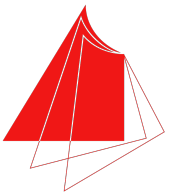


Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

A Study of Seasonal Patterns of Birth for Velke Pole, Slovakia between 1781 and 1900

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science (M.Sc.) in Geomatics

Deepanjana Majumdar

46700

Faculty of Information Management and Media
University of Applied Sciences, Karlsruhe

Advisor:	Prof. Dr.-Ing. Heinz Saler
Co-Advisor:	Prof. Dr. habil. Mark Vetter

January 2016

Declaration

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

The work was done under the guidance of Professor Dr. Heinz Saler, at the Hochschule Karlsruhe - Technik und Wirtschaft, Karlsruhe.

Karlsruhe, 19 Jan 2016 _____

Deepanjana Majumdar

Abstract

This thesis aims to study the birth seasonality patterns in Velke Pole for the period between 1781 and 1900, broken down into periods of 10 years. It also attempts to analyse the relationship between birth seasonality, temperature and precipitation.

To achieve these goals certain statistical methods were chosen to help this research along. Due to the absence of any recorded climatological data for Velke Pole, the geostatistical interpolation technique of ordinary Kriging was used to spatially interpolate temperature and precipitation data for Velke Pole based on nine known sample points for the years 1860-1900. The estimated values retrieved from Kriging were then used to extrapolate temperature data for the years 1781-1859.

Ordinary Least Squares multiple regression statistical technique was used to regress birth data against monthly dummies with and without temperature and precipitation controls to find information on the influence of these factors on the birth seasonality pattern.

It is evident from the results that although precipitation was not a significant enough determinant of birth seasonality, temperature might have a larger role to play as an influencer, but it still cannot be held completely responsible for the seasonal pattern. Multiple factors work together to influence this phenomenon that has been observed in almost all populations.

Acknowledgement

I would like to express my deepest gratitude to my supervisor Prof. Dr. Heinz Saler for his full support, expert guidance, understanding and encouragement throughout this master's degree. Without his incredible patience and counsel this thesis work would have been an overwhelming pursuit. In addition, I would like to extend my appreciation to Prof. Dr. Mark Vetter for co refereeing this thesis.

The completion of this undertaking could not have been possible without Ms. Martina Tomkova and Mr. Svec Marek of Slovakia Hydrometeorological Institute (SHMU) and Ms. Hildegard Kaufmann of Zentralanstalt für Meteorologie und Geodynamik (ZAMG, Austria), for providing the climatological data and for just replying back to the constant barrage of panic stricken emails.

A very big thank you to my fellow students, classes were made fun because of you, those who have already completed, and those still working on their thesis, there is a light at the end of the tunnel.

Thank you to Anirban Majumdar, my family and friends.

Contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Research Problem and Motivation

This master's thesis aims to determine the seasonal pattern of birth in the municipality of Velke Pole, a village located in the central Banska Bystrica region of Slovakia. Along with the seasonal patterns, it is also the task of this thesis to examine whether a relationship exists between seasonality and climate, namely temperature and precipitation, employing various statistical techniques.

Considering this is a community level undertaking, it is deemed to be called a meso-level research. This level of analysis can be useful in bringing to the fore dynamics that exist on a national level or even on a sub-continent level; comparison with macro level studies may reveal whether the pattern forms a subset or has its own distinctive feature.

Birth seasonality has been studied for many years with some of the earliest research dating back to the early 19th Century. Over the last few decades, it has been taken up with renewed vigour as scientific research in the domain of health sciences find connections between prevalence of certain diseases and the month of birth. Apart from such specific findings, birth seasonality has often been found to be a fascinating subject that is a direct consequence of the social fabric and natural environment of a community or nation.

With birth data of Velke Pole readily available from the late 18th Century, it was imperative that they be analysed and evaluated in order to discover more about the land and its people. Lack of social indicators such as education, age of the mothers, etc. dictated the decision to do the analysis on indicators that would have been easier to obtain, temperature and precipitation. Furthermore, as this study is of a historic nature a lot of the inference is based on speculations and conjectures.

Birth Seasonality is dependent on multiple factors at play together, from religious and cultural to economic, as well as climatic conditions.

Influence of climatological factors such as, temperature, length of photoperiod, intensity of light and, in certain specific conditions, even precipitation have been found to be significant on the seasonality of birth, even in the presence of social and economic flux.

Climatological effect on human birth informs the human physiology and activities. The economic activities and seasonal work are often dependent on the month of the year, which implies climate,

and the human body is directly affected by temperature and humidity influencing aspects of childbirth and conception.

In case of this thesis, the lack of data makes it more difficult to assess all the determinants at play in determining causality. There was no record of weather data found for Velke Pole for the period of study, so different techniques have been applied to come close to what the temperature would have been like in the town based on temperature and precipitation records from nearby cities. Records of epidemics and diseases for this time period have been considered to see if there were any changes in the pattern for those particular years. Qualitative research based on studies made by other scholars have also been undertaken in order to draw comparisons with their results and findings.

The flowchart in Figure 1 describes the salient steps undertaken to accomplish this research work. Squares represent an underlying mathematical or statistical process whereas the document shapes indicate an input or output file (as source or intermediate results). Four distinctive techniques have been applied to analyse and extrapolate the available data.

- Orange: The statistical concept of moving averages has been applied to normalise the available monthly birth data in order to remove unwarranted seasonal trends.
- Red: These shapes represent the set of geospatial interpolation techniques, namely Kriging, employed to determine the temperature and rainfall of Velke Pole from adjoining areas.
- Cyan: These shapes indicate the processes undertaken to perform statistical extrapolation techniques based on results obtained from Kriging and temperature data from previous periods with the goal of determining missing monthly temperature values for the period under study.
- Green: Finally, regression analysis has been done on the detrended values of birth rates and the extrapolated temperature data in order to find relationship between dependent and independent variables.

These techniques and the findings are explained in details in Chapters 3 and 4 respectively.

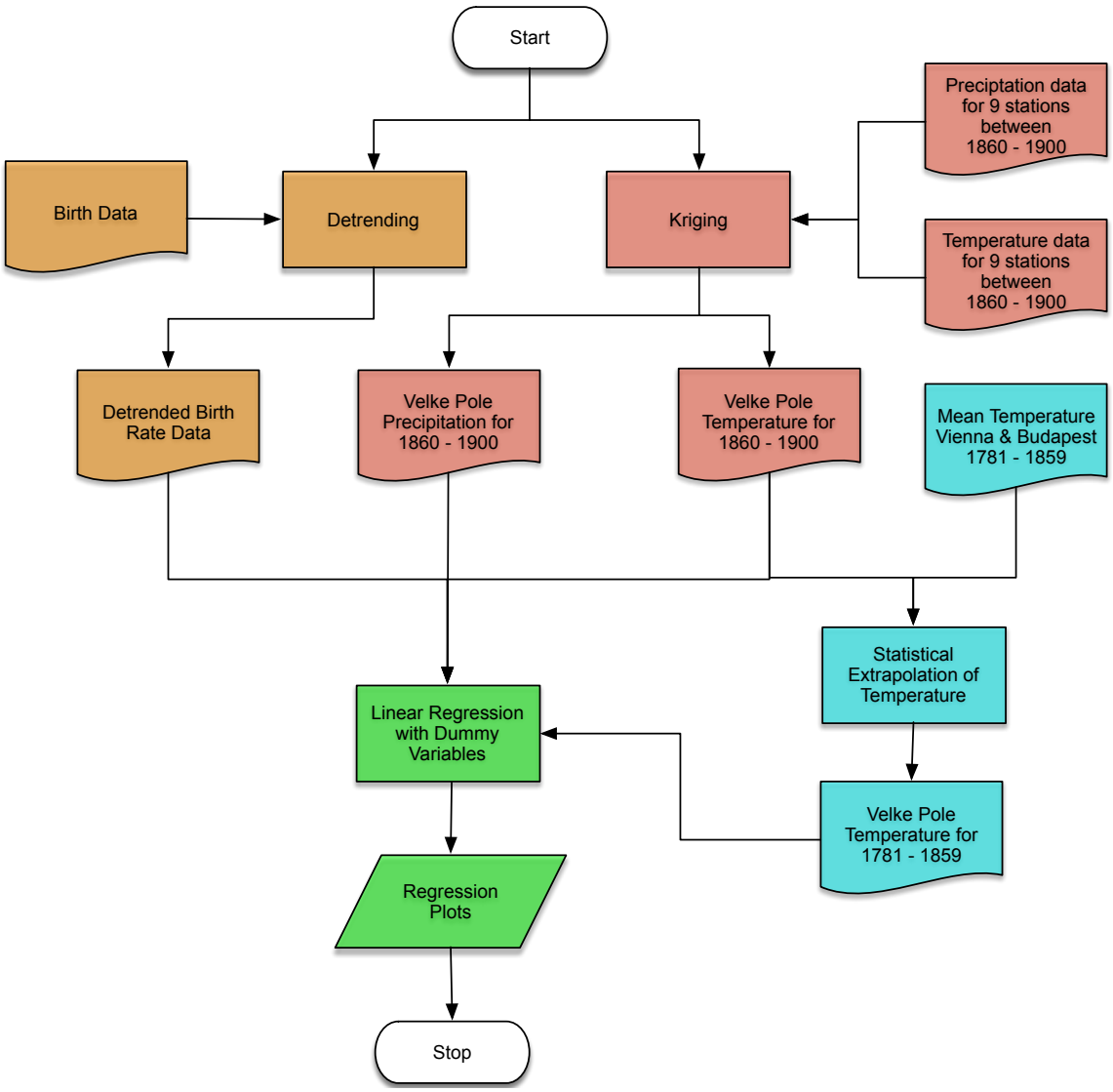


Figure 1: Flowchart describing thesis methodology

1.2 Assumptions and Limitations

- It is assumed that the populations whose activities are going to be analysed were largely agricultural and living off the land, it is safe to postulate that climate would have played a significant role in influencing their behaviour and activities.
- To study the effect of climate on birth seasonality it has to be assumed that they affect at and around the time of conception, so the temperature data is lagged 9 months from the birth month for data analysis, although, it could be said that births do not occur exactly at nine months after conception, again considering the time period of this study it is safe to assume that any premature birth would have had very low to no chances of survival, given the condition of medical science then.

- Another assumption also necessary here was the correctness of the temperature and precipitation data that have been acquired through the use of spatial interpolation and extrapolation techniques. It is apt to say that there is no certain way of telling exactly how correct or incorrect the data might be, but from looking at the interpolated data and the recorded data from the weather stations around Velke Pole, it would be acceptable to say that if any discrepancies exist it would be small enough to not affect its influence on birth seasonality.

1.3 Contributions

- This thesis attempts to document and expose more about the life and times in a part of Europe that is not very widely discussed and very well documented in the English language. The existing literature presented comprehensive information on birth seasonality in North America and the northern and western Europe, but there was a void for the eastern European regions that were and still are conflicted. The general lack of recorded data was also evident when weather data was searched, apart from big, historically important cities not much information could be retrieved, and if there were any data available, they were not in digitized format, therefore inaccessible.
- A record of temperature and precipitation for Velke Pole has now been created 1780 onwards up to 1900, and this can be used for estimations and other research purposes or to interpolate data for other locations nearby.
- The open source programming language R is quite commonly used in the field of Geomatics whenever there are requirements for statistical analysis, this thesis makes way for utilizing Python libraries such as Pandas, Numpy and matplotlib, for data processing, wrangling, analysis and visualization. This can be helpful for geomatics students and professionals who often have prior knowledge of Python due to its widespread use in the field of geoprocessing available in GIS software, such as ArcGIS. This development would render learning a new language from scratch to perform similar tasks unnecessary.
- The data transform functions implemented in Python could be used as templates for processing a wide array of data sets owing to their similarity in data formats and processing requirements

The next few chapters shall explore the works of other scholars on birth seasonality and its determinants, our data and how they have been obtained, the methodology of data processing and analysis and the final deductions.

However, before that, a little background information on the demographic and settlement history of the study region may provide more insight during the interpretation of the data analysis results.

1.4 Settlement History of Velke Pole

The first mentions of the village of Velke Pole was found between 1332-1397, where we discover that its name was "Pratum" meaning meadow and then changed to "Velikapole" (Solcova, 2008).

Germans started settling down in the Hungarian Bars County to mine minerals. The Benedictine monastery had many German visitors and they eventually settled down in the surrounding areas. Colonization by the German population initiated the transition from pastoralism to productive agriculture and the transfer of new technologies in agriculture and mining. In the second half of the 13th century, which was associated with settling of uninhabited areas previously devastated by the Tatar invasions (1241-1242), new settlers played a major role, especially of German ethnicity, in their restoration and development. The king initiated the establishment of the German settlements, which later developed into royal and major mining towns of Slovakia. In addition, the aristocracy played an important role trying to raise their own revenues by settling such areas.

By profession, the Germans were merchants and liked to settle in town, but when the town started filling up, they moved to the nearby villages. Only around 1345 did they start inhabiting towns of Velke Pole and Pila. It was during the rule of Andrew II that German settlers from Thuringia started settling down in Nyitra and Bars County. A new wave of German settlers came in around 15th century, they were called 'Krickerhäuer', many of the town they settled in had the name "Häue" in them, meaning 'to fell', this indicated that the settlers mainly worked in the forests and felled trees to clear the land for settlement. (Ede, 2015).

The presence of the Germans was concentrated in the northern mountainous parts of present Zarnovica District (in the communities Veľké Pole and Píla), which were particularly attractive due to its potential for mining. The main occupation in this village was mining, mainly of silver and lead. Although, during wars with the Ottoman Empire between the 17th and early 18th Centuries the mines were destroyed, Velke Pole had continued to prosper and develop into a town with many craftsmen, as joinery gained prominence.

Velke Pole had belonged to the Bars County till 1922, then to Zólyom County till 1928. Since 1949, it has been incorporated into the Nova Bana district.

The population of Velke Pole had grown steadily until the 18th Century when there was a mass exodus of subordinate people from this region, yet the population grew until the 20th Century. By the end of the 1860s the number of inhabitant were 2300 growing to 3314 in 1930. After the Second World War the population declined rapidly, by the end of 1948 the population had dipped to 731(Church book data via Personal Communication, 2015). The biggest emigration happened during the above-mentioned time when the German population was deported and the population went down rapidly (HsKA Slovakia webpage, 2015). The most current demographic figure is 385 from the year 2014 (www.citypopulation.de).

Chapter 2: Literature Review

In 1826, Quetelet was among the first to have documented the yearly pattern of human birthrate that influenced new waves of studies conducted in this field recognizing birth seasonality as an "ecologically responsive phenomenon" (Condon, Scaglion 1992), meaning that this biological occurrence is intrinsically related to milieu and culture of that population.

Research has shown that seasonal variation in the frequency of births is a nearly universal phenomenon in human populations (Cowgill 1965; Lam & Miron 1991; Bronson 1995). This has sparked curiosity over the years in demographers and population scientists worldwide about the determinant, biological and social, of this phenomenon.

In recent years, the topic of birth seasonality has received renewed attention, because it is being perceived as a major factor influencing medical conditions and human personalities. For scientists to use birth seasonality as a determinant in other social and physiological phenomena, it is important to first establish what affects birth seasonality.

The idea of temperature and other climatic factors playing a significant role in dictating seasonal patterns of birth came from the abundance of research lead by scholars over the year.

In specific case studies, the seasonality of birth is found to be influenced by a combination of factors and most researchers substantiate that theory. Ellison, Valeggia & Sherry (2005, p379) postulate that human birth seasonality can be grouped under three categories: "seasonality due to social factors that influence the frequency of intercourse; seasonality due to climatological factors that directly affect human fecundity; and seasonality due to energetic factors that principally affect female fecundity". Condon & Scaglion (1982) broadly define three important variables, environmental, biological and sociocultural, and their interaction as influences.

Furthermore, the feasibility of this thesis was also confirmed, as evidence was found on studies having been conducted on such an archival level by Lee (1981) and Richards (1983); Lee used 300 years of data, between 1541-1871, for England in order to evaluate the effects of temperature on births. His findings showed that when during hotter summers and colder winters the conception rate went down. Richards (1983), on the other hand, though found similar effects; her estimates were not statistically significant. Her analysis was based on annual data from France between 1740-1909.

The way that temperature apparently affects birth is by influencing coital frequency or more directly by influencing the human physiology.

It has been found by Levine (1991, 1994) that there is an annual variation in sperm quality in men and that it is lowest in July through September and higher in February and march. He found that both heat and photoperiod have significant influence on the seasonal variation on the male reproductive system. There is less direct proof of temperature affecting the female reproductive system.

Seiver (1985), in his research, recorded the birth seasonality pattern in the United States from 1947 to 1976 and in the various geographic regions, then went on to statistically analyse any changes in the seasonality over this period. He dealt with two of the possible determinants of the seasonality pattern, the first being climate and the second hormones. After applying the statistical models he developed for his research, he found certain discrepancies in the pattern between different zones of the country, that there was a significant fall in the number of births in April and May in the south, whereas this was hardly the case in the north and the west. Yet, over the years the scale of this pattern found in the south has gradually diminished. Patterns that exist for the entire nation are, the significant peak in the birth rate in September and no change in the magnitude of this pattern over the time period of the study. Also, a statistically significant rise of births in December has been observed.

Seiver (1985) concluded that the spring trough that occurred in the southern regions of the United States were a result of a dip in conception during the hot and humid summer months, why there is a low conception rate during the summer is based on conjectures and a limited number of materials that studied the causal relationship between climate and fecundity. He also went on to say that the diminishing trough in the recent years is a direct result of the introduction of air conditioning in the southern regions.

Manfredini (2009) studied the Birth Seasonality in Present day Italy from 1993-2005 to investigate the cause of a shift in the trend from a high fertility, high mortality system to low fertility, low mortality system. He observes a shift from a typical European pattern to a bimodal one with peaks in May and September and troughs in June and Nov-Dec. This sort of a radical change has been said to occur in Occidental societies because of an overall change in cultural practices, namely decline in population involved in agriculture to family planning practices involving contraception. He found that rainfall had almost no influence in birth seasonality because Europe does not have a climate characterized by sharply defined dry and wet seasons that affect daily life.

Other countries where agriculture is affected by rainfall, which means availability of food depends on rainfall, can be found to have stronger correlation between birth seasonality and rainfall, for example in the study conducted in Central-Africa by Bailey et al. (1992)

Despite the above findings, this thesis will take a look at precipitation as a determinant, it would be interesting to see if the factor was as inconsequential more than a century ago or not.

Manfredini (2009) concludes that cultural and social setup of the modern times have limited the influences of climate but 40% of the seasonal changes, in Italy, are still the result of photoperiod and temperature during time of conception.

In pre-industrial agricultural populations, the phenomenon of birth seasonality was, generally, more pronounced, but it is observed that seasonality has increased in the 20th century in some high income, low fertility populations such as Sweden.

Lerchl, Simoni & Nieschlag (1993) conducted a research to study the seasonality of birth rate in Germany between 1951 and 1990 where they found that the pattern had changed from peaks in the early months of the year between 1950-1975 to later months 1975 onwards. Peaks in March and April and a so-called shoulder peak mark the seasonal pattern in September is observed in Figure 2, this pattern remained stable till 1975, from then onwards the spring peak gradually turns into a trough and the shoulder peak of September gains magnitude. Many scholars have pointed out the conception period in December to be of consequence; during the Christmas period there is a rise in conception that eventually leads to higher births in September (Cowgill, 1966), despite this claim Lerchl et al., (1993) do not consider it as an all powerful argument for this shift as Roennberg & Aschoff (1990) have observed this shift towards the winter months even in the Southern Hemisphere where it does not correspond with the Christmas period.

Along with a change in seasonality, there was also a significant decline in the overall number of births in Germany 1971 onwards and this coincides with the introduction of oral contraceptives. There was also an overall change in the environmental condition in which humans previously operated and they were now less affected by the changes in temperatures and photoperiodicity, so it can be assumed that there was a transference from climatological factors to more social factors that now ruled birth seasonality, the shift in seasonality may have occurred due to a turn from biological to social influences, these mainly include introduction of birth control and socio-political developments such as legal protection of expectant mothers at work place and more liberal abortion laws introduced in 1968 and 1976 respectively (Lerchl et al., 1993)

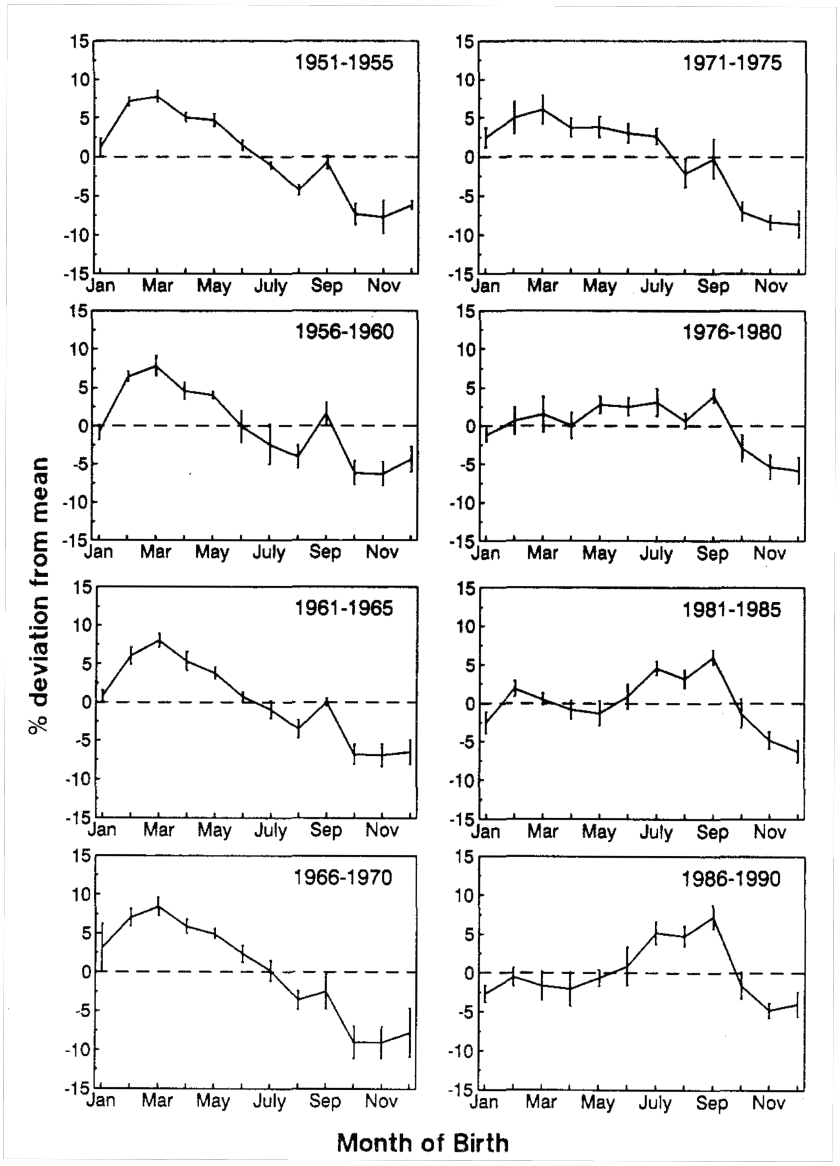


Figure 2: Relative birth rates and their shift in Germany from 1951 to 1990 (Lerch et al, 1993)

Werschler & Halli (1992) investigated why birth seasonality in Canada was different from those of northern United States. As discussed above, birth patterns in United States are characterized by a trough in April-May and a peak in August-September whereas Canada has peaks in April-May and a trough in December-January. One of the conjectures for such a pattern in Canada, is the influence of work cycle and preference for month of birth. The difference in patterns between northern USA and Canada, even in areas with homogeneous climatic and socioeconomic milieu points to the fact that climatic conditions are not the sole influences. The similarity that does exist is the inverse relation between conception and temperature in summer months. In Manitoba, this association was observed and it was also noted that this association decreased with socioeconomic development during the period of 1928-1988, the period of study.

