University of Zurich[UZH]

In the Footsteps of the "Mother of all Pandemics": Spatio-Temporal Analysis of the 1918 Flu Pandemic in the Canton of Berne, Switzerland

GEO 620 Master's Thesis

Author
Corina Leuch
14-721-583

Supervised by
Dr. Oliver Grübner
Dr. Kaspar Staub (kaspar.staub@iem.uzh.ch)

Faculty representative
Prof. Dr. Sara Irina Fabrikant

22.01.2021
Department of Geography, University of Zurich

In the Footsteps of the "Mother of all Pandemics": Spatio-Temporal Analysis of the 1918 Flu Pandemic in the Canton of Berne, Switzerland

Author
Corina Leuch
14-721-583

MSc thesis GEO 620
Department of Geography, University of Zurich
Geographic Information Visualization and Analysis

Supervisors
Dr. Oliver Grübner
Dr. Kaspar Staub

Faculty representative
Prof. Dr. Sara Irina Fabrikant

Date of submission: 31 January 2021

Abstract

The "Spanish" influenza pandemic was one of the most devastating events in recent human history, claiming an estimated 50 – 100 million lives. This thesis aims to study the dissemination of the 1918 influenza pandemic in the case of the Swiss canton of Berne. For that, I received a dataset that contained 143'389 reports of influenza-like illness cases, covering 95% of all the municipalities in the canton of Berne, and dates ranging from July 1918 – December 1918. Furthermore, I recieved various socio-economic data and information on weather and accessability through the railway system.

In a first step, conventional statistical and spatio-statistical methods were applied to characterize the spread of the virus in such a regionally diverse canton as Berne. In a second step, locally specific factors that may have played a role the spread were found for each of the two principal waves of the pandemic. This was achieved by creating a logistic regression model for each wave where the dependent variables were determined using an automated model selection process.

The spatio-temporal analysis confirms that the canton of Berne was struck by two major waves in summer (July/August 1918) and autumn/winter (October 1918 – December 1919). The results of the logistic regression models show a positive association between tuberculosis mortality and influenza incidence in both waves. They also show positive associations between railway access and influenza incidence as well as urbanity and incidence during the first wave. However, it has to be noted that both models suffered from heteroscedasticity.

These findings are consistent with previous literature covering both the study area and the 1918 influenza pandemic in general. Further research efforts should be put into identifying locally specific explanatory factors which may help governments to put emergency plans into place that help contain or at least mitigate future pandemics.

Acknowledgments

The completion of this thesis over the last twelve months would not have been possible without the support and knowledge of various people. First and foremost, I would like to thank my supervisors Dr. Oliver Gruebner and Dr. Kaspar Staub, who always had an open ear in countless virtual meetings. Their continous support during all the steps of my thesis was essential. I am very thankful to them for allowing me to work on this project with such exciting data. Secondly, I would like to express my gratitude to the following people:

- **Manuel Bär** and **Dr. Gereon Kaiping** (GIUZ) for their valuable feedback and inputs following my concept talk.
- **Dr. Hans-Ulrich Schiedt** from the University of Berne for sharing their railway network data with us, for explaining their contents and for their advice on how to best integrate them in my model.
- **Dr. Konstantin Büchel** from the University of Berne for sharing his railway station data with us and for the further inputs given during a meeting.
- **Dr. Magdalena Seebauer** (GIUZ) for her patience, and her helpful inputs which allowed me to share my research in a blog entry on the GIUZ webpage.
- **My fellow students of the Y23-G-19 office** for their support, open ear and motivation, the shared meals and coffee breaks with fruitful discussions, especially during the summer months.
- My friend **Gordon Bühler** for his continuous emotional and technical support, proof-reading and valuable feedback during his free time.
- My parents **Erika and Stefan Leuch** for their support throughout my studies.

Last but not least, I would like to thank everyone not named here, who helped me to stay sane during these crazy times. Thank you for all the discussions, Skype sessions and online game nights, and for listening to me rambling on about researching a pandemic, during a pandemic. This thesis would not be what it is without all of you.

Contents

1 Introduction

In 1918, when World War I was raging across Europe, a second crisis arose that would soon be much deadlier than the war itself: The "Spanish" flu. In today's history, it is often omitted and if it is thematized, it is treated as a mere footnote when covering The Great War. To this day surprisingly little is known about what is sometimes called the "mother of all pandemics" (Morens and Taubenberger, 2018) and cost an estimate of 50 – 100 million lives over the course of two years (Greenberger, 2018), a multiple of the estimated 20 million casualties World War I caused in four years (Royde-Smith and Hughes, 2020).

