The influence of geocoding level and definition of geographic context variables on historical demographic analyses

Karolina Pantazatou
Dep. of Physical geography and Ecosystem Science
Lund University
Sölvegatan 12, SE-223 62 Lund, Sweden
kpantazatou@gmail.com

Finn Hedefalk
Dep. of Physical geography and Ecosystem Science
Lund University
Sölvegatan 12, SE-223 62 Lund, Sweden
finn.hedefalk@nateko.lu.se

Lars Harrie
Dep. of Physical geography and Ecosystem Science
Lund University
Sölvegatan 12, SE-223 62 Lund, Sweden
lars.harrie@nateko.lu.se

Luciana Quaranta
Centre for Economic Demography and Department of Economic History
Lund University
Scheelevägen 15 B, SE-220 07 Lund, Sweden
luciana.quaranta@ekh.lu.se

Abstract

The quality of spatial analysis is highly dependent on the geo-referencing quality and the definition of the geographic context variables used. However, such information is seldom considered in historical demographic and epidemiological research that includes the geographic context. We investigate a suitable geographic level for geocoding of the population and definition of geographic context variables for historical demographic studies. Using longitudinal demographic data combined with unique historical geographic micro-data on residential histories, we compare two geocoding levels (property unit and address unit) and two definitions of distance to wetlands (an indicator of exposure to malaria in the 19th century Sweden). We first statistically compare the differences in the distance to wetlands between the two geocoding levels. Thereafter, by analysing how distance to wetlands affected the mortality of children aged 1-15 living in four rural parishes in Sweden, 1850-1914, we study the effect of different geocoding levels and definitions of context variables on the quality of historical demographic analysis. We find that both the geocoding level and the definition of the geographic context variables strongly influence the results of the analyses. For the distance to wetland variable, decreasing its average positional accuracy by at least 100 meters affects the results. Consequently, the significant differences between the two geocoding levels indicate the importance of considering the geographic level of detail when geocoding.

Keywords: Geocoding, Historical demography, Level of detail, Longitudinal databases, Historical GIS, Survival analysis

1 Introduction

Spatial aspects are important in demographic and epidemiological research for studies both on the micro level (individual level) and macro level (aggregated level) [11]. The development of spatial methods have been triggered by better access to geographic data as well as advanced statistical packages and GIS systems. To perform spatial analysis using modern demographic and geographic data is therefore common today; these analyses are often longitudinal i.e., they cover long time periods.

Also in the field of historical demography several studies have included the geographic context (e.g., [3, 4]), foremost by studying how the environment affects demographic outcomes. However, with some exceptions [6, 10], these analyses have been performed on macro level where individuals have been geocoded to administrative boundaries. Studies using more detailed geographic data (e.g., [3]) have analysed demographic data without a longitudinal component. That is, the developments in demographic and epidemiological research of modern data, of including the geographic context into longitudinal micro-level studies, have just begun to emerge in historical demography. The main reason is the lack of longitudinal databases of historical individual-level data where the individuals are geocoded to detailed physical locations.

During the last decade several longitudinal demographic databases of historical populations on the individual level have been created. Examples in Europe are the Historical Sample of the Netherlands (HSN), the Scanian Economic Demographic Database (SEDD) (Sweden), and the COR sample (Belgium). In most of these databases place names are connected to the individuals, at least on a coarse geographic level. The first systematic geocoding of all the individuals in a database on micro-level has recently been performed for parts of the SEDD database [5].

Furthermore, in epidemiology several studies have addressed how the quality of spatial analyses is dependent on the geo-referencing quality (e.g. [12]). However, such information is seldom considered in epidemiological studies that applies spatial analysis [7]. Additionally, a discussion is often lacking about the appropriateness of the defined variables used to examine the environmental influences on health. In historical demography this may be an even bigger problem because of the later adoption of spatial analysis in this field.

The GIScience field has, at least, three main tasks to enable historical demographic researchers to include the geographic context into their longitudinal analysis. Firstly, a methodology to geocode the longitudinal population is required. A linkage to a modern coordinate system is necessary to enable the use of historical and modern geographic data to compute the geographic context. Secondly, a recommendation of the geographic level (and time resolution) of the geocoding is required. Generally, it is more expensive to make a geocoding on a detailed geographic level since it likely entails more manual work with interpreting written historical sources. Therefore, it is important not to perform the geocoding on a more detailed level than required for the anticipated demographic studies. The third contribution concerns how the geographic context should be included in the longitudinal analysis. That is, how to define and compute measures that describes the geographic context, henceforth denoted as *geographic context variables*.

