specified bandwidth, we can apply a distance function to model the rate at which the decline of influence occurs. Several functions can be applied; a common function is the Gaussian distance function (cf. Fotheringham, Brunsdon and Charlton (2003)). In Paper IV, this Gaussian distance function is used to model the influence of distance for the geographically weighted population density.

A weakness with the spatial frailty model (and other spatial regression models) is their global method of yielding single statistics for the study area; the model is assumed to perform equally well for the entire area with variables that have a constant effect across space (Cromley and McLafferty, 2012). For example, a statistically significant effect of population density on mortality is assumed to be significant in all locations of the study area. This finding may cause misinterpretations about the relationships among variables in specific areas. To

counter this problem, local statistics that produce individual statistics for each location can be applied (Cromley and McLafferty, 2012). For example, social class may affect mortality to a greater extent in some areas in the five parishes studied in this thesis. These statistics are useful for exploratory analyses because they can be illustrated in maps, and, therefore, used to reveal patterns and to develop more accurate global models. The most applied local regression method within GIScience and related subjects is the geographically weighted regression (GWR). The GWR is an extended OLS regression that includes locally varying variables (Fotheringham et al., 2003). As in spatial regression models, GWR utilises spatial weight matrices to model the influence of neighbours when estimating regression variables. A geographical and temporal weighted regression (GTWR) was recently developed to consider neighbouring effects in both time and space (i.e., using spatio-temporal weight matrices) (Fotheringham, Crespo and Yao, 2015).

In Paper V in this thesis, a non-spatial Cox frailty model (using household units as a frailty component) is applied when analysing the effect of soil type on child mortality. Although this research is novel in the sense of adding micro-level geographic factors to longitudinal historical research, the model could be extended to include spatial frailties in future studies. In general, both the adaption of spatial frailty models and local statistics models that are similar to GWR and GWTR (which may also require modification for longitudinal survival analyses), are sparsely used in historical demography. A likely reason for this sparse use is the lack of detailed and longitudinal data that enable the use of spatial frailties and local spatial analyses.

2.6 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.

2.6.1 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 dynamic. 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.

2.6.2 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 II). 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).

2.6.3 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 indications about the productivity and size of a farm, and cadastral dossiers containing ownership information.

2.6.4 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 because this region belonged to Denmark at that time.

Enclosure maps/land survey maps/ (1750 – 1927) are another important resource for geographic data (Figures 2.9a-b). 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 (Figure 2.9c). 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) (Figure 2.9d). 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 (Figure 2.9e). 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 (Figure 2.9f). 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).

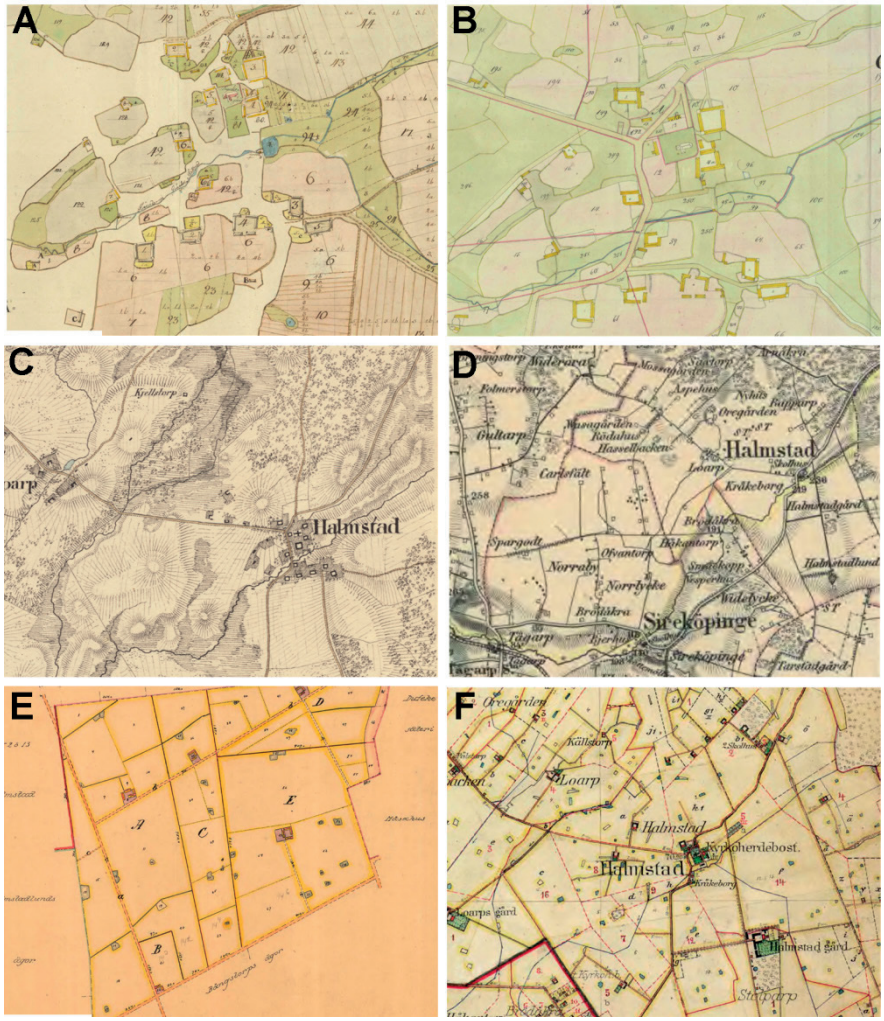


Figure 2.9: Examples of Swedish historical maps covering Halmstad Village. Maps A-C show the village before the land reform, whereas maps D-F show the village after the land reform. A) Land survey map (Inägodelning) 1796. B) Land survey map (Enskifte) 1827. C) Military topographic survey (Skånska rekognosceringskartan) 1815-1820. D) Topographic map (Generalstabskartan) 1860. E) Cadastral dossier in 1913 registering a subdivision. 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 (cf. Paper II).

2.7 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, most databases on the national level contain geographic data in low resolution (e.g., parish boundaries).

Paper II reviews studies using historical demographic data combined with geographic data at various resolutions. An overall conclusion from this review is that there are few historical studies that have constructed or used geocoded longitudinal demographic databases at the micro-level. There are, nevertheless, several studies that use modern geocoded demographic data at the micro-level (although many of them have only studied short time periods or not accounted for the residential histories (cf., Meliker and Sloan, 2011)). This section describes such studies using modern data.

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.

To conclude, using micro-level geographic data 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 dynamic geographic context variables. Consequently, this thesis contributes to this topic by creating a geocoded database that enables the computation of such geographic context variables on the micro-level.

3 Data and study area

A central outcome of Papers II and III is the generation of a longitudinal geodemographic database at the micro-level, which is used in Papers IV and V. Thus, the source data on which this database is based is described.

3.1 Study area

All papers in this thesis 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 who have lived in nine parishes in southern Sweden (Scania) from the 17th century onwards (cf. Bengtsson and Dribe, 1997). Of these nine parishes, 60 historical maps and 150 cadastral dossiers for five rural parishes have been georeferenced and digitised, namely, Hög, Kävlinge, Kågeröd, Sireköpinge and Halmstad. Thus, these five parishes constitute the study area (a total of approximately 130 km^2) (Figure 3.1).

