UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Modelling infectious diseases in a small town using historical official data from 1800 - 1900.

Joyceline Afumbom Songbi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Roel BRAEKERS

2020
2021

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Modelling infectious diseases in a small town using historical official data from 1800 - 1900.

Joyceline Afumbom Songbi

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Roel BRAEKERS

Hasselt University

2nd Year Master of Statistics

Master Thesis

Modelling Infectious Disease

Students:

Joyceline Afumbom Songbi (1850535)

Instructor:

Prof. Roel Braekers

August 24, 2021

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Contents

1 INTRODUCTION **1**

1.1 Background of the Study 1

1.2 Research Objectives 2

2 METHODOLOGY **3**

2.1 Description of the data 3

2.2 Imputation by Predictive Mean Matching (PMM) 4

2.3 Modelling Count Data 5

2.4 Time Series Regression 5

2.4.1 The Poisson Regression Model 6

2.4.2 Assessing goodness of fit 8

2.5 Negative Binomial Regression Model 8

2.6 Logistic Regression Model 8

2.6.1 Assessing goodness of fit 9

2.7 Statistical Software 9

3 RESULTS **10**

3.1 Exploratory Data Analysis 10

3.2 Model Fitting 15

3.2.1 Poisson regression model 15

3.2.2 Negative binomial model 15

3.2.3 Logistic Regression 16

4 DISCUSSION & CONCLUSION **18**

4.1 Discussion 18

4.2 Conclusion 19

List of Figures

1 Linkhout count of death over the years 10
2 Lummen count of death over the years 10
3 Lummen count of death over the years 10
4 Birth Count Per Year-Linkhout 11
5 Birth Count Per Year-Lummen 11
6 Birth Count Per Year- Meldert 11
7 Death Count Per Age group-Linkhout 12
8 Death Count Per Age group-Lummen 12
9 Death Count Per Age group-Meldert 12
10 Linkhout Mortality 13
11 Linkhout Mortality with smoothing spline 13
12 Lummen Mortality 14
13 Lummen Mortality with smoothing spline 14
14 Meldert Mortality 14
15 Meldert Mortality with smoothing spline 14
16 Linkhout Population over the years 21
17 Lummen Population over the years 21
18 Lummen Population over the years 21

List of Tables

1 Dimensions of the data 3
2 Summary statistics of final data sets 4
3 Illustrating top three years with most death per age group 13
4 Table illustrating top ten years in terms of Mortality 15
5 Parameter Estimates for negative binomial model on Linkhout Population 15
6 Parameter Estimates for negative binomial model on Lummen Population 16
7 Parameter Estimates for negative binomial model on Meldert Population 16
8 Parameter Estimates for Logistic regression model on Linkhout Population 17
9 Parameter Estimates for Logistic regression model on Lummen Population 17
10 Parameter Estimates for Logistic regression model on Meldert Population 17

Acknowledgement

First and foremost, praises and thanks to God, Almighty, for His Love and Wisdom throughout this research work in order to complete this research successfully.

Not to mention the Department of Statistics at the University of Hasselt especially my lecturers as well as the coordinators for the thesis for the information they instilled in me during the preceding two years.

I would like to express my deep and sincere gratitude to my research supervisor, Professor Roel Braekers for giving me the opportunity to do the research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under his guidance. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, and great sense of humor.

I am extremely grateful to my lovely parents; Mr Songbi Edward Nsom and Mrs Ndongeh Magdalene Fien for their love, prayers, and sacrifices by educating and preparing me for my future. I am very grateful to my dear sisters; Songbi Linda, Songbi Odette, Songbi Lizette, Songbi Cyria as well as my entire family for their love, understanding, prayers and support towards completing this research.

I would like to say a big thank you to my friends Gerard, Joe, Mr/Mrs Forcha Beloh, Farai, Esthelyne, Dam, Sonita, Christine, Lincoln, Connie, Lovet, Alvine, Chioma,Christopher and Emmanuella for their constant encouragement.

I also want to thank, Faith Worship Centre, Pastor George and family, Pastor Seth and family for their prayers. I could not have made it this far without all your prayers and encouragements.

Abstract

Background: This study focuses within the period 1800 – 1900, when the sanitation and hygiene conditions in Belgium, together with a limited health care system were nowhere near the conditions as nowadays. This period, had several epidemics (mostly Cholera, Typhoid or smallpox) which took place throughout Belgium and made locally a lot of casualties, based on the birth and death dates of the inhabitants of a small rural town in Limburg (Lummen and its sub-municipalities). Therefore, identifying when these when these epidemics took place and whether they were as severe as indicated in the literature on other places in Belgium is of great importance in epidemiological studies.

Objectives: The aim of this project is to identify when epidemics took place and whether they were as severe as indicated in the literature. To determine for each year the number of people who died compared to previous year. Finally, to analyse certain periods in the study where there was excess death based on the age group.

Methodology: A time series linear regression analysis with smoothing splines is used to depict yearly trends in the mortality from the population. Analyses was done to verify which year had the highest mortality between 1800 to 1900. Furthermore, the top three years with the most mortality per age group was analysed. Negative binomial model as well as logistic regression model which are both generalized linear model each year..

Results: The results obtained showed differences in the variability of mortality between municipalities and between different years. Yearly trends were observed in the mortality rate over time with increasing mortality over periods of 10 or 15 years. Also, a handful number of individuals did not live for more than a year. Categorising the observations into four groups based on how long the lived showed that most individuals lived for 15 to 45 years in the various populations.

Conclusion: Overall, the 3 sub-municipalities had similar variability in the death counts, birth counts and mortality between the period of 1800 to 1900. These rates of mortality from the population falls in line with what is reported in history. However, note should be taken with regards to other years within this period which showed quite a high mortality which was not reported in the literature.

Key Words: Keywords: Lummen, Time series regression, Poisson, PMM, Negative Binomial, Deviance.

1 INTRODUCTION

1.1 Background of the Study

The municipality of Lummen is located in the Belgian province of Limburg near Hasselt. On the first of January 2006, Lummen had a total population of 13,691. The total land area is 53.38 km^2 which gives a population density of 256 inhabitants per km^2. Before 1 January 1977, Lummen municipality was divided into 3 sub-municipalities, namely Lummen, Meldert and Linkhout. Today, it is called Lummen. Life expectancy in this region has more than doubled in the last 200 years. According to [Devos, 2020] in Belgium, it rose from 37 years in 1830 to 47 years in 1900, 65 years in 1950 and 81.5 years presently.

The period under study was characterized by industrial revolution, Belgian Revolution of 1830, Belgian independence, and also epidemics. This means, the sanitation and hygienic conditions in Belgium, together with a limited health care system were not as good as today.

In the eighteenth century, the central government became more involved in public health and began to take more preventive measures. Previously, local authorities were in charge of public hygiene. Every city and village took a different approach to this. The most widely used interventions were isolation and quarantine of (potentially) sick patients to prevent disease spread.

The years after the wars were very challenging. The international economic crisis affected the country negatively. This led to the outbreak of epidemics such as cholera, typhoid, smallpox just to name a few. The outbreak of smallpox became the leading cause of death, particularly in children. Epidemics at the time were usually caused by outbreaks of dysentery, also known as "the bloody flux" due to the disease's bloody diarrhoea. There has been no conclusive explanation for the plague's absence, but better quarantine measures, stronger human immunity, declining bacterium virulence, or the disappearance of the black rat, which transmitted the disease (probably) through fleas, have all been suggested as potential reasons. The cause of the decrease in smallpox related deaths, on the other hand, is undeniable.

The invention of a cowpox vaccine by English country doctor in late eighteenth century, the first vaccine ever, meant that the disease could be controlled in the majority of European countries, including Belgium. Smallpox has also been eradicated globally.