In 2019, the World Health Organization (WHO) listed the outbreak of a global pandemic among the top 10 threats to global health: "The world will face another influenza pandemic – the only thing that we don't know is when it will hit and how severe it will be" (World Health Organization, 2019). The world is currently facing a pandemic. An important difference however, is that the presently circulating Sars-Cov-2 virus is a corona virus, and not an influenza virus. However, the symptoms are fairly similar, which might be the reason why the 2020 corona virus pandemic is often compared with the 1918 influenza pandemic. The aim of this master's thesis is to study the spatio-temporal spread of the "Spanish" flu in the Swiss canton of Berne. The study of historic outbreaks can provide information on how viruses spread in the past and therefore help in coping with future outcomes and in improving prevention and interventions. As a case study, historic disease reports of the Swiss canton of Berne are combined with other data sources (e.g. population and transportation data). The disease records provide a unique opportunity to study the dissemination of the virus on a communal level and therefore learn more about the local dynamics of the spread. The first step is an in-depth descriptive analysis of these disease data, to describe the time, location and severity of the pandemic outbreak in the canton of Berne, Switzerland. In a next step, explanatory factors are identified in a data-driven approach of building a linear regression model. Finally, these findings will be effectively visualized to provide a comprehensive overview of the course of the pandemic in the canton of Berne. The topic inherently contains a spatial and a temporal dimension, therefore methods from the field of geographic visualization and analysis are suitable to examine this pandemic outbreak.

2 Related Work

2.1 Important epidemiologic definitions

In order to understand the dissemination of the "Spanish" flu, a knowledge of a few basic epidemiological concepts is essential. Morbidity, mortality and lethality are three of the most important indicators that are used to describe epidemics and pandemics. Their definitions are important and are therefore listed here before the state of the art section.

Definition 2.1.1. *Morbidity.* The morbidity is the incidence of the disease in the total population. Example: If 50 out of 1000 people are infected, the morbidity rate is 5%. Morbidity is often used synonymously with the term incidence (the number of infected people per 100'000 people in a given time period) (Sonderegger, 1991).

Definition 2.1.2. *Mortality.* The mortality is the frequency of deaths through the disease in relation to the entire population. Example: If 5 out of 1000 people in a population die of a disease, the mortality rate is 0.5% or 5 ‰ (Sonderegger, 1991).

Definition 2.1.3. *Lethality.* The mortality is the number of deaths in relation to the number of infected people. This is also often called *Case Fatality Rate (CFR)*. Example: If 5 die out of 50 infected people, the lethality rate is 10% (Sonderegger, 1991).

Understanding the difference between these three concepts is key to understanding an epidemic. Furthermore, it is important to understand how they are related to each other. An infectious disease with a high morbidity is not automatically a problem if not many people show any severe complications or die of it (e.g. many people catch a cold during winter, but as most of them recover within a few days this does not require any further measures). On the other hand, a disease with a lower morbidity can become a real threat if its lethality is high (e.g. in Switzerland, the risk of contracting Ebola is practically 0. However, due to the lethality of the disease, a suspected infection triggers immediate public health measures like isolation and contact tracing (Koch, 2020)). As for

the 1918 influenza pandemic, the morbitidy, mortality and lethality were unusually high for an influenza pandemic (Smallman-Raynor, 2004), which makes it interesting for current research.

2.2 The 1918 influenza pandemic: key facts

The virus that caused the 1918 pandemic was an unusually virulent influenza virus of the strain H1N1, which spread around the globe with unprecedented speed, appearing almost everywhere in the world nearly simultaneously (Morens and Taubenberger, 2018). An influenza virus causes an acute respiratory disease which is colloquially known as "the flu". The symptoms include fever, cough, aches and respiratory complaints which in more severe cases lead to secondary infections such as pneumonia. While influenza was historically seen as mostly unpleasant but leaving no permanent damage, the strain of the influenza virus circulating in 1918 was particularly deadly (Parmet and Rothstein, 2018). It is impossible to determine an exact number of casualties for the 1918 pandemic. The virus that causes influenza was only isolated in the early 1930s and before that, physicians believed that the so-called "Pfeiffer's bacillus" was responsible for the disease. Therefore, the only way of diagnosing an ill person was through clinical procedures (by assessing a patient's symptoms) which is less reliable than modern testing (Van Epps, 2006). Finally, it is always difficult to assign a cause of death for fatal cases. Deaths as a result of cardiovascular disease, pneumonia, or other pre-existing conditions sometimes did have influenza as the immediate cause of death, but in some cases the patient would have died anyway (Patterson and Pyle, 1991). While early research reports an estimated 20 million deaths, more recent studies suggest that the pandemic more likely caused 50 – 100 million deaths globally (Patterson and Pyle, 1991; Johnson and Mueller, 2002). As a comparison: today, an annual average of around 400'000 people die of an influenza infection worldwide, which accounts for around 2% of all respiratory diseases (Paget et al., 2019). Figure 2.1 compares the death toll of a typical flu season with the biggest outbreaks in the last 100 years. It shows that the 1918 pandemic caused more casualties than any other pandemic by orders of magnitude. The two outbreaks 1957 – 58 and 1968 – 69 killed around 2.5 – 4 times as many people as die in an average flu season nowadays. They seem harmless compared to the "Spanish" flu which killed around 125 – 250 as many people. The 1918 influenza pandemic shows to be even more virulent, when one considers that today's population is far larger than it was 100 years ago. In 1918, estimates suggest, the global population was around 1.8 billion