The five parishes vary in their topography and socio-economic characteristics (Dribe and Bengtsson, 1997). With regard to topography, Hög, Kävlinge and Sireköpinge were plain land farming regions (open farmlands), which focused on grain production. Kågeröd was a forest region with large forest areas; Halmstad was a brushwood region with wooded areas in the north and plain lands in the south (Dribe et al., 2011). Sireköpinge, Halmstad and Kågeröd primarily incorporated a manorial system in which tenants leased their farms for a certain time period. In Hög and Kävlinge, freeholders and crown tenants, who owned their land and paid taxes on their land, were more common. Hög and Kävlinge underwent a major enclosure in 1804, in which most of the farmers moved out of their villages and into their property units. Halmstad was enclosed in 1827 and 1844, Kågeröd was enclosed in the period 1839-1842, and Sireköpinge was enclosed in 1849. With the exception of Kävlinge, which developed into an industrial town at the end of the 19th century, all parishes remained rural and maintained a similar economic structure and development, as well as population growth for the entire study period. Although several of the historical maps

originated in the 18th century, the study period is 1813-1914 due to the availability of demographic data; catechetical examination registers that document migration and household compositions have only been available for these parishes since 1813.

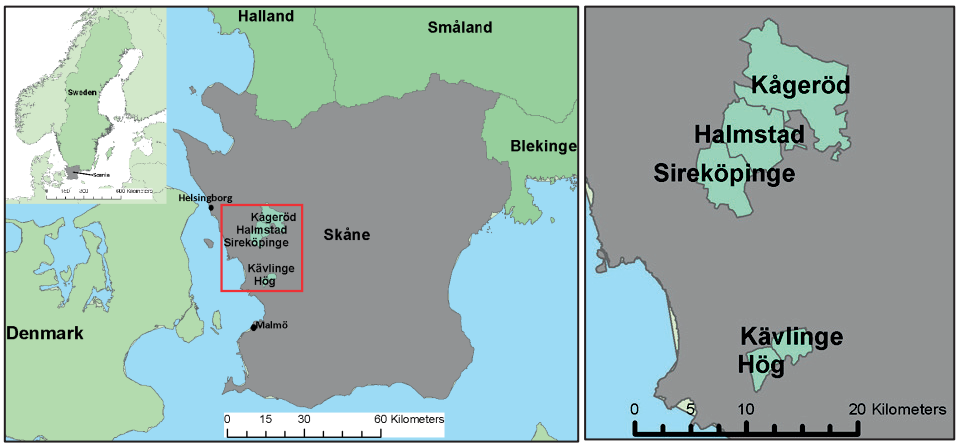


Figure 3.1: The five parishes of Hög, Kävlinge, Sireköpinge, Halmstad and Kågeröd constituted the study area for the papers in this thesis (Source: Paper II).

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, 2015). 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 II.

3.3 Geographic source 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 georeferenced and digitised within an ongoing project (Table 4). So far, approximately 900 property units, 3,000 buildings as well as a substantial amount of roads, railways, streams and wetlands have been digitised.

In Paper III, approximately 150 cadastral dossiers has been digitised and used for the creation of object-lifeline data.

Table 3.1: 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
Cadastral dossiers	1757-1915	150	1:1,000 – 1:8,000

3.4 Geocoding of the SEDD database on different geographical levels

During the geocoding of the SEDD database (Papers II, III), individuals were linked on two geographical levels: the property unit level and the address unit level. This section briefly describes the differences between these levels and aims to facilitate an understanding of the examples presented in Section 3.5.

The geocoding and the creation of object-lifelines required extensive work for the following reasons. Before the land reforms (conducted between 1757 and 1849 in the parishes), most people lived in small villages and cultivated nearby scattered plots. After the land reforms, the self-owned farmers received a cohesive piece of land, which they also moved out to. These lands we denote *property units*. The property units were usually devoted for agriculture, although some of them also contained forest lands. In line with the rapid population growth during the study period, several of the property units were subdivided or partitioned into smaller units. However, the property units did not receive new addresses. Therefore, multiple property units often share addresses (e.g., Figure 3.2). We denote the set of such units an *address unit*. Usually they are close to each other, but not necessarily adjacent (in Figure 3.2, the property units are close to each other, whereas they are adjacent and distant from each other in Figure 3.5).

To perform the geocoding on the address unit level was straightforward because the poll-tax register contains annual information about the address unit for the family head. However, to geocode on the property unit level was more difficult. To geocode individuals to the correct property units (when they shared addresses), we used taxation values in the poll-tax registers combined with textual sources in the maps and cadastral dossiers to separate the units sharing addresses. We also traced the owner histories of the property units when property units shared both taxation values and addresses (cf. Paper II). For several records, extensive manual work was required to achieve a high number of reliable links.

3.5 Study area-specific problems in the geocoding of individuals

Because of the different systems and characteristics within the parishes, different approaches to geocoding need to be applied. The main differences are observed between the manorial systems and freehold systems within the parishes. These differences adhere not only to the geocoding method but also to the availability of historical maps and cadastral dossiers and the extent to which a property unit can

be considered as an appropriate object for determining an individual's residential area.

Hög and Kävlinge parishes (with the exception of the urban area that emerged at the end of the 19th century) were populated by self-owning farmers and freehold and crown tenants, who managed their own property units and participated in farm work (Lundh and Olsson, 2005). For the property units of these farmers, we can anticipate that they spent the majority of their time within the unit. Exceptions are farmers who worked in factories within and around the Kävlinge urban area. Sireköpinge, Halmstad and Kågeröd had a higher share of manorial tenants, large manors (Sireköpinge, Duveke, Bullstofta, and Knutstorp) and large commercial farms that are termed satellite units (Swedish: *Plattgård*). The satellite units were established by the landlords of the manors to increase productivity. Entire villages that had previously used a manorial open-field system, in which the tenants in the village cultivated scattered plots, were merged into one satellite unit with large cohesive farm lands (Olsson, 2008). The previous tenant farmers were evicted, which indicated that many farmers lost their farmland. Some farmers had to become farm workers (Swedish: *Statare*) or servants at the satellite unit. Other farmers remained in their houses (e.g., as crofters), where they were assigned a smaller cohesive plot to cultivate. The size of this plot was often not sufficient for supporting an entire family or household; therefore, farmers also needed to work on the satellite unit (Olsson, 2008). Thus, many of the satellite units and some of the manors, have several smallholdings registered in the EM maps to which individuals can be geocoded.

The availability of historical maps and cadastral dossiers is also related to the system used within the parishes. A major difference is that large parts of the manorial parishes had a single landowner (but many tenants and workers). Therefore, the need for detailed maps for these landlords with large lands were small compared with the many landowners in, e.g., Hög and Kävlinge parishes. That is, landlords with large lands could define the property unit borders for their tenants by themselves, whereas formally defining and documenting the property unit border for a self-owning farmer was more important. Thus, more detailed maps, such as cadastral dossiers, are available for areas with many freeholders and crown tenants compared with areas with a manorial system.

Consequently, based on the different systems used within the parishes, different geocoding procedures have been conducted. Four relatively problematic examples illustrate these procedures.