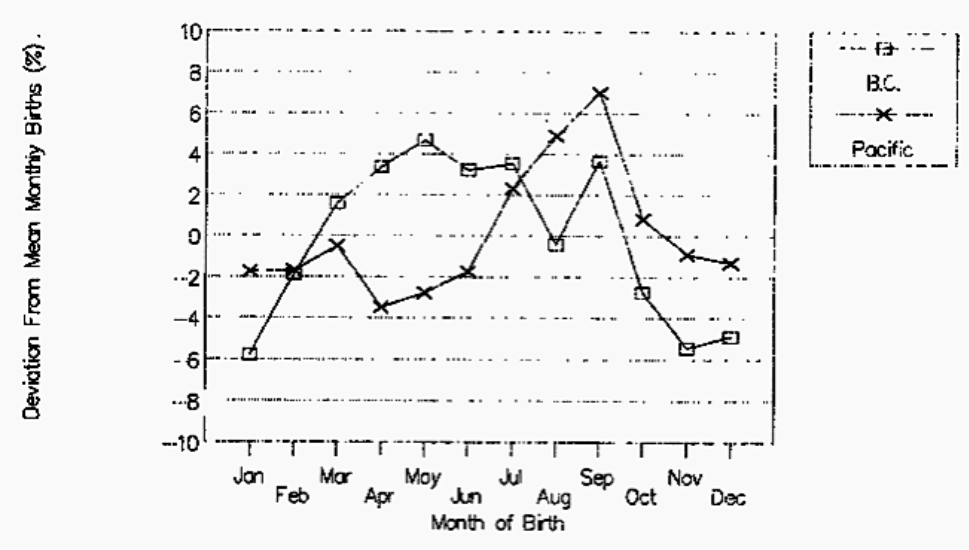


Figure 3: Seasonal birth patterns of British Columbia and Pacific U.S. 1928-1988 (Werschler & Halli, 1992)

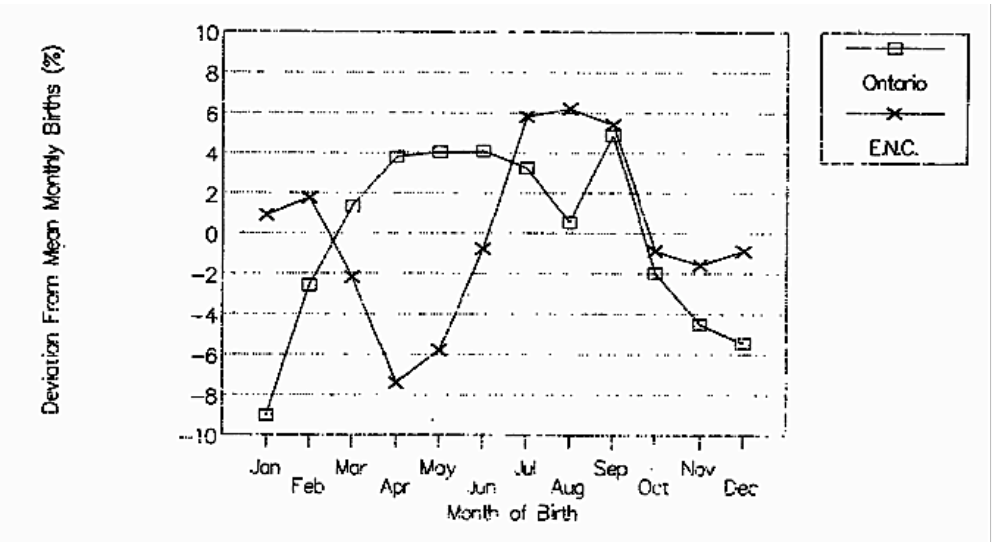


Figure 4: Seasonal birth patterns of Ontario and NE Central U.S. 1928-1988 (Werschler & Halli, 1992)

One general view that has been noticed is that birth seasonality was more pronounced in the pre-industrialization era; in modern urban-industrial societies where food supply is stable and housing conditions are better there exists a curb of the seasonal birth pattern, a more pronounced birth seasonality is observed in low status groups for whom food supplies and housing conditions are not as secure.

Out of the many possible determinants of birth seasonality, photoperiod is believed to be of great importance in high latitude regions and temperate zones (Condon, 1991). Since polar areas experience wide variations in length of day in a year, it is only natural that scientists will be curious to investigate the effects of this phenomenon on birth. There are works explaining some direct physiological changes attributed to photoperiodicity, such as investigations by Villerme (as cited in

Condon (1991)) in France and Argentina where he observed that sexual awareness peaked in spring and early summer and fertility also increases with the increased duration of light (Condon, 1991), despite these findings there was no conclusive proof of photoperiodicity being singularly effective in dictating the seasonality in the findings of Condon (1991) in his studies of Central Canadian Arctic.

In a more recent study on Canadian birth patterns, Cummings (2012) makes case for atmospheric brightness. He stresses that rather than the duration perhaps it is the intensity of light that affects seasonality. A very high, positive correlation coefficient (r≈93%) of birth seasonality and seasonal brightness with a lag of 10 to 11 months makes it a plausible determinant, but the possible presence of confounding variables weakens its case. Atmospheric brightness, measured in lux (1 lux = 1 lumen/sq.m.), is dependent on such environmental factors as latitude, temperature, atmospheric dust, humidity etc. that it would be quite impossible to consider it as a completely independent variable. Despite this counterargument, there is still a strong possibility of luminous brightness being a prominent determinant, because the birth seasonality pattern in Canada is not fully elucidated by the explanations provided by temperature and photoperiodicity, for e.g. according to the temperature explanation July and August should have low conception rate leading to low birth rate in April and May, instead a peak is observed in April.

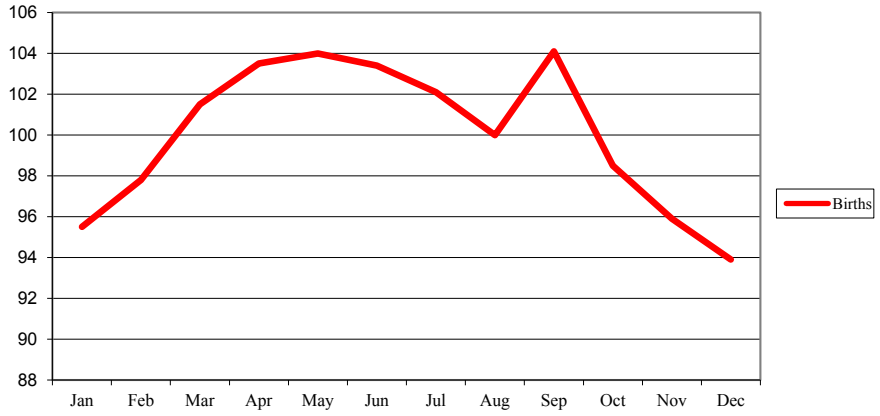


Figure 5: Canadian birth seasonality (1989) from Trovato & Odnyak, 1993 (Cummings, 2012)

Bobak & Gjonca (2001) tried to find a link between seasonality of birth and socio-demographic factors in Czech Republic between 1989-1991. They found that birth seasonality was more pronounced for married, better-educated mothers between the ages of 25 and 34 yrs. and for second and third child whereas the seasonal variation was weak for very young mothers or mothers older than 35, lower educated, unmarried and for their first or fourth child. This sort of pattern was noted in Britain in the 1960s where higher social classes had a stronger birth seasonality variation and also in France. But this is in direct contradiction to findings in the US where birth seasonality is more

pronounced in a population with lower social status. In New York seasonality was more pronounced in non-whites and in illegitimate births (Erhardt et al., 1971).

Lam & Miron (1994) did extensive research studying the global patterns of birth seasonality and came to the conclusion that the most pronounced patterns are found in regions with extreme summer heat and in regions at extreme latitudes. Regions with extreme summer heat, such as the southern United States, have significant drops in conceptions during the hottest months, where births decline substantially in April and May, and in northern Europe, regions of extreme latitudes, where births increase substantially in March and April.

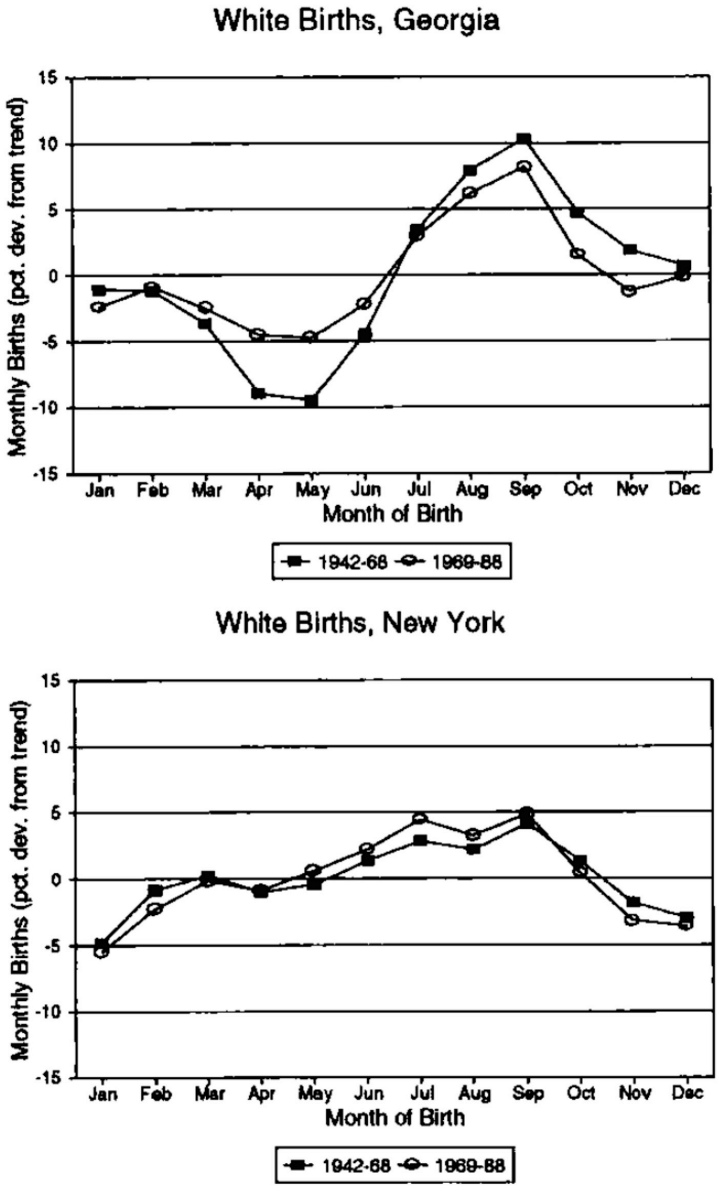


Figure 6: Seasonal birth patterns in Georgia (above) 1942-66, 1969-88 and New York (below) 1942-68, 1969-88

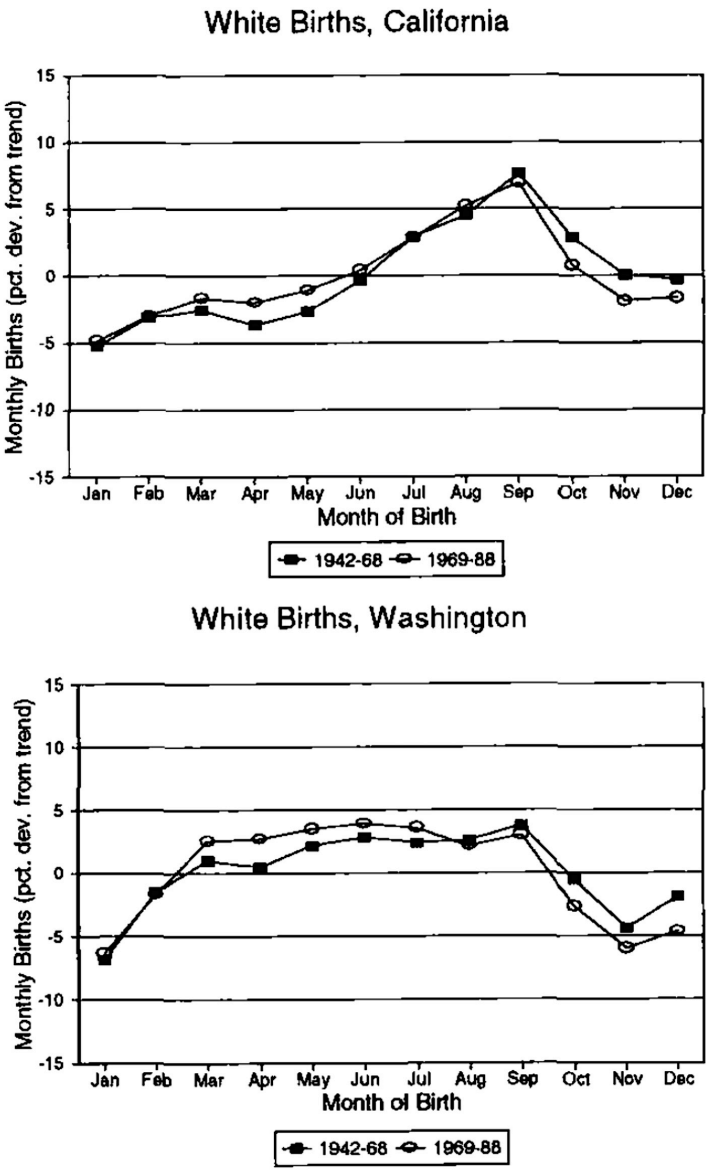


Figure 7: Seasonal birth patterns in 1942-68 and 1969- 88 in California (above) and Washington (below)

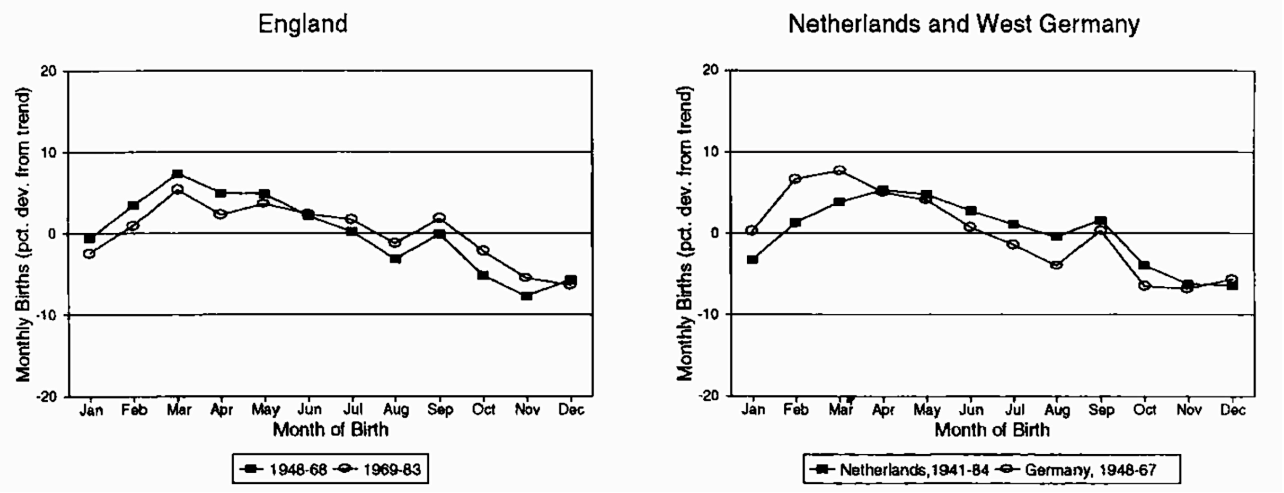


Figure 8: Seasonal birth patterns in England 1948-66, 1969-83(left), Netherlands 1941-84 and Germany 1948-67(right)

As can be seen from the above graphs, the September shoulder peaks are common in both U.S and Europe but there is a distinct spring peak in Europe that is absent in the U.S. Over time, European patterns are becoming more like the United States and losing their spring peaks (Lerchl et al. 1993)

A comprehensive study undertaken by Lam & Miron (1991, 1994, 1996) in examining the complex and multi-faceted relationship between temperature and fecundability, and consequently birth seasonality, yielded statistical models and useful inferences that have been widely cited by other scholars in this field since.

An indication for an annual rhythm of human reproduction is apparent in birth statistics; monthly rates show a clear annual alternation with up to threefold difference between minimum and maximum. Such documentations of birth rhythms go back to the 19th century. Many studies have since been dedicated to the illumination of the causal relationships behind this worldwide phenomenon. In spite of the tremendous indication for the contribution of biological and climatological factors to the annual rhythm of birth, the cause for this phenomenon in humans is still unclear, and many studies claim social indicators to be the factors chiefly responsible.

A look into the above mentioned literature and some more that have not been reviewed here only confirms the view that no one particular determinant can be the sole cause for the patterns of birth seasonality observed all over the world, it is always a combination of factors, and the proportion of the part each one plays keeps changing over time and space.

Chapter 3: Statistical Models

From the literature review, a good collection of data analysis methods has been extracted. The following statistical models were found to be the most commonly used and they will be expanded on.

Before going into the details of the methodology in the next chapter, it is important to define and explain some of the statistical terms that will appear frequently in the following chapters, as well as to explain why these statistical techniques were considered appropriate for this purpose.

3.1. Linear Regression Models

3.1.1. Time Series and Moving Average

The study in this thesis follows a process typically observed in time-series analysis. It is called time-series analysis because the consists of data collected at regular intervals over time and are continuous, for e.g. birth data for each month over 10 years or daily average temperature over a year. A common notation to denote time-series data is shown in Equation (1):

$$\boldsymbol{X} = \{\boldsymbol{X_t}: \boldsymbol{t} \in \boldsymbol{T}\} \quad (1)$$

Where, X is sequence of data points X_t defined over a discrete time spectrum.

The above-mentioned data will most certainly have an inherent structure known as trend or seasonal variation, and it will also suffer from random variation. Reduction or removal of the random variation can more clearly reveal the seasonality or trend in the data. One of the methods applied in reducing this random variation is called *moving average* which is a method for smoothening time-series sequence by average a fixed number of consecutive data points (OECD, 2005).

Given a time-series sequence $\{X_i\}_{i=1}^{N}$, an *n*-moving average is a new sequence $\{MA_i\}_{i=1}^{N-n+1}$ is defined by taking the arithmetic mean of subsequence of *n* terms (Weisstein, 2015):

$$MA_i = \frac{1}{n}\sum_{j=1}^{i+n-1} X_j \quad (2)$$

The moving average sequence moves over time including each time-series data point in the series sequentially.

Moving average has been used extensively by authors to detrend the birth data. To understand detrending, trend must be first defined. A trend in a time-series is a slow, gradual change in some

property of the series over the whole interval under investigation. Detrending, therefore, is the statistical or mathematical operation of removing trend from the series. It is often applied to remove a feature thought to distort or obscure the relationships of interest and is sometimes used as a pre-processing step in statistical techniques such as regression analysis.

In Seiver (1985) presented a model for the detrending of monthly births as follows:

$$MAB_j = \frac{B_j}{\sum_{i=j-6}^{i=j+5} B_i} * 1200 \quad (3)$$

where MAB_j is the moving average births in percent for month *j* and B_j is the births in month *j*. The denominator expresses a 12-term moving average and is centred. MAB_j is expressed in percent. As an example, if births in June, B_j = 30 and the total births between January to December $\sum_{i=j-6}^{i=j+5} B_i = 240$, then the MAB_{June} = 150%.

However, this model had not adjusted for the length of the month. This is easily rectified if B_j is divided by the number of days in the month *j*.

One could also remove trends without using moving averages as He and Earn (2007) (as cited in Dorelien (2013)) have done using the following model:

$$\hat{X}_i = \frac{1}{12}\sum_{j=1}^{12} X_{ij} \quad (4)$$

Here $\hat{X}_i$ is the average number of births in a month of average length in year *i* and X_{ij} is the number of births in month *j* of year *i*.

To correct for the different lengths of each month He and Earn (2007) used the following formula:

$$c_{ij} = \frac{(Days\ in\ the\ year\ i)/12}{Days\ in\ month\ j\ of\ the\ year\ i} \quad (5)$$

The scaled, month-length-corrected monthly values could then be denoted as:

$$Y_{ij} = \frac{c_{ij}X_{ij} - X_i}{X_i} \quad (6)$$

As before, *i* denotes year and *j* the corresponding month. Finally, Equation (7) is the average monthly value achieved by averaging monthly values across all years.

$$Z_j = \frac{1}{N_{yr}} \sum_{j=first\_year}^{last\_year} Y_{ij} \quad (7)$$

Following Seiver (1985), Werschler & Halli (1992) derived a formulation for computing the ratio values of a particular month *i* centred on a 12 month moving average as below:

$$R_i = \frac{B_i}{T_i} * 1200 \quad (8)$$

Where, R_i is the ratio value in month *i*, B_i the number of births in month *i*, and T_i is the total births in a 12-month period centred on month *i*. The ratio values are interpreted as percentage deviation from the mean monthly number of births in a year (Werschler & Halli, 1992). The use of this birth ratio formula has been illustrated in the next chapter for birth trend analysis in Velke Pole.

The mathematical models above serve in data preparation and detrending, which is a prerequisite for doing regression analysis on time-series data.