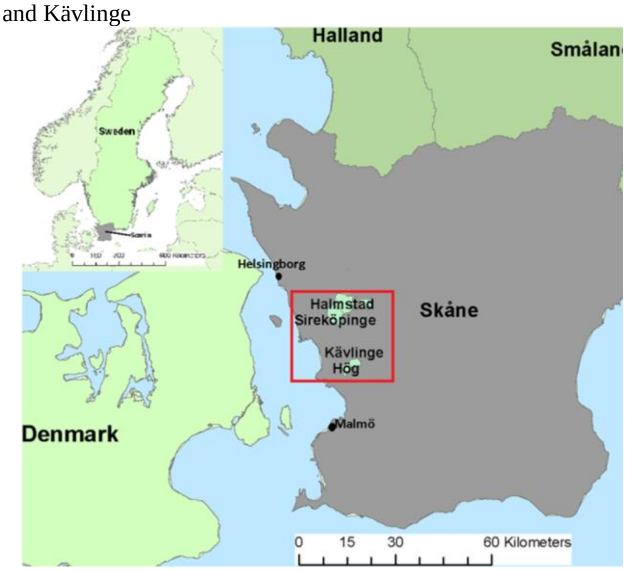
The aim of this study is to analyse the two latter tasks; i.e., to find a suitable geographic level for geocoding of the population and definition of geographic context variables.

2 Background

2.1 Study area and data

We use the longitudinal and individual-level Scanian Economic Demographic Database (SEDD) [2]. The SEDD includes information on all inhabitants that have lived in five rural parishes in southern Sweden from 1646 and onwards (Figure 1). Four of these parishes are considered in this study. Although the study area is relatively small, its long temporal dimension and detailed data makes it suitable for longitudinal analyses. The information is mainly demographic and economic; however, contextual information is also available. Sources for this information primarily include population registers, vital registers and poll-tax registers. The individuals in the parishes were traced from when they were born or in-migrated to when they died or out-migrated.

Figure 1: The four rural parishes: Halmstad, Sireköpinge, Hög and Kävlinge



Source: [5]

The SEDD has been extensively used in historical demographic research. One important demographic outcome that has been studied is mortality. It has, among other things, been noted that large regional differences in both childhood and adult mortality existed until the 20th century, also after controlling for socio-economic factors [1]. Why these regional differences emerge is partly unknown. One hypothesis is that the geographic context caused such differences in mortality; e.g., because of variations in the exposure to infectious diseases. The risk of receiving an infectious disease is, among others, possibly dependent on context variables such as population density and closeness to wetlands/water (malaria was a problem in this area during the 19th century). Other context variables that potentially could influence the mortality are e.g. natural chemical substances in the drinking water, or soil conditions [5].

2.2 Geocoding individuals in SEDD on two geographic levels

To introduce the geographic context in demographic analyses we need geographic information to geocode the individuals as well as for computing the geographic context variables. The geographic information used in this study is based on around 60 historical maps that encompass the four parishes for the period 1757-1914, as well as modern geographic datasets.

A recent project aimed at geocoding all the individuals in SEDD for the time period 1813-1914 (cf. [5]). This geocoding required extensive work for the following reasons. Before the land reforms (conducted between 1757 and 1849 in the parishes), all the individuals 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 forestlands. Their size varied between 0.001 km^2, and 5 km^2 in the study area, with an average size of 0.2 km^2. The average positional accuracy of the digitized property units in this study area is approximately 25 metres. We anticipate that people mainly resided within their property units during day and night, which is a justified anticipation in rural 19th century Sweden.

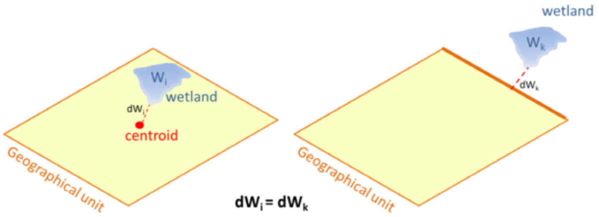
Throughout the period, several of the property units were subdivided or partitioned into smaller units (in line with the rapid population growth). However, they did not receive new addresses; hence, multiple property units often share addresses. We denote the set of such units an *address unit*. Usually they are close to each other, but not necessarily adjacent.

To perform the geocoding on the address unit level is straightforward since the poll-tax register contains annual information about the address unit for the family head. However, the project also aimed to geocode on the property unit level which introduced two main challenges [5]. Firstly, the information about the geometries of the property units is only snapshots in time (from historical maps). To create an object life-line representation of the property units – which is required for the geocoding – the authors had to combine the historical maps with textual sources containing information about changes in property units (mainly cadastral dossiers and poll-tax registers). Secondly, to link each individual to the property unit they lived in, the problem of property units sharing addresses had to be evaded. Thus, they used taxation values in the poll-tax registers combined with textual sources of the maps and cadastral dossiers to separate the units sharing addresses. Therefore, geocoding on the property unit level was more costly than on the address level. Around 57,000 individuals were linked for each year which makes it possible to trace how the individuals have moved within and between the parishes.