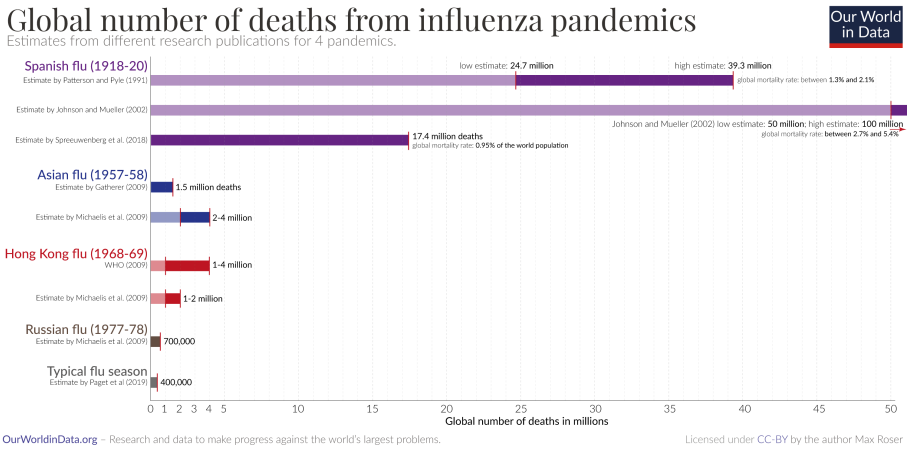

Figure 2.1: Estimated number of deaths of four influenza pandemics in recent history. The chart shows the deaths in absolute numbers not taking poplation size into consideration. Reprinted from Roser (2020)

(Our World in Data, 2020). Therefore, the estimate for the global mortality rate of the 1918 influenza pandemic ranges between 1 – 5.4%. This makes it the deadliest influenza pandemic in recorded human history. With an estimated 7.7 billion people in 2020 (Our World in Data, 2020), a pandemic with a similar moratilty rate would cost between 213 – 427 million lives today. However, when playing with these numbers, one has to consider that today's medicine is far more advanced than 100 years ago and therefore, more lives would be saved (Mills et al., 2004).

Typically, the mortality of influenza by age is described as being "U-Shaped": The very young and the elderly are more likely to die of the disease, while most young adults recover without any damage. However, in the 1918 pandemic the mortality pattern showed a "W-Shape" with a peak in mortality for 30 year olds and lower than expected mortality for 60 – 65 year olds. The reason for this unusual mortality pattern cannot be conclusively determined, but some studies conclude that this was the consequence of earlier H1 infections during an 1889 – 90 influenza outbreak which lead to some cross-immunity in the older population. Young adults, which did not have this prior exposure, were more at risk of dying of the disease (Mamelund, 2011). Furthermore, the pandemic had a high frequency of secondary pneumonia for which young adults, particularly men, seemed to be more suspectible (Jester et al., 2018).

According to Brankston et al. (2007) influenza can be transmitted in the following four ways:

1. **Direct contact**: The transmission occurs as a result of direct physical contact with an infected person (also known as a host).

2. **Indirect contact**: Indirect transmission is the result of contact with a contaminated surface or object (e.g. contaminated surgical instruments, elevator buttons, money, etc.).

3. **Droplets**: Larger droplets are ejected when an infected person coughs or sneezes. Typically these droplets reach a distance of less than one meter in the host's immediate environment. They are too large to remain in the air, therefore special ventilation is not required.

4. **Airborne**: Airborne particles are similar to droplets but much smaller in size and they result from evaporation of droplets. Their small size allows them to remain in the air for a longer period of time and travel further through air currents. They can infect new people if they are inhaled. Controling airborne transmission is the hardest of the four transmission ways because in enclosed spaces it requires control of airflow and special filtering systems.

The 1918 influenza outbreak is colloquially known as "the Spanish flu". Other than the name would suggest, this is not due to the fact that the disease first broke out in Spain. Many countries involved in World War I censored the news of the influenza outbreak out of fear of a declining morale among the population and troops. Neutral Spain did not do so, and therefore Spanish media were the first to report on the outbreak, leaving many thinking that it had originated in Spain (Chowell et al., 2014). Today, this name is often criticized as stigmatizing and incorrect (Hoppe, 2018). The WHO recommends that names of new diseases may not include: geographic locations, people's names, animals or food, cultural, population, industry or occupational references, or terms that "incite undue fear" (World Health Organization, 2015). In this thesis, these reccomendations are accounted for by either speaking of the 1918 influenza pandemic or referring to the disease as the "Spanish" flu, with the word "Spanish" in quotation marks.