3.5.1 Freehold and crown tenants in Hög

Figure 3.2 shows an example in which an initial property unit was gradually subdivided or partitioned into several smaller units. These cases are most common in Hög and Kävlinge parish; however, they occur in the other parishes as well. The figure shows the property units with the address *Hög 14* in 1900 (highlighted in cyan); the aggregated area represents the area of the initial property unit created during the land reform in 1804. Performing a geocoding on an address level is relatively simple, because all units are adjacent to each other. Thus, we can assume that the individuals at the address level resided within the outer borders of the property units. In addition, we can usually assume that the individuals resided within their property units most of the time (however, whether the small holdings in the top left of the figure provided adequate land to support the families who live on the property is questionable).

To link each individual to a property unit is more problematic because of the frequent cadastral procedures that occurred within the address unit during the study period. Cadastral dossiers that register the geometric changes and the owners of each new unit are often available. However, the task of manually investigating the cadastral dossiers, documenting the geometric changes, and linking the registered owners to the poll-tax registers is time-consuming.

Some cadastral procedures (or joint ownerships of property units) were documented in the cadastral dossiers. By tracing the owner histories in the poll-tax registers, the taxation values can be combined to determine the association between the records and the jointly owned property units. Although semi-automatic scripts can be used to combine the taxation values, these procedures require substantial manual work.

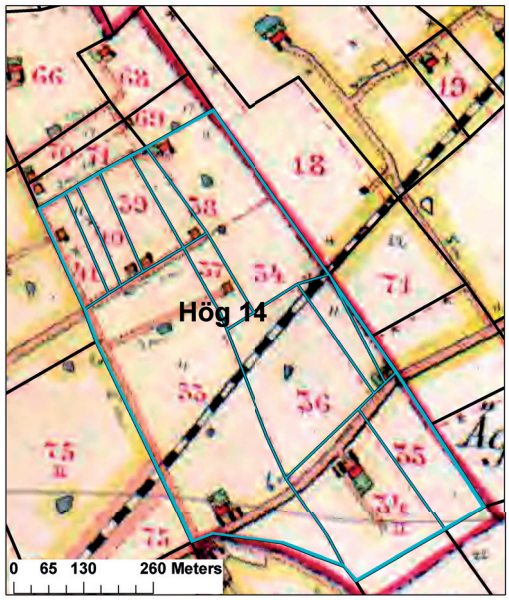


Figure 3.2: Initial property unit Hög 14, which was gradually subdivided or partitioned into several smaller units (highlighted in cyan). The property unit borders are shown for the year 1900, and the Economic Map (EM) is used as the background.

3.5.2 Scarce observations in time and forestlands in Kågeröd

For the manorial parishes Sireköpinge, Halmstad and Kågeröd, fewer observations were obtained during the study period in the form of land survey maps (LSMs) and cadastral dossiers. For example, the geometric shape of the property unit *Finstorp* is observed on an early map from 1800 before the enclosures (in the digital version of this map, information about the association between plots and property units is lacking[4]), military topography survey maps (MTSMs) in 1815-1820, and an EM in 1910-1915 (Figure 3.3). Finstorp is not situated within any of the nearby villages; thus, the extent to which it was affected by any of the land reforms that were conducted within the parish is uncertain. Because the MTSM maps did not document property unit borders, the only geometric observation of Finstorp is obtained for the period 1910-1915. From 1913, a land reform map for Knutstorp (*Knutstorp siftelag*) with the same registered borders as the borders on the EM is also available. How should the lifeline of the property unit be estimated? The taxation value in the poll-tax registers remains unchanged throughout the period; thus, no geometric change is indicated in this register. The only other information that is available is that several villages in Kågeröd were

[4] Information may be available in the paper version of the document; however, collecting and studying this information is beyond the scope of this thesis.

enclosed between 1838 and 1839. However, these enclosures affected areas that were jointly used. Individual units were seldom affected by these enclosures, with the exception, for instance, of shared ownership for forestlands. As changes in the area may have occurred, we estimate the start date of Finstorp to be 1840 and the minimum start date to be 1800 (from the observation in the pre-enclosure map). However, the observed property unit borders in the EM map may have been established in 1913 during the late implemented land reform in Kågeröd (*Knutstorp skiftelag*) (which may have also been conducted by the landlord to specifically define the current borders).

Finstorp and several other property units in Kågeröd parish contain forest lands. As shown in the figure, only the northwest part of the property unit is covered with farm fields. Currently, the farm fields and forest lands have not been separately digitised. In addition, we do not know whether these fields and forests were static in time or whether substantial changes occurred (e.g., if forests had been cleared). A question for proximity analysis is how this topography can affect residents' movement patterns. For example, can we assume that they spent an equal amount of time in the forestlands and farm fields?

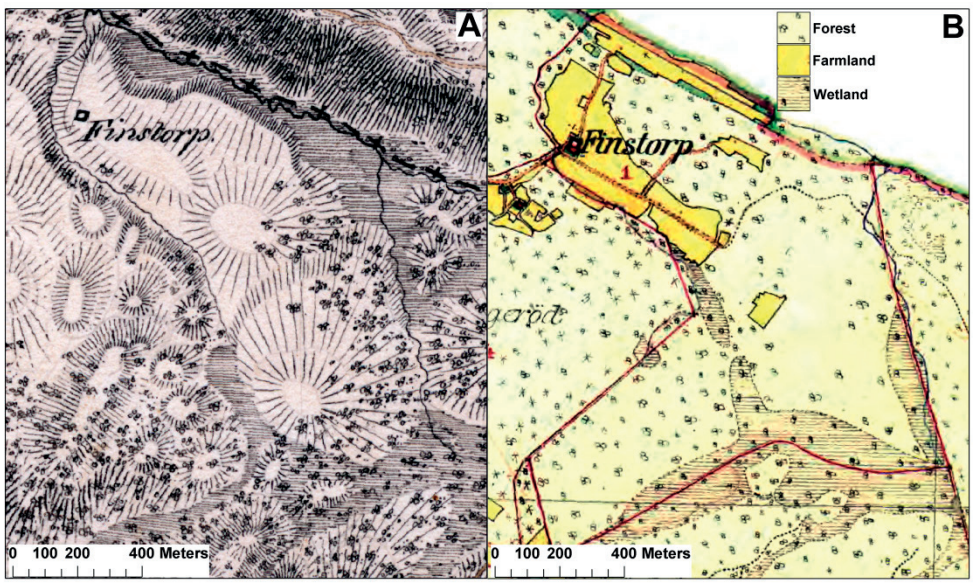


Figure 3.3: Property unit Finstorp. A) Observed in the 1815-1820 MTSM map. B) Observed in the 1910-1915 EM map.

3.5.3 Smallholdings within Böketofta satellite unit (Kågeröd)

Figure 3.4 shows the property unit *Gryttsgård*, which is a smallholding (approximately 12 hectares) under the satellite unit of *Böketofta* (Figure 3.4b). With regard to estimating the lifeline of the property unit, two geometric observations are noted: an observation from an LSM map from 1800 before the implementation of the Böketofta satellite unit in 1838-1839 (Figure 3.4a), and an observation from the EM map in 1910-1915 (Figure 3.4c). A comparison of the LSM and EM maps revealed that the building of Gryttsgård remained, but the scattered plots in the open-field system changed to cohesive plots of lands. No available map registers the land reform in which the plots in the previous Böketofta village were merged to the plots in the Böketofta satellite unit. Thus, we assume that the geometric shape of Gryttsgård is the same for the period 1839-1914.

