Spatial modeling of demographic development in Sweden

Author: Binyan Xia & Yongtao Yu

Supervisor: Lars Rönnegård & Johan Håkansson

D-level essay in statistics, June 2010,

School of Technology & Business Studies, Dalarna University

Abstract

Over the past 200 years, the demography of Sweden has changed dramatically, while the population changes have been different from one area to another. In this thesis, we will focus on discussing the main factors that affect population redistribution parishes in Sweden. It seems that the parish's population, the populations of the nearby districts, the time and whether it is an urban or a rural area are the main variables which the population growth rate is dependent on. In this thesis, the ArcGIS software provides geographical data description regarding the population in Sweden. The empirical results that we get from the gravity model help us to model the relationship between the population growth and the main effect elements. Assuming the data follows Poisson distribution, we wish to obtain the relationship between the factors and the population growth rate based on the generalized linear model, but unfortunately there is over-dispersion. So we choose an alternative mixed linear model in order to fit the data by normal distribution. The final model helps us to understand the formation of big cities and the interrelationship between them.

Key words: Population redistribution; Population growth rate; ArcGis; Generalized linear model; Mixed linear model.

Contents

1 Introduction

1.1 Background

Sweden is a Nordic country on the Scandinavian Peninsula in the Northern part of Europe. Sweden, which is about 450,295 km^2, is the third largest country in the European Union in terms of area, while the total population is only 9.2 million. Hence, Sweden has a low population density of 21 inhabitants per square kilometer. About 85% of the population lives in the urban areas, and it seems that this number will gradually rise as a part of the ongoing urbanization (Yearbook of housing and building Statistics 2007). The demography of Sweden has changed drastically over the 19th and 20th centuries. From 1810 to 2004, the total population of Sweden has increased nearly three times from 2,378,122 to 9,013,396 which is shown in figure 1. On the other hand, the distribution of Sweden's population has changed a lot during the past 200 years. Of course, the growth was different in different parts of the country. The demographic development seems dependent on the population in nearby districts, however we do not know how well the demographic development can be predicted from historical changes.

Stockholm, the capital city, is the largest city in the country. There are 1.3 million people living in the urban area. However, hundreds of years ago, there were only a few people living in Stockholm. While there were a lot of people living in the rural areas. So the questions come up: How did these people move from the countryside to the cities? How did Stockholm become a metropolis in Sweden? Are the people living in Stockholm moving from areas close to Stockholm or from regions far away? Uppsala and Linköping are also big cities in the country which are close to Stockholm. Do the changes of the population in Stockholm affect the growth of these cities or are they independent? While at the same time, do Göteborg, Lund and Malmö have the same phenomenon? If we

know the answers of these questions, we may find out the truth of the formation of the city today.

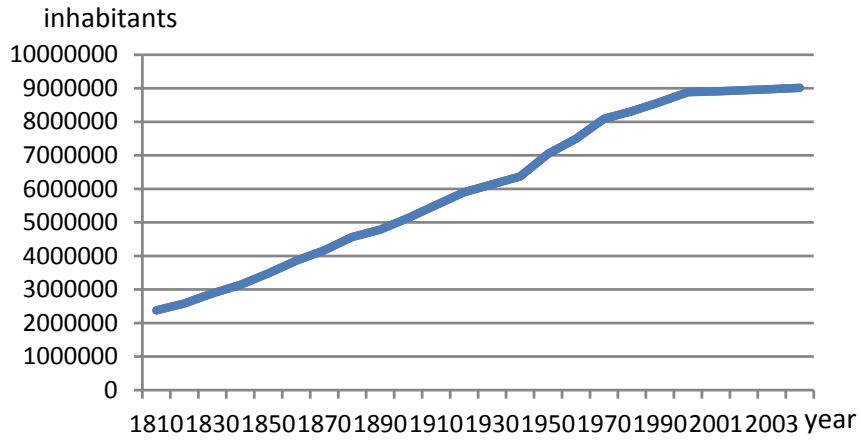


Figure 1: Total population of Sweden from 1810 to 2004

1.2 Previous studies

1.2.1 Previous studies on demography and spatial distribution

The demography of the whole Europe has changed dramatically over the past two centuries. The rapid population growth being accompanied by urbanization and industrialization was followed by a slowing of growth, suburbanization and deindustrialization (Misa and Schot, 2005).

There were a great number of researches about the redistribution of population. The researchers have got several studies about the methods of migration. From the previous research in Sweden about the redistribution of the population, the researchers divided this change into two parts which are concentration and dispersion; and they denoted this change both on local and regional levels. There are several ways which may lead to the concentration or dispersion.

Firstly, an important way is colonization. The colonization may lead to new residences in some outlying areas (Enequist, 1937, Hoppe, 1945, and Rudberg, 1957). For the original place, plenty of people moved out. Under this situation, it can be considered that the population was dispersed. Secondly, urbanization is also a significant way which causes the redistribution of population. The urbanization contains two processes which are countryside urbanization and urbanization. Under

this way, the residence attracted many people to live or find job (Godlund, 1964, Jakobsson, 1969). Both these two processes of the urbanization led to the population being concentrated locally. Thirdly, there are also other kinds of urbanization: the suburbanization and counter urbanization. Because cites were too crowded, people moved to the beautiful and comfortable suburban areas. (Lewan, N. 1967) The larger towns experienced outmigration when more and more people moved into smaller towns. The suburbanization and counter urbanization lead to the dispersion of population. Finally, some studies are dealt with over longer periods (Andersson, 1978&1987, Bäcklund, 1999). These factors may affect the population redistribution in form of association in such a long period. After these various change, it obtains the final structure of today's city.

In the former studies, some researchers hold that the spatial distribution of population is determined by four different kinds of factors: 1) distance variables, 2) demo-graphic variables, 3) housing construction and 4) level of industrialization. Consequently, they chose the change in population as the dependent variable and distance, resident population, population density, excess of births, urban population, housing construction, type of housing and level of industrialization as independent variables (Aldskogius, 1970). Of course, we can draw inspiration from these studies that we may consider that the distance and population density affect the change of population significantly.

1.2.2 Gravity model

The gravity model of migration is a model in urban geography root in Newton's law of gravity, and used to predict the degree of interaction between two areas (Slack, 2009). Newton's law states that: "Any two bodies attract one another with a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them."

When used geographically, the words 'bodies' and 'masses' are replaced by 'locations' and 'importance' respectively, where significance can be measured in the the case of population numbers, gross domestic product, or another appropriate variables. In consequence, the gravity model of migration is based on the idea that as the importance of one or both of the location increases, there will also be an increase in movement between them. The larger the distance between two locations is, however, the movement between them will be less. This phenomenon is known as famous distance decay.