There are several hypothesis on where the disease first broke out. Possible starting places include central Spain; Étaples (France), the French countryside or Camp Funston (Kansas, USA), where several soldiers fell ill within a short period of time (Smallman-Raynor, 2004). To this day, it is not certain where the disease originated. While it is not clear where the disease broke out first, research suggests that the strain responsible for the pandemic is an avian flu. This means the virus originated in wild water fowl which would explain why

it broke out in different places at almost the same time – the birds with their migratory behaviour spread it all over the world (Morens and Taubenberger, 2018). In the Northern Hemisphere, the pandemic struck in several waves. The first wave arose in spring 1918. It had a very high morbidity but the mortality was within normal limits. After this mild first wave, the second by far more deadly wave followed in fall/winter 1918/1919. This wave killed millions of people and some places saw a mortality that was three times higher than in the first wave. Some places saw a third, mild wave in spring 1919, before the pandemic was over (Smallman-Raynor, 2004).

In 1918, the medicine was far less developed than today: anti-viral medicine was first developed in the 1950s (Field and De Clercq, 2004). Even if the patient was not dying from the virus itself, but from secondary bacterial pneumonia, their chances of survival were not much better, as penicillin was only discovered ten years later by Alexander Fleming (Fleming, 1929). Remedies included basic supportive medications such as aspirin, quinine, ammonia, turpentine, salt water, or topical rubs (Jester et al., 2018). Therefore, the main strategy for the mitigation of the pandemic consisted of nonpharmazeutical interventions. These interventions were quite similar to what would today be known as "social distancing" and included: school, restaurant and church closures, ban of public gatherings, mandatory mask wearing and disinfection/hygiene measures (Morens and Taubenberger, 2018). Bootsma and Ferguson (2007) found a negative correlation between the date of the implementation of social interventions and mortality for cities in the US. This means the earlier the measures were implemented, the bigger their effect was. However, most cities relaxed the measures a few weeks after their implementation which led to an increase in case numbers and eventually a second wave (Bootsma and Ferguson, 2007).

The pandemic had quite an impact on the demographic structure of the world. In absolute numbers, it is the biggest demographic shock in human history (Sonderegger, 1991), despite the fact that other events killed higher percentages of the population at risk (e.g. the Black Death pandemic killed an estimated 30 – 50% of the European population between 1347 – 1351 (DeWitte, 2014)). However, Smallman-Raynor (2004) concludes that the "Spanish" flu was (1) world-wide rather than regional; and (2) concentrated in time (most deaths occured in a period of just six months). To this day, the reason why the pandemic was so deadly remains unknown. Some theories suggest wartime deprivation as a reason for the high amount of deaths. However, these theories have to be challenged as bot the morbidity and the mortality in North

America, outside of the European war theatres, were similar to the ones in Europe (Smallman-Raynor, 2004).

2.3 Situation in Switzerland and in the canton of Berne

With its location in the heart of Europe, Switzerland was not spared the horrors of the 1918 influenza pandemic. Within Switzerland, the canton of Berne was one of the most severely hit regions. This section gives an overview of the course of the flu pandemic, for Switzerland as a whole and for the canton of Berne specifically.

Studies suggest, that the 1918 influenza pandemic caused around 2 million infections in Switzerland (Sonderegger, 1991). In 1920, Switzerland had a population of around 4 million (Bundesamt für Statistik, 1921), therefore this estimate means the morbidity of the "Spanish" flu in Switzerland was around 50% or more simply: 1 in 2 people were infected. In the years before 1918, influenza deaths were relatively stable with an average of 750 deaths per year. This number rose to around 25'000 in the two years of the outbreak (Sonderegger, 1991). Considering the population, this means the mortality was around 6.25 ‰ and the lethality was around 12.5 ‰. These numbers show the impact of the pandemic in Switzerland. Even though they seem really high, compared to the world-wide numbers discussed in the previous section (see section 2.2), on average, Switzerland seems to have been less affected. The canton of Berne was among the most severely hit places in Switzerland. It accounted for 4'658 deaths out of the almost 25'000 deaths, while the second ranked canton only had half as many deaths (Staub et al.). This means that roughly one in five people that died of influenza during the 1918 pandemic died in the canton of Berne. Again, considering the population of the canton of Berne this translates to a mortality of around 7 ‰ . These numbers show that the canton of Berne was on average more affected than average Switzerland. As similarly described in other regions, the mortality pattern was W-shaped, killing many adults between 20 and 40 years, while the mortality pattern for the other age cohorts remained relatively similar to the years before. Furthermore, men were more likely to die than women (Kohli, 2018). Figure 2.2 shows the number of deaths per age cohort for the years of 1917, 1918, and 1919, (Kohli, 2018). It shows the unusual mortality pattern of the 1918 influenza pandemic with the exceptionally high death numbers in the 20 – 49 age cohortes. This lead to a demographic shock also in Switzerland (Sonderegger, 1991). In 1917, the life expectancy in

Switzerland was around 55.4 years. In 1918, it decreasased to 46.3 years, before it again rose to 55 years in 1919 (Kohli, 2018). These numbers further illustrate the impact that the "Spanish" flu had on the population.