In the poll-tax registers, a possible indication of a geometric change can be observed in 1871. Until this year, the property unit has a taxation value of 1/4 *mantal*. The taxation value is split into two *mantal* with a value of 1/8 for each *mantal*. One part is registered for the Böketofta satellite unit, whereas the other part is registered for the tenant. No individuals are present in the household of the owner from the Böketofta satellite unit, perhaps because this part of the property unit was directly managed and taken over by the satellite unit. In the tenant's property unit, several families and households are registered. In this case, the areas named b and b2 were digitised as Grytsgård, and all individuals linked to this property unit were geocoded to this area. However, a question is how much of the area for Gryttsgård observed in the EM map actually belongs to the tenant and the extent that Böketofta directly manages the area.

Another issue with this property unit and other units that belong to satellite units is the accuracy of the assumption that the individuals who are linked to this property unit also spent most of their time within the unit. Evicted tenants often had crofts assigned with land, with which they could only partly support their family and household. Therefore, tenants also had to work on the satellite unit. Thus, in proximity analyses, there is the question of whether they should be partly linked to the land of the satellite unit.

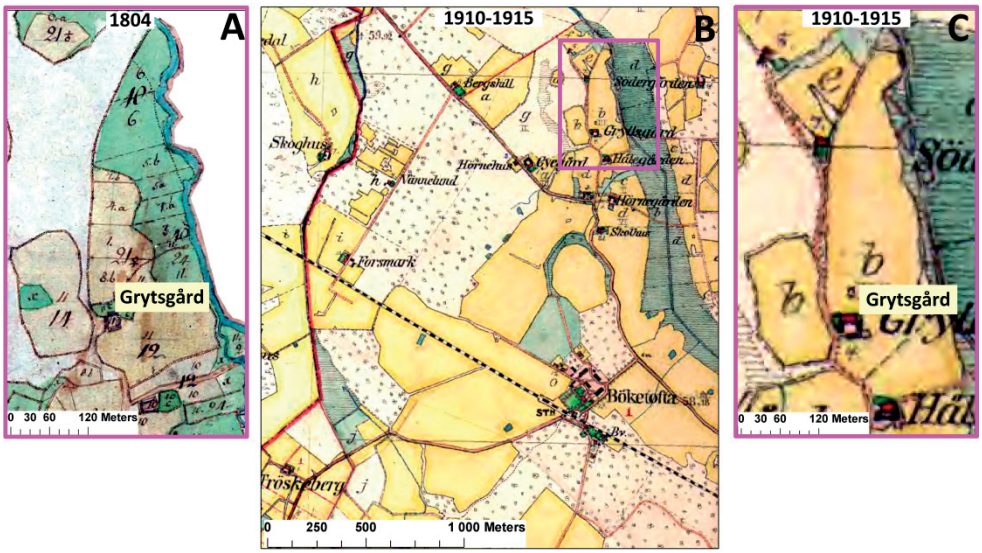


Figure 3.4: Property unit Grytsgård and the Böketofta satellite unit. A) Grytsgårds house observed in a LSM map.B) Böketofta satellite unit observed in the EM map. C) Böketofta property unit (fields b and b2 belong to this unit) observed in the EM map.

3.5.4 Leasing arrangements and different environments in Kävlinge

In Kävlinge, a geocoding complexity is the presence of several non-adjacent property units that share an address. A specific problem is related to the area of *Gryet* in northern Kävlinge (top right corner in Figure 3.5). Until the second part of the 19th century, Gryet had been an outfield area that was partly covered with wetlands and jointly owned by several property units within the parish. Many of the wetlands were drained and began to be cultivated. Some of the property units in Gryet had small plots of farmlands, whereas other property units were probably residences for industrial workers. Because the property units in Gryet were leased by some of the larger property units in Kävlinge, they also share addresses with these larger property units. Figure 3.5 shows an example of this problem, in which the highlighted property units have the address *Kävlinge 02*. The smaller units in the top right corner are located in Gryet, and the two larger units are located in the bottom left corner (these two units were created during a subdivision of the initial property unit Kävlinge 02). According to the poll-tax registers, the property units in Gryet belong to the right-most of the two larger property units.

Because the living conditions and environment likely differed markedly between the property units in Gryet and the larger property units, the individual should be linked to the correct property unit. However, this linking is problematic for several reasons. For instance, the property units in Gryet seldom have any taxation value registered in the poll-tax registers; therefore, it is not possible to use the taxation

value as an identifier. It is, nevertheless, possible to first link the tenant/owner documented in the EM map with the corresponding person registered in the poll-tax registers. When available, use of the information in the poll-tax register about the property unit area to trace the history of the owners is possible. However, linking other families in the poll-tax registers that resided in the property unit but in different households than the owner/tenant is challenging. These families can be linked to the property unit where they resided using information about the holding number of the property unit (which corresponds to the owner of the unit). However, because the owner is the family who resides in a larger property unit, this information cannot be used to identify whether a family lived in a property unit in Gryet or in the larger property unit. Sometimes comments in the poll-tax registers can be used to locate where those families likely resided; however, this manual task is time-consuming.

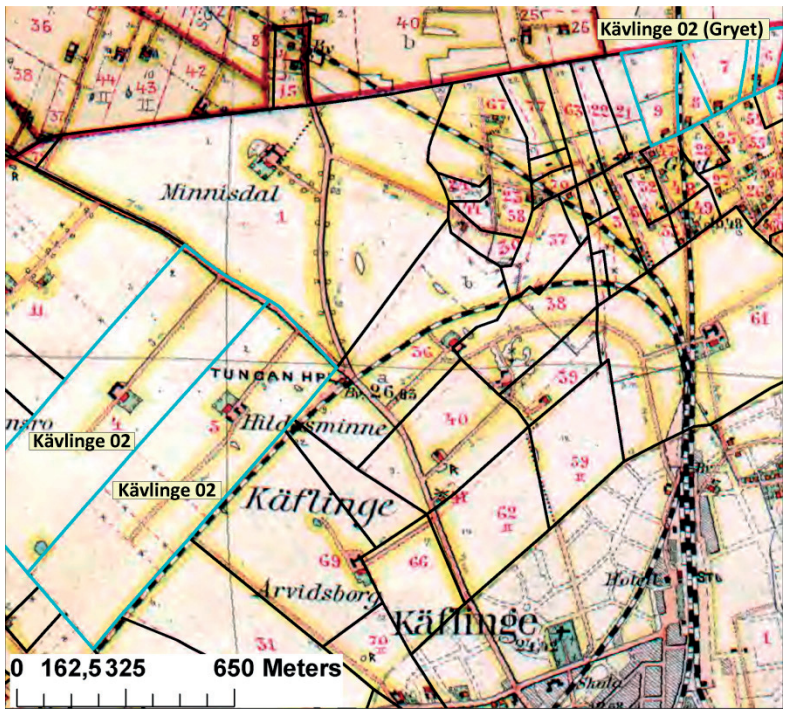


Figure 3.5: Property units with address Kävlinge 02 (highlighted in cyan). Bottom-left: two relatively large rural property units; top-right: small property units in Gryet.

3.5.5 Concluding remarks

The five parishes encompass a heterogeneous area that experienced extensive economic, social and geographic changes for the majority of the study period. As illustrated in the examples, these characteristics complicate the geocoding of the longitudinal demographic data. However, they also present possibilities for studying demographic and geographic changes and detecting patterns and trends in time and space. We are now able to ask questions such as how certain innovations or structural changes affected people's lives and how this effect differs between groups and systems. A population in a homogenous and static environment would be simpler to geocode, but the data would be narrower and shallower and provide fewer opportunities to conduct extensive and substantive research.