The gravity model is used widely that it can't only be used to estimate variance aspect: traffic flow, migration between two areas and the number of people likely to use one central place, but also be used to determine the sphere of influence of each central place by estimating where the breaking point between the two settlements will be. For example, the point at which customers find it preferable can be determined by this, based on distance, time and expense considerations, to travel to one center rather than the other. In 1931, William J. Reilly expanded the gravity model into Reilly's law of retail gravitation. He used gravity model to calculate the breaking point between two places where customers will be drawn to one or another of two competing commercial centers.

1.3 Aim

The main aim of this thesis is to model Swedish population data with a gravity model using a generalized linear model.

On one hand, we can analyze what elements the population growth depends on. Making last year's population as an offset, we can make a generalized linear model that the population change ratio is related to the population density (as a dummy variable), the time (as a class variable) and gravity which is dependent on the interaction of nearby areas. On the other hand, we may also add parish as a random effect and fit the data into a mixed linear model to avoid over-dispersion.

1.4 Outline

In this article, we firstly give a background of Sweden population's development, some previous studies about the population redistribution and the gravity model. In the section 2, the data is described and processed by ArcGis software in order to show the relationship between the data. In the third section, we introduce the methodology relevant to our analysis. In fact, we focus on make generalized linear model and mixed linear models based on gravity model to fit the data. In the last section, we give the results of our analysis. Then we obtain the conclusion on what we have done and discussion about the problems need for further research.

2. Description of data

2.1 Summary of data

The data we use is from "spatial population redistribution in Sweden 1810-1999" by Johan Håkansson (2000). It is quite smart data that it divided Sweden into 1840 small areas which is more convenient for us to analyze using statistics method. These small areas are defined by the different parishes in Sweden. With the help of the agency of the church, priests recorded the daily activities of people in every parish, for example, birth and death. Actually, we can also find information including where people lives from tax department. Together with the population, this information was also recorded at the parish levels. The latest data series of the population in every parish we can get is from 1810 to 2004. Firstly, the data has a code for each area respectively which is so useful in the model analysis that we can use it as a random effect in the mixed linear model. Secondly, using the X-coordinate and the Y-coordinate, we can easily orientate where the parish is. And we can also know the distance between each area. There is one thing we should know is that the data we have here is not surface data but point data. This means we only know the position but we have

no idea of whether one is adjacent to another. So we can't make a matrix to show the relationship from the connection. Thirdly, based on the population of each year and the area accordingly, we can obtain the population density. All of these 6 variables given to us are so important for the model foundation, then, we should have an overview of the data.

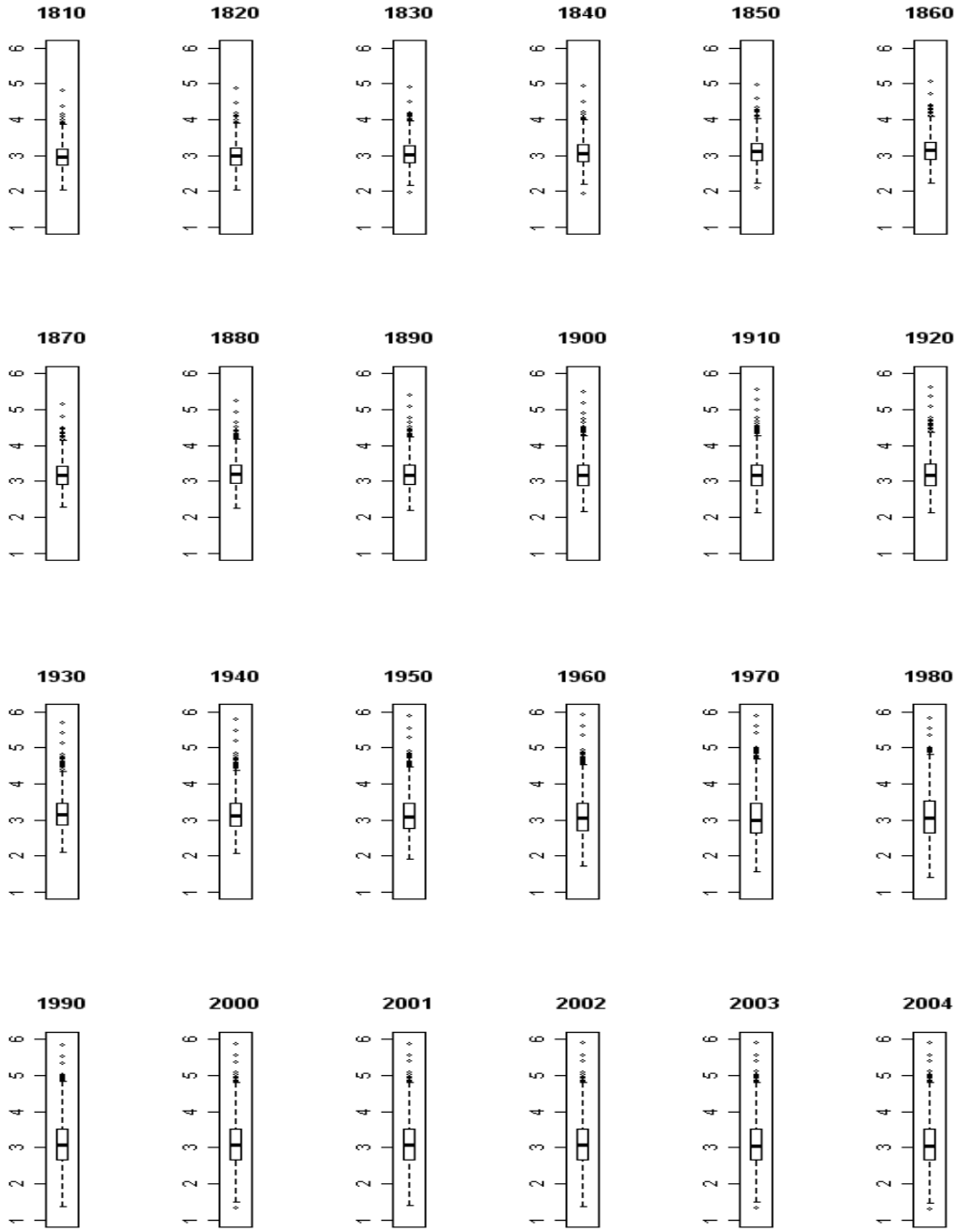


Figure 2: Boxplot of the log 10 transformation of parish population sizes in Sweden from 1840 to 2004

We can get the details of Sweden's total population in each year by summary the data using R. From the mean value of the population, it can be seen that the population of Sweden increase year by year. However, the median value gives different information (Figure 2). It rose from 1810 to 1890, whereas it decreased gradually after 20th century. By analyzing these two data, we consider that the population widely distributed in various regions in 19th century. But there was a trend of population concentration. It means the big cities had a growing number of people. On the contrary, the population in sparsely populated areas was fewer and fewer. This result can also be obtained from the minimum and maximum value. After 1900, the minimum value has been getting smaller and smaller, however, the maximum value has a trend of growing at the same time.