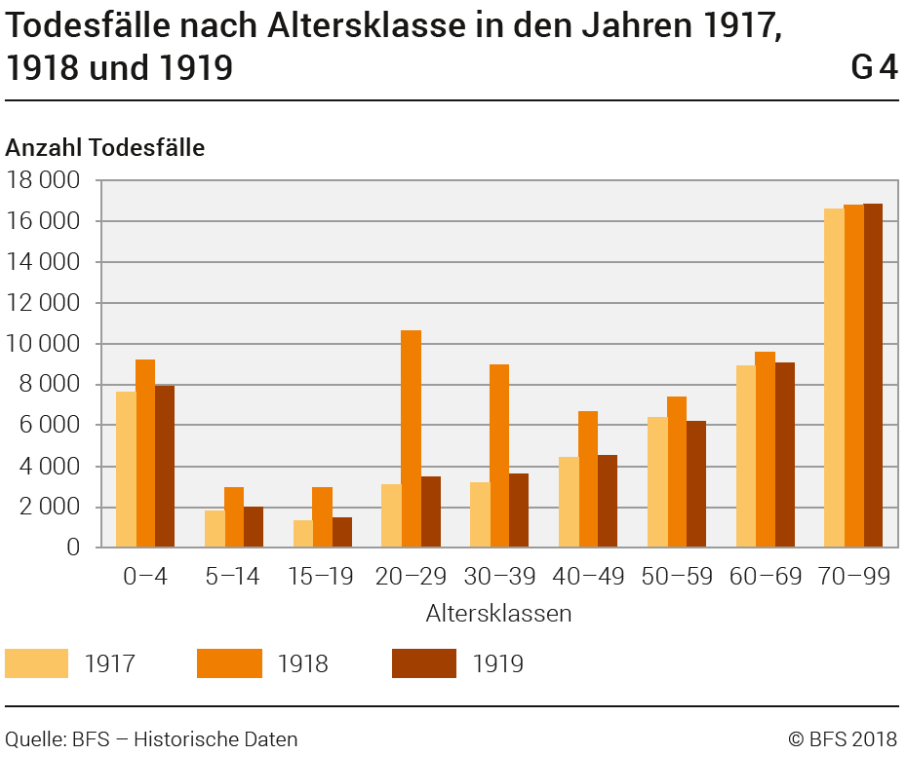

Figure 2.2: Number of deaths in Switzerland by age cohorts 1917 – 1919. The chart shows the spike of deaths in the medium age categories (20 – 49) in 1918. Reprinted from Kohli (2018).

Similar to other places in the Northern Hemisphere, Switzerland was hit by a first, milder wave in July/August 1918, where the mortality remained within the usual limits for an influenza outbreak. It was followed by the much more severe autumn wave from around October – December 1918 that showed a much higher mortality. In this context, the "Landesstreik" (Swiss general strike) has to be mentioned. Fueled by the collapse of the old order and the rise of the international worker's movement, Swiss workers gathered for a nationwide strike in November 1918. During this strike, around 250'000 workers from all over the country crowded together to protest for better working conditions, causing an intervention by the Swiss army (Degen, 2012). This strike could have been the reason for a second peak in the second wave (for weekly infection numbers, see figure 2.3). After two principal waves, a few local outbreaks occured, but

generally the infection numbers kept decreasing and reached an endemic level in July 1919 (Sonderegger, 1991).

As for the spatial dimension, it is unknown how the disease was introduced to Switzerland. One thesis suggests that the disease reached Switzerland from a northern and northwestern direction, where the virus was spreading in the trenches of World War I, and spread via Basel. However, this thesis is not confirmed by mortality rates at the time of the outbreak (Sonderegger, 1991). A further thesis suggests a spread of the virus through Switzerland from west to east: western Switzerland was hit harder by the first wave, while central and eastern Switzerland recorded more deaths during the second wave (Sonderegger, 1991). Finally, cities were affected first, leaving the assumption that they were the starting point of the epidemic (Sonderegger, 1991). According to Sonderegger (1991), in Berne the course was similar to the cantons in western Switzerland (big impact of the first wave), with the exception of high mortality rates in November 1918. The western part and the Jura region were hit harder by the first wave, the Alpine regions were hit harder by the second wave.