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 1.1.

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, will aid data sharing and facilitate 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: 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 micro-level geographic factors to longitudinal historical demographic analysis (first objective in Section 1.2). We achieve this goal by combining several techniques into a universal scheme. The methodology consists of three process steps: (1) creating geographic snapshot data by scanning, geocoding, and digitizing historical maps; (2) transforming the snapshot data into an object-lifeline representation using supplementary longitudinal data; and (3) linking individuals in the demographic data to the geographic objects. In the first step, multiple historical maps covering the same area over a long period are scanned and geocoded into modern geodetic reference systems. Thereafter,

geographic objects are digitised from the historical maps. These objects have two purposes: (1) to geocode individuals to a physical location; and (2) to enable the computation of geographic context variables in longitudinal demographic analyses. The first step results 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 these individuals to geographic objects (i.e., geocode demographic databases) over long time periods, we need to model when the geographic object was created, changed and ceased to exist. Thus, the second step is to transform the digitised objects in the temporal snapshots into an object-lifeline data model. The reason for this is to maintain the analytical advantage of using longitudinal demographic data. This transformation can be performed by combining the historical maps with supplementary longitudinal register data (e.g., poll-tax registers) and snapshot data (e.g., cadastral dossiers). The third step is to link the individuals in the demographic data to, depending on how they moved, one or several geographical objects. Which geographic object types to use inherently depends on the available historical sources. From tax registers, church books, and cadastral dossiers, we often have data for the names of the property units where the individuals lived. By identifying these names in the historical geographic data, we can link the individuals to a property unit. In some cases, we can even obtain higher resolution results for the linkage by utilising information about the buildings in which individuals lived; such information is sometimes available in land enclosure maps and cadastral dossiers.

The methodology is evaluated in a case study using longitudinal individual-level data from the Scanian Economic Demographic Database (SEDD) (Bengtsson et al., 2014). 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 approximately 53,000 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 evaluated using a spatio-temporal relationship query. The evaluation demonstrates that the methodology is feasible and that the resulting geocoded database can be used in demographic studies that include micro-level geographic factors.

The scientific contribution lies in both the methods for geocoding longitudinal demographic databases at the micro-level and in the database resulting from the case study. The methodology integrates a set of different techniques into a single scheme that encompasses all parts of the process. This allow for the inclusion of the micro-level geographic context to longitudinal historical demographic

analyses. To our knowledge, the result is a unique contribution in terms of geocoding individuals over such long time periods.

4.3 Paper III: A Longitudinal Integrated Demographic and Geographic Database on the Micro-Level

This paper is a data descriptor of the geocoded data developed in Paper II, but which also contains: (1) an improved version of the geocoding; (2) an inclusion of wetlands stored in object-lifeline time representations; and (3) a data quality evaluation. The datasets describe the geography over the five rural parishes and the geocoded population (at the property unit level) for this area for the time period 1813-1914. As stated in Paper II, the population is a subset of the Scanian Economic Demographic Database (SEDD). This paper adds approximately 150 digitised cadastral dossiers which have been used to improve the lifeline estimation of the property units as well as the geocoding of the individuals.

In addition to the property units, the historical wetlands are stored in temporal representations of longitudinal object-lifelines (cf. Worboys 2005). The reason for storing wetlands in object-lifelines is to enable accurate proximity analyses for the whole study period. Proximity to wetlands and still water may have indicated exposure to various diseases such as water-borne diseases and malaria (until the 20th century, malaria was a problem in Sweden and Europe, and wetlands likely provided habitats for mosquitoes that transmit malaria). These wetlands were gradually drained during the 19th and 20th centuries; thus, their lifelines were estimated to enable estimations of dynamic geographic variables. To create an object-lifeline representation of the wetlands in the study area, we combined information about 1) digitised historical wetlands from the MTSMs from 1812-1820; 2) modern wetland and water data; 3) soil type data; 4) the EM from 1910-1915; and 5) data regarding registered joint drainage units before 1920 within the five parishes.

As a quality evaluation for the datasets, the positional accuracy and the geocoding match rate are estimated for the datasets, and a spatio-temporal topology is implemented.

We determined the positional accuracy of the digitised property units using modern property unit borders as the reference data. We were only able to collect a sample of the historical property units that had remained unchanged until modern time. The sample (n=86) represented 7.5% of the historical property units. Two measures of positional accuracy were used. The first measure is a quality value called QBOM (Harig et al., 2016), which is based on the Buffer Overlay Method (Tveite and Langaas, 1999). The QBOM evaluates how similar each historical

property unit and its corresponding modern property unit are to each other. The mean and median QBOM are estimated to 0.90 and 0.92 respectively (0.92 represents a 92% similarity). The second measure is estimated from the Euclidean distance between the historical property unit centroid and the centroid of the corresponding modern property unit. Here, the Root Mean Square Error (RMSE) and median error are estimated to 16.7 and 10.3 meters, respectively. For the wetland data, the RMSE of the georeferenced 1812-1820 MTSMs, from which most of the wetlands had been digitised, was estimated to be 30 metres.

Match rates for the geocoding on the property unit and address unit level are measured in percent person-years for the individuals in SEDD. For all parishes, 37% and 45% of the person-years (with a total of 113,548 years) in SEDD are linked to property units and address units, respectively, for the period 1813-1848. For the period after all land reforms had been implemented; i.e., 1849-1914, 67% and 92% of the person years (in total 339,595 years) are linked to property units and address units, respectively. The match rates for the period 1813-1848 are low because some of the land reforms had not been implemented in this period. Thus, people still lived within the villages and cultivated nearby scattered plots

Finally, the spatio-temporal topology checks that the property unit polygons do not overlap in space and time, and that no empty spaces appear in the study area.

We know that the largest share of the individuals that have not been geocoded (for a portion of or their entire life) to a property unit were poor and landless individuals; therefore, they were difficult to geocode. However, we have not performed statistical tests to evaluate possible bias in the linking related to certain variables (e.g., social class, gender, age).

Nonetheless, as for Paper II, the scientific contribution of these data is the opportunities they present to address novel research questions. We have annual information about the shape of each property unit, which enable us to accurately trace the residential histories of individuals in time and space. Combined with the historical geographic data and other modern geographic data, such as elevation and soil conditions, the datasets can be used for spatial analyses and spatio-temporal analysis, as well as the inclusion of geographic factors in longitudinal analyses of historical demographic data. They enable, for the first time, inclusion of geographic micro-level variables into historical longitudinal analyses of a rural area. As a proof of concept, the database enables the analyses carried out in Paper IV and V.

4.4 Paper IV: Importance of the geocoding level for historical demographic analyses: A case study of rural parishes in Sweden, 1850-1914

Paper IV analyses the methodological issues related to including a micro-level geographic context in longitudinal demographic analyses.

Historical studies using macro-level or micro-level geographic context factors have contributed to the understanding of how demographic outcomes are affected by the environment. The recent and ongoing geocoding of databases on the micro-level can be used to account for the variation of context variables on a more local scale. However, performing geocoding on a highly detailed level for long time periods and accounting for changes in geography involve high costs. Finding an optimal geocoding level that balances cost and applicability requires an understanding of the scale at which the geographic factors operate. Moreover, the most logical quantification methods for certain types of analyses must be determined. Otherwise, the high spatial resolution of the geocoding cannot be properly utilized and the demographic models might produce unreliable results. However, the appropriateness of the geocoding level, both in the spatial and temporal aspect (accounting for geographical changes), and the appropriateness of the quantification of geographic context variables have received little attention in historical demographic research.