2.2 Data processing

After summarizing of the data, we may do some basic data handling using R. Firstly, we can get the population density (noted as pop.density) calculated from the population and the area size of each parish in km^2. Secondly, we can define Standard metropolitan statistical areas (SMSA's) which is expresses by means of a dummy variable. Dummy variable is defined as one that takes the values 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to change the outcome. Of course, this SMSA may not be the real cities, but in the former studies many statistics took it as a kind of empirical criterion. So in this thesis, we also regard SMSA as cities. In our case, we can make a hypothesis that the top 25% of the areas are urban areas in accordance with the population density. Cities usually have a larger population density than the rural areas. That is to say when the population density is bigger than the 75% quantile, we represent it as a numerical value 1; when the population density is smaller than the 75% quantile, we represent it as a numerical value 0 instead. Thirdly, we can calculate the interaction:

$$\mathbf{I_i} = \sum_{\mathbf{j=0, i\neq j}}^{\mathbf{n}} \frac{\mathbf{pop.density_j}}{\mathbf{distance_{ij}}^2}$$

where i and j indicates different parish and the gravity:

$$\mathbf{grav_i} = \mathbf{I_i} \times \mathbf{pop.density_i},$$

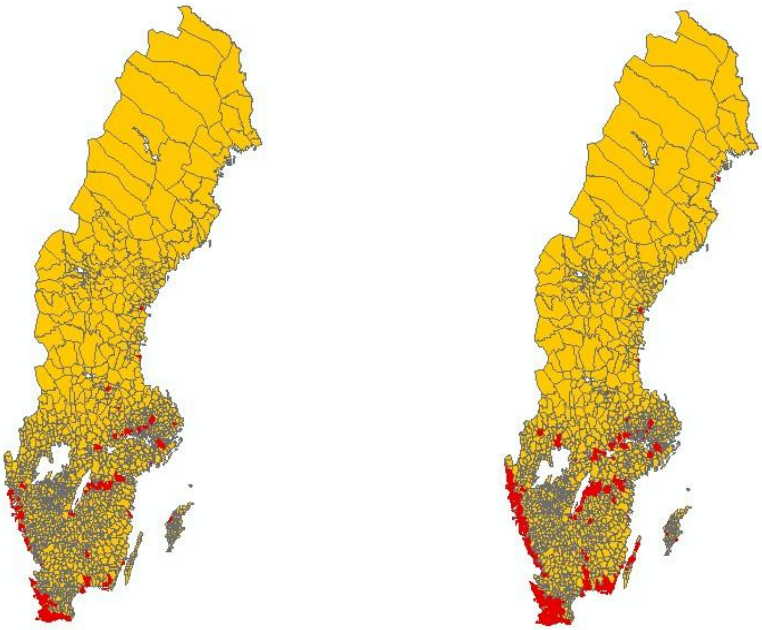
which indicates the power of influence between the nearby cities. And as it shows, the bigger distance between the two cities is, the smaller gravity is; the larger populations of the surrounding cities are, the bigger gravity is.

2.3 Analysis by using ArcGIS

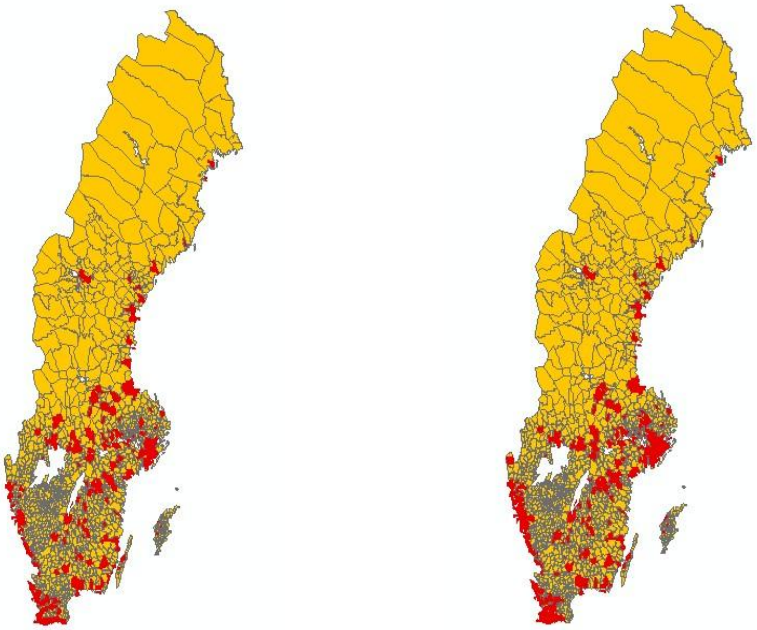
In such a long period, people learn about the world using models, such as maps and globes. In recent years, with the widespread use of computer, it has become possible to analyze these maps in computers. Combining the models with a number of tools, it makes up a geographic information system (GIS). In GIS, you can find all the information of a map, such as land, climate zones, population density, mineral resources and other things of interest to you. You can also find the location of a point feature from the map if you have the X-coordinate and Y-coordinate.

ArcGIS is a series including a group of geographic information system (GIS) software products which are produced by ESRI. This software can be used in viewing spatial data, creating layered map and doing the basic spatial analysis.

The population density varies from different area. As is shown above, we take population density as the criterion for dividing urban and countryside. We may simply define the place with population density over 30inhabitants/km^2 as Standard metropolitan statistical areas, because 75% quantile of the population density is 29.917person/km^2. After this, we can plot these cities on the map of Sweden map by using ArcGIS which is shown in the figure 3. The red areas are Standard metropolitan statistical areas where the population density is over 30person/km^2. The yellow areas are other areas where we think it is the rural part.



(a) (b)



(c) (d)

Figure 3: Standard metropolitan statistical areas for the years 1810(a), 1870(b), 1930(c), 2004(d)

Here, we choose the maps of the year 1810, 1870, 1930 and 2004. By analyzing the map, we can obtain the changes of cities. In the early 19th century, there were only a few the cities which were concentrated in the south, southwest and some middle areas. After 60 years, around the 1810's SMSA, the south, southwest and middle areas appeared more cities. Of course, the increase and

change of population related to the original cities. The big cities attracted more people to move to the nearest place. So the population distribution showed concentration at that time. In the next map, we can observe that more Standard metropolitan statistical areas appeared. However, these Standard metropolitan statistical areas did not distribute as before. They tended to disperse from each other. By comparing the map of 1930 with the map of 1860, we get that some south and southwest cities disappeared. The people moved to other place and more urban residence grew up in the middle of Sweden. In the recent 50 years, the people in the south and southwest concentrated again. And the other Standard metropolitan statistical areas did not disappear compared with early 19th century. On the other hand, the number of Standard metropolitan statistical areas in 2004 is obviously larger than that in 1810. When we consider the grow rate of the SMSA and rural areas, there are also great differences between them. From figure 4, we can see the grow rate is more stationary in the rural area, while in the SMSA, it appears increase tendency, even in some place really high.