Regression analysis is a statistical study of the relationships between dependent and independent variables. It is widely used for prediction and forecasting and in this research, regression analysis has been used to gauge the extent to which the dependent variable (birth rate) is influenced by the effects of two independent variables (temperature and precipitation).

The detrending formula from Lam & Miron's (1996) research on the effect of temperature on human fertility has been used for regression analysis in this study. The formula has been slightly adjusted to correctly reflect the effect of moving average calculation:

$$b_t = \frac{B_t}{\sum_{m=t-5}^{t+6} B_m} * 12 \quad (9)$$

where B_t is the number of births in month *t* divided by the number of days in the month, B_m is the actual number of births in a month, and b_t is the detrended births per day in month *t*.

Following this, the natural logarithmic values of b_t are used as the dependent variable in a multiple regression model using climatic factors such as temperature and precipitation data as independent variables:

$$\ln b_t = \sum_{s=1}^{12} \alpha_s d_t^s + \beta_1 T_{t-9} + \beta_2 T_{t-9}^2 + \beta_3 P_{t-9} + \beta_4 P_{t-9}^2 \quad (10)$$

where, ln b_t is the logarithm of the detrended monthly births in month *t*. d_t represents the dummy variable for month *s* whose coefficient is α_s. T and P are temperature and precipitation at the moment of conception, considering a 9-month lag, hence $t - 9$. The quadratic expressions used have been so used to magnify the effects of the climatic parameters.

Equation (10) is a modification of Lam & Miron's (1996) regression formula in two ways:

1. It does not use a 10-month lag, as used by Lam & Miron for temperature independent variable; and
2. The precipitation independent variable and its quadratic form are introduced to study the effect of both temperature and precipitation as opposed to the usage of only T in Lam & Miron

Lam & Miron (1996) noted that the use of monthly birth data instead of monthly conceptions complicate interpretation of the dependencies. The assumption used there is also the basis of this research, i.e., births in month *t* correspond to conceptions in month $t - 9$.

3.1.2. Dummy Variables

The use of 'dummy variables' in Equation (10) is important for determining seasonality. A dummy (indicator) variable can only take on values 0 or 1 (Garvaglia & Sharma, 1998). They are used for seasonality analysis in time-series data and act as numeric stand-in for a qualitative or categorical variable such as a month. In the regression model of Equation (10), a dummy variable with a value of 0 will cause coefficients to disappear. A value of 1 will cause the corresponding coefficient to function as a "supplemental intercept". With the introduction of dummy variables in the model, it is possible to define subsets of observations for seasonality analysis that have different slopes and intercepts from observations controlling for temperature and precipitation as independent variables without requiring the change of regression model. The following example (Wadsworth, 2015) illustrates the use of dummy variables.

Let there be a dummy/categorical variable called *Gender* and a linear regression model be expressed as:

$$Salary = \alpha + \beta_0 Age + \beta_1 Gender \quad (11)$$

Equation (11) expresses a model for computing the salary gap based on gender of a person.

The dummy variable *Gender* would be 1 if a person is a male and 0 for female. For men, the predicted salary would be:

$$\widehat{Salary_{men}} = \hat{\alpha} + \widehat{\beta_0}\, Age + \widehat{\beta_1} * 1 \quad (12)$$

Whereas, for women, the predicted salary would be:

$$\widehat{Salary_{women}} = \hat{\alpha} + \widehat{\beta_0}\, Age + \widehat{\beta_1} * 0 \quad (13)$$

So, the predicted difference of salaries between men and women is:

$$\widehat{Salary_{men}} - \widehat{Salary_{women}} = \hat{\alpha} + \widehat{\beta_0}\, Age + \widehat{\beta_1} - \hat{\alpha} - \widehat{\beta_0}\, Age \quad (14)$$

Which results in $\widehat{\beta_1}$ - the coefficient of the male dummy variable.

The constant β_0 is the intercept of the default group of women with age set to zero and $\beta_0 + \beta_1$ is the intercept for men.

Generalizing the observation, the coefficient of dummy variables measures the estimated difference between the categorical variables coded with the value 1 and the variables group coded with 0 (the default group).

As will be explained in the next chapter, the months (January to December) are coded as dummy variables in our analysis in order to compute the seasonality of observations with and without controlling the independent variables of temperature and precipitation.

Before concluding this treatise on regression techniques, it is important to highlight the significance of using logarithmic scale to detrend birth rate in Equation (10). Using logarithmic variables in regression models is useful when non-linear relationships exist between dependent and independent variables (Benoit, 2011).

3.1.3. Logarithmic Transformations

Equation (10) depicts a log-linear model because the left-hand side of the equation uses a logarithmic detrended birth rate whereas the independent variables on the right-hand side use natural estimated coefficients.

Logarithmic transformations are especially useful when a non-linear relationship exists between the dependent and independent variables, in such cases a linear model just does not cut it and a multiplicative form of modelling achieved by a log dependent variable can work out much better.

We often have a dependent variable that is more than or equal to 0, this is called a limited dependent variable, such as birth. Although, in practice it may not make much of a difference, theoretically, the limited dependent variable can be dealt with using a log linear model. ln b_t satisfies quite different properties as opposed to just b_t, ln b_t has maximum and minimum values of $\infty \; and \; -\infty$, respectively and is, therefore no longer limited.

Revisiting Equation (10), it can be said that one unit increase in T_{t-9} will produce an expected increase in lnb_t by β_1 units. In terms of the detrended birth rate b_t, this means that the expected value of b_t is multiplied by e^{β_1}, where $e \approx 2.718$. Therefore, in terms of the effects of independent variables temperature (T) and precipitation (P) on the dependent variable b_t, it can be said that:

- Each unit increase in T or P multiplies the expected value of b_t by e^{β_n}, where β_n is the coefficient corresponding to the independent variable in Equation (10); and
- The effect of an m unit increase in T or P can be calculated by multiplying the expected value of b_t by $e^{m*\beta_n}$.

3.2. Geostatistical Interpolation Methods

In order to apply the above-mentioned statistical analysis technique to study birth seasonality and climatological phenomena, it was first imperative to determine the monthly average temperature and precipitation of the area of interest. Due to the absence of data, it was decided that geostatistical interpolation method of Kriging would be applied on data sets gathered from nearby weather stations to interpolate for Velke Pole.

In the following sections the essence of geostatistics and Kriging will be discussed to better under the method and its salient features.

3.2.1. Definition

In its most basic form, geostatistics is used to explore the spatial patterns of distribution and interpolate the value of a feature of interest at an unsampled location (Goovaerts, 2000a)

"Geostatistics offers a way of describing the spatial continuity of natural phenomena and provides adaptations of classical regression techniques to take advantage of the continuity" (Isaaks & Srivastava, 1989)

Spatial continuity, fundamentally, is the nature of data sets pertaining to earth sciences to continue over space and sometimes over time. So, it is usually the case that two sets of data close to each other will be similar and will have similar value as opposed to two data sets that are further away from each. Therefore, it can be said that this is the principle on which the subject called Geostatistics is based, whereby an array of statistical techniques has been brought together and used in conjunction to each other in order to fulfil its purpose.

3.2.2. Interpolation techniques: deterministic and geostatistical

The following interpolation techniques are elucidated adopting the categorization followed by ESRI documentation on Geostastical Analyst, it was so decided as ESRI's ArcGIS 10.3 is the software used to run the Kriging technique.

- Deterministic: These techniques create surfaces from measured points based on the extent of similarity between points or the degree of smoothing. Deterministic interpolation techniques comprise the following four methodologies:
    - Inverse Distance Weighted (IDW): Is one of the most popular interpolation techniques owing to its simplicity of implementation. It is based on the assumption that the likelihood of similarity between geospatial points that are close to each other is higher than points that are farther apart. The measured values closest to the prediction location, therefore, have the most influence. It assigns greater weight to points closest to the prediction location and the weights diminish as a function of distance. While IDW is quite simple to use, it does not provide prediction standard error on the predicted value(s) and this is considered as a major drawback of the technique.
    - Global Polynomial Interpolation (GPI): Fits a smooth surface that is defined by a polynomial function to given sample points. The result from a GPI is a smooth surface representing gradual trends over the area of study. It is most useful when the interpolation surface varies slowly from region to region in the area of interest. A first-order GPI fits a single plane through the data; a second-order GPI fits a surface with a bend in it, allowing surfaces representing valleys; a third-order allows for two bends. A GPI, therefore, is not able to interpolate surfaces with varying shapes.

- Local Polynomial Interpolation (LPI): In contrast to GPI, LPI fits many polynomials within specified overlapping neighbourhoods. To overcome the limitation of GPI, LPI uses varying polynomial order to interpolate points within each of the neighbourhoods. A criticism of LPI in the literature is that if the neighbourhood sizes are small then many empty areas can be created in the prediction.
- Radian Basis Functions (RBF): Are a special case of splines and defined as a collection of exact interpolation functions such as spline (variations), multi-quadratic function, inverse multi-quadratic function, etc. RBFs produce a smooth surface from a large number of data points. They produce good results for gently varying surface. RBFs are inadequate when the surface values change abruptly over short.

- Geostatistical: These interpolation techniques leverage statistical properties of the measured points and try to quantify the autocorrelation among them. Although there are different geostatistical interpolation techniques, kriging is the most popular and widely used technique. The advantage of kriging is that it not only produces surface predictions but also provide a measure of accuracy and errors in the predictions. Kriging assumes that the distance and direction between measured points reflect spatial correlation that can be used to compute the variation in surface. It is the preferred methodology when the measured data locations are evenly dense and uniformly distributed throughout the investigation area.

3.2.3. Comparison of geostatistical and deterministic interpolations

Goovaerts (2000a) argued that geostatistical interpolation techniques (Kriging and its family of variations) perform significantly better than their deterministic counterparts, because the latter ignore the pattern of spatial dependence for variables such as temperature, precipitation, and photoperiod. Furthermore, it is possible to correlate secondary information such as elevation to a kriging process in order to improve the prediction results. Deterministic interpolation techniques do not offer this range of flexibility.

Kriging also compensates the effort of "data clustering" which is a phenomenon when a concentrated region of points are assigned more weights than the general distribution of data points.

As outlined above, Kriging furthermore offers estimation error (kriging variance) in addition to the prediction itself. The estimation error can be used for stochastic simulation of the prediction value.

Based on these advantages, kriging has been chosen as the methodology of choice for geospatial interpolation in this contribution. In the following, a summary of the most important concepts of kriging is presented from Goovaerts (2000b).

3.2.4. Kriging methodology

Let $z(u_\alpha)$ be a set of observations at n different locations. The observation set can be formulated as:

$$\{z(u_\alpha),\ \alpha = 1, \dots, n\} \quad (15)$$

Then the estimated observation $z^*(u)$ for estimation point u can be defined as:

$$z^*(u) - m(u) = \sum_{\alpha=1}^{n(u)} \lambda_\alpha [z(u_\alpha) - m(u_\alpha)] \quad (16)$$

where,

- u is the location vector for the estimation point (expressed as spatial coordinates);
- u_α is the location vector of neighboring data point;
- $n(u)$ is the number of data points in the local neighbourhood;
- $m(u)$ is the expected mean of $z(u)$;
- $m(u_\alpha)$ is the expected mean of $z(u_\alpha)$;
- $\lambda_\alpha(u)$ is the kriging weight assigned to a data point $z(u_\alpha)$ for location u.

The goal of kriging is to find out the matrix of kriging weights that minimizes the variance of the estimator:

$$\sigma^2(u) = Var\{z^*(u) - z(u)\} \quad (17)$$

The kriging weights are derived from covariance function or semivariogram ($\hat{\gamma}(h)$) which outlines a measure of dissimilarity between observations. The semivariogram is formulated as half the average squared difference between a data pair separated by lag vector h:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{n=1}^{N(h)} [z(u_\alpha) - z(u_\alpha + h)]^2 \quad (18)$$

where $N(h)$ is the number of pairs of data locations at distance h apart and $z(u_\alpha + h)$ is the lagged version of the variable.

It is worth noting that the semivariogram function looks like a mirror opposite of the covariance function as illustrated in the Figure 9 below (Bohling, 2005):

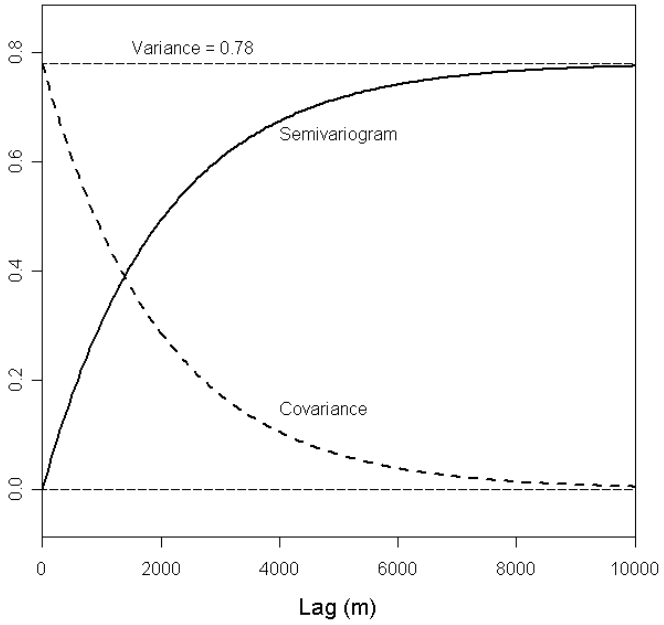


Figure 9: Relationship between semivariogram and covariance

The three important attributes derived from an estimated semivariogram graph (Figure 10) are:

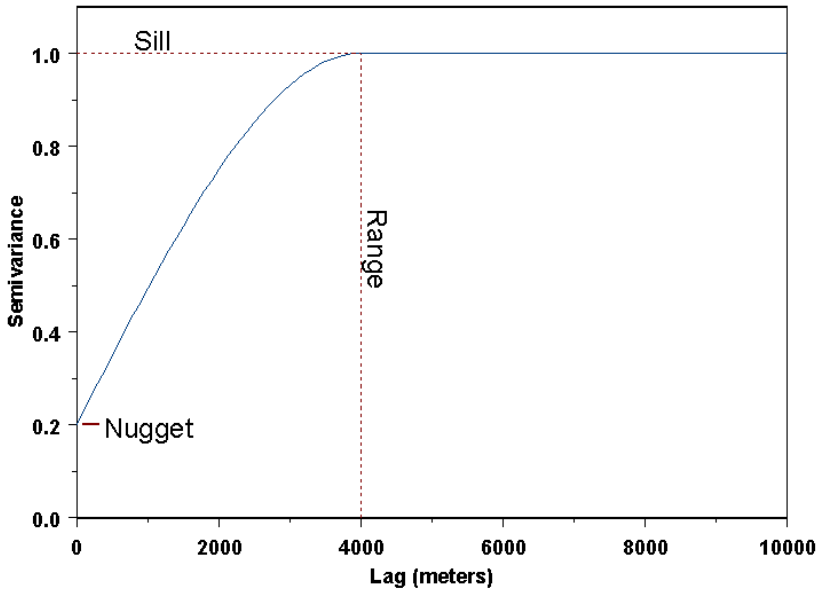


Figure 10: Semivariogram characteristics depicting sill, range, and nugget

- Sill: The Y-axis value where the semivariogram reaches a plateau;
- Range: The lag distance where the semivariogram reaches the sill value; and

- Nugget: The semivariogram value at origin (0 lag). This value should theoretically be zero.

Once the estimated semivariograms are generated for a particular kriging dataset, the next logical step is to model the semivariogram. The reason is that the kriging function will have to use a stable semivariogram model for all possible lag vectors among data points and that the model has to obey certain properties in order for the equation to be solvable.

Some frequently used models are:

Nugget:

$$g(h) = \begin{cases} 0 & if\ h = 0 \\ c & otherwise \end{cases} \quad (19)$$

Spherical:

$$g(h) = \begin{cases} c * \left(1.5\left(\frac{h}{a}\right) - 0.5\left(\frac{h}{a}\right)^3\right) & if\ h \leq a \\ c & otherwise \end{cases} \quad (20)$$

and, Gaussian:

$$g(h) = c * \left(1 - exp\left(\frac{-3h^2}{a^2}\right)\right) \quad (21)$$

where, h is the lag vector, a is the range, and c represents sill. A comprehensive treatment of all semivariogram models is presented in Goovaerts (2000a).

The next chapter will demonstrate the application of these statistical techniques for analysing the relationship between monthly birth figures in Velke Pole with temperature and precipitation.

Chapter 4: Methodology & Implementation

This chapter deals with the steps involved in planning the study, executing it, analysing and interpreting the results. The emphasis is on the implications of decisions made in the planning of the research for subsequent analysis and elucidation of results.

4.1. Research Design

Empirical research is undertaken to answer questions that often take the form of whether and to what extent several variables for interest are related. To begin formulating this research it was necessary to define the kind of study based on the research question.

4.1.1. Approach

The thesis follows a historical correlational design where the relationship between our independent and dependent variables without manipulating them, in their natural state, was examined. This kind of research can be useful to predict future events or explain present conditions.

This work falls in the category of what is known as observational research as opposed to experimental research. In an observational research, the independent variables are not manipulated but observed, and their effect on the dependent variable is analysed. The dependent variable is the measure of change in the independent variable. The independent variable causes change in the dependent variable and not the other way round.

4.1.2. Data Collection

In this study, there are three independent variables, out of which temperature and precipitation are continuous variables and the month is a categorical variable for which dummies or indicators are used. The dependent variable is birth rate.

The study is also conducted on data over 120 years starting from 1781 to 1900 therefore it falls in the domain of time series problem. The birth data is available from 1758 to 1900 and a decision was made to start the research from the turn of the 18th Century and continue until the end of 19th Century mainly because the data between 1958 and 1980 is the combined data for Velke Pole and Pila.

As introduced in the last chapter, geostatistical interpolation technique of Kriging was applied to determine the temperature and precipitation data of Velke Pole between 1860 and 1900 using nine

weather stations around the area of interest; they were Hohenpeissenberg, Regensburg and Augsburg in Germany, Vienna Hohe-Warte and Linz in Austria, Bratislava and Oravsky Podzamok in Slovakia, Budapest-Lorinc in Hungary and Brno-Turany in Czech Republic (see Figure 11).

For the years between 1781 and 1859 the temperature and precipitation data were extrapolated based on the kriging data and data for Vienna and Budapest, because recorded data was available for only these stations for the aforementioned years.

4.1.3. Data Sources

In order to go ahead with this thesis, it was necessary to collect the data on which statistical techniques would be applied, so that one may reveal vital information regarding the study area. The following data sources all provided data sets in digital format, either in comma separated value (CSV) or Microsoft Excel format (XLS).

- The primary data source used for this thesis is the church book or parish register of Velke Pole, which by rule recorded the number of baptisms performed. This baptism data is considered equivalent to birth data as is evident from the literature review that had also used parish register records as birth rates. There is enough evidence that babies were baptized as soon as possible from the following citation found in a prayer book in England around the 1500s, "The pastors and curates shall oft admonish the people that they defer not the Baptisme of Infants any longer than the Sunday, or other Holy day next after the child be borne, unless upon a great and reasonable cause declared to the Curate." (Basten, 2015).

- To study the influence of environmental factors, average monthly temperature and precipitation data were particularly necessary. Such data for Velke Pole did not exist; therefore, data was collected from cities as close to the position of the study area as possible. The temperature and precipitation data were accessed through HISTALP (Historical Instrumental Climatological Surface Time Series of the Greater Alpine Region) (HISTALP, 2015). This initiative started within the climate division of the Austrian weather service (ZAMG). Data from stations in Germany, Hungary, Czech Republic, Slovakia, and Austria have been collected to perform Kriging in order to interpolate the data for Velke Pole. The Period for which the data was available for the all the necessary weather stations was 1860 to the present time. To cover for the rest of the years previous to 1860 extrapolation methods have been performed.

Figure 11: Weather stations around Velke Pole

- For the abovementioned purpose it was necessary to collect data from as many stations as possible, to this end the meteorological services of Slovakia, Hungary and Poland were contacted, out of which only Slovak Hydrometeorological Institute (SHMU, 2015) could provide data for a weather station (Oravsky Podzamok) which would prove to be most useful.

4.2. Data Preparation

Although there are a wide range of programming languages and stand-alone tools available for data analysis, it has been assessed that Python offers the most number of advantages and flexibility when it comes to comprehensibility of code, ease of use and availability of additional libraries. Most notably Python solves the "two-language" problem, which can be summarized as follows (McKinney, 2014): In many organizations, it is a common practice to research, develop prototype and testing ideas using more statistical language and tools such as R and MATLAB and then later transform the algorithms in more web-friendly languages such as Java or .NET for visualization. Python enables both explorative development of data analysis algorithms (because it is interpretive in nature) and also has a good suite of extension libraries for visualization of results.

As stated above, the data sources were available in Comma Separated Value (CSV) or Microsoft Excel (XLS) format. All data preparation (cleaning, munging, reshaping) steps were performed using Python and its suite of libraries. Some of the repeatable data preparation steps and their code snippets are detailed below:

4.2.1. NumPy and pandas

pandas (citation) is an open source library for data preparation, analysis and modelling. It provides rich data structures and functions for efficiently working with structured data. It provides a concept called DataFrame, which is also a common feature in R, as a rich data type to hold two dimensional tabular column oriented data structure. Through this abstraction, pandas is able to perform aggregation and subset selection of data in an efficient manner without requiring a relational database to store the data.

Just like relational databases, a DataFrame can consist of a tabular data structure containing an ordered collection of columns (McKinney, 2014). Each column can have a different data type. Furthermore, a DataFrame has both a row and a column index through which data can be queried and referenced. Since analysis was done on a modern Intel Quad-Core machine with 16 GB of RAM, it was possible to buffer DataFrames and perform all computations in-memory.

The high-level data manipulation functions in pandas are built on top of another library called NumPy, which stands for numerical Python and is a package for scientific computing. Amongst other things NumPy offers: fast array processing capabilities; functions for performing element-wise computations with arrays or mathematical computations between arrays; tools for reading and

writing large data sets; and linear algebra operations, Fourier transform and random number generation.

The first step towards using pandas is to import the library in a Python program as shown below:

```
import pandas as pd
```

After which the variable **pd** contains a reference to the pandas library. In order to load an entire dataset of births corresponding to the Velke Pole village, a statement similar to one below has to be executed:

```
df = pd.read_csv('../data/VPBirth.csv', sep=";")
```

where **df** refers to a DataFrame variable containing the tabular values of the csv file named **VPBirth.csv** where the entries are delimited by semicolon.

It is worth noting that the pandas library offers a set of convenience functions to read (write) data from a wide range of source (sinks). **read_csv** is once such convenience function which enables syntactic parsing of csv files. It also supports I/O to Microsoft Excel (via **read_excel**), sql data sources/sinks, json, html, and Hadoop file system.