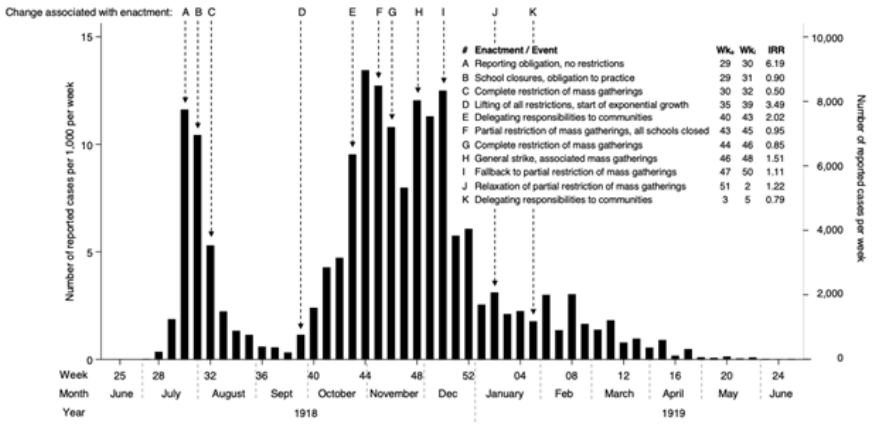

Figure 2.3: Number of cases over time and the interventions in the canton of Berne. The chart shows that measures were both taken and repealed relatively fast during the first wave. During the second wave, the ban on assemblies was kept longer. Furthermore, the graphic shows the general strike and the subsequent second peak of the second wave. Reprinted from Staub et al..

Like everywhere, the measures taken by the Swiss government consisted of nonpharmazeutical interventions. Cantons were given the right to ban all public assemblies in July 1918. Furthermore, it became mandatory to report cases to the federal authorities in October 1918, something which many physicians saw as

an unneccessary bureaucratic step (Sonderegger, 1991; Staub et al.). Therefore, the above-mentioned morbidity and lethality rates remain mere estimates with high dark figures. In the canton of Berne specifically, the government took action to contain the spread of the virus, as shown in figure 2.3.

2.4 Spatio-temporal analysis of influenza pandemics

Several studies on the transmission of the 1918 influenza pandemic exist. One early paper was written by Patterson and Pyle (1991). They show the global spread of the pandemic and report the onset of the disease for different regions of the world and the demographic consequences the pandemic had. They conclude (1) that the world had become one epidemiological unit (i.e. the disease spread with unpreccedented speed and reached almost every corner of the world) and (2) that the fall wave was responsible for the majority of deaths. Mills et al. (2004) focus on the general transmission of the virus and try to estimate a reproduction number of the virus. Their findings show that with the right measures, antiviral medicine, and generally improved health care, today, a similar virus could be mitigated even though increased air travel would speed up the long-distance transmission.

There are several studies that focus on specific locations. Chandra and Kassens-Noor (2014) describe the spread, mortality and evolution of the virus in India. They conclude that the pandemic slowed down with time and that the virus became less virulent. Furthermore, they suggest that weather could have an influence in the spread of influenza: places with less monsoon-rain than usual were more severely hit by the pandemic. Reyes et al. (2018) also focus on India. They collected different demographic and environmetal variables to find factors for the spread and conclude that long-distance travel via railroad was an important driver in the spread of the virus.

Olson et al. (2005) focus on the age-specific mortality rate in New York City by calculating the excess mortality compared to a baseline mortality from monthly and weekly deaths. Their data show a strong evidence that an early wave of the pandemic was already present in February – April 1918. These results are supported by Yang et al. (2014), who focus on the outbreak duration and calculated the reproductive number (the average number of people one patient infects). Their findings show that the mortality among young adults was generally higher than among other age groups and that school children might have been an important driver in the spread of the pandemic in New

York City. Gog et al. (2014) focus not on the "Spanish" flu but on the H1N1 influenza pandemic of 2009 which is colloquially known as the "swine flu". They received a dataset containing doctor's visits for influenza-like illnesses and fitted a transmission model using these data. They conclude that in 2009, short-distance spread played a dominant role rather than long-distance events. This short-distance spread was further catalyzed by the opening of schools. This highlights the dominant roll children play in transmitting influenza. In another study Eggo et al. (2011) fitted a range of city-to-city models to mortality data in England, Wales, and the US. Their findings show that long-distance spread played a big role in the beginning of the outbreak and as the disease became more widespread, short-distance transmission played a major role in disseminating the virus. The reason for this seems to be that with time, the disease became more widespread and measures were taken to prevent the spread. Therefore, long-distance contacts were shut down, and local transmission became the main driver for the disease (Eggo et al., 2011). A second study with a focus on England and Wales was conducted by Chowell et al. (2008). They studied death rates, transmissibility and various geographic and demographic indicators in English and Welsh cities, towns and rural areas by estimating the reproductive number using the deaths as a proxy for incidence. They found varying death rates, where rural areas were affected more. By contrast, they found no association between population density and death rate. What they did find was a low correlation between household size and death rates, both for rural and urban areas for the winter wave. Furthermore, they found a correlation between urbanness and onset of the pandemic, where the onset in more urbanized areas was earlier than in rural areas. They call for further geographic studies to explore the pattern of the influenza outbreak. Smallman-Raynor et al. (2017) study the pandemic on both the national and local level in the United Kingdom. Their national results show the characteristics of the waves. The mild summer wave spread relatively fast from north to south. The second wave on the other hand was – apart from its much higher deadliness – slower and moved from south to north. The third wave had a similar spatial pattern as the first one: faster and moving from north to south. The spread of all of the waves was characterized by a clear spatial contagion. The local study focuses on Cambridge, a city which, apart from the famous university, was also home to a large number of naval troops. This presents the unique opportunity to study the transmission patterns on local level among a diverse range of groups with different demographic characteristics. The starting point of the outbreak was returning naval cadettes but the disease

spread quickly among the general population. Chowell et al. (2008)'s findings further show a peak in incidence in the 15 – 35 age group.