During the nineteenth century, tuberculosis was the leading cause of death, but epidemics were also caused by other diseases. Belgium was ravaged by cholera seven times; in 1832-1833, 1848-1849, 1854, 1859, 1866, 1883-1885 and 1891-1895. With around 43,400 victims, the epidemic of 1866 was the most severe; it hit young active population the hardest. According to [Devos, 2020], typhoid, smallpox and

influenza also caused mortality crises in 1846-1847, 1871 and 1918-1919 respectively. Several mortality peaks, particularly during that period showed a correlation with wartime. Meaning that when war, high prices, and epidemics all coincidentally happened at the same time, the most severe mortality crises occurred. That was the case in 1690, when there was an outbreak of infectious diseases such as typhus and dysentery in Western Europe as a result of a failed harvest, resulting in devastating death rates [Devos, 2020].

A study on the bio-demography of human ageing by [Vaupel, 2010] concluded that, age is among the strongest risk factors for mortality, even in the absence of a pandemic, with an exponential increase in death risk with age. A systematic review and meta-analysis from [Galbadage et al., 2020] showed that men are more likely to die from COVID-19 than women. Furthermore, studies carried out by [Dana et al., 2020, Lodi et al., 2020], to explain the cause of high mortality in males compared to females reveals that socio-behavioral and cultural aspects are some of the factors explaining higher mortality in men compared to women.

In recent years, there has been an increasing amount of literature on modeling clinical and epidemiological data. However, statistical methods have been developed to deal with such issues. These methods are extensively discussed in the following chapters.

1.2 Research Objectives

This work is aimed to answer the following objectives

- Identify when epidemics took place and whether they were as severe as indicated in the literature
- Determine for each year the number of people who died compared to previous year
- Is there a certain period in the study where there was excess death based on the age group

2 METHODOLOGY

2.1 Description of the data

The data used in this research is an old demographic data collected by the municipalities at the time between 1750 and 1900. Then, Lummen was divided into three communities namely; Lummen proper, Meldert and Linkhout. For each community, 5 data-sets were recorded. These data-sets had overlapping information, which made it possible for them to be merged. As a result, 3 data-sets were produced, one for each municipality. Of more importance were the date variables since the study is focused on how long people about that time lived. However, for some subjects, these were not recorded. Table 1 shows the dimensions of the data and the proportion of missing observations.

Municipality	N	Variables	Missing Obs
Linkhout	4272	23	60.4%
Lummen	17709	23	59.2%
Meldert	5061	23	61.4%

Table 1: Dimensions of the data

As can be seen in Table 1, Linkhout had 60.4%, Meldert had 61.4% and Lummen had 59.2% missing observations. Before applying modelling techniques, other methods had to be devised in order to minimize the amount of missing observations.

For subjects who had no date of death recorded, but had children that were part of the study, the children's birth dates were taken as the parent's "last_known" value. If the individual had multiple children, then the date of birth of the youngest child was considered as the "last_known" date for the parent. Some individuals might not have had children, but were married. If their dates of death were not recorded, but their dates of marriage recorded, then date of marriage was taken as a "last_known" value.

Some individual's dates of death were earlier than their dates of births and so these observations were flagged as errors in data capturing and were removed.
Some observations had neither birth dates nor death dates and so were not considered for this analysis. For those having either of this dates, predictive mean matching which is a multiple imputation approach was used to impute these missing values.

The "last_known" column, was necessary in order to determine how long an individual lived and this in turn would allow us know when they were present in the population in a given year. The total years lived for an individual was taken to be the difference between the "last_known" date and the date of

birth of that individual.

Finally, there were some subjects who married or were married more than once and recorded for each of the occasions. This resulted in duplicates and hence these observations were removed from the analysis.

A few assumptions were made regarding the data. Firstly, we assumed that individuals born in a community stayed there until the end of the study period. Secondly, for those born towards 1900 were still alive at 1900. Thirdly, the study considered only individuals who lived and died between 1800 to 1900. Finally, individuals born before 1800 but died between the study period were not considered in the population.

Municipality	**N**	**Median**	**Mean**	**Std Dev**	**Minimum**	**Maximum**
Linkhout	1571	48	36.74	23.45	0	98
Lummen	7655	44	35.11	21.81	0	92
Meldert	2293	37	31.06	20.15	0	92

Table 2: Summary statistics of final data sets

2 shows a summary statistics of the final data used for this analysis. Comparing Table 1 and Table 2, we can see that more than half of the data could not be used for analysis after fixing the data.

Separate analysis was performed for three communities in order to take in to account their population size since the higher the population, the higher the death numbers.

2.2 Imputation by Predictive Mean Matching (PMM)

In this analysis, the method implemented to cater for missing observations was Imputation by Predictive Mean Matching. This was applied for those individuals with missing dates of births or deaths only. This technique has been around for a long time [Little, 1988], but only recently has it become widely available and practical to use. PMM is a convenient way to perform multiple imputation for missing data especially for normally distributed quantitative variables.

Generally, imputed values will be discrete if the real values were discrete, when the original variable is skewed, the imputed values will also be skewed. The reason for this as proposed by [Allison, 2015], is that the imputed values are real values which are "borrowed" from the individual's original data. PMM

assumes that the data are missing at random.

A review about PMM for imputation by [Rubin, 2004] revealed that the main disadvantage of the technique is that there is no mathematical theory to justify it as a result we have to rely on Monte Carlo simulations. However, no simulation can study all the possibilities.

[Little, 1988] suggested that, PMM does almost as well as parametric methods for a correctly specified model, and a little better than parametric methods in certain mis-specified models. So the current consensus seems to be that this is an acceptable and potentially useful method.

2.3 Modelling Count Data

2.4 Time Series Regression

Time series regression is widely used in environmental epidemiology. It is useful for extracting meaningful statistics and other characteristics of the data. This time series methodology has been largely used in measuring short term associations between exposures such as air pollution, weather variables or pollen and health outcomes such as mortality [Bhaskaran et al., 2013].

While it is a regression method, it predicts the dependent variable based on the outcome values at prior points in time rather than independent factors [Yang and Berdine, 2015]. Therefore, using this methodology, the relationship between the mortality per year is analysed and patterns can be found over the time period of 1800 to 1900.

A time series is usually a sequence of data points recorded at regular intervals. This implies, a time series was observed which illustrates the counts of death per year. However, the counts of deaths per year changes and the underlying population per year is also of interest. It becomes meaningful to use the mortality per year.

In order to capture the trends or periodic behaviour over time with respect to the count of deaths of the various municipalities, a time series linear regression analysis with smoothing splines was used. By using smoothing splines, patterns are captured which depicts possible trends in the counts of death per year in every municipality while taking into account the population present.

To fit a linear regression in the time series context, a dependent time series x_t for $t = 1, ...n$, is being influenced by an independent series $z_{t1}, z_{t2}, z_{t3}, ...z_{tq}$ where the inputs are assumed fixed and known. This relation can be expressed using the model;

$$x_t = \beta_0 + \beta_2 z_{t2} + ... + \beta_q z_{tq} + w_t$$

In order to smooth the data, a polynomial regression in terms of time is fitted. A cubic polynomial of time would have $x_t = m_t + w_t$ where

$$m_t = \beta_0 + \beta_1 t + \beta_1 t^2 + \beta_3 t^3$$

Furthermore, in order to smooth the data, a method of smoothing spline was used which minimizes a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^{n} [x_t - m_t]^2 + \lambda \int (m_t^{''})^2 dt$$

where m_t is a cubic spline with a knot at each t and primes denote differentiation. The degree of smoothness is controlled by the parameter λ. From the equation above, if $\lambda = 0$, this leads to $m_t = x_t$ which are not smooth. If $\lambda = \infty$, the curve becomes constantly smooth. Therefore, larger λ values leads to smoother fits. This implies λ controls the smoothness of the linear regression (from completely smooth) and the data itself(no smoothness). In order to fit smooth the λ from the smooth.spline package in R is set to 0.5

2.4.1 The Poisson Regression Model

The Poisson Distribution

The Poisson distribution is a discrete distribution which describes the number of events occurring in a fixed time interval. The distribution is bounded by zero and infinity and has only one parameter μ which is equal to the mean and variance. The distribution is given by the formula:

Assumptions

To employ a Poisson regression, like with many other regression models, numerous assumptions must be made.