2.3 Defining geographic context variables

Often the definition of geographic context variables can be challenging. Population density, for example, requires a decision of level of detail used in definition and also if some kind of weighting depending on closeness should be introduced. Also seemingly simple geographic context variables for closeness require awareness with the definition. Here we focus on one such variable: distance to wetlands, which is a possible indicator for exposure to malaria and hence important in demographic and epidemiological research. To compute this variable we need information about the geometric extent of property units and wetlands. Simple approaches are based on either the borders and/or centroid of the polygons (Figure 2). Depending on how the distance is defined different values are, of course, obtained (cf. [8]). We use two such simple measures of distance to wetlands: *centroid-to-border* (Figure 2a) and *border-to-border* (Figure 2b).

Figure 2: Variation of distance to wetland from border of wetland polygon to either border or centroid of a property unit.



Source: [8]

3 Method

We first compare the differences in the distance to wetlands between the two geocoding levels. Here, the centroid-to-border measure is used describe the geographic context variable. Thereafter, by analysing how distance to wetlands affected the mortality of children aged 1-15 living in the study area, we study the effect of the different geocoding levels and definitions of context variables on demographic survival analysis. Children are selected because they were sensitive to environmental factors. We study the period after the large land reforms had taken place in the parishes; i.e., 1850-1914. Before these reforms there was little variation in the distances to wetlands (because everyone lived within the village).

3.1 Statistical analyses of geocoding levels

We examine if and how much the distance to wetland variable differs whilst computed over different geocoding levels. First, the absolute difference between every property unit's distance-value and its corresponding address unit distance-value is calculated. Then, the mean and the standard deviation of all calculated absolute differences are computed. Because the values of the absolute differences do not necessarily follow a normal (Gaussian) distribution, we use Chebyshev's Inequality, which can be applied to arbitrary distributions, to estimate the probability of how many absolute-difference values lie within k standard deviations of the mean:

$$Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \quad ,k > 0 \qquad (2)$$

where X represents a randomly chosen absolute difference value from the computed dataset, μ is the mean of all absolute-difference values, σ is the standard deviation and *k* is any real number higher than zero.

3.2 Survival analysis

Survival analysis is a standard approach to study how living conditions and other factors affect the likelihood of experiencing demographic outcomes such as mortality. Survival analysis includes several methods that focus on questions related to duration until an event occurs (e.g. an individual's death). The models include the hazard rate (*h(t)*), which is the conditional probability that an event occurs at a particular time *(t)*. One method to model the hazard function is the Cox proportional hazard model. The general expression of the hazard function is [9]:

$$h_i(t) = h_0(t)e^{\beta_1 x_{i1}+\ldots+\beta_k x_{ik}} \qquad (1)$$

where $h_0(t)$ is the baseline hazard function, x_i the independent variables that affects the hazard and β are parameters describing the influences of the variables. By using the Cox model, we analyse if and how the distance to wetlands influences mortality of children for the period 1850-1914. We estimate separate models for the two geocoding levels as well as the two measures for the distance to wetlands.

4 Results

4.1 Statistical analyses of geocoding levels

Table 1 shows the interval (μ-√2σ, μ+√2σ) or (μ-2σ, μ+2σ) within which the absolute difference between the proximity-to-wetland values of the two geographic units belonging to different geocoding levels might lie in, with a probability of 50% or 75% respectively. Here, quite large differences between the two geocoding levels can be observed.

Table 1: Differences in shortest distance to wetlands between property units and address units. Centroid-to-border distance measure.

	Address Unit – Property Unit (Absolute difference in m)			
Parish	Mean	St. Dev.	Interval (50%)	Interval (75%)
Halmstad	81.5	98.3	0 – 220.5	0 – 278.1
Hög	103.4	78.4	0 – 214.3	0 – 260.2
Kävlinge	228.9	172.5	0 – 472.8	0 – 573.9
Sireköpinge	100.4	110.1	0 – 256.1	0 – 320.6
All parishes	120.2	127.8	0 – 300.9	0 – 375.8

4.2 Survival analysis

Table 2 shows the distribution of the children's time at risk in percentage among the variable distance to wetlands. Tables 3-4 show the impact of distance to wetlands on mortality for children aged 1-15, 1850-1914. Controls are included for parish of residence, social class, gender and birth year (the results of these parameters are not reported here). When using the property unit level and the centroid-to-border distance measure, children residing at least 400 meters from wetlands experience a 19% lower risk of death (hazard ratio) compared to children living within 400 meters from a wetland (Table 3) (*P<0.1*). For the address level, as well as when using the border-to-border distance measure, no significant effects from distance to wetlands are found.