An important step of explorative data analysis is to inspect the data at every step of processing. pandas provides a convenient function to inspect the first few rows of the DataFrame through the **head()** method:

```
print df.head()
```

will result in the following output (top 5 rows of the Velke Pole birth numbers):

Table 1: Velke Pole birth DataFrame

	Year	1	2	3	4	5	6	7	8	9	10	11	12
0	1758	13	15	14	5	5	5	2	4	7	13	9	11
1	1759	11	10	7	4	3	6	11	5	11	15	10	13
2	1760	19	15	18	5	13	6	10	9	13	14	8	13
3	1761	14	16	14	9	9	11	10	10	12	14	15	20
4	1762	13	12	12	8	5	3	8	10	6	10	10	9

For table manipulation (joins, transforms) however, it is not helpful if entries in the DataFrame are not indexed. For all analyses in this research work, the year was used as the index. Setting an index in pandas is quite straightforward:

```
df.set_index('Year')
```

When `print df.head()` command is again executed, the output this time will be:

Table 2: DataFrame indexed on Year

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Year												
1758	13	15	14	5	5	5	2	4	7	13	9	11
1759	11	10	7	4	3	6	11	5	11	15	10	13
1760	19	15	18	5	13	6	10	9	13	14	8	13
1761	14	16	14	9	9	11	10	10	12	14	15	20
1762	13	12	12	8	5	3	8	10	6	10	10	9

After this step, the DataFrame **df** can be used for set-theoretic operations such as transpose:

```
df1 = df.transpose()

df1.head()
```

would result in the following output (truncated):

Table 3: Transposed DataFrame

Year	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767
Jan	13	11	19	14	13	8	16	13	13	13
Feb	15	10	15	16	12	19	13	18	20	4
Mar	14	7	18	14	12	12	16	9	17	9
Apr	5	4	5	9	8	4	8	15	7	7

Although setting index is the first step towards DataFrame manipulation, it is still not sufficient for performing element-wise operation on the data. This is illustrated by two data manipulation queries that have been used quite frequently for calculating Lam & Miron's detrended birth rate (Equation (9) of Chapter 3).

The first is the numerator, which denotes monthly births divided by the total number of days in that month. The second relates to a centred moving average on the monthly birth rates.

For both these queries, the first requirement is to **stack** the DataFrame using the statement:

```
result = df.stack()
```

Which results in a pivot of the column labels resulting in a **result** DataFrame having a hierarchical index with a new innermost level of row labels. The above statement will result in:

```
     Year
Jan  1758    13
     1759    11
     1760    19
...

Dec  1871    15
     1872    11
     1873    15
...
```

Once this is done, it is relatively straightforward to implement the two queries. The division of monthly births divided by the number of days in that month can be achieved in the following few steps:

```
dates = pd.to_datetime([' '.join(item) for item in result.index])
```

Joins each month of each year into a comma-separated DataFrame **dates** where each element is of data type **date**:

```
(['1758-01-09', '1759-01-09', '1760-01-09', '1761-01-09',
  '1762-01-09', '1763-01-09', '1764-01-09', '1765-01-09',
  '1766-01-09', '1767-01-09', … ])
```

The DataFrame **dates** could then be parsed by the following higher-level function to reveal the number of days in the month:

```
days = dates.days_in_month
```

Finally, the numerator of Lam & Miron's detrended birth rate formula can be obtained from:

```
result = (result / days).unstack()
```

The final **result** DataFrame corresponding to numerator looks like (truncated for brevity):

Table 4: Truncated numerator of detrended birth rate formula of Lam & Miron

Year	**1758**	**1759**	**1760**	**1761**	**1762**
Jan	0.419355	0.354839	0.612903	0.451613	0.419355
Feb	0.535714	0.357143	0.517241	0.571429	0.428571
Mar	0.451613	0.225806	0.580645	0.451613	0.387097

The denominator of the detrended birth rate formula could be obtained by applying pandas' higher-order function of centred rolling mean:

```
movAvg = pd.rolling_mean(result.stack(), window=12, center=True)
```

Which gives the following result:

```
     Year
Jan  1758         NaN
     1759         NaN
     1760         NaN
     1761         NaN
     1762         NaN
     1763         NaN
     1764    0.424731
     1765    0.422043
     1766    0.440860
     1767    0.413978
     1768    0.424731
     1769    0.448925
     1770    0.473118
…
```

It is worth noting that the first six and last six elements of the **movAvg** DataFrame are NaN (Not a Number) since the six predecessor and six successor values are not present in the original dataset.

4.2.2. Python Data Visualization

Once the explorative analysis of data is completed a set of 2-D charts were produced via the library

Matplotlib, which is good for making two- and three-dimensional plots of arrays. Matplotlib is written in Python and makes heavy use of NumPy. It emulates the graphic commands of MATLAB and lets Python programmers create simple plots with few commands.

Matplotlib charts can be triggered inline through Python function calls. An example below, which illustrates the plot of the variation in moving average values in Velke Pole birth rates is show below:

```
result[['1769','1770','1771','1800']].plot()
```

Produces the chart in Figure 12.

In the subsequent sections, the detailed methodology of geospatial interpolation using Kriging and regression analysis is provided. Both methodologies rely on processed data through these data preparation and exploration steps.

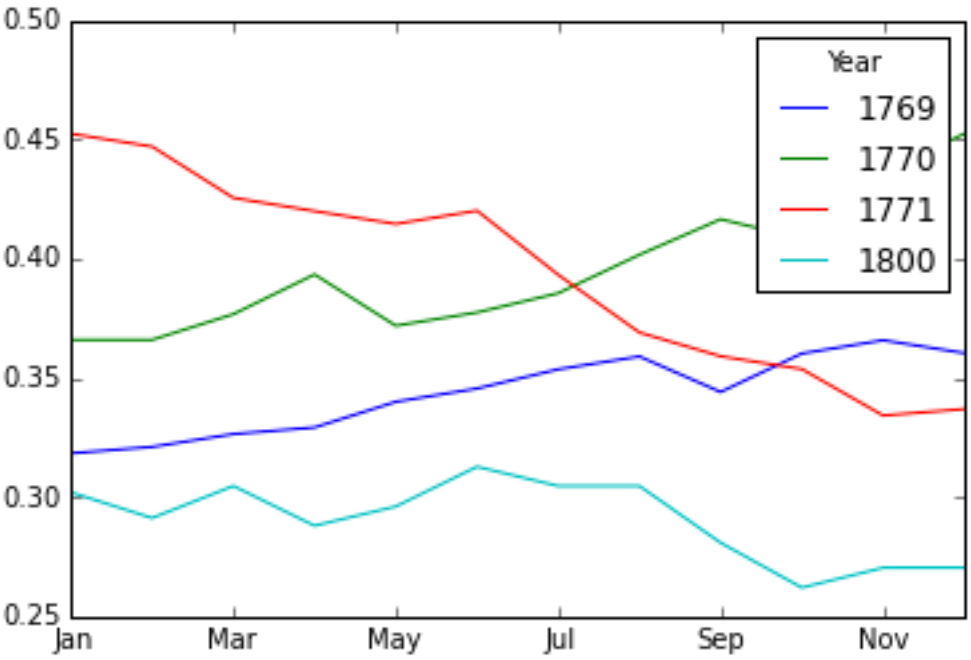


Figure 12: Moving average variation of four years based on birth data

4.3. Spatial Interpolation and Extrapolation Methodology

4.3.1. Kriging

Separate CSV files were created for each year containing temperature or precipitation data, X-Y coordinates and elevation in meters of all the chosen weather stations. Once the CSV files were prepared, they were imported into ArcMap and point features were created from the coordinate

information in the files. The projection system used is WGS 84 Web-Mercator with spatial reference as EPSG 3857 used widely in web-based mapping programs such as Google and OpenStreetMap.

After the features have been drawn they are exported as shapefiles to have objectIDs assigned to the attribute tables, so that they can be edited in the future. To the point features a map of countries were added, they were extracted from TM-World-Border shapefile, downloaded from Sandvik (2015). The borders of countries of interest were selected and exported as a separate layer, the projection system matched to that of the weather station point feature.

Once the data has been prepared geostatistical analyst has to be activated in ArcMap to run the program.

The first step is to explore the data. Histograms and QQ plots can be used to check for normality as well as the best transformation fit, one can choose between log and box-cox to make the model normal, in case they are skewed. In this thesis, one of the assumptions is that the data is normal so this step has been skipped, though one of the other data exploration methods to identify global trends was performed. It was recognised that the trend would be best fit by a second-order polynomial; however, the trend removal method did not work for the data set, probably because of the very small size of sample, which is also the reason why no outliers were removed.

The following steps describe the workflow:

1. The process begins with choosing the data set, in this case it is the year and the data field is the month. For each year, the process was run twelve times for all the twelve months.
2. The second step consists of selecting the Kriging type; in this incident, it was Ordinary Kriging.
3. In this step the semi-variogram modelling can be specified, out of Gaussian, spherical etc. Stable model was chosen for this case. Other settings such as nugget calculation can also be specified in this case, it was set to default for this research (Figure 13).
4. In this step, adjustments can be made to the neighbourhood search field, the radius could be changed and weights could be assigned to the data points. Again, here everything was set to default. In this step, by entering the coordinates of the area of interest it is possible to extract the predicted value (Figure 14).
5. The last step is the cross validation step. Here the actual values and the predicted values can be compared. The values calculated on the prediction errors give an indication on whether

the model is appropriate for making maps or predictions. To judge if a model provides accurate predictions, the following verification steps can be taken (Figure 15):

• The predictions are unbiased, indicated by a mean prediction error close to 0.

• The standard errors are accurate, indicated by a root-mean-square standardized prediction error close to 1.

• The predictions do not deviate much from the measured values, indicated by root-mean-square error and average standard error that are as small as possible.

The Cross Validation dialog box also allows you to display scatterplots that show the error, standardized error, and QQ plot for each data point (ESRI ArgGIS, 2015).

6. At the end of this workflow a raster surface is created only for the extent of the data points, it is expanded to fit the extent of the layer containing the countries and clipped to its shape. (See temperature prediction map of Figure 16 and precipitation prediction map of Figure 17).

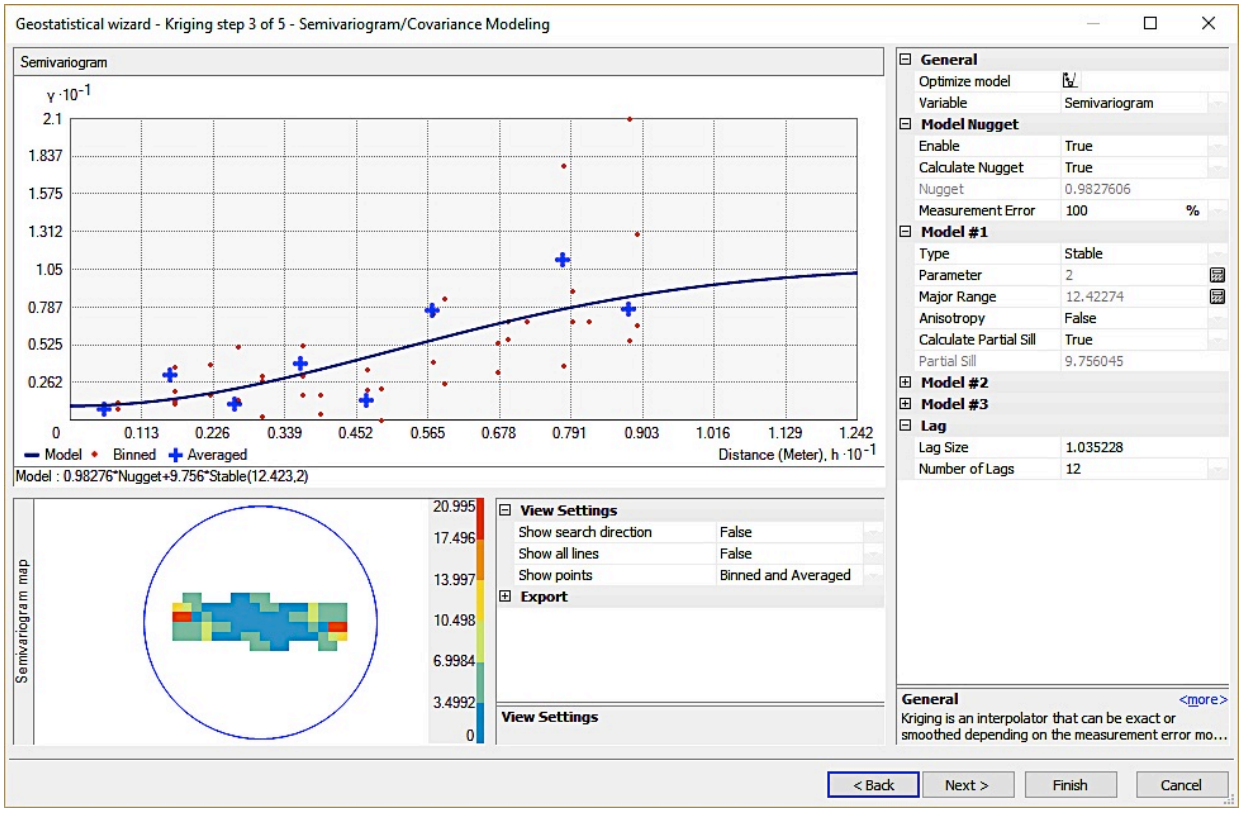


Figure 13: Semivariogram modelling for temperature prediction for May 1880

Ordinary Kriging was preferred over other Kriging types, because it is a model that has flexibility and can be used for different applications, though it requires proficiency and time to achieve a good end result. This model is deemed appropriate for use with datasets that are scant, so it seemed apposite that it be used for this study.

Cokriging is another method that was experimented with; it is similar to ordinary kriging except that it uses supplementary variables in order to improve accuracy. Elevation values were used s auxiliary variable for this research as it is considered an important factor in modelling temperature and precipitation predictions, but the method may or may not increase accuracy (Burrough & McDonnell, 1998), and indeed in this case, it did not increase accuracy as was evident from cross validation statistics, and therefore abandoned.

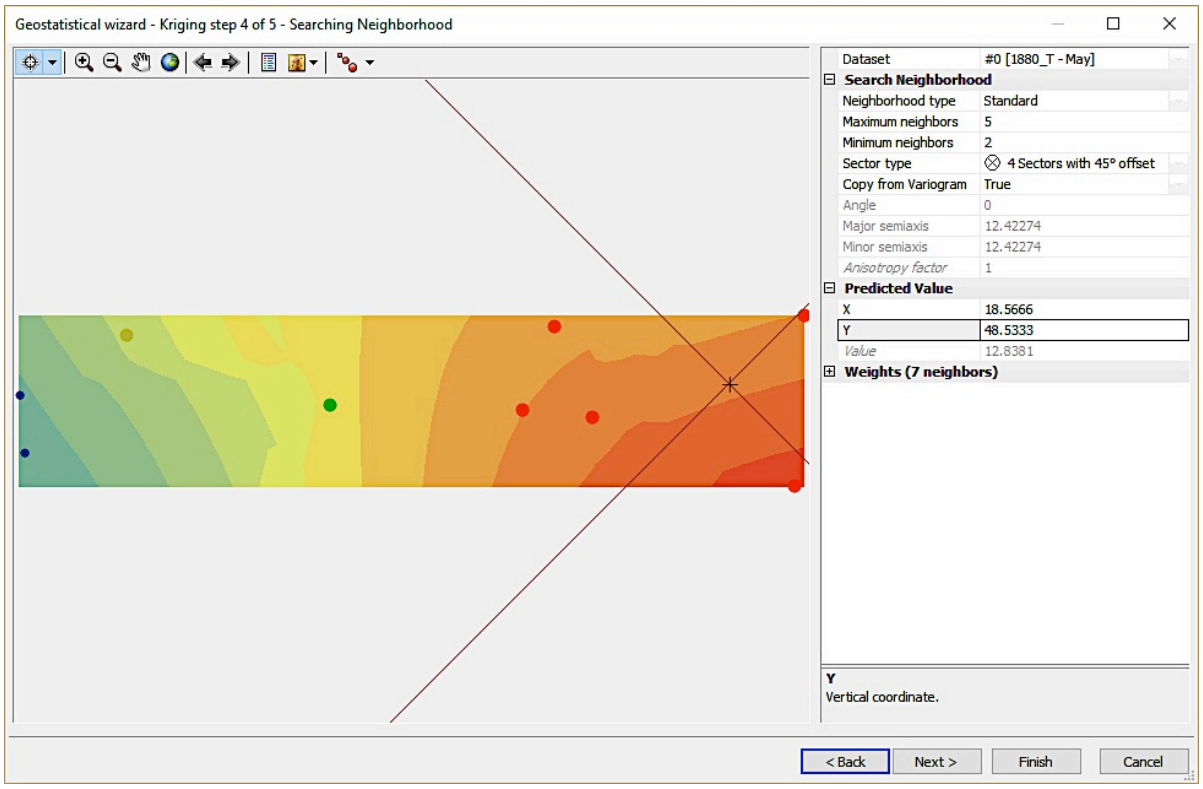


Figure 14: Neighbourhood search parameters and prediction value for specific X-Y coordinates

The geostatistical layer, as the raster surface is referred to by ArcGIS, is represented as filled contours in the case of the temperature and precipitation prediction maps; in this display it is assumed that points located within a category will have the same values.

The data classification scheme according to which the contour categories are classified is geometric interval. This algorithm is aimed at accommodating continuous data, therefore apt for use in this situation. "This classification scheme creates class breaks based on class intervals that have a geometrical series. The geometric coefficient in this classifier can change once (to its inverse) to optimize the class ranges. The algorithm creates geometric intervals by minimizing the sum of squares of the number of elements in each class. This ensures that each class range has approximately the same number of values with each class and that the change between intervals is fairly consistent." (ESRI ArcGIS, 2015).

The prediction values that make up the temperature and precipitation data for Velke Pole between 1860 and 1900 are extracted manually by entering the coordinate location in the dialogue box at step 4 after running the diagnostics in step 5.

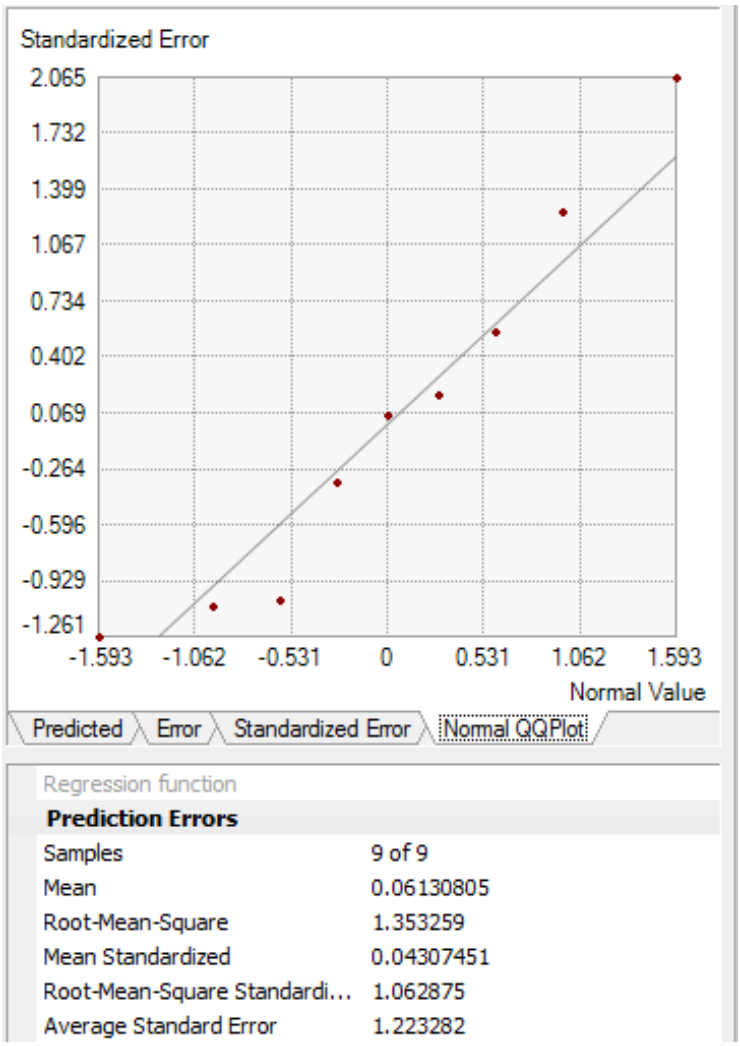


Figure 15: Cross-Validation diagnostics for temperature predictions for October 1870

4.3.2. Statistical Extrapolation

After the Kriging data is generated for every month of each year, it is assumed this is the accurate data for Velke Pole between 1860 and 1900; this is then used in conjunction with temperature records from Budapest-Lorinc and Vienna Hohe-Warte to extrapolate for data between 1781-1859.