Another country which, compared to others, is well-studied is Spain, which is one of the countries that experienced a very high mortality burden (Chowell et al., 2014). Spain was struck by three pandemic waves with varying timing and intensity where the most severe one was the second wave in October – November 1918. Chowell et al. estimated the excess death rates from respiratory deaths in the provinces of Spain. Then, they explored the associations between the excess deaths and different socioeconomic factors. They found a north-south gradient in excess mortality rates with higher mortality in the north. Their model included latitude, population density and the proportion of children and explained about 40% of the geographic variation. However, this geographic variation can be attributed to different factors in each wave. The substantial unexplained percentage suggests that other factors (that were not included in their model) played an important role in the dissemination of the disease. These could for example be co-morbidities, climate, or background immunity. A more recent study by Cilek et al. (2018) assessed the severity on the three pandemic waves in Madrid by looking at the age-specific excess death rates for respiratory diseases and other causes. The findings do not support the before-mentioned "W-Shaped" mortality pattern but instead a high excess mortality rate among the youngest and oldest in the population.

A few studies with a focus on Switzerland also exist. Zürcher et al. (2016) focus on the city of Berne and on Switzerland as a whole. They use Poisson regression models to quantify the excess pulmonary tuberculosis deaths attributable to influenza. Their data show that yearly PTB mortality increased during the "Spanish" flu. Furthermore, several studies exist on the canton of Geneva. Chowell et al. (2006) try to estimate the reproductive number of the "Spanish" flu. Ammon (2002) focuses on studying the socioeconomic burden of the disease. She studies the disruptions the virus caused, e.g. the frequent school closures and the overcrowding of the hospitals. According to her, one of the biggest problems in the 1918 outbreak was the inconsistency of the measures taken, which further contributed to a climate of insecurity. Furthermore, there are a few studies that focus on specific cantons in Switzerland. Sonderegger (1991) has a focus on the canton of Berne. Besorger (2018) focusses on the canton of Zug in his analysis. The results show that the second wave was far more virulent in the canton of Zug than the first wave. Further studies with a focus on Switzerland cover the cantons of Basel Stadt/Land (Tscherrig, 2016), Nidwalden (Tscherrig, 2018), and Valais (Marino, 2014). These studies are all

rather qualitiative and show the different courses of the epidemic in the different cantons. However, a quantitative approach for a canton as regionally diverse as Berne is still lacking.

Finally, Staub et al. used the same data as used in this project to calculate relative incidence rate ratios to assess the change in incidence connected to public health interventions. The findings show that during the first wave, school closures and restrictions on mass gatherings were associated with a reduction of new cases. During the second wave, the cantonal authorities hesitated to take measures, and instead delegated the responsability to the municipalities and the association between the interventions and the reduction in cases was less distinctive.

2.5 Possible determinants of spread

In the literature part so far (section 2), several studies were presented that identified locally specific factors that explain the spread of the influenza pandemic. Below, some factors that might help explain the dissemination of the disease are summarised. In a next step, these factors will be used in a logistic regression model to see if they help explainig the spread of the 1918 influenza strain in the canton of Berne.

The following socio-economic and physical factors have been included in previous studies and found to contribute to explaining the spread of the 1918 influenza pandemic:

(a) **Proportion of people working in agriculture**: Bengtsson et al. (2018) found notable differences in excess mortality between social classes. They analyze the risk of death for white collar workers, high-skilled manual workers, low-skilled manual workers, unskilled manual workers, and farmers. While they found no notable differences between the first four social classes, people working in agriculture had a significantly lower risk of death (Bengtsson et al., 2018).

(b) **Tuberculosis mortality**: Several studies suggest that TB also played a role in the 1918 influenza pandemic and could have influenced both morbidity and mortality (Mamelund, 2011; Zürcher et al., 2016). Areas that have been hit harder by tuberculosis might also have been hit harder by influenza, because the risk of severe cases or contracting bacterial pneumonia could have been higher, therefore leading to more severe cases (Mamelund, 2011).

(c) **Urban vs. rural settings**: Evidence suggests that the panemic first reached the cities and from there spread to the surroundings (Sondereger, 1991). Chowell et al. (2008) found no association between population density and death rates but concluded that the onset of the pandemic was earlier in urban areas. Since there are no studies that found an association between population density and transmission, urbanness rather than population density is considered in this thesis.

(d) **Accessability**: In their study, Reyes et al. (2018) found that the number of passengers travelling on railway lines was an important factor in explaining the spread of the pandemic in India.