Thus, this study contributes to the literature by offering insights into the importance of using appropriate quantification methods and choosing the geographic level so that the most suitable geocoding is used for longitudinal demographic analyses on the micro-level. The overall aim is to study is how the geographic level of the geocoding and the choice of quantification methods for the geographic context affect the results of historical demographic analyses. In a novel case study we use geocoded database from Paper II and III to analyse geographic factors that may affect mortality in the five parishes for the period 1850-1914. We selected two geographic context variables that are related to exposure to infectious diseases: population density and proximity to wetlands. The latter is considered a possible indicator of exposure to malaria, which was a problem in this area during the 19th century (Lindgren and Jaenson, 2006).

The method is performed in three steps: (1) Quantify the geographic factors; i.e., create geographic context variables; (2) compare the results of the geographic variables computed on the different geographic levels; and (3) analyse how the geographic levels, as well as the definitions of the geographic variables, affect the results obtained from survival analyses when measuring the impact of these variables on mortality. We analyse three geographical levels of the geocoding: *property units*, *object-lifeline address units*, and *snapshot addresses units* (from

the economic map in 1914). For the distance to wetland measures, we test both snapshot addresses with object-lifeline wetlands, and snapshot addresses with snapshot wetlands from 1914. We use two common quantification methods for each geographic variable. For the proximity to wetlands variable, the *centroid method* and *random-points methods* are used. The centroid method calculates the shortest Euclidean distance between the centroid of the geographic unit and the nearest wetland. The random-points method calculates the median distance of 100 points randomly distributed over the geographic unit and the shortest Euclidean distance between every point and the nearest wetland. For the population density, we use an *unweighted population density* and a *geographically weighted population density* (GWPD). The unweighted population density represents the population divided by the area of the geographic unit. In the GWPD population densities of neighbouring units are also considered in the calculation (and their influence units decay with distance).

Even relatively small differences between the property units and the coarser geographic levels influenced both the magnitude of the effect and the statistical power in the survival analyses. Also the choice between the common methods used to quantify the geographic context variables substantially influenced the results of the survival analyses, which was occasionally greater than the influence of the geographic level. In addition, the results show the importance of accounting for geographic changes over time. Finally, the case study indicated that proximity to wetlands affected the mortality of women, which might indicate exposure to malaria mosquitoes. In conclusion, for rural historical areas, geocoding to the property unit level is likely necessary for fine-scale analyses at distances within a few hundred metres. We must also carefully consider the quantification methods that are the most logical for the geographic context and the type of analyses.

This study contributes to the literature by offering insights into the importance of using appropriate quantification methods and choosing the geographic level so that the most suitable geocoding is used for longitudinal demographic analyses on the micro-level. Moreover, a novelty of this study is the use of longitudinal and micro-level demographic and geographic data. Previous studies on the effect of geocoding level have mostly used data that cover a short period that does not encompass substantial geographic changes. Hence, in this study, we are also able to analyse how different temporal models affect the results of the demographic analyses. Finally, although the main purpose of this study was not to conduct a complete analysis of mortality, the case study is, to the best of our knowledge, one of the first longitudinal studies to: 1) analyse the effect of proximity to wetlands in a historical context; and 2) to use a geographically weighted population density method for spatio-temporal micro-level data. Thus, the case study offers some insights into how the geographic context affected mortality in the study area.

4.4 Paper V: Unequal lands: Soil type, nutrition and child mortality in southern Sweden, 1850-1914

Paper V uses the geocoded database to conduct a longitudinal demographic analysis with geographic micro-level factors. The paper focuses on the role of nutrition in historical societies by analysing the effect of soil type on child mortality in the five parishes between 1850 and 1914.

Child mortality differed greatly among regions in preindustrial Europe; not only between rural and urban areas, but also between close by rural areas (Bengtsson and Dribe, 2010, 2011; Claësson, 2009; Gregory, 2008; Van Poppel, Jonker, and Mandemakers, 2005). Not much is known about the role of nutrition in such geographic differences, and about the factors affecting the nutritional level and hence the resistance to diseases. Because large differences in mortality between the rural areas have been found also after controlling for various socio-economic factors, there are indications of other factors in play; e.g., such related to exposure to diseases that affected the mortality. However, for historical datasets in general, and for our study area specifically, there is a lack of longitudinal and individual-level economic information on income and farm-level productivity. Measures based on taxation information and type of land can be used, but such information is sometimes an inaccurate measure of the farm-level productivity (cf. e.g., Svensson (2001)). Therefore, as a new measure of nutrition and indicator of agricultural productivity, we use information on the underlying soil type for each property unit.

We analyse the effects of soil type on mortality of children aged 1-15 for the period 1850-1914. Our main hypothesis is that soil type affected the farm-level agricultural productivity. This, in turn, affected the nutritional level of the children and thus their likelihood of dying. We expect that children living on farms covered by relatively large areas of fertile soils, such as clayey till or clay, experienced lower mortality than children living on farms with less fertile soils. Moreover, it should primarily affect the mortality of farmers, i.e., children to individuals owning or leasing land. Labourers, which were mostly paid in money or in kind, and large-scale farmers, on the other hand, should be less affected by the soil type. We expect stronger effects of soil type on mortality from non-virulent and nutrition dependent diseases compared to highly virulent airborne-infectious diseases. As an alternative hypothesis, soil type may instead be a measure of exposure to virulent diseases. If so, we expect soil to affect the mortality of all social groups equally. In addition, there should be stronger effects of soil on mortality from highly virulent airborne-infectious diseases.

The effect of soil type on the mortality risks are analysed considering as outcome all-cause mortality and mortality from non-airborne and airborne infectious

diseases. The main independent variable is the soil type of the property units in which the children resided. This is a categorical variable that, for each property unit, represent large spatial coverage of the soils: 50-75% or 75-100% of clayey till, 50-100% of either clay-till/clay soils or sandy soils, and mixed soils (foremost constituted of clayey till, clay-till/clay and sandy soils). We control for parish of residence, taxation value, social class, sex, and birth year. We estimate separate models for children to labourers (n=12,900) and small and medium-scaled farmers (n=3,235). For the all-cause mortality analysis, Cox proportional hazard models are used, whereas a competing-risks regression model is used for the mortality analysis from non-airborne and airborne infectious diseases. All Cox proportional hazard models also included a shared frailty component to measure the proportion of unobserved characteristics that are shared between members of the same household

Soil type primarily affected the mortality of children to farmers. Particularly, these children experienced relatively lower mortality when living in property units covered by very high proportions of clayey till (75-100% coverage). Labourers' children were also affected by soil type, but in an opposite way from the farmers' children; they had a lower mortality when residing in property units covered by mixed soils. However, the mixed soil type is the smallest category and may be heterogeneous. It may also correlate with other unobserved factors which have not been considered in the models. Therefore, the results indicate that soil type is a measure for nutrition and resources, through agricultural productivity, for children to farmers. Moreover, we found no support for our second hypothesis, which predicted that soil was instead a measure of exposure to virulent diseases, because soil types did not affect the mortality of the two groups equally. Consequently, for the farmers' children, the results indicate a relatively important role of nutrition as a mortality predictor, which is in line with previous research on the link between nutrition and mortality in pre-industrial societies (e.g., McKeown, 1976; Fogel 1994; 2004; Puleston and Tuljapurkar, 2008; Floud et al., 2011). Lastly, on the previously found geographic differences on child mortality between the five parishes in the study area, these were not found to be related to soil type through agricultural productivity and hence nutrition.