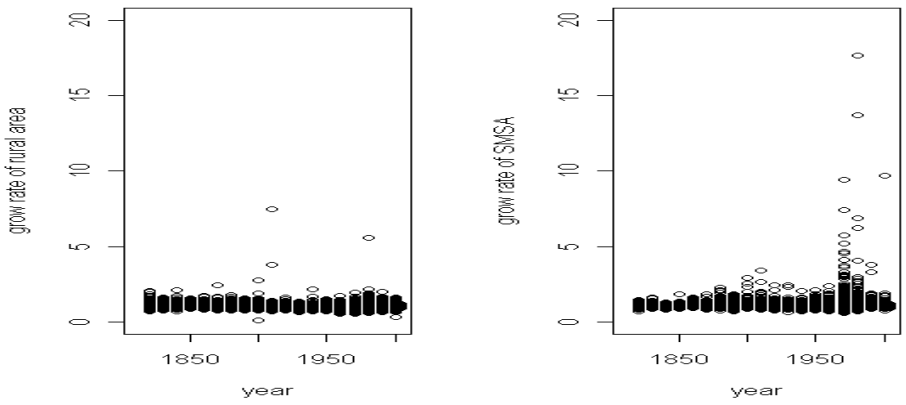
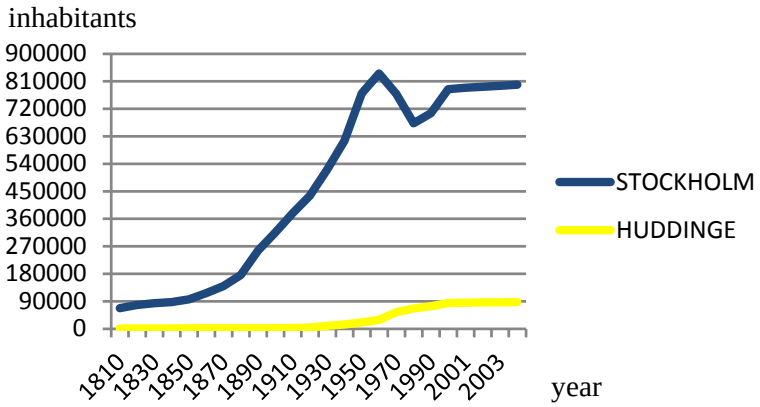
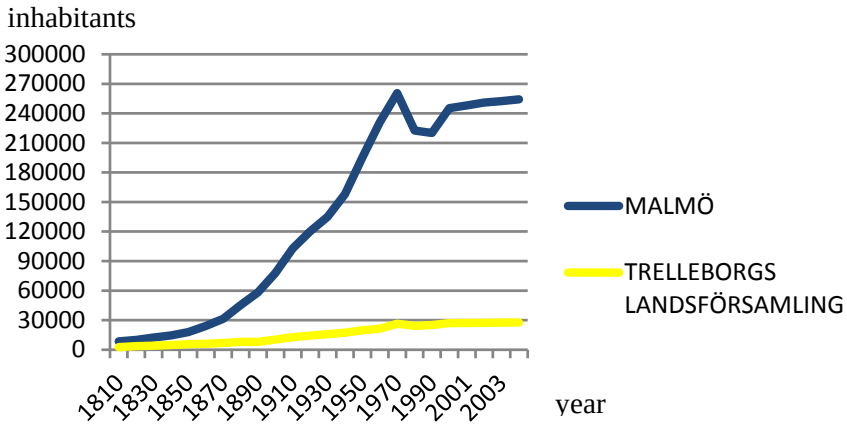


Figure 4: Comparison of the growth rate between SMSA and rural areas

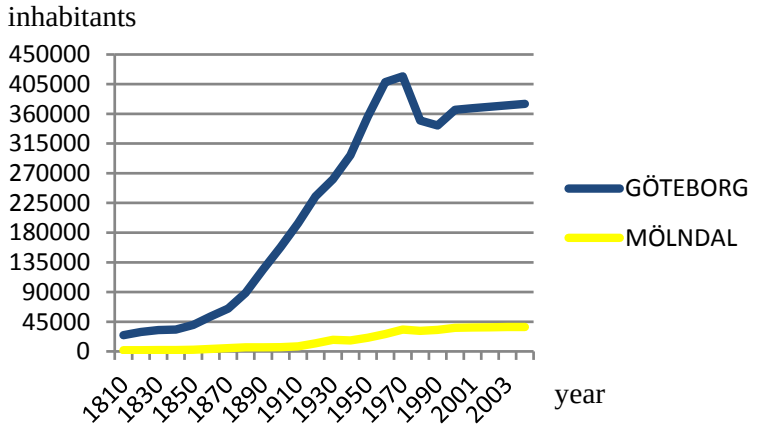
On the other hand, we may also choose three biggest cities in Sweden which are Stockholm, Malmö and Göthenborg. These cities have largest population density, so they are typically Standard metropolitan statistical areas. We also choose the nearest cites to them to make comparison.



(a)



(b)



(c)

Figure 5: Population of three metropolises and theirs nearest cities in Sweden from 1810 to 2004. Stockholm (a), Malmö (b), Göthenborg (c)

Figure 5 shows that the three metropolises have the same trend of population growth. All of their population increased from 1810 to about 1960. From 1960 to 1975, the number decreased rapidly. The reason for this

phenomenon may be suburbanization and counter urbanization. A growing number of habitations were built up at the fringe areas of metropolitans in 1960s. Concurrently, the suburban areas were also built up outside urban areas. These fringe areas and suburban areas combined with each other and attracted a lot of people to reside (Lewan, 1967). In 1970s, the large cities experienced outmigration and the countryside experienced immigration (Ahnström, 1980). After 1980, the population increase again. The small towns near the big cities also have the same tendency with other small towns. However, the small towns did not change with the metropolises at the same time. We may think that there is no significant relationship between metropolises and its nearest towns. This will be discussed in the following chapter.

3 Methodologies

3.1 Spatial econometrics

The terminology of spatial econometrics was invented by Jean Paelinck in the early 1970s. There was an increasing amount of papers on regional science dealing with estimation and testing for spatial problems in economics (Anselin, 1988).

In general, the sample data collected for regions or points in space are positively spatially dependent rather than independent from each other. In other words, observations from one location tend to exhibit values similar to those from nearby locations. Spatial regression models give us the permission to explain dependence among the observations. Usually, the observations are collected from points or regions located in space. Orientated by latitude-longitude coordinates, each observation is linked to a location which in the case of point-level samples. For region-level observations we can depend on latitude-longitude coordinates of a point located in the region, perhaps a centroid point.