- The occurrence of an event does not affect the occurrence of a subsequent event (independent events).
- The probability of an event to occur in a certain time interval should be the same for every other time interval of that same length (the rate at which events occur is constant).
- Rate changes resulting from the combined impact of many explanatory factors are multiplicative.
- At each level of the independent variable, the number of deaths has variance equal to the mean
- Independent error terms

The probability for each individual being examined is assumed to be the same, so a parameter λ is defined to be the average number of deaths during each time increase. Considering N is the population of that year, then:

$$\lambda = Np$$

where p is the probability that a given individual dies. By rearranging the formula;

$$p = N\lambda$$

The probability that exactly K deaths will be observed during a time interval follows the binomial distribution expressed as:

$$P(k) = p^k(1-p)^{N-K} = (\frac{\lambda}{N})^K(1-\frac{\lambda}{N})^{N-K}$$

As the population N increases, the limit of this expression is the Poisson distribution:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

A unique characteristic of a Poisson distribution is that the mean (average expected value) and the variance of a Poisson distribution are both equal to λ. However, count data are all non-negative integers hence the mean value of the counts is always greater than zero.

In addition, the distribution of count data is skewed to the right, and the variance of count data tends to rise as the number of counts increases.

Poisson regression models are a type of generalized linear model in which the systematic effects are multiplicative, the error distribution is Poisson, and the link function is the natural log [Yang and Berdine, 2015]. The log link ensures that all the fitted values are positive.

It is useful for modeling a count variable Y, by counting the number of times that a certain event occurs during a given time period. In this analysis, the dependent or outcome variable (Y) is the number of deaths obtained in each each during the study period, described by a set of explanatory variables $X_1, X_2, ..., X_k$. Therefore, time is treated as an independent covariate.

Poisson regression works by fitting a regression equation on the observed data which accurately models the expected value of the dependent variable Y, $E(Y)$ as a function λ on a set of independent variables X_1 X_2,..., X_k and β regression parameters [Kuhn et al., 1994]. Considering Y is the number of deaths in a subgroup, and N is the population size, then

$$E(Y) = N\lambda(X, \beta)$$

The general form of the likelihood function is obtained as:

$$L(Y;\beta) = \prod[N\lambda(X,\beta)]^y e^{-N\lambda(X,\beta)}/y!$$

The regression coefficients are estimated by maximizing the likelihood function. It is important to specify the function λ, which is commonly considered to be log-linear, in order to use the likelihood function.

It is described as a linear function of the X predictor variables as the natural log of the predicted rate of deaths Y.

$$\lambda(X;\beta) = exp(\beta_1 X_1, \beta_2 X_2 + ... + \beta_k X_k)$$

The exponent of a Poisson regression coefficient is a rate ratio which corresponds to a one unit difference in the independent variable variable. Considering the number of deaths every year in the study period, we may see that each person is followed for a different amount of time. In this situation, the goal is to model rates (counts per unit of time).

2.4.2 Assessing goodness of fit

The Pearson chi-squared and deviance test statistics can be used to measure the model's goodness-of-fit. The deviance is assumed to follow a chi-square distribution with degrees of freedom equal to the model residual in this case. To check for overdispersion, this test works by comparing the deviance with its degrees of freedom. The deviance is a measure of the difference between observed and fitted values, compared to its degrees of freedom. The data are considered to be overdispersed if the observed variability exceeds that anticipated by the Poisson distribution. If there is no overdispersion, the ratio will be close to 1; if it is greater than 1 and less than 1 result in over-dispersion or under-dispersion respectively. Large values of these statistics, as well as small P-values are indication that the model does not fit the observed data.

2.5 Negative Binomial Regression Model

This regression model is used to model over dispersed count data that is, when the variance of the data is higher than the mean. Because it has the same mean structure as Poisson regression and an extra parameter to describe over-dispersion, it can be regarded a generalization of the Poisson regression.
The confidence intervals for Negative binomial regression are likely to be bigger than those for a Poisson regression model if the conditional distribution of the outcome variable is over-dispersed.
In this study we want to model the count of deaths per year. The model has a less restrictive property in that the variance is not equal to the mean.

2.6 Logistic Regression Model

Logistic regression is a technique for modelling the probability of a discrete outcome given an input variable. It is a particular case of generalized linear models (GLM), in which the dependent variable is dichotomous. Individuals are assigned to either one of two classes). In this case, a binary regression model is considered which models the probability for an individual to dies yes or no for a given year and because the logistic model's dependent variable can only take two values (0 or 1), the probability predicted by

the model must also fall within that range. The probability approaches 0 when X (independent variable) takes on smaller values. As X increases, however, the probability approaches 1.

This model is important in that firstly, it is an extremely flexible and easily used function and also it lends itself to clinical meaningful interpretations [Hosmer et al., 2000]. If the conditional mean of Y given x is denoted as $\pi(x) = E(Y|x)$, the form of the logistic regression model used will be;

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

which gives us the estimated probability to die. An appropriate logit transformation of $\pi(x)$ will result in;

$$g(x) = ln[\frac{\pi(x)}{1-\pi(x)}] = \beta_0 + \beta_1 x$$

where, $\pi(x)$ is the probability to die. As this probability increases from 0 to 1, the logit rises from $-\infty$ to $+\infty$ with this transformation (resulting in a sigmoidal shape).

Fitting this model requires that the regression coefficients has to be estimated using maximum likelihood approach. Maximum likelihood works by finding the smallest possible deviation between the observed and predicted values. This value is called the deviance (-2 Log Likelihood). This requires constructing a likelihood function which expresses the probability of the observed data as a function of the unknown parameters. The log-likelihood is defined as;

$$L(\beta) = ln[l(\beta)] = \sum_{i=1}^{n}\{y_i ln[\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)]\}$$

Differentiating $L(\beta)$ with respect to the regression coefficients and setting the expression equal, we obtain the value for β (maximum likelihood estimate) that maximizes $L(\beta)$.

2.6.1 Assessing goodness of fit

The log-likelihoods of having observed the true outcome, given the predicted probability of that outcome, are connected to the deviance residuals.

2.7 Statistical Software

All statistical analysis was done using SAS software, version 9.4 of the SAS system and RStudio version 1.4.1717.

3 RESULTS

3.1 Exploratory Data Analysis

The exploratory data analysis was done based on the imputed data.