A scientific contribution of Paper V is the added new findings to the literature about the importance of nutrition and agricultural productivity regarding child mortality in a preindustrial society. The results also deepen our understanding of the reasons for the geographic mortality differences. In addition, the study illustrates that the geographic variables may be used as proxies for other demographic characteristics. Finally, this study is, to our knowledge, the first longitudinal study on the micro-level that analyses the effects of soil type on mortality in a historical rural society. Thus, it shows how new knowledge can be added to demographic research by including micro-level geographic factors in longitudinal historical analyses.

5 Conclusions and future studies

5.1 Conclusions

Studying how micro-level geographic factors have influenced human living conditions over long time periods provides many new and important insights into various research fields. Within demographic and historical studies, an expanding field of research has focused on geographical aspects. However, these studies have been limited to the macro level because of a lack of individual-level and longitudinal geocoded historical databases to enable studies at the micro-level. Thus, an essential dimension has been missing in longitudinal analyses of historical demographic data.

The aim of this thesis is to improve historical demographic research by adding the micro-level geographic context to longitudinal historical analysis. This aim consists of three research objectives: (1) to enable the inclusion of micro-level geographic factors to longitudinal historical analyses; (2) to study how the geographic geocoding level affects the results of demographic analyses; and (3) to use geographic information to improve and extend the knowledge of demographic change. The first objective was achieved by developing methods for geocoding longitudinal historical demographic databases at the micro-level (Paper II), by the resulting geocoded database (Paper III), and by contributing to the standardisation of longitudinal historical databases that include geographic data (Paper I). The second objective was primarily accomplished using the resulting geocoded database to study the effect of different geographic levels and methods for quantifying geographic factors on survival analyses (Paper IV). The third objective was achieved by performing longitudinal demographic analyses that include geographic micro-level factors (Paper V). Thus, the first two objectives focused on methodology, whereas the third objective focused on adding new knowledge to demographic research.

The following conclusions can be drawn from this thesis:

Research objective 1: *To enable the inclusion of micro-level geographic factors in longitudinal historical analyses*

- By extending the standardised data model IDS to include geographic data with the creation of IDS-Geo, we provided the possibility of integrating detailed geographic data within IDS by adding standardised geometric

data types to permit the distribution of geometric representations. The option to include geographic data in a common model using standardised exchange formats will aid data sharing and facilitate the development of extraction and transformation programs (Paper I).

- To create geocoded longitudinal demographic databases and maintain the longitudinal data's analytical advantage, the use of longitudinal object-lifeline (or similar) time representations is necessary. This thesis has also shown how supplementary textual sources can be used to transform snapshots of geographic objects (digitised from historical maps) to object-lifelines (Papers II, III).
- Although some parts of the methodology developed in Paper II are not new in historical and geographic research, the integration of a set of different techniques into a single scheme, which encompasses all parts of the process, enables the inclusion of the micro-level geographic context to longitudinal historical demographic analyses (Paper II).
- A main outcome of Papers II and III is the geocoded subset of the Scanian Economic Demographic Database (SEDD) and the additional geographic data. Approximately 53,000 individuals who lived in the five studied rural parishes are linked to the property units in which they lived for the period 1813-1914. The geographic database includes historical objects: property units, wetlands, buildings, roads and railroads. The property units and wetlands are stored in object-lifeline time representations, whereas the other objects are stored as snapshots in time. Thus, this database presents one of the first opportunities to study historical spatio-temporal patterns at the micro-level.

Research objective 2: To study how the geographic geocoding level affects the results of demographic analyses

- Even relatively small differences between the property units and the coarser geographic levels influenced both the magnitude of the effect and the statistical power in the survival analyses. Therefore, the property unit level is likely needed for fine-scale analysis when analysing distances less than approximately 350 metres (Paper IV).
- Substantial changes in both the geographic data that are used to geocode the population and the external data used to compute the context variables, such as the drainage of wetlands, should be considered (Paper IV).
- The methods used to quantify the geographic context variables substantially influenced the results of the survival analyses compared with the geographic level. Therefore, the use of appropriate quantification methods to properly utilise the geographic level is important (Paper IV).

- This study is one of the first to use longitudinal geographic micro-level data to analyse the effect of the proximity to wetlands and geographically weighted population density on mortality in a historical context. We observed a strong effect of proximity to wetlands on the mortality of adult females. The close proximity to wetlands may have increased the exposure to malaria-transmitting mosquitoes (Paper IV).

Research objective 3: To use the geographic information to improve and extend the knowledge of demographic change

- Using the geocoded database for the period 1850-1914 to conduct a longitudinal demographic analysis with geographic micro-level factors, we discovered that nutrition served a relatively important role as a predictor of child mortality in a rural setting. Certain soil types seem to have influenced the agricultural productivity, which in turn affected the nutrition of the farmers' children and their likelihood of dying (Paper V).
- Regarding the geographic differences in child mortality among the five parishes in the study area, these differences were not related to soil type via agricultural productivity and nutrition. Thus, the new results from Paper V deepen our understanding of the reasons for the geographic mortality differences. In addition, the study illustrates that the geographic variables may be used as proxies for other demographic characteristics (Paper V).

To summarise, this thesis has addressed a set of important issues related to the micro-level geographic context that have currently been missing in historical demography and related fields. The thesis has contributed to historical demographic research by including and studying micro-level geographic factors in longitudinal historical analyses. The thesis also adds new findings to the debate regarding the importance of nutrition and agricultural productivity to child mortality in preindustrial Sweden. The geocoded longitudinal demographic database has proven to be an important and unique resource for historical demographic research (Paper IV, V). In addition, the geocoded database is openly and freely available to other researchers (cf. Paper III). Consequently, the results and subsequent knowledge gained from this thesis represent a starting point for additional research regarding how the geographic context affects demographic outcomes.

5.2 Future studies

This thesis generates potential for additional research, including 1) studies that use the current version of the database and 2) the extension of the geocoded database.

Using the current version of the database, additional insights into how geographic factors affect human living conditions at the micro-level are possible. For example, Paper IV revealed the strong mortality effect of proximity to wetlands on adult females, which may indicate that close proximity to wetlands increased the exposure to malaria-transmitting mosquitoes. Before drawing conclusions on the casual mechanisms of this relationship, additional research is needed. Environmental information about factors associated with the malaria risk, such as elevation, population density, temperature, precipitation, soil types, topographic wetness indices (TWIs), and forestlands, can be added. Moreover, a longitudinal social network analysis, for example, can be performed by computing network centrality measures to estimate central individuals in parishes to determine the effect of this centrality on their demographic outcomes.