Table 2: Distribution of the time at risk in person-years on the variable distance to wetlands (ages 1-15).

	Centroid-to-Border		Border-to-Border	
	Property unit	Address unit	Property unit	Address unit
	%	%	%	%
Dist. to wetlands				
<400 m	57.4	63.5	47.1	63.1
>=400 m	42.6	36.5	52.9	36.9

Subjects: 6792; deaths: 393; years at risk: 36837.52

Table 3: Impact of distance to wetlands on mortality (ages 1-15). Centroid-to-border measure.

	Property unit		Address unit	
	Haz. ratio	P>z	Haz. ratio	P>z
Dist. to wetlands				
<400 m	1.00	rc	1.00	rc
>=400 m	0.81	0.07	0.97	0.78

Subjects: 6792; deaths: 393; years at risk: 36837.52

rc = reference category.

Table 4: Impact of distance to wetland on mortality (ages 1-15). Border-to-border measure.

	Property unit		Address unit	
	Haz. ratio	P>z	Haz. ratio	P>z
Dist. to wetlands				
<400 m	1.00	rc	1.00	rc
>=400 m	1.00	0.99	1.06	0.63

Subjects: 6792; deaths: 393; years at risk: 36837.52

rc = reference category.

5 Concluding remarks

We find that both the geocoding level and the definition of the geographic context variables strongly influence the results of the demographic analyses. Only when the property unit level is used, significant effects are observed for the exposure to wetlands on child mortality. When using a slightly coarser geographic level such as the address unit, no significant effects are found. Thus, for the distance to wetland variable, decreasing its average positional accuracy by at least 100 meters affects the results (Table 1). Consequently, the significant differences between the two geocoding levels indicate the importance of considering the geographic level of detail when geocoding. Further studies are, nevertheless, needed to more accurately find out what geocoding level that is needed to produce reliable results in demographic analysis.

Moreover, using a correct definition of the geographic context variables is also a crucial aspect. This can be seen in the very different results between the two distance measures at the property unit level (Tables 3-4); i.e., using the centroid-to-border measure produce significant effects whereas the border-to-border measure do not.

In this study we only used two simple distance measures to estimate exposure to wetlands. However, in future studies we aim to extend this section by testing several other ways of estimating the exposure to wetlands as well as including other geographic context variables and geocoding levels (i.e., buildings).

References

[1] T. Bengtsson and M. Dribe. Quantifying the Family Frailty Effect in Infant and Child Mortality by Using Median Hazard Ratio (MHR). *Historical Methods,* 43(1):15-27, 2010.

[2] T. Bengtsson, M. Dribe, L. Quaranta and P. Svensson. *The Scanian Economic Demographic Database, Version 4.0 (Machine-readable database), C.f.E.D.* Lund University, Lund, 2014.

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

[4] I. N. Gregory. Different places, different stories: Infant mortality decline in England and Wales, 1851–1911. *Annals of the Association of American Geographers,* 98(4):773-794, 2008.

[5] F. Hedefalk, L. Harrie and P. Svensson. 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*, 48(3):153-173, 2015.

[6] F. Hedefalk, L. Quaranta and T. Bengtsson. The influence of micro-level soil factors on child mortality in southern Sweden, 1850-1914. In *IUSSP Seminar on Spatial Analysis in Historical Demography: Micro and Macro Approache*s, Quebec City, Canada, 2015.

[7] G. M. Jacquez. A research agenda: Does geocoding positional error matter in health GIS studies? *Spatial and spatio-temporal epidemiology,* 3(1):7-16, 2012.

[8] K. D. Pantazatou. *Issues of Geographic Context Variable Calculation Methods applied at different Geographic Levels in Spatial Historical Demographic Research - A case study over four parishes in Southern Sweden* (Master's thesis). Lund University, Lund, Sweden. 2016.

[9] T. M. Therneau and P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer-Verlag, New York, 2000.

[10] C. B. Villarreal, Bettenhausen, E. Hanss and J. Hersh. Historical Health Conditions in Major US Cities: The HUE Data Set. *Historical Methods*, 47(2):67-80, 2014.

[11] P. R. Voss. Demography as a spatial social science. *Population Research and Policy Review*, 26(5-6):457-476, 2007.

[12] P. A. Zandbergen. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *Bmc Public Health*, 7(1):1-13, 2007.