Commonly, we have two kinds of data: surface data and point data. Surface data is that kind of data which shows whether one is adjacent to another or not. On the other hand, for the point data we only have the position, but we do not know whether they are connected with each other. When discussing about the surface data, it is

popular to use the following model: Firstly, using matrix, we can write the model in (1) and (2), where **y** is an n-by-1 vector containing the dependent variable observations, **W** is an n-by-n spatial weight matrix that identifies the connectivity or neighbor structure of the sample observations, **X** is the n-by-k matrix of explanatory variables which may include an intercept term. The n-by-1 vector **ε** is a normally distributed disturbances which represents zero mean, constant variance, zero covariance. (e.g. $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I_n})$,) And we use $\mathbf{I_n}$ to denote an n-by-n identity matrix. The scalar parameter **ρ** and the k-by-1 vector **β** along with the scalar variance parameter $\boldsymbol{\sigma^2}$ are model parameters to be estimated. The associated DGP for this model which we label SAR is shown in (3).

$$\mathbf{y} = \boldsymbol{\rho}\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} \quad \textbf{(1)}$$

$$\mathbf{y} = (\mathbf{I_n} - \boldsymbol{\rho}\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I_n} - \boldsymbol{\rho}\mathbf{W})^{-1}\boldsymbol{\varepsilon} \quad \textbf{(2)}$$

$$\mathbf{E}(\mathbf{y}) = (\mathbf{I_n} - \boldsymbol{\rho}\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} \quad \textbf{(3)}$$

(Fisher and Getis, 1997)

Unfortunately, in this paper we have the point data, so we do not use this popular model. Instead, we should use a kind of model that focuses on analyzing the point data. Furthermore, a random effect model (variance components model) is a sort of hierarchical linear model. This kind of model assumes that the dataset being analyzed consists of a hierarchy of different populations whose differences relate to that hierarchy. In econometrics, random effect models are widely used in the analyses of hierarchical or panel data when one assumes no fixed effects (individual effects). (Christensen, R. 2002)

3. 2 From GLM model to mixed LM model

In statistical modeling practice, we usually formulate a general linear model in the following way:

$$\boldsymbol{y_i} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_{1i}} + \boldsymbol{\beta_2 x_{2i}} + \cdots + \boldsymbol{\beta_k x_{ki}} + \boldsymbol{\varepsilon_i} \quad \textbf{(4)}$$

or in matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \textbf{(5)}$$

We also make the following assumptions.

1. ε_i are iid $N(0, \sigma^2)$
2. $x_{ji} \perp \varepsilon_i \forall i, j, k$

The first assumption tells us that all the observations are independent. However, in reality, we often observe the same individual or subject over repeated occasions. Sometimes, the observations are clustered that is we observe individuals in the same family, same community or same school, so it is very reasonable to assume that the individuals within the same cluster are correlated. In that case, we may violate assumption 1. For an estimate of the regression parameters becomes inefficient and the standard error estimate of the regression parameters are wrong, any inference based on them are incorrect.

When we know the basis of the correlation between observations, a solution to the above problem can be given in the following way: Let us assume that there are *p* clusters (subjects, communities, or schools, etc.) in the data and the observation in clusters are correlated but the observations between clusters are uncorrelated, then equation (4) can be modified as

$$\boldsymbol{y_{ji} = \beta_0 + \beta_1 x_{1ji} + \beta_2 x_{2ji} + \cdots + \beta_k x_{kji} + u_j + \varepsilon_{ji};\ i = 1, 2, \ldots, n_j; j = 1, 2, \ldots, p.} \quad \textbf{(6)}$$

when $u_j \sim iid\ N(0, \sigma_u^2)$, $\varepsilon_{ji} \sim iid\ N(0, \sigma_e^2)$, $x_{ji} \perp u_j$, $\varepsilon_{ji} \forall i, j$ and $u_j \perp \varepsilon_{ji} \forall i, j$. The above model is then called a linear mixed model. It is called mixed model because it has some fixed (non-random) effects parameters, $\beta_0, \beta_1, \ldots, \beta_k$, and some random effects $\boldsymbol{u_j}$.

We do not need to assign any mean to the random components since the mean can be absorbed in the common intercept term. In matrix notation we write equation (6) as

$$\boldsymbol{y = X\beta + Zu + \varepsilon} \quad \textbf{(7)}$$

where **X** is the design matrix associated with $\boldsymbol{\beta}$, **Z** is also a design matrix associated with the random vector **u**.

According to formulation (6), the correlation between two observations in group j can be calculated as

$$\boldsymbol{\rho_{y_{ij},y_{jk}} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}, i \neq k} \quad \textbf{(8)}$$

while $\rho_{y_{ij},y_{i'k}} = 0 \; \forall j \neq j'$. This is interpreted as the intra-class correlation. (Pawitan, 2001)

A model containing only the fixed effect term is called a fixed effects model. On the other hand, a model containing both the fixed effects and the random effect is called a mixed effects model. The random effects models are also known as multilevel models (in the context Structural Equation Models) and hierarchical models (mainly in Bayesian Statistics). In most of the cases, the repeated observations are taken over time, so the study of the above phenomena (repeated measurement) is also called logitudinal data analysis.

4 Result

4.1 Model analyses

4.1.1 Generalized linear model

The choice of a distribution from the Poisson family is often dictated by the nature of the empirical data. For example, it is commonly to use Poisson regression analysis to model count data. The data series we get is obvious count data, so we can make the hypothesis that the model follows Poisson distribution.

As discussed above, we choose the population growth rate as a dependent variable. This kind of rate data can be modeled as $log\left(\frac{\mu}{t}\right) = X\beta$, where the adjustment term *log(t)* is called an offset. In this case, the last year's population (denote as $\boldsymbol{pop_{i-1}}$) seems to be a good offset choice. As discussed above, the SMSA should be a key variable which influences the population growth rate. Cause the global urbanization makes immigration to the cities a common phenomenon. Moreover, the gravity model gives us strong support to make the gravity an independent variable in our model. Time is also an important variable

that affects both the population and the population density significantly (Figure 6).