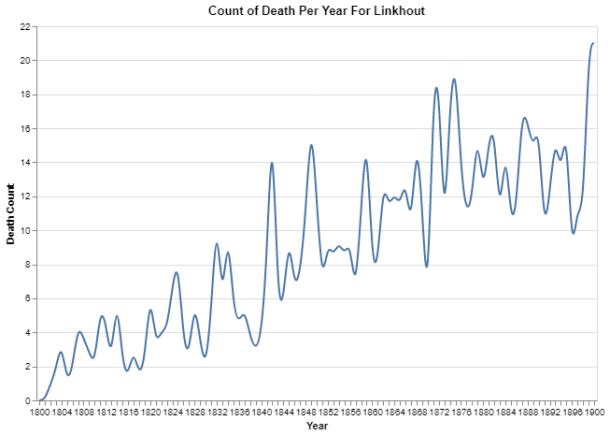

Figure 1: Linkhout count of death over the years

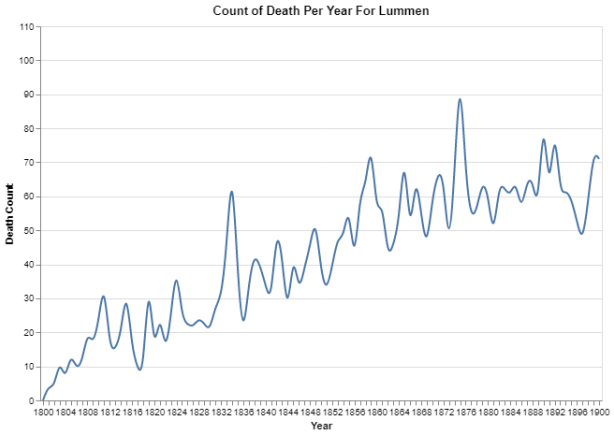

Figure 2: Lummen count of death over the years

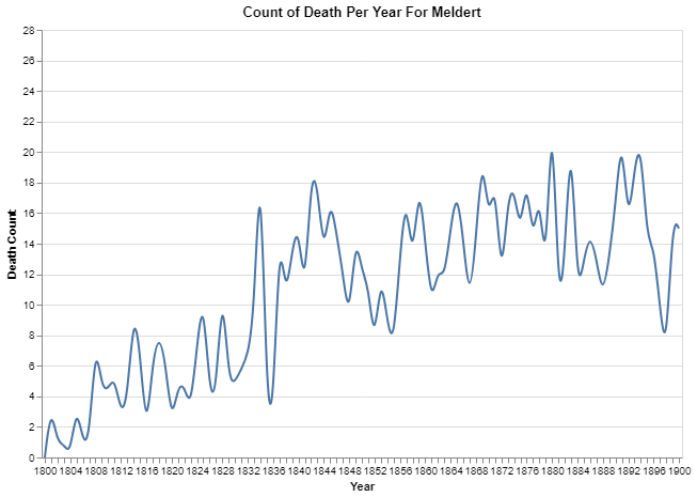

Figure 3: Lummen count of death over the years

Figure 1, shows the count of death per year for Linkhout. It reveals that the deaths increases steadily with the highest death count of 21 in the year 1898. Figure 2 represents the count of deaths in Lummen. The deaths increases with time. The highest record was 89 deaths in 1876. Figure 3 represents the count of deaths in Lummen. The deaths increases with time. The highest record was 20 deaths in 1880. All three communities experience almost similar pattern of deaths which shows an increase over time.

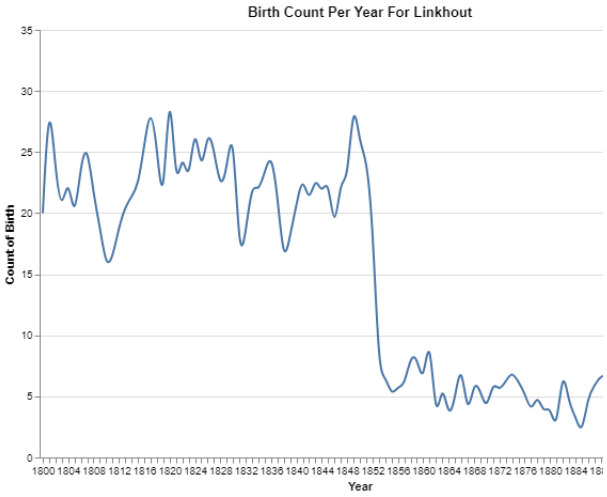

Figure 4: Birth Count Per Year-Linkhout

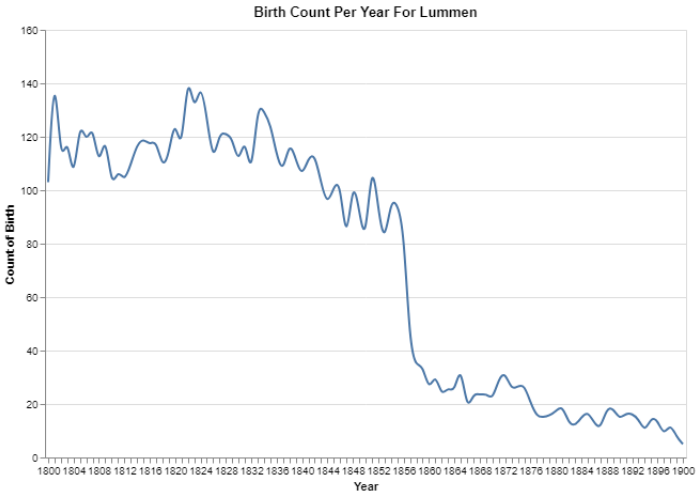

Figure 5: Birth Count Per Year-Lummen

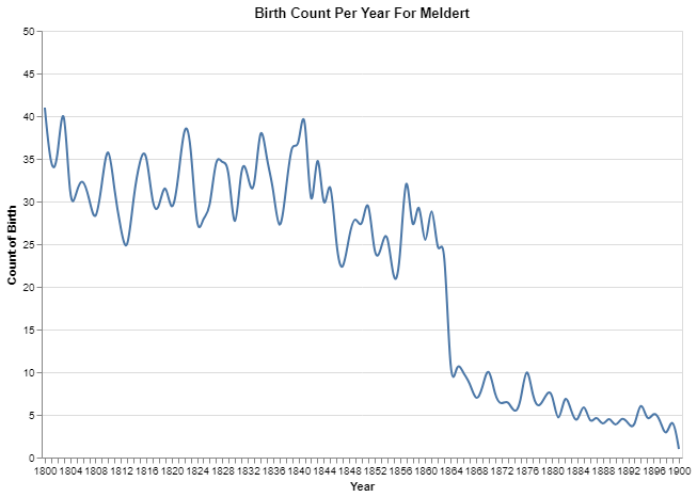

Figure 6: Birth Count Per Year- Meldert

Figures 4, 6 and 5 shows the birth count per year over the study period. However, after sometime there was a steep decrease in deaths and it remained steady. This shows that there was epidemic at during that year which concided to 1852, 1844 and 1864 for Linkhout, Lummen and Meldert respectively.

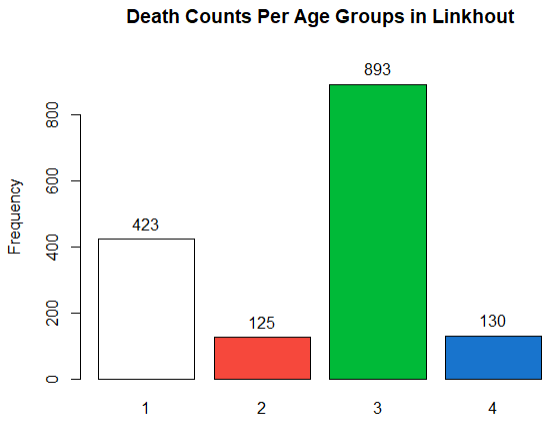

Figure 7: Death Count Per Age group-Linkhout

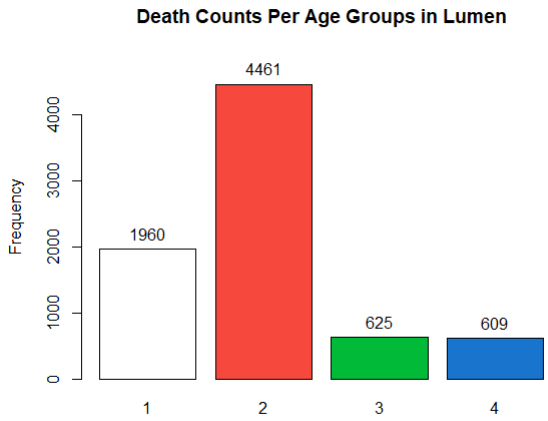

Figure 8: Death Count Per Age group-Lummen

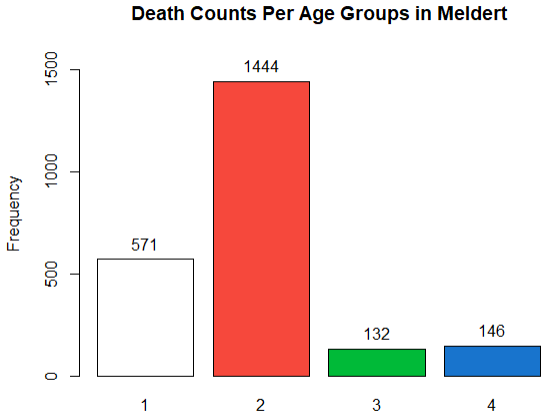

Figure 9: Death Count Per Age group-Meldert

Age groups were created based on the years lived of the individual. Another variable derived is the age group which consist of grouping the individuals based on the years lived. As a result, four groups were created as follows: under 15 years, 15 to below 45 years, 45 to below 65 years and above 65 years. Figures 7, 8 and 9 illustrates the count of individuals in the various age groups. Figure 7 individuals in group 3 with the highest death counts. However, Lummen and Meldert illustrates that individuals in group 2 have the highest counts of death.