Out of the nine weather stations, only Vienna and Budapest had records for the above years and a methodology was developed for data extrapolation, which is elucidated below: (H. Saler, email communication, October 2015)

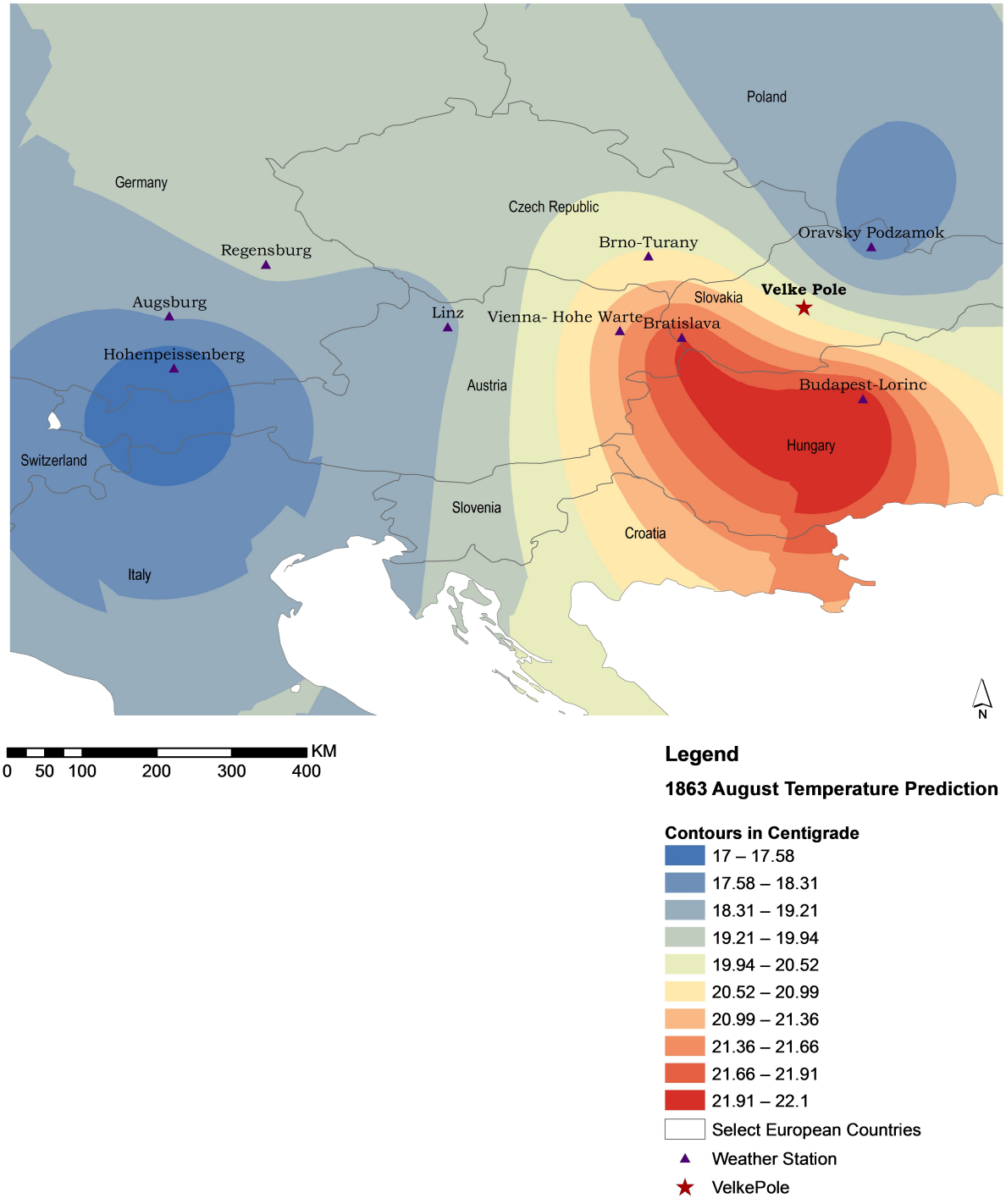


Figure 16: Temperature prediction map of August 1863

1. Mean temperature values were calculated for Vienna and Budapest from 1860 to 1900.
2. The Kriging values were subtracted from the mean temperature values indicating temperature for Velke Pole for the same years.
3. Twelve graphs were plotted for the values of the difference for each month over the years. A trend line was drawn through each graph to determine the regression coefficient (Figures 19 & 20).

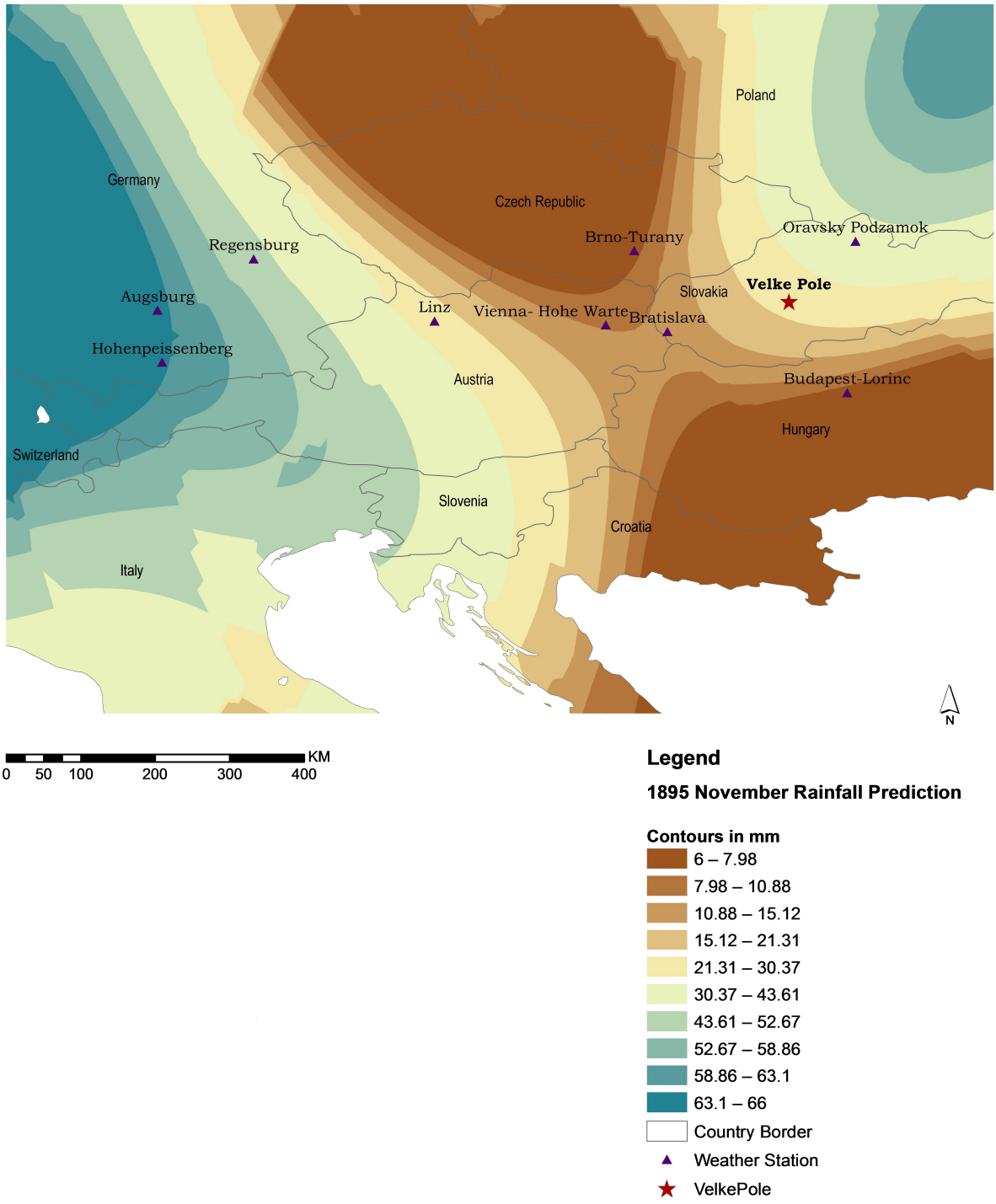


Figure 17: Precipitation prediction map of November 1895

4. The regression coefficients (R-square) were tested for significance with F-Test, they were found to be not significant (see Table 6 and Figure 21). The F-statistic was calculated using the following formula:

$$F = \frac{R^2}{1 - R^2} \times (n - 2)$$

5. The decision was taken to consider constant corrections in the form of the mean difference values for each month to make a homogenous model.
6. The mean difference values were adjusted with the mean temperature values of Vienna and Budapest for the years 1781-1859 to generate temperature data for Velke Pole.

The steps are outlined in the flowchart of Figure 18.

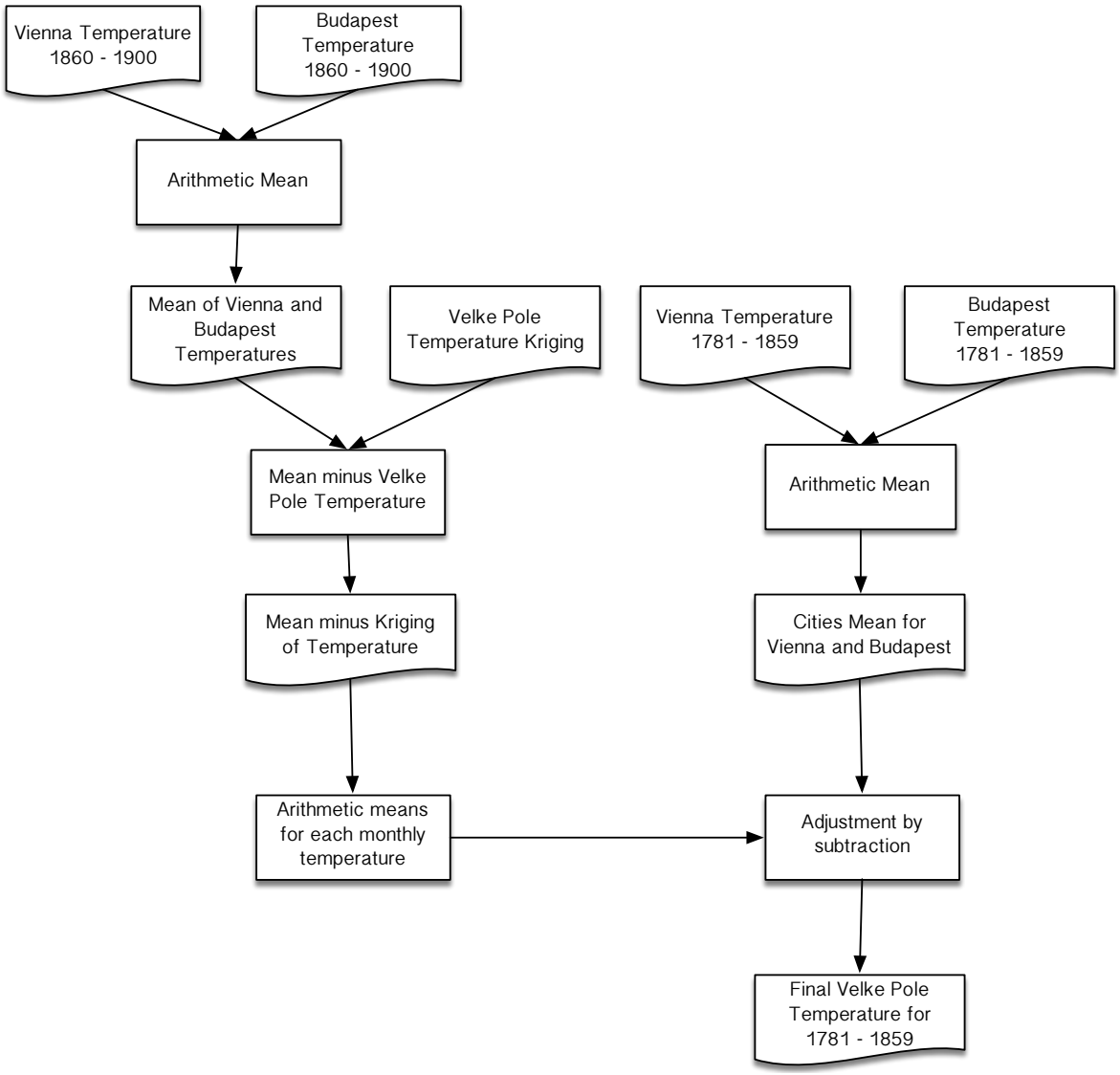


Figure 18: Flowchart describing steps for temperature extrapolation

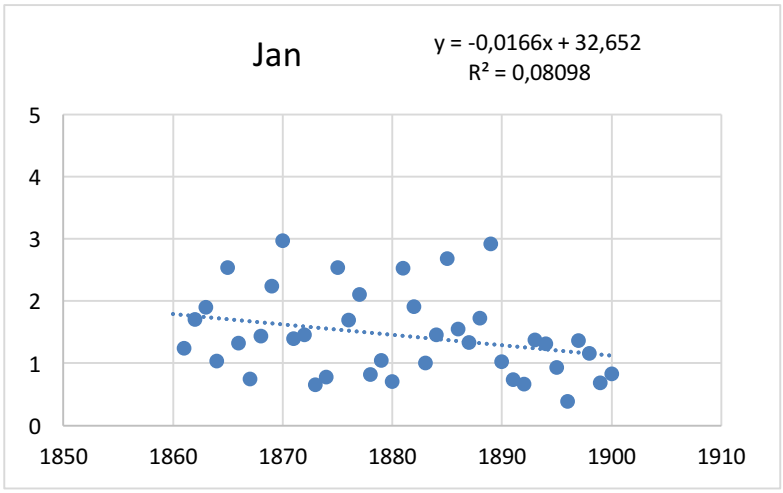


Figure 19: Difference plot with trendline for January 1860 – 1900

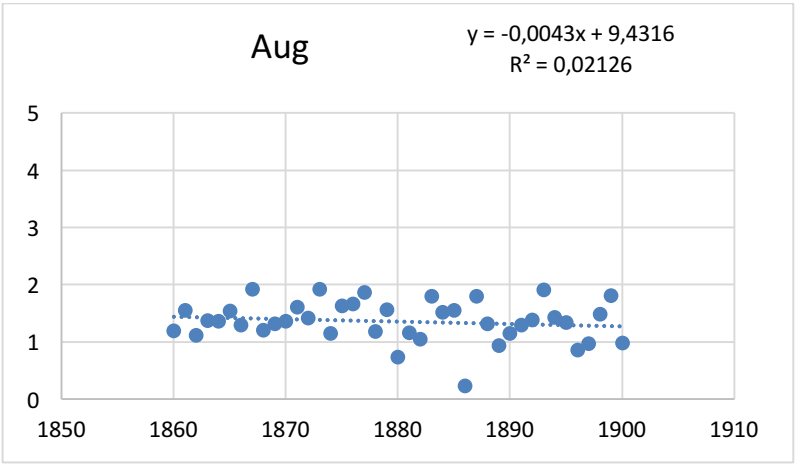


Figure 20: Difference plot with trendline for August 1860 – 1900

The first two steps were implemented in Python using the pandas library. The concept of DataFrame, as explained earlier in the chapter, has been used to load the data from CSV files and buffer in the memory for further computation.

```
vienna=pd.read_csv('HISTALP_AT_WIE_T01_1860_1900.csv', sep=";")
```

```
budapest=pd.read_csv('HISTALP_HU_BUD_T01_1860_1900.csv', sep=";")
```

vienna and **budapest** are two DataFrames that store the temperature data of the cities of Vienna and Budapest respectively. As explained earlier, computation on the DataFrames can only be performed once they are indexed. For that, the **set_index** method of pandas was called:

```
vienna = vienna.set_index('year')
```

```
budapest=budapest.set_index('year')
```

As in relational databases, in order to do pairwise arithmetic mean by years, the two indexed DataFrames were first concatenated:

```
cities_concat = pd.concat((vienna, budapest))
```

Which results in a union DataFrame called **cities_contact** containing yearly pairwise values. Another intermediate step is to perform **groupby** operation on this DataFrame to club together entries of corresponding years:

```
by_row_index = cities_concat.groupby(cities_concat.index)
```

The final step is to use Python's **mean()** function to compute the means of the **by_row_index** dataframe:

```
cities_means = by_row_index.mean()
```

The **cities_mean** dataframe contains the result of the arithmetic means of the two cities. A `print cities_means.head()` statement reveals:

Table 5: cities_mean DataFrame

year	jan	feb	mar	apr	may	jun	jul	aug	sep
1860	0.8	-0.85	2.55	10.00	16.05	18.80	17.70	19.20	15.95
1861	-4.5	3.05	5.15	7.85	12.15	19.85	20.50	21.35	16.40
1862	-3.3	-0.80	6.60	12.75	16.60	18.10	20.65	19.15	16.65
1863	2.0	2.15	6.75	9.50	17.00	18.85	19.95	21.70	17.05
1864	-7.7	-0.50	6.10	6.85	12.15	18.70	18.20	17.35	15.50

Once this is obtained, the process reading the table prepared Kriging and subtracting it from the arithmetic means of Vienna and Budapest form are straightforward:

```
kriging=pd.read_excel('VelkePole_Kriging_T.xlsx')
```

```
difference = two_cities_means – kriging
```

The DataFrame **difference** is then exported as a Microsoft Excel spreadsheet for further processing.

Table 6: R-Square significance test

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
R-Square	0.0938	0.0006	0.0255	0.0868	0.2699	0.0341	0.0094	0.0213	0.0080	0.1029	0.0625	0.0035
F- Statistic	4.0365	0.0226	1.0197	3.7063	14.4184	1.3773	0.3688	0.8472	0.3148	4.4737	2.5983	0.1357
F(1, 39) α = .05, .01	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33	4.09, 7.33
Significance	not significant	not significant	not significant	not significant	99% significant	not significant	not significant	not significant	not significant	95% significant	not significant	not significant

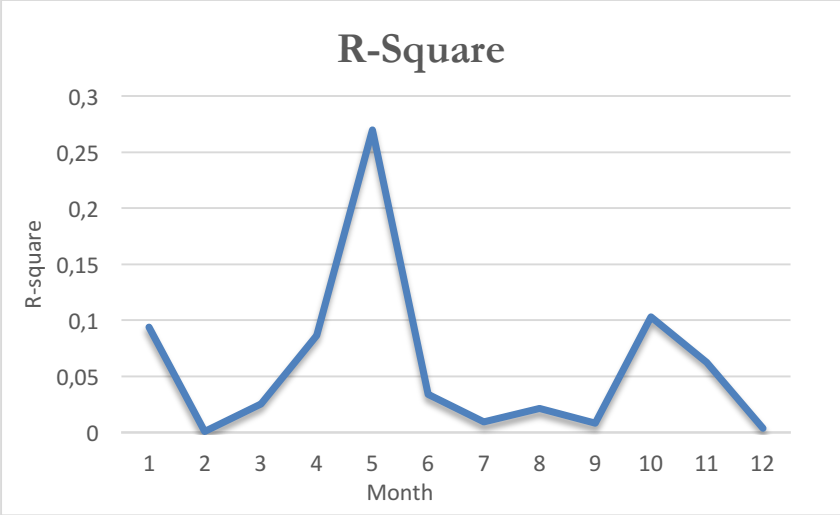


Figure 21: R-Square plot

4.4. Regression Analysis

The past chapters and sections have described the methods and processes necessary to reveal the pattern of birth seasonality and causality, this section is a culmination of those processes and will demonstrate the step by step procedure of the regression analysis to draw inference based on the data, which has been collected, manipulated and plotted.

This section is divided further into subsections that look into, detrended monthly birth, multiple regression analysis method and the results of the analysis.

4.4.1. Detrended Monthly Birth

Adhering to the detrending methodology as described by Daniel A. Seiver (1985), it is used to reduce or remove any trend contamination in the seasonality using a 12 – month centred moving average and it is normalized to a hundred percent. To implement using the pandas library of Python, a simple DataFrame named **Ri** is defined to store the percent detrended births as follows:

```
Ri = (birth/moving_avg) * 100
```

Where, **birth** and **moving_avg** are DataFrames containing the values of monthly births for Velke Pole and a centred moving average of the monthly births respectively.

The resulting plot between years 1758-1900 is shown below:

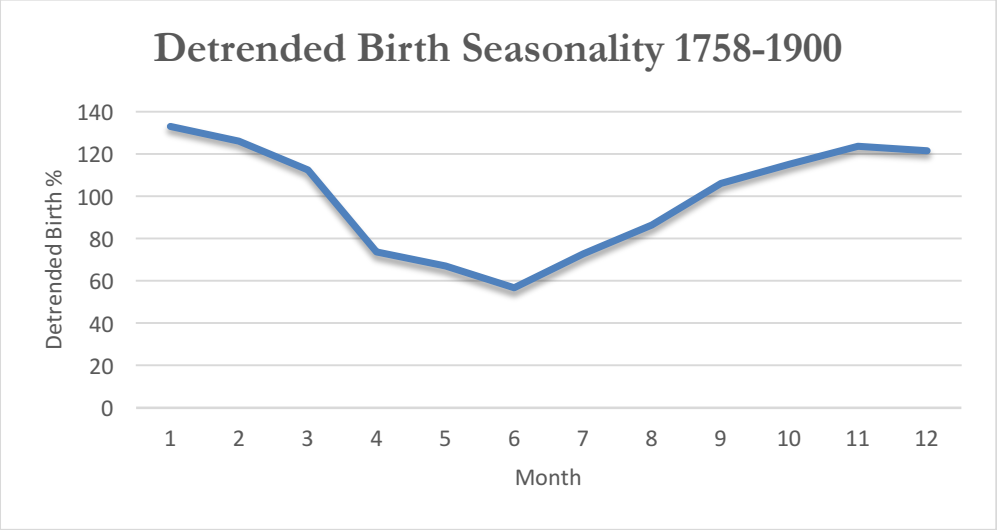


Figure 22: Average birth seasonality for Velke Pole, 1758-1900

The above plot clearly shows a dip in birth in June and peaks in November to January. This, of course, is the average plot for almost 150 years, so the years have been, eventually, grouped together

by decades to look at the seasonality more meticulously and to see if there have been any changes in the pattern over the years.

The decade-wise break down of the seasonality pattern and the regression results with and without controlling for temperature and precipitation are discussed in the next subsections.

4.4.2. Multiple Regression Analysis Method

The regression analysis was applied using the Lam & Miron (1996) approach as described in Chapter 3 using Equations 9 and 10.

The detrended birth values are calculated using average daily births in a month as opposed to average monthly births used by Seiver (1985) in the previous section. Section 4.3.1 describe the procedure for calculating the average daily birth in a month using Python. The detrended birth values are logarithmically transformed using natural log, this value is referred to as lnbt by Lam & Miron and it is the dependent variable in the regression analysis.

The dummy variables are assigned to each 11 months instead of 12 in keeping with the *n-1* rule for regression models that contain the intercept, which has been used in this regression analysis.

For the years between 1781-1859, the regression is applied for the dummies with and without controlling for temperature alone, as can be seen in the data arrangement in Table 7. Table 8 demonstrates the arrangement of the data for the years 1860-1900 for which precipitation values were also available and the regression is performed with controls for temperature, as well as precipitation.

Explanations of the data columns of Tables 7 and 8:

- **Year** – The year for which the analysis is performed
- **Month-Birth** – The month of birth
- **Birth-9 Months** – The month lagged 9 months to the month of birth to represent the month of conception
- **ln(Birth)** – This denoted the detrended and logged daily birth for a month as described equation 10
- **Temp** – This column contains the temperature corresponding to the conception month (Birth-9 Months), it is one of the independent variables in degree centigrade

- **TempSq** – This column denotes the square of the temperatures and is used in order to model curvature when relationships are thought to be non linear. It also magnifies effects of extreme temperatures
- **Jan-Nov** – Columns denoting the dummy assignments, December is indicated by the intercept, as no dummy value has been assigned to it
- **Rain** - This column represents precipitation in millimetres for month of conception (Birth – 9 Months) and is one of the other independent variables
- **RainSq** - Like TempSq, it is used to model non linear relationships and highlight any change in extreme high or low precipitation

Tables 7 and 8 contain temperature and precipitation in degrees centigrade and millimetre, respectively. Before running the regression, the above data was further processed to obtain zero-mean normalised coefficients in the results. This was achieved by standardising or normalising the dependent and independent variables before applying regression analysis.