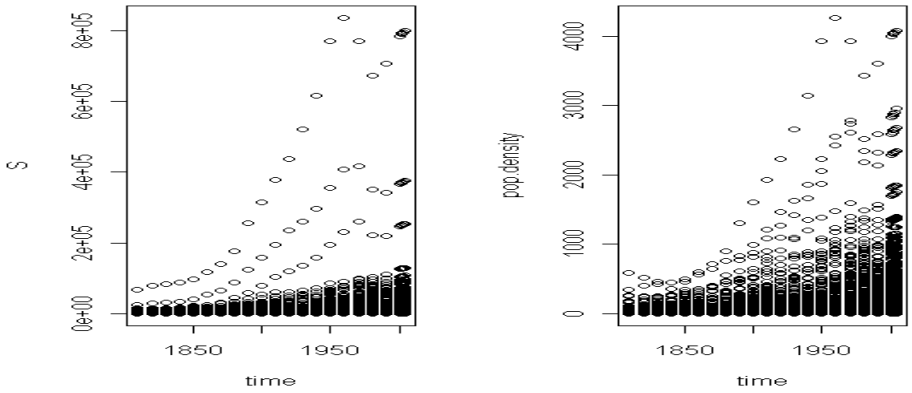


Figure 6: The population sizes and the population density (pop.density) changed with time

After having decided the four independent variables and the dependent variable, we may easily get to the GLM model function:

$$\boldsymbol{log\left(\frac{pop_i}{pop_{i-1}}\right) = \beta_1 gravity_i + \beta_2 year + \beta_3 SMSA + \varepsilon}$$

Let's have a look at the plots according to this model (Figure 7). The first panel shows residuals versus fitted values. Although the data comes from a large sample space, the trend of heteroskedasticity is evident, so we can not ignore it. The second one is a Q-Q normal distribution plot of standardized residuals. This plot shows that our hypothesis is unreasonable. So the data might follow a distribution other than Poisson distribution. Notice that we have both residuals and standardized residuals. The difference between them is that the latter one has been corrected for difference in the SD of residuals depending on their position in the design. The third plot is of the square root of the absolute value of the standardized residuals which reduces the skewness of the distribution and makes it much easier to find out if it tends to be dispersion. Of course, as we see, it shows a sign of over-dispersion. And we might consider adding random effects. The last one is residuals versus leverage and "cook's distance" which is a measure of the influence of each observation on the regression coefficients. The leverage of observations on the fitted value $\hat{\mu}_i$ is the derivative of $\hat{\mu}_i$ with respect to y_i (Dalgaard, 2002).

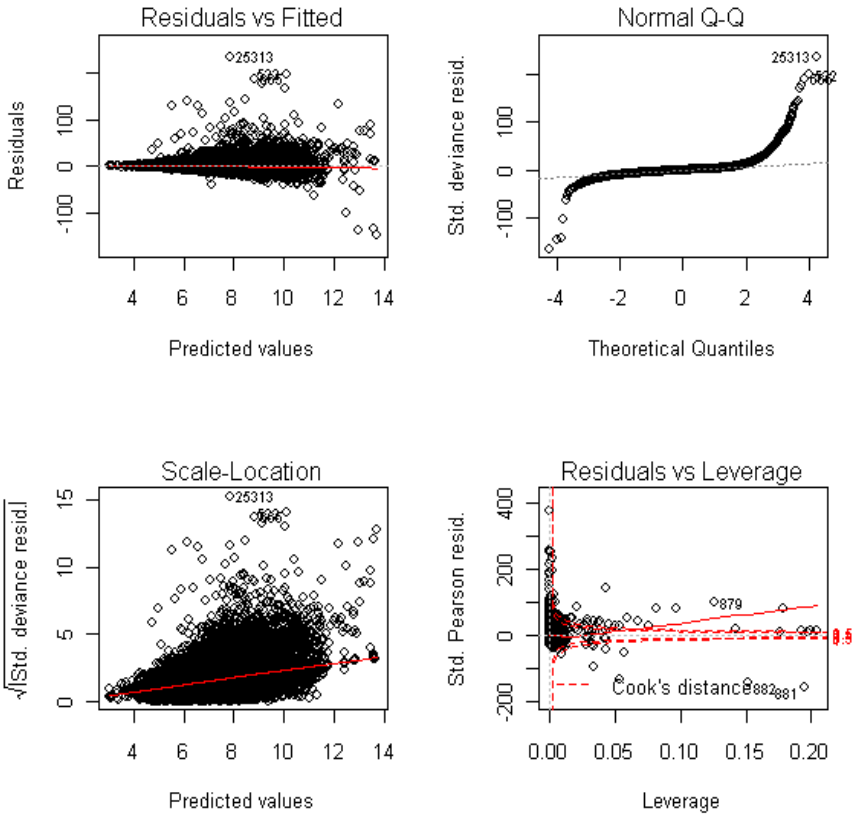


Figure 7: Plot of the GLM assuming Poisson distribution

When it comes to the deviance, as we known, if the model for the mean is correctly specified a value of the Pearson-statistic X^2, substantially larger than its expectation indicates over-dispersion. The generalized Pearson statistic is:

$\mathbf{X^2}$ **=2012796** for 42316 degrees of freedom.

Hence, $\widehat{\emptyset} = \frac{\mathbf{2012796}}{\mathbf{42316}} \approx \mathbf{47.56584}$, which clearly indicates over-dispersion.

The model we discuss above takes time as a continuous variable. We can also use it as a factor. We would get nearly the same result.

$\mathbf{X^2}$ **= 1952439** for 42295 degrees of freedom

$\widehat{\emptyset} = \frac{\mathbf{1952439}}{\mathbf{42295}} \approx \mathbf{46.16241}$. It also indicates over-dispersion as the former model.

4.1.2 Mixed linear model

When fitting very simple parametric models such as Poisson distribution, over-dispersion is often encountered. The Poisson distribution has only one free parameter and does not allow for the variance to be adjusted independently of the

mean (Olsson, 2002).

If over-dispersion is a feature of the Poisson model, we may overcome it in two ways. On the one hand, we can try Quasi-Poisson distribution. On the other hand, an alternative model with additional free parameters may provide a better fit. In the count data model, a Poisson mixed model like the negative binomial distribution which has two parameters can be used instead where the mean of the Poisson distribution can itself be thought of as a random variable drawn.

Firstly, we can try the Quasi-Poisson distribution.

$\mathbf{X^2 = 1952439}$ for 42295 degrees of freedom

$\widehat{\emptyset} = \frac{\mathbf{1952439}}{\mathbf{42295}} \approx \mathbf{46.12641.}$

So using the Quasi-Poisson distribution dose not solve the over-dispersion problem. In this case, the different parishes can be considered as a random effect, so we repeat one parish for 24 times in order to consider different parish respectively. We investigate three different models with growth rate as response, with log growth rate as response and with inverse growth rate as response:

$$\frac{\boldsymbol{pop_i}}{\boldsymbol{pop_{i-1}}} = \boldsymbol{\beta_1 gravity_i + \beta_2 year + \beta_3 SMSA + b_1 parish + \varepsilon}$$

$$\boldsymbol{log}\left(\frac{\boldsymbol{pop_i}}{\boldsymbol{pop_{i-1}}}\right) = \boldsymbol{\beta_1 gravity_i + \beta_2 year + \beta_3 SMSA + b_1 parish + \varepsilon}$$

$$\left(\frac{\boldsymbol{pop_i}}{\boldsymbol{pop_{i-1}}}\right)^{\boldsymbol{-1}} = \boldsymbol{\beta_1 gravity_i + \beta_2 year + \beta_3 SMSA + b_1 parish + \varepsilon}$$