Linkhout			Lummen			Meldert		
year	age_group	% death	year	age_group	%death	year	age_group	%death
1842	1	3.5	1834	1	3.3	1834	1	4.2
1875	1	3.1	1875	1	2.6	1842	1	2.8
1832	1	2.8	1859	1	2.5	1859	1	2.5
1900,1842,1841	2	4.8	1895,1879	2	2.2	1880	2	2.4
1834	2	4.0	1889	2	2.1	1876	2	2.3
1899	2	3.2	1886	2	2.0	1896	2	2.1
1895	3	3.1	1900	3	3.4	1893,1880,1871,1862,	3	5.3
1897	3	3.0	1878	3	3.2	1865	3	4.5
1899	3	2.9	1890	3	3.0	1891	3	2.8
1895	4	8.5	1900	4	5.6	1883	4	8.2
1899,1889,1881	4	6.9	1888	4	10.7	1892,1890	4	6.2
1898	4	6.9	1894	4	15.6	1894	4	5.5

Table 3: Illustrating top three years with most death per age group

Furthermore , table 3 illustrates the top three years with the most death counts in every age group per town.

Yearly mortality rate per 1,000 people of the population was calculated to determine the rate of occurrence of death.

$$\text{Yearly Mortality Rate} = \left(\frac{\text{Number of deaths/ year}}{\text{Population}}\right) * 1000$$

The value of population was calculated by taking a cumulative count of the number of births each year will removing the number who die in a given year so that it excluded those that had died in previous years.

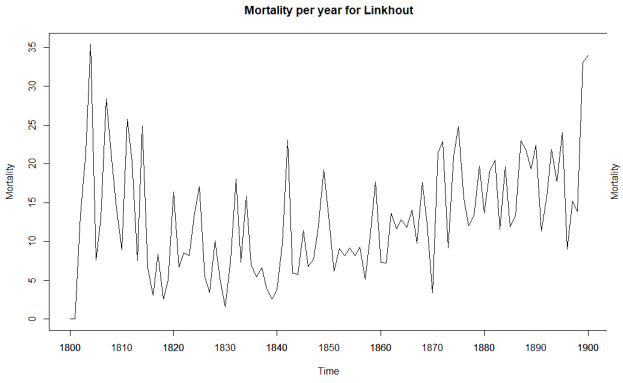

Figure 10: Linkhout Mortality

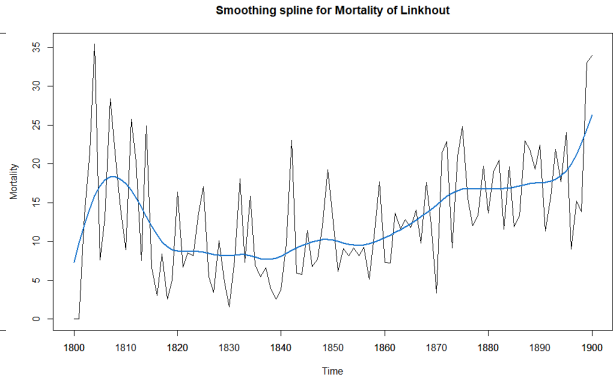

Figure 11: Linkhout Mortality with smoothing spline

Figure 10 shows the mortality rate calculated per thousand of the population of a given year. It shows how the mortality rates were changing over the years in the study in that town. There were more deaths per 1,000 in the early 1800, which decreased over time and then increased steadily.

A time series regression was fitted using smoothing splines to show trends in the mortality as shown in figure 15. For Linkhout community, the trends revealed that there was an increase in the mortality for the first ten years after which it decreased and remained constant and then increased gradually.

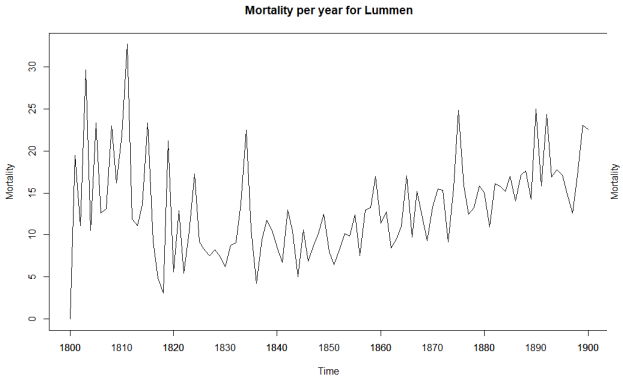

Figure 12: Lummen Mortality

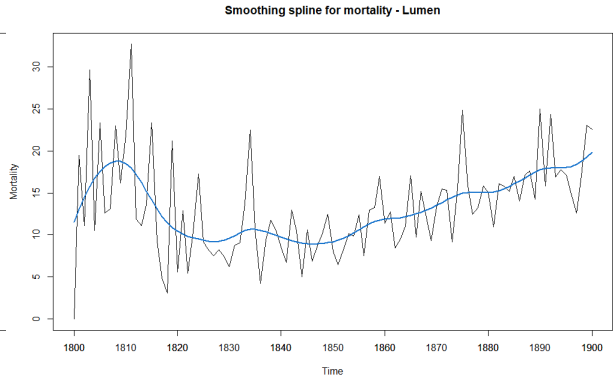

Figure 13: Lummen Mortality with smoothing spline

Lummen community had a very similar pattern of mortality just like Linkhout. There were more deaths per 1,000 in the early 1800, which decreased over time and then at around 1834 there was a sharp increase. This was again followed by a steady increase.

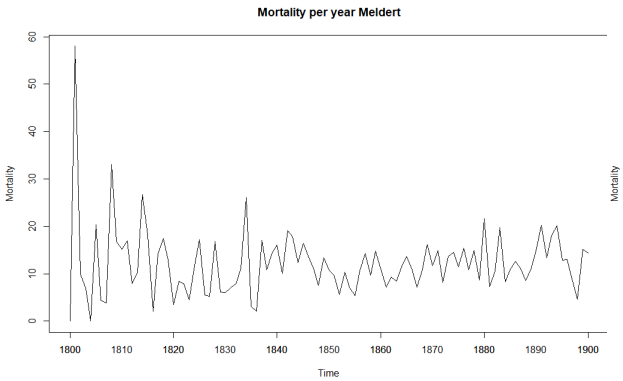

Figure 14: Meldert Mortality

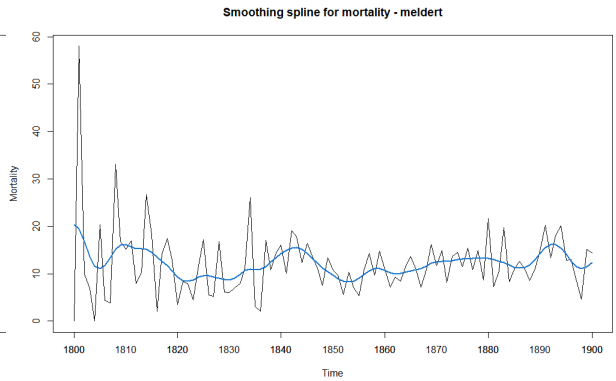

Figure 15: Meldert Mortality with smoothing spline

Meldert showed a small difference in the mortality. There were more deaths per 1,000 in the early 1800, which decreased over time. It decreased at the beginning but later shows similar patterns in every 10 years.

	Linkhout		Lummen		Meldert	
Ranking	Year	Mortality	Year	Mortality	Year	Mortality
1	1804	35.39	1811	32.67	1801	57.97
2	1900	33.98	1803	29.59	1808	33.08
3	1899	33.01	1890	25.01	1814	26.63
4	1807	28.4	1875	24.81	1834	26.00
5	1811	25.75	1892	24.36	1880	21.61
6	1814	24.91	1805	23.35	1805	20.30
7	1875	24.85	1815	23.32	1891	20.21
8	1895	24.06	1899	23.02	1894	20.07
9	1842	23.03	1808	23	1883	19.70
10	1887	22.94	1900	22.5	1842	19.07

Table 4: Table illustrating top ten years in terms of Mortality

Above is a follow up on the mortality plots for the communities. The periods with peaks showed years with high mortality.