The temporal dimension of the geocoded database can be extended both backward and forward in time to address the period of time before the enclosures to the present. First, the scattered agricultural plots from the open-field system, which were used before the enclosures, can be digitised from the georeferenced historical maps; thereafter, the individuals can be geocoded to their scattered plots. For this period, longitudinal and detailed production series for grain and livestock for each farm for the five parishes can be incorporated from the Historical Database of Scanian Agriculture (Olsson and Svensson, 2016). Second, the geocoding can also be extended forward in time. The period 1914-1968 may likely require digitisation and standardisation of addresses. From 1968 and onwards, standardised addresses are available, which will facilitate the geocoding. Finally, the buildings in temporal snapshots can be transformed to object-lifelines and linked to their corresponding property unit, which will enable a geocoding on the building level. In addition, other geographic objects, such as road networks, can be transformed to object-lifelines, and land use can sometimes be digitised. Although several of the abovementioned improvements may be time-consuming, they will facilitate new research.

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.

Banerjee, S., Wall, M. M., and Carlin, B. P. 2003. Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota.*Biostatistics*, 4(1): 123-142.

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

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., and Dribe, M. 2010. Quantifying the Family Frailty Effect in Infant and Child Mortality by Using Median Hazard Ratio (MHR). *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 43*(1): 15-27.

Bengtsson, T., and Dribe, M. 2011. The late emergence of socioeconomic mortality differentials: A micro-level study of adult mortality in southern Sweden 1815-1968. *Explorations in Economic History, 48*(3): 389-400.

Bengtsson T. Dribe M., Quaranta L., and Svensson P. 2014. *The Scanian Economic Demographic Database, Version 4.0 (Machine-readable database)*. C.f.E.D. Lund University, Lund http://www.ed.lu.se/databases/sedd/sedd-public-access.

Bernardinelli, L., and Montomoli, C. 1992. Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in medicine*, 11(8): 983-1007.

Berners-Lee, T. 2006. Design issues: Linked data. Retrieved September, 2016, from http://www.w3.org/DesignIssues/LinkedData.html.

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/.

Centre for Economic Demography (CED). 2015. *SEDD—The Scanian Economic Demographic Database*. http://www.ed.lu.se/databases/sedd (accessed August, 2016).

Claësson, O. 2009. *Geographical differences in infant and child mortality during the initial mortality decline: evidence from southern Sweden, 1749-1830*. Department of Economic History, Centre for Economic Demography, Lund University.

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.

Cromley, E. K., and McLafferty, S. L. 2012. *GIS and public health*, 2nd edition. Guilford Press.

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

Darmofal, D. 2009. Bayesian spatial survival models for political event processes. *American Journal of Political Science*, 53(1): 241-257.

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.

Dibben, C., Sigala, M., and Macfarlane, A. 2006. Area deprivation, individual factors and low birth weight in England: is there evidence of an "area effect"? *Journal of epidemiology and community health*, 60(12):1053-1059.

Dribe, M., and Lundh, C. 2005. People on the move: Determinants of servant migration in nineteenth-century Sweden. Continuity and Change, 20(1): 53–91.

Dribe, M., Olsson, M., and Svensson, P. 2011. Production, prices and mortality: Demographic response to economic hardship in rural Sweden, 1750–1860. In *Paper for the meetings of the European Historical Economics Society, Dublin*.

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.

Floud, R., Fogel, R. W., Harris, B., and Hong, S. C. 2011. *The changing body: Health, nutrition, and human development in the western world since 1700*: Cambridge University Press.

Fogel, R. 1994. Economic growth, population theory, and physiology: the bearing of long-term processes on the making of economic policy. *The American Economic Review*, 84(3), 369-395.

Fogel, R. 2004. *The escape from hunger and premature death, 1700-2100: Europe, America and the Third World*. Cambridge: Cambridge University Press.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, A. S., Crespo, R., and Yao, J. 2015. Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, 47(4):431-452.

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.

Harig, O., Burghardt, D. and Hecht, R. 2016. A supervised approach to delineate built-up areas for monitoring and analysis of settlements. *ISPRS International Journal of Geo-Information*, 5(8):137.

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.

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.

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 August, 2016, from http://www.lantmateriet.se/sv/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Lantmateristyrelsens-arkiv---LMS/Geometriska-jordebocker/.

Lantmäteriet. 2014b. *Skifteskartor*. Retrieved August, 2016, from http://www.lantmateriet.se/Kartor-och-geografisk-information/Historiska-kartor/Arkiven-som-ingar/Lantmateristyrelsens-arkiv---LMS/Skifteskartor/.

Lantmäteriet. 2014c. *Häradsekonomiska kartan*. Retrieved August, 2016, 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 August, 2016, 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.

Lee, D., Ferguson, C., and Mitchell, R. 2009. Air pollution and health in Scotland: a multicity study. *Biostatistics*, kxp010.

Lindgren, E., and Jaenson, T. G. 2006. Fästing-och myggöverförda infektionssjukdomar i ett kommande, varmare klimat i Sverige. *Ent. Tidskr*, 127:21-30.

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

Lundh, C., and Olsson, M. 2005. Contract-workers in Swedish Agriculture in the Nineteenth and Twentieth Centuries: A Comparative Study of Standard of Living and Social Status. In *Paper for the 20th International Congress for the Historical Sciences, Sydney 3-9 July 2005*.

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.

McKeown, T. 1976. *The modern rise of population*. London: Arnold.

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.

Olsson, M. 2008. Storjordbruk, statare och andra, in *Statarliv: i myt och verklighet,* edited by C. Lundh*,* and M. Olsson. Gidlunds förlag: Hedemora.

Olsson, M., and Svensson, P. 2016. *Historical Database of Scanian Agriculture*, version 2.0. Lund. Department of Economic History, Lund University.

Peuquet, D. J., and Duan, N. 1995. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of geographical information systems*, 9(1):7-24.

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.

Portele, C. 2012. OGC Geography Markup Language (GML)–Extended schemas and encoding rules. In *Open Geospatial Consortium Inc*.

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.

Puleston, C. O., and Tuljapurkar, S. 2008. Population and prehistory II: Space-limited human populations in constant environments. *Theoretical Population Biology, 74*(2):147-160.

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.

Quaranta, L. 2015. Using the intermediate data structure (IDS) to construct files for statistical analysis. *Historical Life Course Studies*, 2:86-107.

Quaranta, L. 2016. STATA Programs for Using the Intermediate Data Structure (IDS) to Construct Files for Statistical Analysis. *Historical Life Course Studies*, 3:1-19.

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.

Schade, S., and Cox, S. 2010. Linked Data in SDI or How GML is not about Trees. In *Proceedings of the 13th AGILE International Conference on Geographic Information Science-Geospatial Thinking* (pp. 1-10).

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.

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.

Smith, D. S. 1983. Differential mortality in the United States before 1900. *The Journal of Interdisciplinary History*, 4:735.

Stead, W. W., Hammond, W. E. and Straube, M. J. 1982. A Chartless Record – Is it Adequate? *Proceedings of the Annual Symposium on Computer Application in Medical Care*, Nov 2, 89–94.

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.

Tveite, H. An accuracy assessment method for geographical line data sets based on buffering. 1999. *International journal of geographical information science*, 13(1):27-47.

Van Poppel, F., Jonker, M., and Mandemakers, K. 2005. Differential infant and child mortality in three Dutch regions, 1812-1909. *Economic History Review*, 58(2):272-309.

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 August, 2016 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.

World Wide Web Consortium (W3C). 2016. *Linked Data*. Retrieved September, 2016, from https://www.w3.org/standards/semanticweb/data.

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).