The process of standardisation of coefficients is attained using the following two steps:

1. Calculate the mean and standard deviation of the variables.
2. Create a new standardized version of each variable. To get it, create a new variable in which you subtract the mean from the original value, then divide that by its standard deviation.

This process is carried out in order to get variances for both dependent and independent variables as 1 and the change can be expressed as a change in 1 standard deviation as opposed to the unit of measurement.

Following the above steps the regression analysis is executed in Microsoft Excel using the data analysis tools plugin. Running this generates a worksheet containing regression statistics, analysis of variance and predicted values and many more such information. An example of the worksheet for 1791-1800 with controls for temperature is illustrated by the following figure 23.

Table 7: Data arrangement for a year with temperature as independent variable

Period	Year	Month-Birth	Birth-9 Months	ln(birth)	Temp (°C)	TempSq	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
1	1781	Jan	Apr	0.070215983	6.366097561	40.52719816	1	0	0	0	0	0	0	0	0	0	0
2	1781	Feb	May	0.177601531	14.12268293	199.4501731	0	1	0	0	0	0	0	0	0	0	0
3	1781	Mar	Jun	-0.175788255	16.07341463	258.354658	0	0	1	0	0	0	0	0	0	0	0
4	1781	Apr	Jul	-0.073998245	18.32731707	335.8905511	0	0	0	1	0	0	0	0	0	0	0
5	1781	May	Aug	-0.5511332	18.34195122	336.4271745	0	0	0	0	1	0	0	0	0	0	0
6	1781	Jun	Sep	-0.518343377	13.62365854	185.6040719	0	0	0	0	0	1	0	0	0	0	0
7	1781	Jul	Oct	-0.244614141	9.487073171	90.00455735	0	0	0	0	0	0	1	0	0	0	0
8	1781	Aug	Nov	0.15001979	3.195609756	10.21192171	0	0	0	0	0	0	0	1	0	0	0
9	1781	Sep	Dec	-0.126929225	-3.58097561	12.82338632	0	0	0	0	0	0	0	0	1	0	0
10	1781	Oct	Jan	-0.410663215	-3.684146341	13.57293427	0	0	0	0	0	0	0	0	0	1	0
11	1781	Nov	Feb	0.484006239	-1.390487805	1.933456336	0	0	0	0	0	0	0	0	0	0	1
12	1781	Dec	Mar	0.206093958	3.946097561	15.57168596	0	0	0	0	0	0	0	0	0	0	0

Table 8: Data arrangement for a year with temperature and precipitation as independent variable

Period	Year	Month-Birth	Birth-9 Months	ln(birth)	Temp (°C)	TempSq	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Rain (mm)	RainSq
1	1861	Jan	Apr	0.39677284	8.21	67.4041	1	0	0	0	0	0	0	0	0	0	0	70.23	4932.3
2	1861	Feb	May	0.32661418	14.45	208.8025	0	1	0	0	0	0	0	0	0	0	0	61.05	3727.1
3	1861	Mar	Jun	0.73746818	17.65	311.5225	0	0	1	0	0	0	0	0	0	0	0	69.04	4766.5
4	1861	Apr	Jul	-0.3701324	16.64	276.8896	0	0	0	1	0	0	0	0	0	0	0	62.62	3921.3
5	1861	May	Aug	-1.0392105	18.01	324.3601	0	0	0	0	1	0	0	0	0	0	0	56.54	3196.8
6	1861	Jun	Sep	-2.1350386	14.43	208.2249	0	0	0	0	0	1	0	0	0	0	0	33.26	1106.2
7	1861	Jul	Oct	-0.7719529	6.71	45.0241	0	0	0	0	0	0	1	0	0	0	0	39.17	1534.3
8	1861	Aug	Nov	-0.5488093	1.25	1.5625	0	0	0	0	0	0	0	1	0	0	0	51.25	2626.6
9	1861	Sep	Dec	-0.5160195	-1.73	2.9929	0	0	0	0	0	0	0	0	1	0	0	47.17	2225
10	1861	Oct	Jan	0.79162374	-6.2	38.44	0	0	0	0	0	0	0	0	0	1	0	39.37	1550
11	1861	Nov	Feb	0.49899117	2.28	5.1984	0	0	0	0	0	0	0	0	0	0	1	14.13	199.66
12	1861	Dec	Mar	0.1836081	3.83	14.6689	0	0	0	0	0	0	0	0	0	0	0	46	2116

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.736618664
R Square	0.542607056
Adjusted R Squ	0.486511695
Standard Error	0.716580983
Observations	120

ANOVA

	df	SS	MS	F	Significance F
Regression	13	64.57023967	4.966942	9.6729399	5.48E-13
Residual	106	54.42976033	0.513488		
Total	119	119			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.957633973	0.362672861	2.64049	0.00952915	0.2385998	1.6766682
Temp	0.248857761	0.295776577	0.841371	0.40203437	-0.3375481	0.8352636
TempSq	0.058061711	0.29195044	0.198875	0.84274131	-0.5207584	0.6368819
Jan	-0.312782495	0.4014674	-0.779098	0.43765782	-1.1087307	0.4831657
Feb	-0.13307946	0.577754982	-0.230339	0.81827237	-1.2785349	1.012376
Mar	-1.172729618	0.72353837	-1.620826	0.10802612	-2.6072148	0.2617556
Apr	-2.251443687	0.831292229	-2.708366	0.00788509	-3.8995614	-0.603326
May	-1.745727543	0.833241261	-2.095105	0.03854426	-3.3977094	-0.0937457
Jun	-2.806631111	0.585930825	-4.790038	5.4456E-06	-3.968296	-1.6449663
Jul	-1.477882398	0.40315122	-3.665826	0.00038713	-2.2771689	-0.6785959
Aug	-0.94545225	0.320791616	-2.947247	0.00394348	-1.5814528	-0.3094517
Sep	-0.399386292	0.350285437	-1.140174	0.2567836	-1.0938612	0.2950887
Oct	-0.136530029	0.378056257	-0.361137	0.71871595	-0.8860633	0.6130033
Nov	-0.109962796	0.335449991	-0.327807	0.74370431	-0.775025	0.5550994

RESIDUAL OUTPUT

Observation	Predicted ln(birth)	Residuals
1	0.561969871	-0.074666243
2	1.063940825	-0.719398077
3	0.193159046	-0.031100674
4	-0.962382361	1.18265517
5	-0.390462869	-0.066586348

Figure 23: Screenshot of Regression Analysis Summary Output on Microsoft Excel

The most important parameters in Figure 23, the output of the linear regression function in Excel, is described below:

Multiple R - This is the correlation coefficient. This determines how strong the linear relationship is. For example, a value of 1 means a perfect positive relationship and a value of zero means no relationship at all. It is the square root of r squared.

R squared - This is the Coefficient of Determination. This defines how many data points are on the regression line. For example, 80% means that 80% of the variation of y-values around the mean is explained by the x-values.

Observations - Number of observations in the sample.

For the purpose of this study the analysis of variance (ANOVA) table is not of importance but it will be good to clarify the degrees of freedom specified in this table.

Total df - n – 1 (120-1 = 119)

Regression df - The number of independent variables (13)

Residual df - The total df – regression df (119 -13 =106)

The predicted values shown below are calculated from the coefficients following the linear regression formula. They are predicted logged daily birth rate generated separately once regressed against monthly dummies showing the pure seasonality, and once regressed against monthly dummies with temperature controls. As they are all standardised values the plots appear as deviation from trend.

Chapter 5: Results

This chapter looks into the outcome of the regression analysis performed on Microsoft Excel. To recapitulate, regression was performed twice for the period 1781-1859, once on the dummy to show the pure seasonality and once with controlling for temperature lagged 9 months from birth; and three times for the period 1860-1900, once with just the dummy variables, once with controlling for temperature and once controlling for precipitation, both lagged 9 months from birth. The purpose of this exercise was to estimate the effect of temperature and precipitation on birth seasonality.

The predicted values for the standardised logged birth/day values were retrieved from the summary output, regressed once against dummies and once with temperature controls, and plotted for each decade.

"The inclusion of lagged temperature should drive the monthly dummy coefficients to a zero if seasonality in births is caused entirely by temperature", that is the assumption of this method as defined by Lam & Miron (1996)

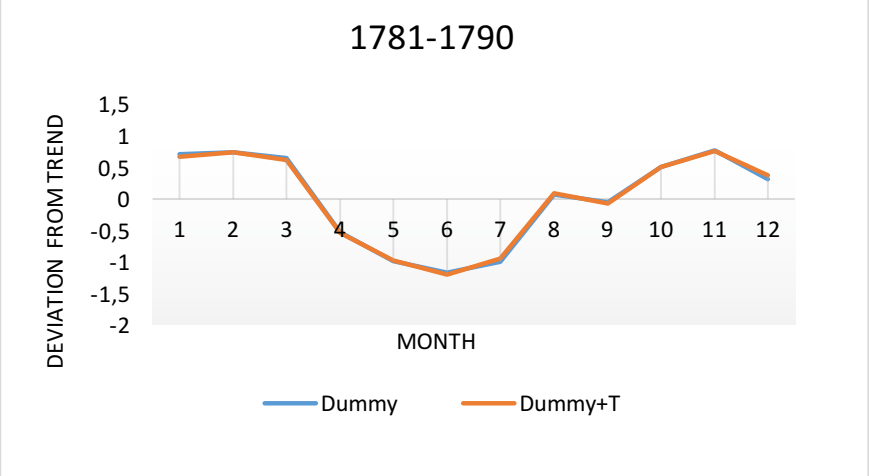
Another salient feature of this method is that the effects of temperature are based on regressions that also include monthly dummies, this essentially means that the nonseasonal effects of temperature are identified here. It is clear that the seasonal variation in temperature will be much more than the nonseasonal variation, but there is evidence that there is a significant nonseasonal variation in temperature, as well.

The way to interpret the graph plots, presented in the following subsections, is to see if the predicted value plots for the dummies with temperature controls are driven towards the zero-mean, this would imply that temperature explains a lot of the variation observed. If they do not show any change and remain similar to the plots of the predicted values of only the monthly dummies, the seasonal variation either is not at all influenced by temperature or are "due to systematic differences in the effects of seasonal versus nonseasonal variation in temperature" (Lam & Miron, 1996)

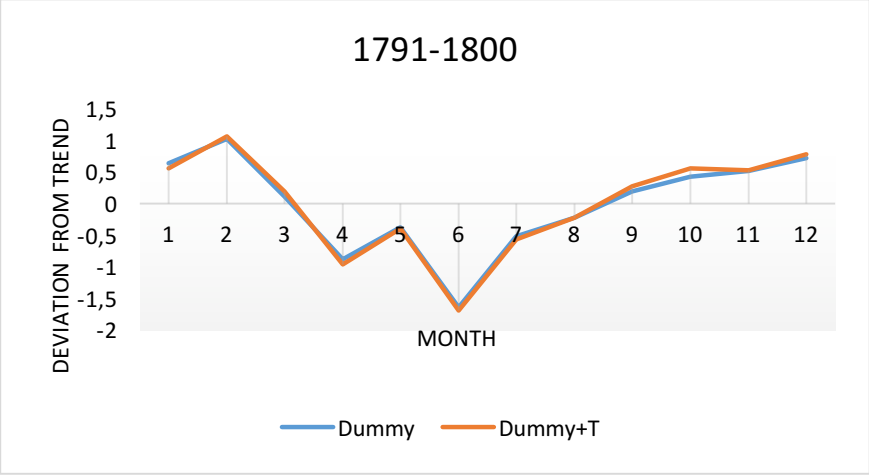
Over the next few subsections, the different plots over the 12 decades will be reviewed for temperature and precipitation effects on the seasonality of birth and inferences will be made to understand the seasonal patterns and their determinants.

5.1. Decade-wise regression charts with Dummies and Temperature Controls

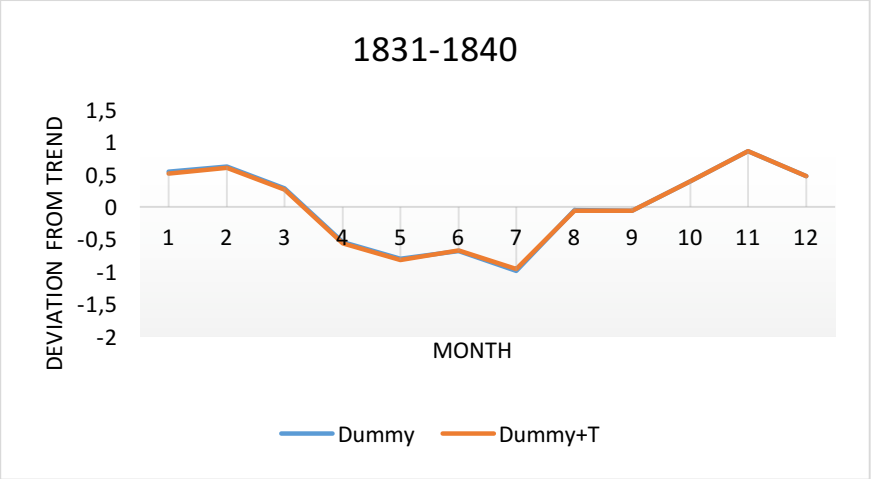
In this section the regression charts obtained from running the multiple regression using Microsoft Excel using the Dummies and Dummies with temperature controls would be analysed. The first five charts show the decades where no significant magnitude differences are found between the dummies and the dummies with controls for temperature, subsequently the other decades where changes are noticed will be elaborated on.



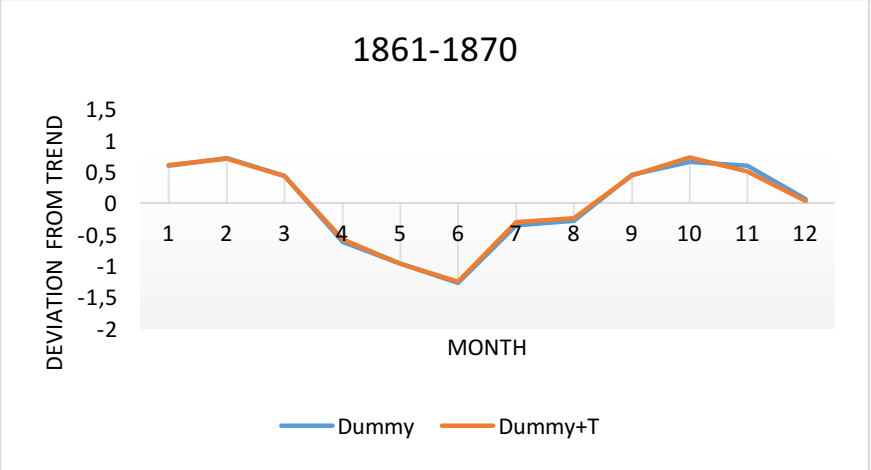
(a)



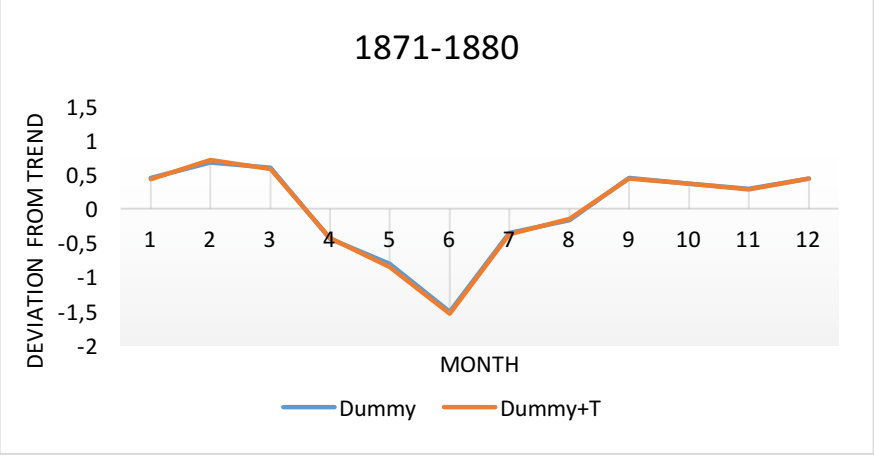
(b)



(c)



(d)



(e)

Figure 24(a - e): Seasonal birth patterns with and without controls for temperature

Figure 24 (a-e) show little to no difference in the predicted values for the monthly dummies and the dummies with temperature controls, suggesting that temperature controls do not explain much of the seasonal patterns displayed here. The temperature controls neither reduce nor increase the magnitude of birth. It is evident from these results that there are other factors that have more influence in determining the birth seasonality patterns than temperature, nevertheless it would not be prudent to dismiss temperature as a determinant completely. It is possible that the seasonal variation of temperature explains more of the pattern than the nonseasonal temperature variations.

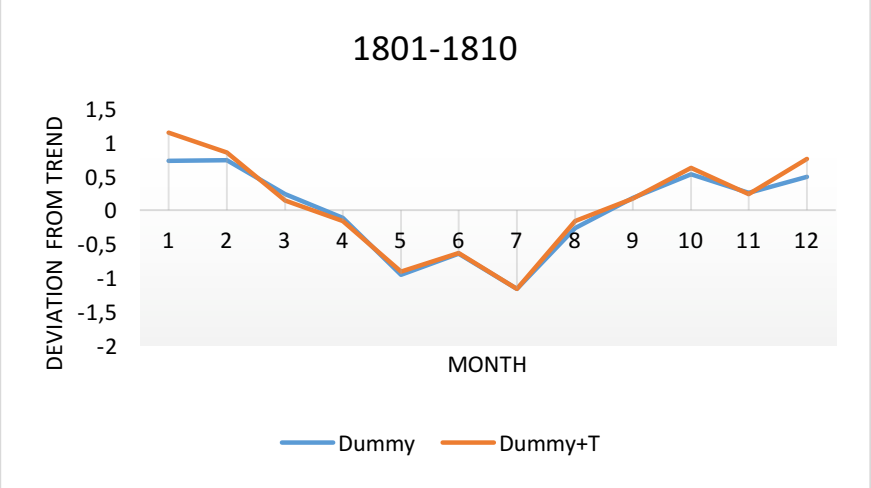


Figure 25: Seasonal birth patterns with and without controls for temperature 1801 – 1810

Figure 25 reveals a peak in January, October and December births, there is an overall trough for the months of May, June and July, with a slight peak in June. Adding temperature controls increase the magnitude by about 30 % for the months of January and December.

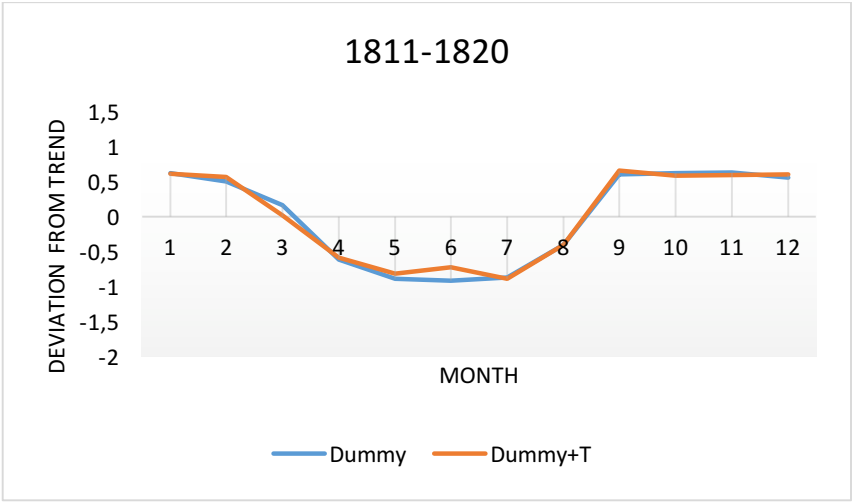


Figure 26: Seasonal birth patterns with and without controls for temperature 1811 – 1821

Figure 26 shows a clear trough for the summer months without controls for temperature, after introducing temperature controls there is a 20% movement towards the zero mean for the month of June. However slight, the nonseasonal effect of temperature is prominent in this case.

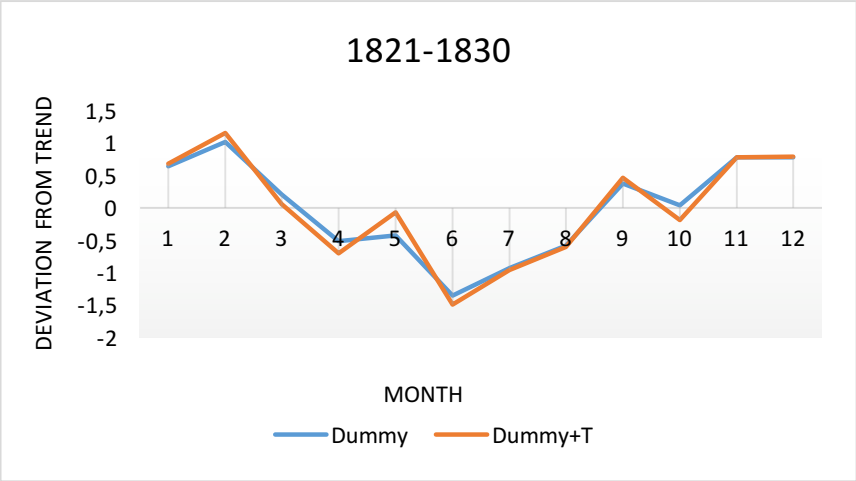


Figure 27: Seasonal birth patterns with and without controls for temperature 1821 – 1830

Figure 27 displays many fluctuations in the pattern; the clear summer month trough is punctuated with a peak towards the mean for the month of May with the introduction of temperature controls by about 30% increase signifying a positive effect of temperature, whereas it drives the value just below the mean for the month of October.

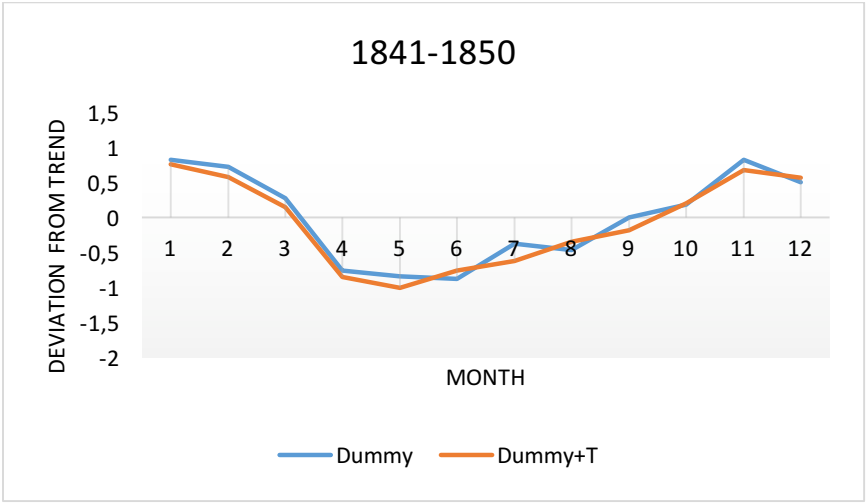


Figure 28: Seasonal birth patterns with and without controls for temperature 1841 – 1850

Figure 28 shows a sharp dip below the mean from March to April, then through some fluctuations finally rises above the mean in September-October and peaks in November. With the introduction of temperature control, the curve has an overall smoothing effect although it generally follows the trajectory of the seasonal pattern without temperature control.