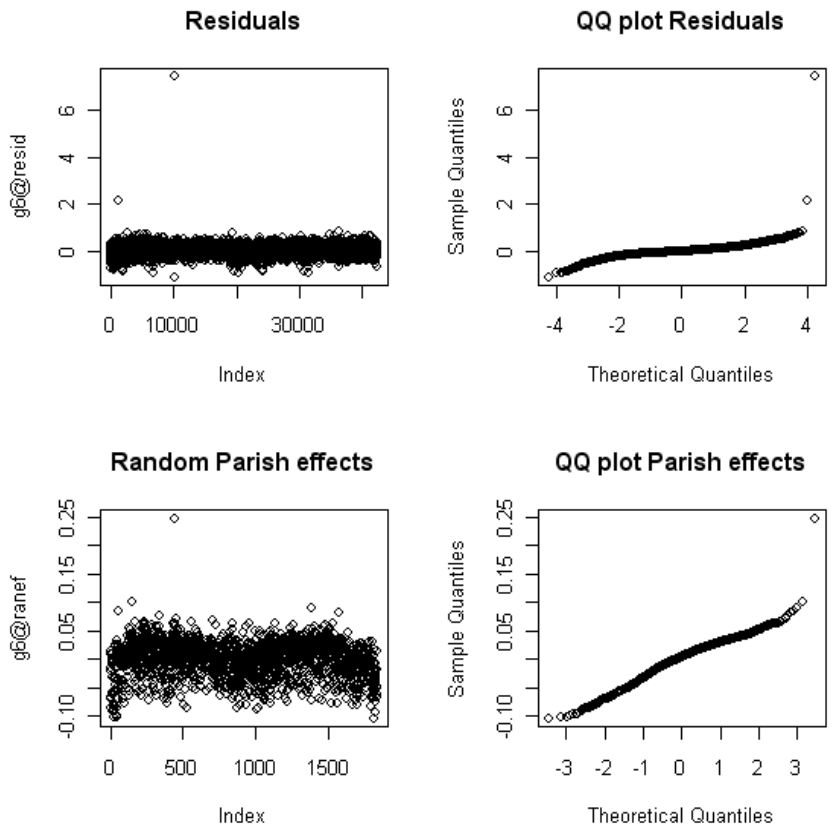


Figure 8: Diagnostics plot of a mixed linear model with inverse growth rate as response variable

We try the simple model, log transform and inverse transform one by one. Compare them with each other, then, we finally discovered that the inverse model shows the best fit of the data (Figure 8).

From our observation, there are only two outliers in the inverse model. After calculating, we can get they are the parish NÄSHULT of the year 1900 and the parish BJÖRKÖ-ARHOLMA of the year 2000. Then, we can cut them down and get a new regression in the same inverse mixed linear model. The new model shows a better fit that the QQ plot in figure 9 appears significant linear. Finally, we obtain the inverse mixed linear model.

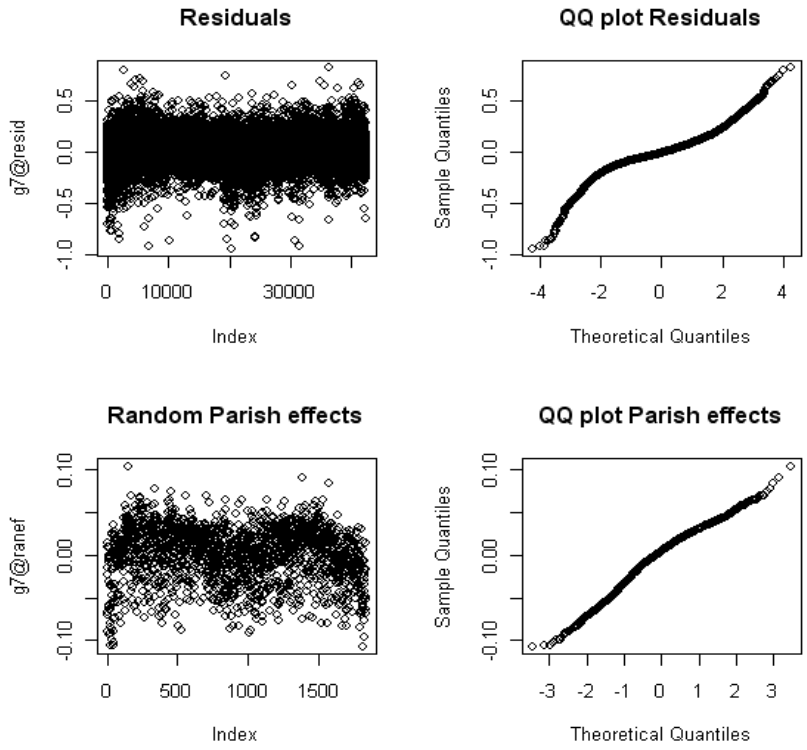


Figure 9: Diagnostics plot of a mixed linear model with inverse growth rate as response variable after cutting down two outliers

From the regression result, we get the intra-class correlation is $\boldsymbol{\rho_{y_{ij},y_{jk}}} = \frac{\mathbf{0.0014026}}{\mathbf{0.0014026 + 0.0118495}} = \mathbf{0.1058}$ which means how much the random effect (parish) can explain the model. From table 1, we can see that if the year increases 1, the inverse of the grow rate increases 5.6496×10^{-4}; if the gravity increases 1, the inverse of the grow rate decreases 7.9126×10^{-8}; and if the SMSA increases 1, the inverse of the grow rate decreases 3.0266×10^{-5}.

Table 1: The result of the inverse mixed linear model

	Estimate	Std. Error	t-value
(Intercept)	-9.536×10^{-2}	1.6836×10^{-2}	-5.67
Year	5.6496×10^{-4}	8.7196×10^{-6}	64.79
Grave	-7.9126×10^{-8}	2.8776×10^{-8}	-2.75
SMSA	-3.0266×10^{-5}	1.3776×10^{-3}	-0.02

4.2 Conclusion

From the data description section, we can see that the population growth ratio is related to the parish's population, the nearby parish's population, the time and also whether it is a city. In the model section, we firstly make a generalized linear model assuming the data follows Poisson distribution, but unfortunately it shows over-dispersion. So then we can alternatively choose a mixed linear model setting parish as a random effect. Of the three mixed linear model, the inverse model shows the best fit, so we choose that as the final model and also cut down the outliers. After obtaining the final model, we can come to the conclusion: 1) If the gravity increases, the population growth ratio will increase. This is easy to explain that the gravity increase means the population of the surrounding area increasing or the distance between the parishes is decreasing, then, the population of the focus parish is increase because of the nearby parishes' influence; 2) The later the time is, the smaller the population growth ratio is. This may because of with the time increasing, the growth is more steady, people would like to stay in a certain place in the modern life, so the migration increases not so fast as before; 3) If it is a city or to say a SMSA, the population growth ratio tends to be larger. That is to say the cities have a larger population grow rate than the rural areas for big city always attract more people to live in.