3.2 Model Fitting

3.2.1 Poisson regression model

The most popular method for analyzing rates is Poisson regression. However, because the over-dispersion parameter was above one, a negative binomial model which models count over dispersed data was preferred.

3.2.2 Negative binomial model

The results are presented in the tables below;

Maximum Likelihood Parameter Estimates

Parameter	DF	Est	Std Error	Wald 95% CL		Chi-square	Pr >Chi Sq
Intercept	1	-28.476	3.2107	-34.769	-22.183	78.66	<.0001
Years	1	0.016	0.0017	0.0126	0.0194	86.44	<.0001
sex(f)	1	-0.0977	0.0875	-0.2692	0.0739	1.24	0.2646

Table 5: Parameter Estimates for negative binomial model on Linkhout Population

Table 5 shows the result of the Poisson regression which models the death counts per year using gender and year as the covariates. One unit change in the years, the difference in the logs of expected counts of deaths is expected to change by 0.0113, given the other predictor variables (gender) in the model are held constant.

Maximum Likelihood Parameter Estimates

Parameter	DF	Est	Std Error	Wald 95% CL		Chi-square	Pr >Chi Sq
Intercept	1	-39.232	2.126	-43.399	-35.066	340.54	<.0001
year	1	0.0226	0.0011	0.0204	0.0249	392.15	<.0001
sex(f)	1	-0.1322	0.0559	-0.2418	-0.0227	5.6	0.018

Table 6: Parameter Estimates for negative binomial model on Lummen Population

Table 6 shows the result of the binomial regression model for Lummen. The difference in the logs of expected death counts is expected to be 0.1322 unit lower for females compared to males, while holding the other variables constant in the model. Furthermore for a one unit change in the years, the difference in the logs of expected counts of deaths is expected to change by 0.0113, given the other predictor variables (gender) in the model are held constant.

Maximum Likelihood Parameter Estimates

Parameter	DF	Est	Std Error	Wald 95% CL		Chi-square	Pr >Chi Sq
Intercept	1	-19.414	2.9483	-25.193	-13.635	43.36	<.0001
Years	1	0.0113	0.0016	0.0082	0.0144	50.53	<.0001
sex(f)	1	-0.1038	0.0846	-0.2696	0.062	1.51	0.2199

Table 7: Parameter Estimates for negative binomial model on Meldert Population

Table 7 shows the result of the poisson regression model. Gender was not significant. However, results shows that for a one unit change in the years, the difference in the logs of expected counts of deaths is expected to change by 0.0113, given the other predictor variables (gender) in the model are held constant.

3.2.3 Logistic Regression

The results for the three communities are presented below;

Parameter Estimates						
Parameter	Est	Std Error	95% CL		Chi-square	Pr >Chi Sq
Intercept	-16.753	3.245476	-23.16	-10.44	-5.162	2.44e-07
year	0.0067	0.0017	0.003	0.010	3.862	0.000112

Table 8: Parameter Estimates for Logistic regression model on Linkhout Population

Table 8 shows the results of modelling the probability to d die in Linkhout per year. For every one-unit change in years, we expect a 0.006 increase in the log-odds of deaths, holding all other independent variables constant.

Parameter Estimates						
Parameter	Est	Std Error	95% CL		Chi-square	Pr >Chi Sq
Intercept	-21.523394	1.459245	-24.394	-18.674	-14.75	<2e-16
year	0.009318	0.000781	0.0077	0.0108	11.93	<2e-16

Table 9: Parameter Estimates for Logistic regression model on Lummen Population

Table 9 shows the results of modelling the probability to die in Lummen per year. For every one-unit change in years, we expect a 0.009 increase in the log-odds of deaths, holding all other independent variables constant.

Parameter Estimates						
Parameter	Est	Std Error	Wald 95% CL		Chi-square	Pr >Chi Sq
Intercept	-7.313	2.684	-12.60	-2.076	-2.725	0.00644
year	0.001	0.001	-0.001	0.004	1.148	0.25096

Table 10: Parameter Estimates for Logistic regression model on Meldert Population

Table 10 shows the results of modelling the probability to die in Meldert per year. The variable *year* was insignificant.

4 DISCUSSION & CONCLUSION

4.1 Discussion

This research considered individuals in the time frame between 1800 and 1900. Based on existing literature, Belgium was ravaged by cholera seven times; in 1832-1833, 1848-1849, 1854,1859, 1866, 1883-1885 and 1891-1895, with the epidemic of 1866 the most severe. Also, typhoid, smallpox and influenza also caused mortality crises in 1846-1847, 1871 and 1918-1919 respectively.

During the period under study, it is believed that there were years which encountered epidemics. Therefore, first objective in this study was to identify when epidemics took place and whether they were as severe as recorded in literature. Interestingly, all these epidemics fall in the the peaks of the time series curve which trends in every ten years in that period.

Another interest was to find out the mortality rate of each year compared to the previous year. Since the population was between 1800 to 1900, a very high mortality rate was observed in the beginning of the study followed be sharp decrease. These were largely due to small samples and probably poor living conditions. The mortality rate based on age groups revealed that children between 0-15 years had higher mortality compared to other age groups. This is attributable to high infant mortality.
High mortality could be supported by the coincidence of wars, high prices, happening at the same time.

Based on the nature of the data, generalized linear modelling techniques were implemented. A Poisson regression model was fitted to model the counts of deaths for each year. However, since the data was over-dispersed, a negative binomial regression model was fitted. Three different models were fitted for the three towns respectively in order to take in to account the population dynamics. Overall, year was significant to model the count of deaths in the three towns. However, gender was only significant in the modelling counts of death in Lummen.

A logistic regression model which models the probability to die each year was also fitted. Three different models were fitted for the three towns respectively to take in to account the population dynamics in the towns.
The models for Linkhout and Lummen showed that, *year* was significant to model the probability to die in the towns. However, this was not the case for Meldert. This could be supported by the time series plots wherein Meldert did not show much difference in their deaths.

4.2 Conclusion

The study was focused on studying the population of individuals in 3 municipalities, which have since been merged into 1 large municipality called Lummen in a particular period of time. Because these data were recorded more than 100 years ago, a lot of errors and incomplete records were realized. This called for extensive data cleaning and the use of survival techniques to deal with missing observations. Imputation based on predictive mean matching was implemented. Based on the data, time series regression which uses smoothing splines, captured possible trends in the deaths per year in every municipality while taking into account the population present at the time. The study identified epidemics in the time period but because of limited population size conclusion can not be made about how severe they were. The study concluded that the rates of mortality from the population falls in line with what is reported in history. However, note should be taken with regards to other years within this period which showed quite a high mortality which was not reported in the literature.

The limitation of the study include; firstly, a lot of missing data were observed for almost all the covariates leading to loss of information, and secondly the age of the participants were not recorded and lastly there was need for other meaningful variables to predict a person's probability to die.

References

[Allison, 2015] Allison, P. (2015). Imputation by predictive mean matching: Promise & peril. *Statistical Horizons*.

[Bhaskaran et al., 2013] Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., and Armstrong, B. (2013). Time series regression studies in environmental epidemiology. *International journal of epidemiology*, 42(4):1187–1195.

[Dana et al., 2020] Dana, P. M., Sadoughi, F., Hallajzadeh, J., Asemi, Z., Mansournia, M. A., Yousefi, B., and Momen-Heravi, M. (2020). An insight into the sex differences in covid-19 patients: what are the possible causes? *Prehospital and disaster medicine*, 35(4):438–441.

[Devos, 2020] Devos, I. (2020). From plague to coronavirus. a very short history of epidemics in flanders and belgium.

[Galbadage et al., 2020] Galbadage, T., Peterson, B. M., Awada, J., Buck, A., Ramirez, D., Wilson, J., and Gunasekera, R. S. (2020). Systematic review and meta-analysis of sex-specific covid-19 clinical outcomes. *Frontiers in medicine*, 7:348.

[Hosmer et al., 2000] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2000). *Applied logistic regression*. Wiley New York.