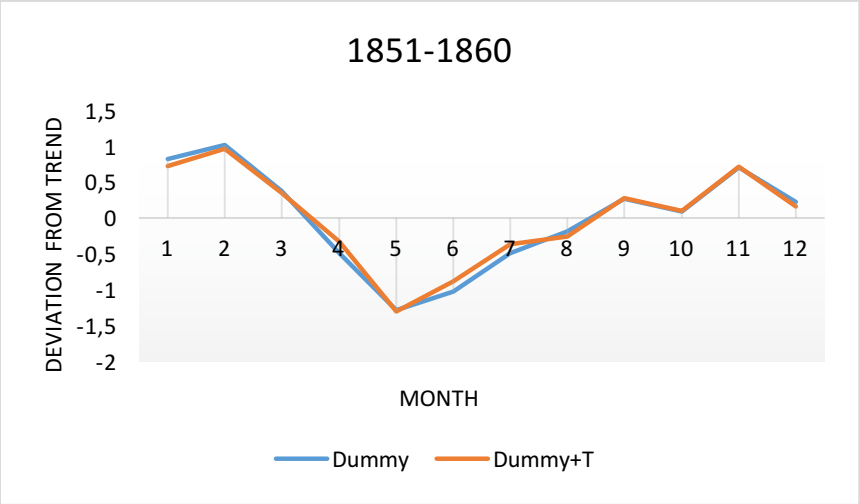


Figure 29: Seasonal birth patterns with and without controls for temperature 1851 – 1860

Figure 29 does not show much of a change in trend with controlling for temperature except the months of June and July where it is very slightly closer to the mean than the pure seasonality, a 15% difference is noticed for July.

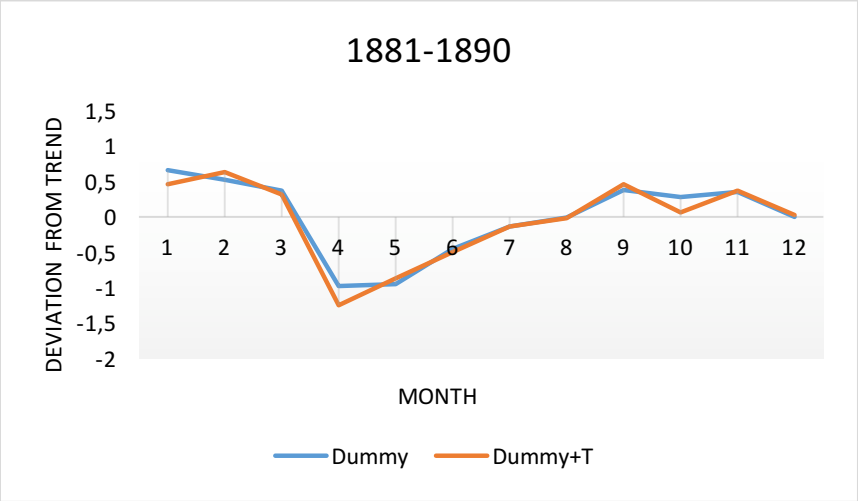


Figure 30: Seasonal birth patterns with and without controls for temperature 1881 – 1890

In Figure 30, there is a decrease in the magnitude of birth above mean overall with a dip just below the mean the mean in December. After introducing temperature controls the deviation moves further away from the mean in April. In October, the predicted value moved towards the mean by almost 30% indicating some nonseasonal effects of temperature.

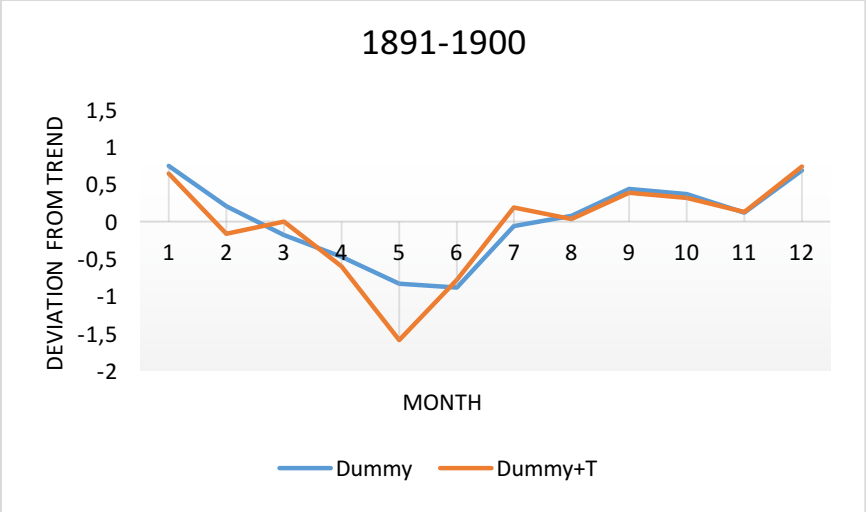


Figure 31: Seasonal birth patterns with and without controls for temperature 1891 – 1900

For the last decade, the magnitude of change with temperature controls is the maximum of all the other decades, it can be observed in Figure 31. There is a dip below the mean for February, in March, there is a rise towards the mean, but a sharp dip in the month May is very significant, the temperature controls drives the dip by almost 70%, in July it again rises slightly above the mean compared to the seasonal variation.

5.2. Decade-wise regression charts with Dummies and Precipitation Controls

In this section, decades from 1861 to 1900 are discussed; only for these years were precipitation data available, the predicted logged birth data obtained from regression analysis against monthly Dummies and Dummies with precipitation controls are plotted as deviation from trend.

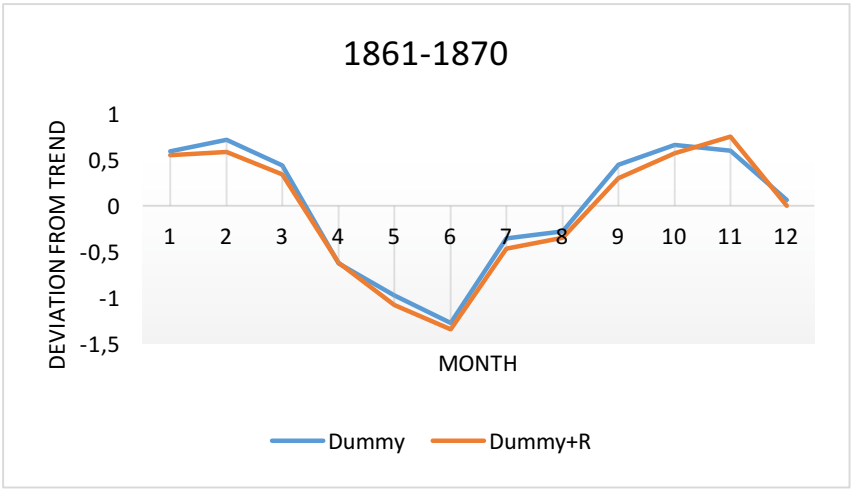


Figure 32: Seasonal birth patterns with and without controls for precipitation 1861 – 1870

In the Figure 32, it can be observed that the introduction of precipitation controls follows slightly below the values for pure seasonality except the month of November when the magnitude increase slightly and creates a peak before sharply falling in December.

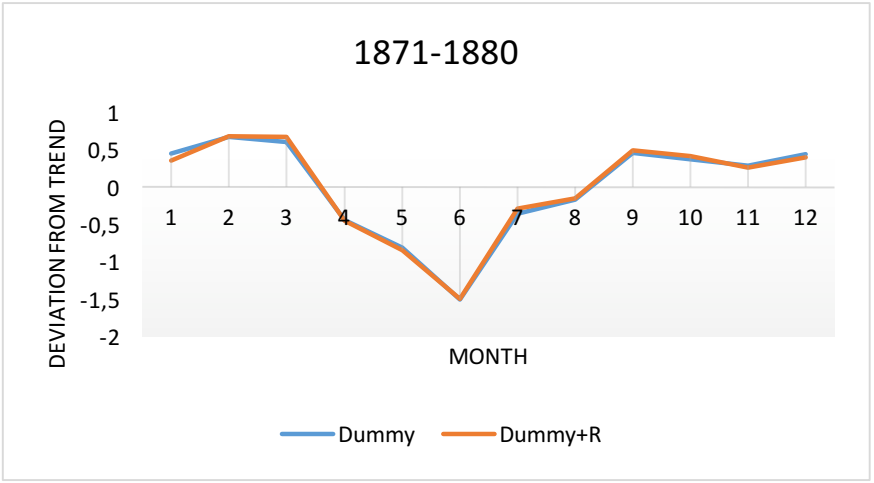


Figure 33: Seasonal birth patterns with and without controls for precipitation 1871 – 1880

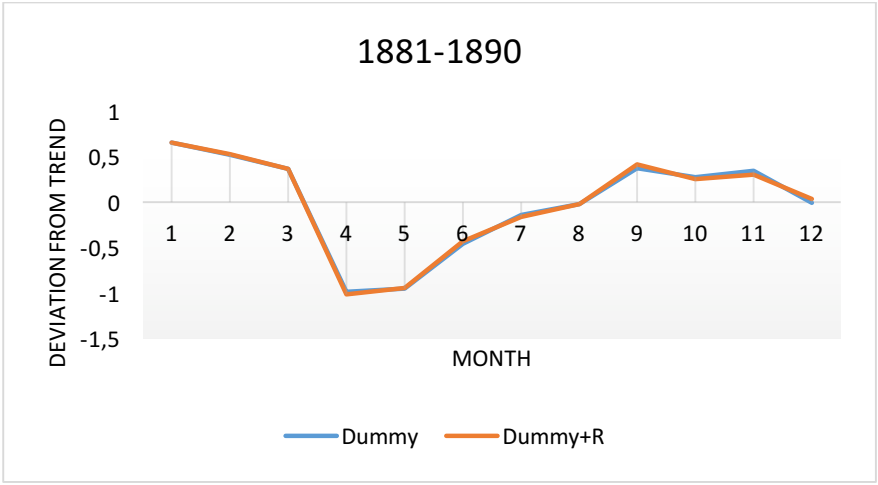


Figure 34: Seasonal birth patterns with and without controls for precipitation 1881 – 1890

Figures 33 and 34 do not show any change in the seasonal pattern after controls for precipitation are introduced therefore it can be said that any nonseasonal variation in precipitation do not affect the seasonality patterns significant enough to be discussed.

Of the 4 decades that looked into the effects of precipitation on birth seasonality the decade between 1891-1900 (Figure 35) shows the most differences between pure seasonality plots based on dummy variables and introduction of precipitation controls. The plot is stable between July and August, when the seasonal variation rises above the mean and the temperature-controlled values remain below the mean.

Some of the nonseasonal variations observed in the last few figures will be explored in the next subsection through the use of scatterplots between conception and temperature variation within a month.

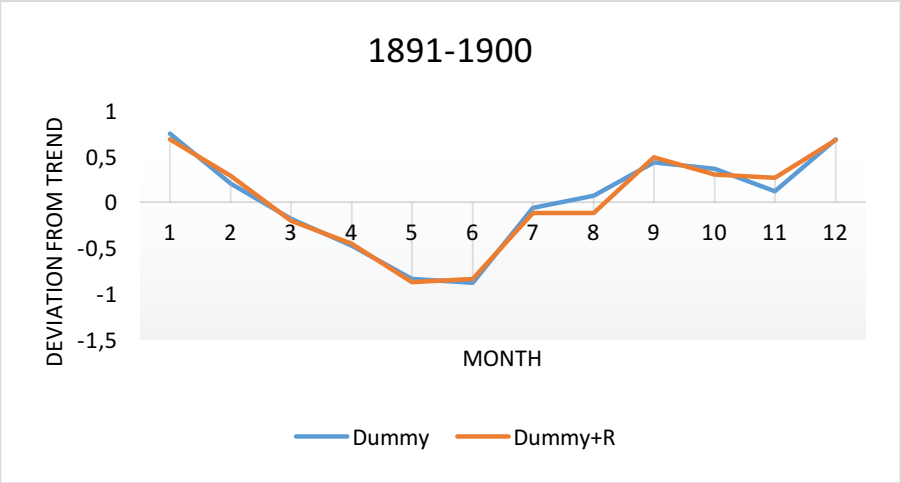


Figure 35: Seasonal birth patterns with and without controls for precipitation 1891 – 1900

5.3. Nonseasonal Variation of Temperature and its Relationship with Conception

In this section, careful attention is paid to the empirical analysis of the time-series data on temperature and conception for particular months.

The figures show scatterplots with monthly mean temperatures on the horizontal axis and detrended (as measured with a nine-month lag of the detrended monthly births) conceptions on the vertical axis.

The plots show the range of the temperature difference within a months and the relationship between conception and temperature.

The variation in temperature within a month is called the nonseasonal variation and for some months, the difference is quite significant.

Among the few chosen scatterplots that are displayed to demonstrate the effects of nonseasonal variation of temperature on conception, it is evident where there is a significant positive or a negative correlation.

For Figure 36(a) the variation in temperature is 6°C and there is an increase in the conception as temperature rises. Figure 36(b) has a similar effect for August, compared to that Figure 36(f), also

plots for August, shows a definite change in the trend, from a positive to a negative correlation for the same temperature condition with variation of 4 and 5°C and maximum temperature of 20°C. This change in trend matches the trend that is observed elsewhere around the world that with increase in temperature there is a decrease in conception; hence, hotter regions observe a dip in spring births.

Although, March 1840-1849 (Figure 36(c)) has a temperature variation of 7 degrees and a below freezing minimum temperature, the conception rate is highest at that temperature and the trend shows an inconsiderable negative relationship between the two variables.

October 1840- 1849, (Figure 36(d)) has a temperature variation of 7 degrees and there is a significant positive correlation with the conception rate increasing steadily with the temperature

Figure 36(g) shows the month of May between the years 1890-1899, this month of this decade was chosen to investigate the sharp fall in the birth from its trend observed in Figure 31. The result of the plot is quite unsatisfactory as the nonseasonal variation of temperature is not great enough and the conception seems quite random with no clear indication of a systematic relationship existing between temperature and conception.

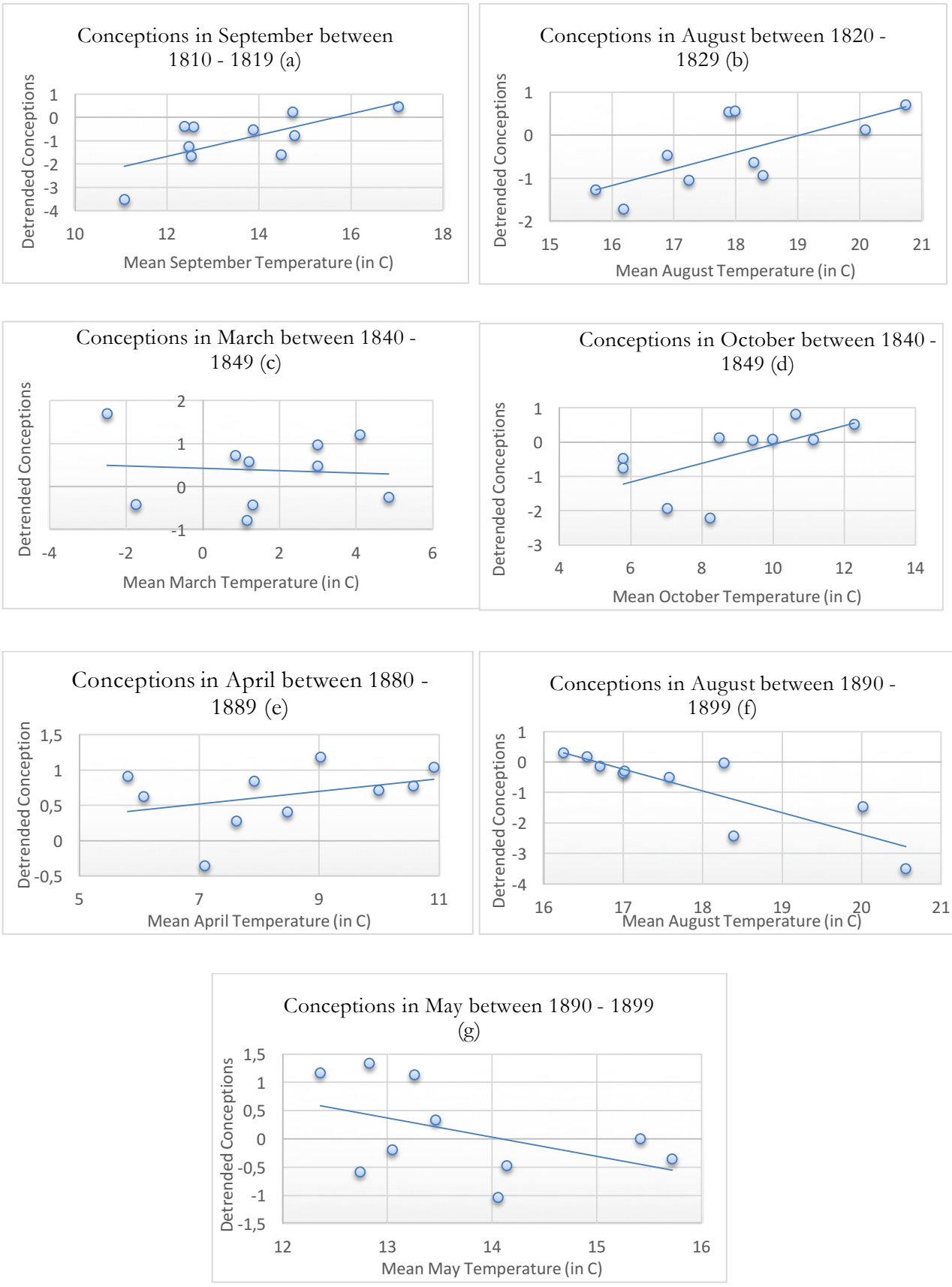


Figure 36 (a-g): Nonseasonal temperature variation and conception

5.4. Seasonality Patterns – A comparative study

A general spring summer trough is constant over almost all the decades, which matches with patterns observed in southern and south-western United States, but this pattern stands in sharp contrast to western European countries, such as Sweden, Finland, England, Luxembourg, where a peak in the spring months have been observed in the post-war period (Lam & Miron, 1994). Another common feature between these western European countries is a peak in September, this peak can also be observed in certain decades (Figures 33- 35) in Velke Pole.

Another feature of the seasonality pattern that has been observed is a February peak, this sort of peak is also quite distinctive when compared to post-war seasonality patterns around the world, this sort of peak has not been observed elsewhere although Australia and New Zealand show a slight upward trend in March.

November and sometimes October show a secondary peak in birth for Velke Pole, against quite distinctive and no equivalent trend has been found in other countries. Most states in the U.S. instead show a peak in September. September rise in births were observed in the 1811-1821, 1821-1830, 1871-1880, and 1881-1890. The magnitude of these peaks was not great, but they were, nonetheless, observed as increases above the mean.

The decade of 1891-1900 observes some discrepancies in the pattern; the February and November peaks, which were characteristic features of the seasonal pattern for Velke pole, were replaced with a dip towards the mean.

The above-applied statistical analysis fails to show any real relationship of consequence to rule that climatological factors were major determinants of birth seasonality even in a time when they should have been a major influence on the life and activities of a population. Although, a strong seasonal pattern, which has remained largely unchanged for the length of the studied time- period exists, it fails to give any indication of major causes of its existence.

Previous studies dealing with this subject has stated that conception can be low in summer due to high temperatures when it directly affects the male physiology lowering sperm count, but this reason cannot be used in this study as the climate is of temperate/continental in nature and do not experience extreme heat or cold.

Chapter 6: Conclusion

6.1. Summary

As this thesis draws to a close, it is an opportune time to look back on the some of the methods performed to answer our basic research questions, what pattern did the seasonality of birth in Velke Pole follow, what does the statistical analysis reveal about the causality and what is the conclusion one draws from the final results?

The work began with data collection of temperature and precipitation records and where data was missing it was generated using spatial interpolation technique of Kriging and extrapolation based on the kriging and recorded data.

The focus on climatological data was driven by the fact that the time period under scrutiny was of a historical nature, and the numerous literatures that have emphasized on the importance of climatological factors on the seasonality of birth.

The method of ordinary least squares multiple regression was performed to find evidence whether temperature and precipitation were in fact major determinants of the seasonal patterns of birth in Velke Pole. Based on the results derived from rather complex series of statistical processes, inferences were drawn and knowledge of the history of the people, their mode of subsistence, religious affinity and other sociological circumstances influenced the deductions made

6.2. Deductions

The last chapter presented the results of the statistical analyses in details and to what extent the climatological factors studied have impacted the seasonality. The answers were not satisfactory and no clear conclusion could be drawn about temperature and precipitation being the major influences on the seasonality pattern.

Therefore, it was necessary to look into the sociological aspects of the population to better understand what could have worked in conjunction with the natural environment to develop a pattern quite distinctive from the post-war global patterns of birth as presented by Lam & Miron (1994).

One factor that could have a major contribution on the birth seasonality is religion, the population of Velke Pole was of a catholic denomination and it is reasonable to assume that religious practices and rituals affected the way life was lead in that era, much more than it does now. In keeping with

observations made by other researchers (Seiver, 1985) the holidays during the period of Christmas may have affected conception rate to be higher, leading to higher births during the fall months, and could be a contributing factor for a secondary peak observed in September for certain decades, as discussed previously. Amidst high conception rates in the early months of the year, a sharp dip in March is observed in some decades (Figure 37), one possible reason for this could be the period of lent among Christians when they observe fasting and abstinence. The other reason is the spring seeding time, which saw the work force, mainly the men, go away for work, although they did come back intermittently to prepare their own fields (Maday, 1984). Another major reason of conception rates falling in March is because this was peak flu season in the 19th century, especially in the temperate climate zone. Women falling sick or the immunity being generally low could have prevented conception even in healthy adults.