5 Further discussions

In this thesis, we have already discussed the redistribution of Sweden's population over time by statistics models. But actually both the generalized linear model and the mixed linear model do not show a perfect fit. So maybe there are still factors we do not detect or the data itself has some missing part. So in the further study, we should continue to simulate Sweden's population by using the statistics model. The researchers may try other models and other influence factors we do not discuss in this thesis in order to predict the population change from the model result in the future studies.

Reference

[1] Aldskogius M, (1970) *Population change and urban growth-A study of large and medium-sized urban regions in Sweden*, Geografiska Annaler. Series B, Human Geography, Vol. 52, No. 2, pp. 131-140

[2] Ahnström, L. (1980) *Turnaround-trenden och de nordiska huvudstadsregionernas utveckling efter 1950*, NordREFO 1980, no. 3-4.

[3] Andersson, R. (1987) *Den svenska urbaniseringen.* Kontextualisering av begrepp och processer. Uppsala universitet, Geografiska Regionstudier, no. 18, Kulturgeografiska institutionen.

[4] Anselin L. (1988) *Spatial Econometrics: Methods and Models*, departments of Geography and Econometrics, University of California, Santa Barbara ISBN 90-247-3735-4

[5] Bäcklund, D. (1999) *Befolkningen och regionerna – Ett fågelperspektiv på regional ekonomisk utveckling i Servige från 1820 och framåt*, Östersund, Swedish institute for Regional Research, Rapport 100.

[6] Borgegård, L. E., Håkansson, J. & Malmberg, G. (1995) *Population Redistribution in Sweden –Long Term Trends and Contemporary Tendencies*, Geografiska Annaler 77B, pp. 31-45.

[7] Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, Journal of the American Statistical Society 88, 9-25.

[8] Christensen, Ronald (2002) *Plane Answers to Complex Questions: The Theory of Linear Models (Third ed.)*. New York: Springer. ISBN 0-387-95361-2.

[9] Dalgaard P. (2002) *Introductiory Statistics with R*, ISBN 0-387-95475-9, 183

[10] Enequist,G. (1937) *Nedre Luledalens byar,* Uppsala, Geographica, no. 4.

[11] Faraway J.J. (2002) *Practical Regression and Anova using R,*

[12] Fisher M.M. and Getis A. (eds.) (1997) *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, DOI 10.1007/978-3-642-03647-7_18

[13] Godlund, S. (1964) *Den Svenska urbaniseringen,* Medelande från Göteborgs universitets geografiska institution, no. 76.

[14] Håkansson, J. (2000) *Spatial Population Redistribution in Sweden 1810-1990*, Department of Social and Economic Geography, Umeå University.

[15] Hoppe, G. (1945) *Vägarna inom Norrbottens län*. Studier av den trafikgeografiska utvecklingen från 1500-talet till våra dagar, Uppsala, Geographica no. 16.

[16] Jakobsson, A. (1969) *Omflyttningen i Sverige 1950-1960. Komparativa studier av migrationsfält,* flyttningsavstånd och mobilitet, Meddelande Från Lunds Universitet Geografiska Institutionen, Avhandlingar LIX.

[17] Lee, Y. and Nelder, J. A. and Pawitan, Y. (2007), Generalize Linear Models with Random Effects: United Analysis Via H-likelihood, Chapman and Hall/CRC.
models. Biometrika 73, 13.22.

[18] Lewan, N. (1967) *Landsbebyggelse i förvandling.* Meddelande från Lunds universitets geografiska institution.

[19] McCulloch, C. E. and Searle, S. (2002), Generalized Linear and Mixed Models,

Wiley:NY.
[20] Misa, T. J. and Schot, J. (2005) *Inventing Europe: technology and the hidden integration of Europe*. Hist. Technol.,21, 1–19.
[19] Olsson U., *Generalized Linear Models, An Applied Approach*, (2002), ISBN 91-44-04155-1, 61-62
[21] Orsby T, Napoleon E, and Burke R. *Getting to Know ArcGIS Desktop*. Esri Press; 2nd,Updated edition.
[22] Pawitan, W. (2001), In All Likelihood, Claredon Press, Oxford.
[23] Rodrigue, J.P., Comtois, C., Slack, B. (2009) *The Geography of Transport Systems*. London, New York: Routledge. ISBN 0-415-88324-7.
[24] Rudberg, S. (1957) *Ödemarkerna och den perifera bebyggelsen i inre Nordsverige Uppsala*, Geographica, no. 33.
[25] Statistics Sweden. *Yearbook of housing and building Statistics 2007*. Statistics Sweden, Energy, Rents and Real Estate Statistics Unit, 2007. ISBN 978-91-618-1361-2

Appendix

1. Data sample

Forskod	Forskod 2000	Y-koordinat	X-koordinat	area(km)	1810	1820	1830	…
11402	HAMMARBY	1620	6604	21	420	425	444	…
11403	FRESTA	1622	6601	23	404	356	391	…
11501	VALLENTUNA	1628	6603	64	1074	990	1081	…
11502	MARKIM	1627	6611	25	376	399	406	…
11503	ORKESTA	1630	6611	22	358	326	345	…
11504	FRÖSUNDA	1633	6613	42	629	550	620	…
11505	KÅRSTA	1637	6617	57	598	606	686	…
11506	VADA	1635	6609	20	263	261	261	…
11507	ÖSSEBY-GARN	1638	6607	110	1364	1017	1008	…
11508	ANGARN	1633	6604	18	230	275	237	…
…	…	…	…	…	…	…	…	…

2. Models comparison

model	R code
1	g1<-glm(pop_t ~ grave + year + SMSA_t, family=poisson(link=log), offset=log(pop_t_1), data=new_1)
2	g2<-glm(pop_t ~ grave + factor(year) + SMSA_t, family=poisson(link=log), offset=log(pop_t_1), data=new_1)
3	g3<-glm(pop_t ~ grave + factor(year) + SMSA_t, family=quasipoisson(link=log), offset=log(pop_t_1), data=new_1)
4	g4<-lmer(pop_t/pop_t_1 ~ grave + year + SMSA_t + (1|parish), data=new_1)
5	g5<-lmer(log(pop_t/pop_t_1) ~ grave + year + SMSA_t + (1|parish), data=new_1)
6	g6<-lmer(1/(pop_t/pop_t_1) ~ grave + year + SMSA_t + (1|parish), data=new_1)