[Kuhn et al., 1994] Kuhn, L., Davidson, L. L., and Durkin, M. S. (1994). Use of poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American journal of epidemiology*, 140(10):943–955.

[Little, 1988] Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

[Lodi et al., 2020] Lodi, E., Scavone, A., Reggianini, L., and Modena, M. G. (2020). Covid-19: a gendered disease? possible interpretations and knowledge limitations. *Giornale italiano di cardiologia (2006)*, 21(8):570–574.

[Rubin, 2004] Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

[Vaupel, 2010] Vaupel, J. W. (2010). Biodemography of human ageing. *Nature*, 464(7288):536–542.

[Yang and Berdine, 2015] Yang, S. and Berdine, G. (2015). Poisson regression. *The Southwest Respiratory and Critical Care Chronicles*, 3(9):61–64.

Appendix

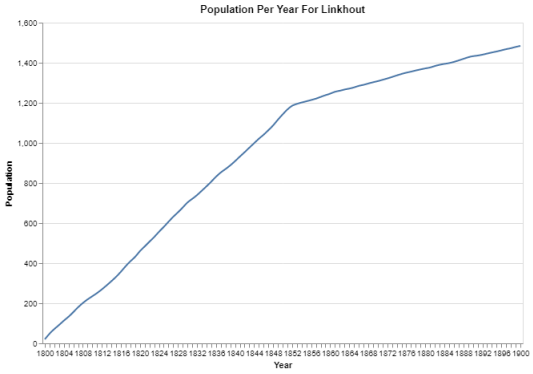

Figure 16: Linkhout Population over the years

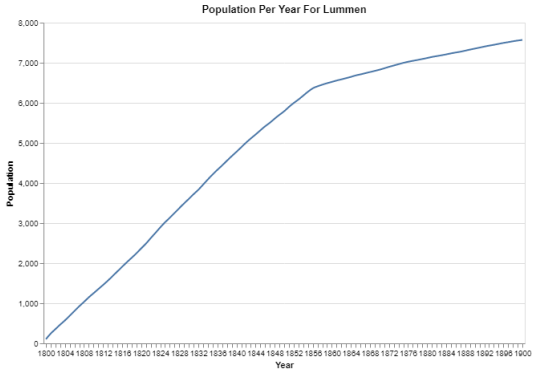

Figure 17: Lummen Population over the years

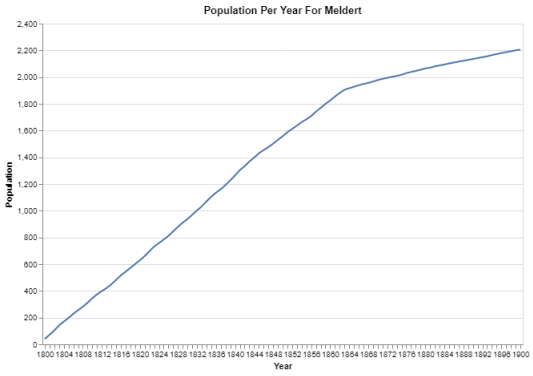

Figure 18: Lummen Population over the years

Codes

```
## install packages
install.packages("reticulate")
install.packages("altair")
reticulate::conda_create("r-reticulate")
install.packages("lubridate")
install.packages("lavaan")
install.packages("mice")
install.packages('plyr')
install.packages('epiDisplay')
install.packages('gmodels')

# load packages
library("lubridate")
library(lavaan)
library(mice)
library("altair")
library("lubridate")
library("plyr")
library("epiDisplay")
library("gmodels")

## importing dataset
linkhout = read.csv("C:\\Users\\joyce\\Linkhout_cens_data.csv")

######converting to time-date formats
linkhout$birthdates <- dmy(as.character(linkhout$Birthdate))
linkhout$deathdates <- dmy(as.character(linkhout$DateDeath))

## extract years only
linkhout$birthdates <- year(linkhout$birthdates)
linkhout$deathdates <- year(linkhout$deathdates)

# add the age variable
#linkhout$age <- linkhout$deathdates - linkhout$birthdates
str(linkhout)

## selecting only observations that were born in Linkhout and died in Linkout
## selecting observations with either birthdate or deathdate  known
## this will ensure no observation is present with both missing values on birth and death date

linkhout1 <- subset(linkhout,
                    (Birthplace == "Linkhout" & PlaceDeath == "Linkhout")|(DateDeath != "NA" | Birthdate != "NA")
                    )
## selecting some variables
```

```
sel <- c("NrLi","Sex","Birthplace","PlaceDeath","NrChild","birthdates","deathdates","age","DEATH","DIED")
linkhout = linkhout[sel]

## descriptive statistics on birthdate and deathdate
summary(linkhout1$birthdates)
summary(linkhout1$deathdates)

#########################################################################
################## IMPUTATION USING PREDICTIVE MEAN MATCHING

sel2 <- linkhout[,c("birthdates","deathdates")]

# returns a tabular form of missing value present in each variable in a data set.
md.pattern(sel2)

## now let's impute
imputed_Data <- mice(sel2, m=1, maxit = 50, method = 'pmm', seed = 500)
summary(imputed_Data)

## this step replaces all the "NA's" in the rows of the variables deathdates and birthdates to form a complete dataset.
complete_dates <- complete(imputed_Data ,1)

## now will add this complete observations to the original dataset and call the new dataset Linkhout_comp.

linkhout_comp <- data.frame(linkhout1$NrLi, linkhout1$Sex, linkhout1$Birthplace, linkhout1$PlaceDeath,
                            linkhout1$NrChild,complete_dates$birthdates,complete_dates$deathdates,linkhout1$DEATH,
                            linkhout1$DIED)

# calculating the years lived of all participants using the imputed birth and death dates.
linkhout_comp$age <- complete_dates$deathdates - complete_dates$birthdates

## deleting observations where age is negative
linkhout_comp <- subset(linkhout_comp, (age > 0 & complete_dates.deathdates <= 1900 & complete_dates.birthdates >= 1800))

# descriptive analysis of imputed values
summary(linkhout_comp$age)
summary(linkhout_comp$complete_dates.birthdates)
summary(linkhout_comp$complete_dates.deathdates)

## importing dataset IMPUTED datasets
linkhout = read.csv("C:\\joyce\\joyce\\New folder\\LINKHOUTPLOT.csv")

######################### MORTALITY  FOR THE VARIOUS TOWNS ########
#########################################################################
#### plotting for Mortality for Linkhout
```

```
linkhout$NEW_YEAR <- as.character(linkhout$NEW_YEAR)
#
alt$Chart(linkhout)$
  mark_line(interpolate = 'bundle')$
  encode(
    alt$X('year(NEW_YEAR):O',title='Year', axis=alt$Axis(labelAngle=0)),
    alt$Y('MORTAL_R:Q',title= 'Mortality Rate')
  )$
  properties(title='Mortality Rate For Linkhout',
             width=600,
             height=400)

#### plotting for LUMMEN
lumen$NEW_YEAR <- as.character(lumen$NEW_YEAR)
alt$Chart(lumen)$
  mark_line(interpolate = 'bundle')$
  encode(alt$X('year(NEW_YEAR):O',title='Year', axis=alt$Axis(labelAngle=0)),
    alt$Y('MORTAL_R:Q',title= 'Mortality Rate'))$properties(title='Mortality Rate For lumen',width=600, height=400)

#### plotting for Meldert
meldert$NEW_YEAR <- as.character(meldert$NEW_YEAR)
alt$Chart(meldert)$
  mark_line(interpolate = 'bundle')$encode(
    alt$X('year(NEW_YEAR):O',title='Year', axis=alt$Axis(labelAngle=0)),
    alt$Y('MORTAL_R:Q',title= 'Mortality Rate'))$
  properties(title='Mortality Rate For Meldert', width=600,height=400)

################DEATH PER YEAR FOR THE VARIOUS TOWNS ###############################################################################
#plotting for Linkhout
linkhout$NEW_YEAR <- as.character(linkhout$NEW_YEAR)
#
alt$Chart(linkhout)$mark_line(interpolate = 'bundle')$
encode(alt$X('year(NEW_YEAR):O',title='Year', axis=alt$Axis(labelAngle=0)),alt$Y('TOTALDIE:Q',title= 'Death Count'))$
properties(title='Count of Death Per Year For Linkhout',width=600, height=400)
################BIRTH COUNTS#################
linkhout$TOTALBORN
# birthcout
alt$Chart(linkhout)$
  mark_line(interpolate = 'bundle')$
  encode(