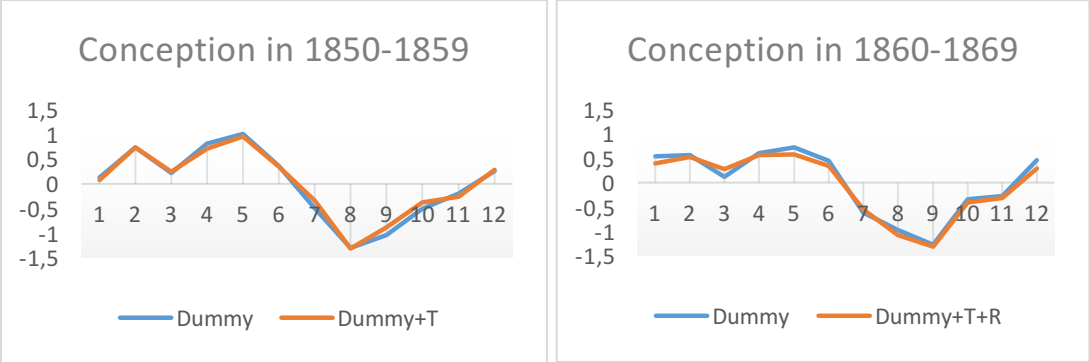


Figure 37: Conception rates between 1850 - 1869

Among various other reasons, some scholars have studied the effects of photoperiod and photo intensity as determinants of seasonal patterns of birth, which are said to affect hormonal concentration, sexual activity and sperm quality (Manfredini, 2009; Cummings, 2012). Without data and statistical analysis this influence of photoperiod should not be ruled out completely.

Agricultural practices and seasons of harvest could play a role in dictating seasonality. In the last couple of decades of our study namely 1880-1900, there was presence of railways in the region, the closest in Zarnovica (15km) built in 1896 and the farthest built in 1873 35 km away to the north (H. Saler, email communication, Nov 2015). The presence of railway would mean that the main provider of the household, mainly the man, could have been away for lengths of period for work in areas away from the hometown. It has already been established that the population consisted mainly of miners, joiners/carpenters and farmers; therefore, the possibility of people working during the harvesting seasons away from their homes cannot be overlooked. The agricultural practices of this

population as indicated by Maday (1984) consisted of spring and autumnal seeding, harvesting and working with crops such as corn, sugar-beet and potato; this would explain the dip in conception during July, August and September (also Figure 37), as winter wheat, a major crop grown in the region (SARC, 2008), is harvested through summer to early fall and hard work in the field along with direct influence of temperature on the human physiology could result in the overall summer time trough in conception.

The general consensus among the community of scholars who have been involved in studying seasonality of birth over the years is that no one determinant is all powerful in influencing the patterns single-handedly, it is always a combination of factors that help in designing this phenomenon observed in almost all populations. Sociological, religious, economic practices over regions that are also, in part, dictated by the environmental factors, help in creating unique patterns in every region. It is safe to say that explanations for naturally occurring phenomena are always confounding to a degree and one may never find out in entirety their causes, one can do nothing much but speculate and fit together results of experiments carried out on subsets in order to understand nature.

6.3. Limitations

The three primary limitations of the research were:

1. The lack of large birth samples due to the small size of settlement produce large percentage differences, which may be misleading at a first glance, and the results would not be comparable to large scale studies where big differences would only constitute a small percent of the entire population.
2. Absence of environmental data from the immediate vicinity of Velke Pole for the required time period under study makes investigations into the causal relationship difficult and uncertain.
3. Absence of a well-formed body of work on socio-economic conditions of Velke Pole and Slovakia, in general, in international literature and the English language deterred efforts to validate the findings of this research.

6.4. Implications of this work on future research

Some of the avenues in which future research can be carried out are elucidated below:

- The set of techniques or processes made use of in this thesis can be applied to other settlements or geographic locations around the world to reveal information on causality of birth seasonality.
- Benchmarking the accuracy of temperature and precipitation predictions by comparing results obtained from different deterministic as well as geospatial interpolation techniques, such as Inverse Distance Weighting, Splines, Triangulation and use of resampled DEM with high accuracy in Cokriging, this could help in determining whether improvements can be made in obtaining results with higher accuracy.
- Development of full-fledged Python libraries for performing geospatial and deterministic interpolation, including semi-variogram analysis. Although a couple of open-source libraries exist for Kriging in Python they are still in a rudimentary phase.
- The regression model can be further expanded using more independent and categorical variables especially for socio-economic indicators such as education, age group of mothers, and birth order among various others.
- Interactive tools can be developed for visualization and analysis purposes so the user could select different independent variables to inspect their degree of influence on the phenomenon
- The techniques presented in this research have only been used for analysing the past, but extensions could be made to do forecasting of the future birth patterns based on changing environmental and sociological parameters.

References

Bailey, R. C., Jenike, M. R., Ellison, P. T., Bentley, G. R., Harrigan, A.R., and Peacock, N. R. (1992). "The ecology of birth seasonality among agriculturalists in central Africa". Journal of Biosocial Science, 24, pp 393-412.

Bobak M., Gjonca A., (2001), "The seasonality of live birth is strongly influenced by socio-demographic factors", Human Reproduction, 16(7), pp. 1512-1517.

Bohling, G. (2005). "Introduction to Geostatistics and Variogram Analysis". Lecture Notes C&PE 940. Kansas University.

Bronson, F.H. (1995), "Seasonal Variation in Human Reproduction: Environmental Factors." Quarterly Review of Biology 70(2): Pp. 141-64.

Burrough, P. A. and McDonnell, R. A. (1998). "Principles of Geographical Information Systems.". Oxford University Press. New York.

Cowgill, U. M. (1965). "Season of birth in man, contemporary situation with special reference to Europe and the Southern Hemisphere". Ecology 47(4): Pp. 614-423.

Condon, R. G. and Scaglion, R. (1982). "The ecology of human birth seasonality". Human Ecology, 10, 495–510.

Condon, R. G. (1991). "Birth Seasonality, Photoperiod, and Social Change in the Central Canadian Arctic". Human Ecology. Vol. 19. No. 3. Pp. 287 – 321.

Cummings, D. R. (2012). "Canadian birth seasonality and its possible association with seasonal brightness". Canadian Studies in Population 39. No. 1 – 2. Pp. 45 – 62.

Dorelien, A. M. (2013). "A time to Be Born: Birth Seasonality in Sub-Saharan Africa". Population Studies Center Research Report. No. 13 – 785.

Ede. R. (2015). "The history of Bars County (Kingdom of Hungary)". Personal Communication, 2015.

Ellison, P.T., Valeggia, C.R., and Sherry, D. S. (2005) "Human birth seasonality". Seasonality in Primates: Implications for Human Evolution. Diane K. Brockman and Carel P. van Schaik (Eds). Cambridge Studies in Biological and Evolutionary Anthropology, Cambridge University Press, pp. 379 -400.

Erhardt, C., Nelson, F. G., and Parker, J. (1971). "Seasonal patterns of conception in New York City". Americal Journal of Public Health. No. 11. Pp. 2246 – 2258.

Garavaglia, S. and Sharma A. (1998) “A smart guide to Dummy variables: Fourapplications and a macro”. Proceedings of the Northeast SAS Users Group Conference.

Goovaerts, P. (2000a). “Geostatistics for Natural Resource Evaluation”. Oxford University Press.

Goovaerts, P. (2000b). “Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall”. Journal of Hydrology 228 (2000). Pp. 113–129

Isaaks, E. H. and Srivastava, R. M. (1989). “Applied Geostatistics”. Oxford University Press.

Lam, D.A. and Miron. J. A. (1991). "Seasonality of Births in Human Populations." Social Biology 38(1-2): Pp. 51-78.

Lam, D.A. and Miron, J. A. (1994). "Global Patterns of Seasonal Variation in Human Fertility." Human Reproductive Ecology: Interactions of Environment, Fertility, and Behavior, edited by K.L. Campbell and J. J. Wood. New York: New York Academy of Sciences. Pp. 9-28.

Lam, D, and Miron, J. A. (1996). "The Effect of Temperature on Human Fertility." Demography, 33(3): Pp. 291-305.

Lee, R. (1981). “Short-term Variation: Vital Rates, Prices, and Weather”. The Population History of England, 1541- 1871: A Reconstitution. Ed. E. A. Wrigley and R. Schofield, Cambridge, MA. Harvard University Press. Pp. 356 – 401.

Levine, R. J. (1991). “Seasonal Variation in Human Semen Quality.”. Temperature and Environmental Effects on the Testis. Eds. A.W. Zorgniotti. New York: Plenum. Pp. 89 – 96.

Levine, R. J. (1994). “Male Factors Contributing to the Seasonality of Human Reproduction”. Human Reproductive Ecology: Interactions of Environment, Fertility, and Behavior. Eds. K. L. Campbell and J. J. Wood. New York: New York Academy of Sciences. Pp. 29 – 45.

Lerchl A, Simoni M, and Nieschlag E. (1993): "Changes in seasonality of birth rates in Germany from 1951 to 1990". Naturwissenschaften; 80: Pp. 516–518

Maday, J. (1984). “Hochwies Paulisch - Das Siedlungsgebiet um den Reichberg."

Manfredini, M. (2009). "Birth seasonality in present day italy, 1993-2005". Human Ecology, 37: pp.227–234.

McKinney, W. (2014). “Python for Data Analysis”. O'Reilly Media.

Richards, T. (1983). “Weather, Nutrition, and the Economy: Short-Run Fluctuations in Births, Deaths, and Marriages, France 1740 – 1909”. Demography 20. Pp. 197 – 212.

Roenneberg, T. and Aschoff, J. (1990). "Annual rhythm of human reproduction: I. Biology, sociology, or both?". Journal of Biological Rhythms. Fall 5(3). Pp. 195- 216.

Seiver, D. (1985). "Trend and Variation in the Seasonality of U.S. Fertility, 1947 to 1976." Demography 22: Pp.89-100.

Solcova, L. (2008). "Historical development of settlement villages Veľké Pole, Píla and Radobica in the Novobanská štálová area". Geography in Czechia and Slovakia (Theory and Practise at the Onset of 21st Century), Svataňová at al., Masaryk University , Brno, 2008, ISBN 978-80-210-4600-9.

Weschler, T. & S. Halli. (1992). "The seasonality of births in Canada: A comparison with the northern United States". Population and Environment 14(1). Pp 85–94

SARC (2008). "Country Report on the State of Plant Genetic Resources for Food and Agriculture". Second Slovak National Report on Conservation and Sustainable Utilization of Plant Genetic Resources for Food and Agriculture.

Internet Sources:

Basten, S.:

https://familysearch.org/learn/wiki/en/Birth-baptism_intervals_for_family_historians, 20.10.2015.

City Population:

http://www.citypopulation.de/php/slovakia-banskobystrickykraj.php?cityid=517348, 10.10.2015

ESRI ArcGIS Help Portal: http://resources.arcgis.com/en/help/, 01.11.2015

HISTALP: www.zamg.ac.at/histalp/, 02.07.2015

Settlement Development in Velke Pole and Pila, Slovakia: http://www.imm.hs-karlsruhe.de/slovakei/history_velkepole.php , 12.06.2015

NumPy: http://www.numpy.org/, 07.07.2015

Pandas : http://pandas.pydata.org/, 07.07.2015

Sandvik, B.: www.thematicmapping.org/, 14.06.2015

SHMU: www.shmu.sk/, 02.07.2015

StatsModels: http://statsmodels.sourceforge.net/, 07.07.2015

Wadsworth, J.:

http://personal.rhul.ac.uk/uhte/006/ec2203/Lecture%2013_Use%20and%20Interpretation%20of%20Dummy%20Variables.pdf, 01.11.2015

Appendix

Python: Calculation of detrended birth rates

```
import numpy as np

import pandas as pd

df1 = pd.read_csv('VPBirth.csv', sep=";")

df1 = df1.set_index('Year')

df1 = df1.transpose()

result = df1.stack()

result
```

```
     Year
Jan  1758    13
     1759    11
     1760    19
     1761    14
     1762    13
     1763     8
     1764    16
     1765    13
     1766    13
        … (truncated)
```

```
dates = pd.to_datetime([' '.join(item) for item in result.index])

days = dates.days_in_month

dates
```

```
DatetimeIndex(['1758-01-16', '1759-01-16', '1760-01-16', '1761-01-16',
               '1762-01-16', '1763-01-16', '1764-01-16', '1765-01-16',
               '1766-01-16', '1767-01-16',
               ... (truncated)
               '1891-12-16', '1892-12-16', '1893-12-16', '1894-12-16',
               '1895-12-16', '1896-12-16', '1897-12-16', '1898-12-16',
               '1899-12-16', '1900-12-16'],
              dtype='datetime64[ns]', length=1716, freq=None, tz=None)
```

```
result = (result / days).unstack()

df2 = result.transpose()

df3 = pd.rolling_mean(df2.stack(), window=12, center=True)

df3
```

```
Year
1758  Jan         NaN
      Feb         NaN
      Mar         NaN
      Apr         NaN
      May         NaN
      Jun         NaN
      Jul    0.283532
      Aug    0.278155
      Sep    0.263274
      Oct    0.244457
      Nov    0.241679
      Dec    0.236303
1759  Jan    0.239081
```

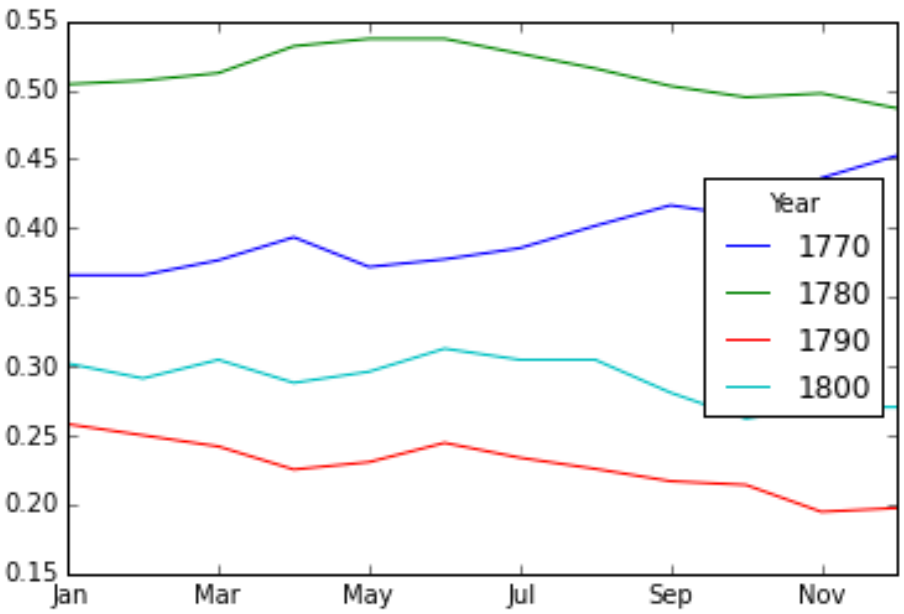
```
Feb    0.263274
Mar    0.265963
Apr    0.277074
May    0.282450
Jun    0.285228
Jul    0.290604
Aug    0.312110
Sep    0.325451
Oct    0.355021
Nov    0.357799
```

… (truncated)

```
df4 = df3.unstack().transpose()
```

```
%matplotlib inline
```

```
df4[['1770','1780','1790','1800']].plot()
```



```
birth = df1.transpose()
```

```
moving_avg = pd.rolling_mean(birth.stack(), window=12, center=True)
```

```
Ri = (birth/moving_avg) * 100

moving_avg = moving_avg.unstack()
```

```
moving_avg_per_day = pd.rolling_mean(df2.stack(), window=12, center=True)

moving_avg_per_day = moving_avg_per_day.unstack()
```

```
bt = df2/moving_avg_per_day
```

```
lnbt = np.log(bt)

lnbt = lnbt.stack()

lnbt
```

```
Year
1758  Jul   -1.480409
      Aug   -0.768118
      Sep   -0.120729
      Oct    0.539677
      Nov    0.216170
      Dec    0.406548
1759  Jan    0.394861
      Feb    0.304939
      Mar   -0.163678
      Apr   -0.731431
```

```
May   -1.071121

Jun   -0.354971

Jul    0.199701

Aug   -0.660148

Sep    0.119241

Oct    0.309641

Nov   -0.070828

Dec    0.086304

  ... (truncated)
```

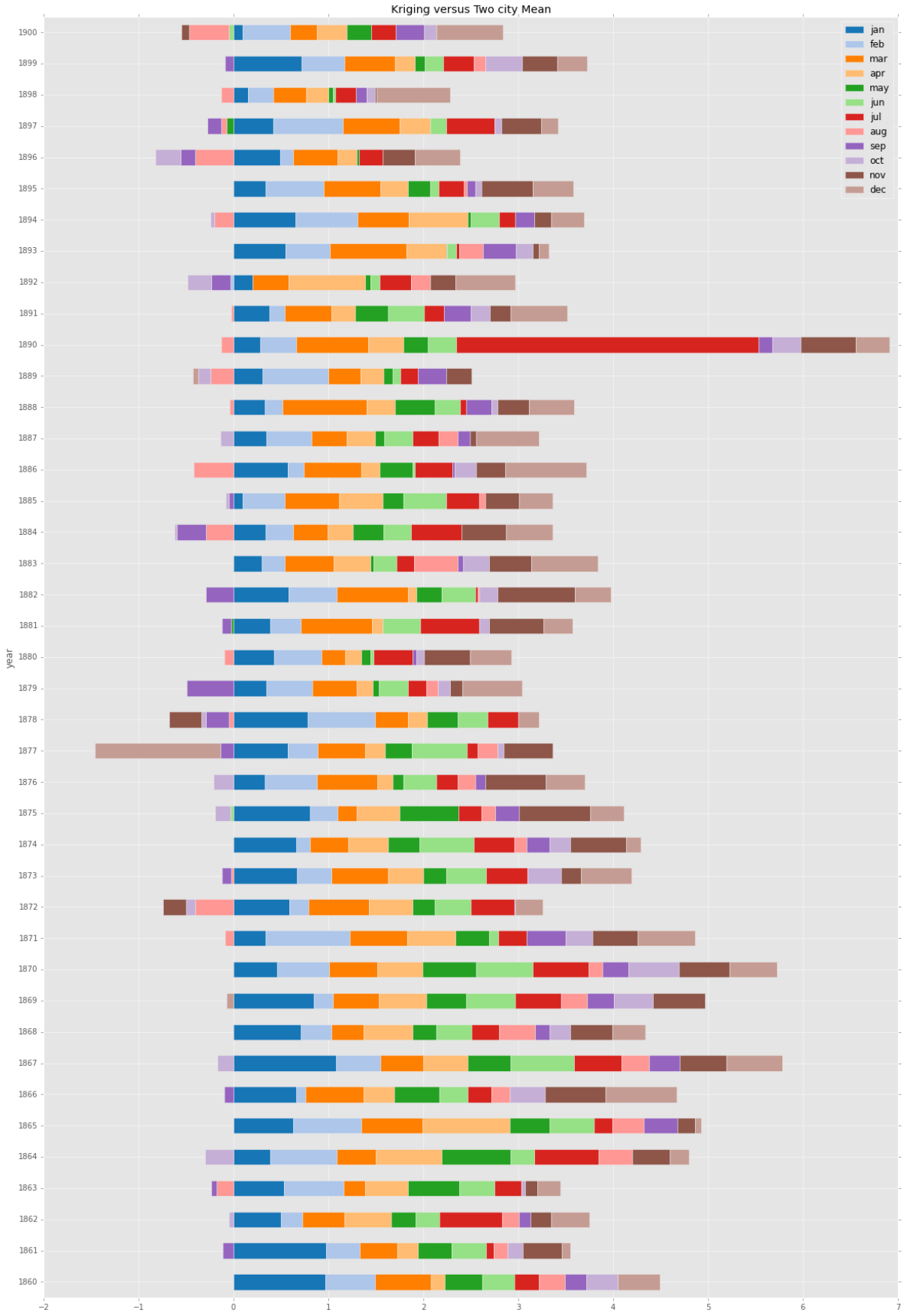
```
stacked_Ri = Ri.stack()
```

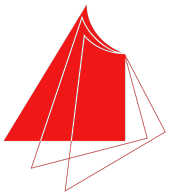
```
Ri.to_excel('Ri.xlsx')

bt.to_excel('bt.xlsx')

lnbt.to_csv('lnbt_naturalLog.csv')

stacked_Ri.to_csv(' stacked_Ri.csv')
```

Python: Comparison of means versus Kriging values

```
import numpy as np

import pandas as pd

%matplotlib inline

import matplotlib

matplotlib.style.use('ggplot')

colorTable = [(31, 119, 180), (174, 199, 232), (255, 127, 14), (255, 187, 120
),

              (44, 160, 44), (152, 223, 138), (214, 39, 40), (255, 152, 150),


              (148, 103, 189), (197, 176, 213), (140, 86, 75), (196, 156, 148)
,
```

```
            (227, 119, 194), (247, 182, 210), (127, 127, 127), (199, 199, 19
9),
            (188, 189, 34), (219, 219, 141), (23, 190, 207), (158, 218, 229)
]


for i in range(len(colorTable)):
    r, g, b = colorTable[i]
    colorTable [i] = (r / 255., g / 255., b / 255.)


two_cities_means=pd.read_csv('cities_means_v_b.csv', sep=";")


two_cities_means=two_cities_means.set_index('year')


kriging=pd.read_excel('VelkePole_T.xlsx')


kriging=kriging.set_index('year')


diff = two_cities_means - kriging


kriging_OP=pd.read_excel(' VelkePole_T_OP.xlsx')


kriging_OP=kriging_OP.set_index('year')


diff_OP = two_cities_means - kriging_OP


kriging_difference = kriging - kriging_OP


diff.plot(kind='barh', figsize=(20, 30), stacked=True, title="Kriging versus
Two city Mean", color=tableau20)
```


Kriging versus Two city Mean
year
jan
feb
mar
apr
may
jun
jul
aug
sep
oct
nov
dec
1900
1899
1898
1897
1896
1895
1894
1893
1892
1891
1890
1889
1888
1887
1886
1885
1884
1883
1882
1881
1880
1879
1878
1877
1876
1875
1874
1873
1872
1871
1870
1869
1868
1867
1866
1865
1864
1863
1862
1861
1860
−2
−1
0
1
2
3
4
5
6
7



Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

A Study of Seasonal Patterns of Birth for Velke Pole, Slovakia between 1781 and 1900

Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science (M.Sc.) in Geomatics

Deepanjana Majumdar

46700

Faculty of Information Management and Media
University of Applied Sciences, Karlsruhe

Advisor:	Prof. Dr.-Ing. Heinz Saler
Co-Advisor:	Prof. Dr. habil. Mark Vetter

January 2016