    alt$X('year(NEW_YEAR):O',title='Year', axis=alt$Axis(labelAngle=0)),
    alt$Y('TOTALBORN:Q',title= 'Count of Birth'))$
  properties(title='Birth Count Per Year For Linkhout',width=600,height=400)

#####################################################################
```

```
############GROUPING INDIVIDUALS BY YEAR LIVED ###################
#####
linkhout_imp <- read.csv("C:\\Users\\joyce\\complete\\linkhout_imp.csv")
lumen_imp <- read.csv("C:\\Users\\joyce\\complete\\lummen_imp.csv")
meldert_imp <- read.csv("C:\\Users\\joyce\\complete\\meldert_imp.csv")

### seletion only observations between 1800 and 1900
linkhout_imp <- subset(linkhout_imp, (NEW_YB != "NA" & NEW_YD != "NA") & (NEW_YB >= 1800 & NEW_YD <= 1900))
lumen_imp <- subset(lumen_imp, (NEW_YB != "NA" & NEW_YD != "NA") & (NEW_YB >= 1800 & NEW_YD <= 1900))
meldert_imp <- subset(meldert_imp, (NEW_YB != "NA" & NEW_YD != "NA") & (NEW_YB >= 1800 & NEW_YD <= 1900))

##delete observations with negative years lived
linkhout_imp <- subset(linkhout_imp, IMP_YSL >= 0)
lumen_imp <- subset(lumen_imp, IMP_YSL >= 0)
meldert_imp <- subset(meldert_imp, IMP_YSL >= 0)

## summary statitics
summary(linkhout_imp$IMP_YSL)
summary(lumen_imp$IMP_YSL)
summary(meldert_imp$IMP_YSL)
sd(linkhout_imp$IMP_YSL);sd(lumen_imp$IMP_YSL);sd(meldert_imp$IMP_YSL)

## create age groups for complete data linkhout;
linkhout_imp$age_group[linkhout_imp$IMP_YSL >= 0 & linkhout_imp$IMP_YSL < 15 ] <- 1
linkhout_imp$age_group[linkhout_imp$IMP_YSL >= 15 & linkhout_imp$IMP_YSL < 45 ] <- 2
linkhout_imp$age_group[linkhout_imp$IMP_YSL >= 45 & linkhout_imp$IMP_YSL < 65 ] <- 3
linkhout_imp$age_group[linkhout_imp$IMP_YSL >= 65 ] <- 4

## create age groups for complete data lummen;
lumen_imp$age_group[lumen_imp$IMP_YSL >= 0 & lumen_imp$IMP_YSL < 15 ] <- 1
lumen_imp$age_group[lumen_imp$IMP_YSL >= 15 & lumen_imp$IMP_YSL < 45 ] <- 2
lumen_imp$age_group[lumen_imp$IMP_YSL >= 45 & lumen_imp$IMP_YSL < 65 ] <- 3
lumen_imp$age_group[lumen_imp$IMP_YSL >= 65 ] <- 4

## create age groups for complete data meldert;
meldert_imp$age_group[meldert_imp$IMP_YSL >= 0 & meldert_imp$IMP_YSL < 15 ] <- 1
meldert_imp$age_group[meldert_imp$IMP_YSL >= 15 & meldert_imp$IMP_YSL < 45 ] <- 2
meldert_imp$age_group[meldert_imp$IMP_YSL >= 45 & meldert_imp$IMP_YSL < 65 ] <- 3
meldert_imp$age_group[meldert_imp$IMP_YSL >= 65 ] <- 4

##plots of number of deaths lived
tab1(linkhout_imp$age_group,main = "Death Counts Per Age Groups in Linkhout")
tab1(lumen_imp$age_group,main = "Death Counts Per Age Groups in Lumen")
tab1(meldert_imp$age_group,main = "Death Counts Per Age Groups in Meldert")
```

```
################## LINKHOUT
# select only age group == 1
linkhout_group1 <- subset(linkhout_imp, age_group == 1)

# Find the year with the most deaths for age group 1
link_grp1_death <- tab1(linkhout_group1$NEW_YD,graph = TRUE,sort.group = "decreasing")

# select only age group == 2
linkhout_group2 <- subset(linkhout_imp, age_group == 2)
# Find the year with the most deaths for age group 1
link_grp2_death <- tab1(linkhout_group2$NEW_YD,sort.group = "decreasing")
link_grp2_death

# select only age group == 3
linkhout_group3 <- subset(linkhout_imp, age_group == 3)
# Find the year with the most deaths for age group 1
link_grp3_death <- tab1(linkhout_group3$NEW_YD,graph = FALSE,sort.group = "decreasing")
link_grp3_death

# select only age group == 4
linkhout_group4 <- subset(linkhout_imp, age_group == 4)
# Find the year with the most deaths for age group 1
link_grp4_death <- tab1(linkhout_group4$NEW_YD,graph = FALSE, sort.group = "decreasing")
link_grp4_death

#####################################################################
################## TIME SERIES ANALYSIS#######################

### Linkhout
# creating time series object for MORTALITY
link_ts <- ts(linkhout_series$MORTAL_R,start = 1800, end = 1900, frequency=1)

### Smoothing Splines
plot(link_ts,ylab = 'Mortality',main="Smoothing spline for Mortality of Linkhout")
axis(1,at=seq(1800,1900,10),lwd = 0)
lines(smooth.spline(time(link_ts),link_ts, spar =0.6), lwd=2, col=4)
### Lummen
# create a time series object for mortality
lumen_ts <- ts(lumen_series$MORTAL_R ,start = 1800, end = 1900, frequency=1)
## plot
plot(lumen_ts, type="l", ylab="Mortality",main="Mortality per year for Lummen")
axis(1,at=seq(1800,1900,10),lwd = 0)
## Smoothing Splines
plot(lumen_ts,ylab = 'Mortality',main="Smoothing spline for mortality - Lumen")
axis(1,at=seq(1800,1900,10),lwd = 0)
```

```
lines(smooth.spline(time(lumen_ts),lumen_ts, spar =0.6), lwd=2, col=4)

#### SORT TABLES TO DETERMINE TOP THREE MORTALITY YEARS  ##########################################################
proc sort data =linkhout_plot;
by MORTAL_R NEW_YEAR;
run;
/*sort lummen*/
proc sort data =lummen_plot;
by MORTAL_R NEW_YEAR;
run;
/*sort meldert*/
proc sort data =meldert_plot;
by MORTAL_R NEW_YEAR;
run;

### Fitting Negative binomial Model ###################################
## Linkhout
proc genmod data = count_linkhout;
  class sex  /param=glm;
  model  count = deathdates sex / type3 dist=negbin ;
run;

/*does the model fit the data properly*/
/* If the test had been statistically significant, it would indicate that the data do not fit the model well.*/
data pvalue;
  df = 170; chisq = 180.0574;
  pvalue = 1 - probchi(chisq, df);
run;
proc print data = pvalue noobs;
run;

## Fitting logistic regression Model ###################################
##logistic regression for linkhout
logis_linkhout <- pcount_linkhout <- glm(Frequency_death/Cumbirth ~Death_date ,combined_linkhout,family=binomial,weights= Cumbir
summary(logis_linkhout)
confint(logis_linkhout)

##logistic regression for Lummen
logis_lumen <- glm(Frequency_death/cumbirth ~Death_date ,combined_lummen,family=binomial,weights= cumbirth)
summary(logis_lumen)
confint(logis_lumen)

##logistic regression for Meldert
logis_meldert <- glm(Frequency_death/Cumbirth ~Death_date ,combined_meldert,family=binomial,weights= Cumbirth)
summary(logis_meldert)
```

```
confint(logis_meldert)
str(combined_meldert)
```