(e) **Weather**: Roussel et al. (2016); Chandra and Kassens-Noor (2014) and Reyes et al. (2018) found that weather had an influence on the spread of the 1918 influenza pandemic. In the case of Reyes et al. (2018) the rainy season in India showed a negative association with influenza cases. Chandra and Kassens-Noor (2014) found that places with below-average monsoon were more severely affected by the pandemic. Roussel et al. (2016) analyzed different climatic factors such as sunshine, duration or humidity, and found out that climate may play a role in the spread of influenza.

This list is by no means absolute, as several potentially important factors are missing (e.g. literature further states that the age pattern is an important explanatory factor (Mamelund, 2011; Jester et al., 2018)). These are not included in this list, as there was no data available on a sub-national level.

2.6 Research gaps and research questions

In general, there are not many studies that focus on the spatio-temporal analysis of the 1918 influenza pandemic as many of the above-presented papers are merely descriptive studies that estimate numbers of cases and deaths and the pandemic's impact on public life. The spatial component of the dissemination is often omitted in these studies, especially if they focus on Switzerland. Secondly, most studies focus on death records instead of incidence reports. Therefore, the following research gaps were identified:

Research Gap 1: Missing information on how the virus spread in the canton of Berne (and Switzerland)

So far, no quantitative and geographically comprehensive study of the dissemination of the "Spanish" flu in the canton of Berne exists. Accounts on this area are mainly qualitative descriptions of the course of the pandemic. Furthermore, many studies focus on national or cantonal entities, or on cities. To date, there are no studies that focus on such a small scale as the municipality level. This thesis is a first step to overcome this gap by conducting a case study on the municipality level for the canton of Berne.

Research Gap 2: Missing information on what locally specific factors contributed to the spread

The influence of different factors (weather, transportation, space, socio-economic factors, etc.), have been described in various studies in other countries for example by Chowell et al. (2014); Eggo et al. (2011); Reyes et al. (2018), etc. (see section 2.5). Knowing which locally specific factors play a role in the dissemination of airborne diseases can help to identify measures that help preventing and mitigating future outbreaks (e.g. by putting an emergency concept into place as reccomended by the World Health Organization).

Research Gap 3: No effective communication of research

Communication research results is often seen as not that important and therefore often largely omitted (Sonderegger, 1991), which leads to the problem that research results are barely noted outside of the scientific community. With the "Spanish" flu being such an important part of human history (and an increased interest due to the ongoing Sars-Cov-2 pandemic), I argue that communication and presentation of research is of importance. The presentation will be easily accessible requiring no special skills to understand it. There will be no research question associated with this research goal, however throughout the answering of the first two research questions an emphasis will be put on visualisations that support an effective communication of the results.

Based on these research gaps, the thesis will address the following research goals and questions:

Research Goal 1: Descriptive spatio-temporal analysis of the influenza data

The first part of the project is to conduct a spatio-temporal analysis of the 1918 influenza pandemic in the canton of Berne. This part is mostly descriptive and includes steps like determining the incidence rates and observing how they change over time. Furthermore, the global and local Moran's I statistics is calculated to take spatial distribution into account. This part is a first step to fill the knowledge gap of how the pandemic spread in such a regionally diverse canton as Berne.

> **Research question 1**
> According to which spatio-temporal patterns did the 1918 influenza spread in the canton of Berne?

Research Goal 2: Finding determinants of spread

The second part of the analysis is to create a data-driven epidemiologic model to determine the spread of the influenza pandemic of each of the two waves in the canton of Berne. Based on various explanatory variables, which are taken from existing frameworks, contributing factors for the spread are identified.

> **Research question 2**
> Which locally specific factors determine the spread of the 1918 influenza pandemic in each of the two principal waves?

Research Goal 3: Effective visualisation of results

There is no research question associated with the last research goal. However, special emphasis will be put on the visualisation of the findings in order to facilitate a better understanding of the results. If maps are created, cartographic standards will be followed and an emphasis will be put on making the visualisations more accessible (e.g. in resprect to color vision defficencies).

Based on previous literature (section 2), a hypothesis is created for each research question.

Hypothesis 1

In the canton of Berne, the national patterns also hold true. This means the epidemic generally spread from the west towards the east.

Hypothesis 2

The spread of the virus can be explained with locally specific factors such as socio-economic (e.g. percentage of people working in agriculture, urbanity, etc.) or physical (e.g. accessibility, weather) factors. For each factor, a sub-hypothesis that describes the relationship between the outcome and each explanatory factor is formulated:

(a) The higher the proportion of people working in agriculture, the lower the incidence of influenza in a municipality.

(b) The higher the average TB mortality, the higher the incidence rate.

(c) The incidence was higher in the urban spaces than in rural areas.

(d) The better a municipality's access to the railway system, the higher the incidence.

(e) The higher the precipitation, the lower the incidence.

3 Methodological approach

3.1 Data sources and data preprocessing

3.1.1 The canton of Berne – a short geography

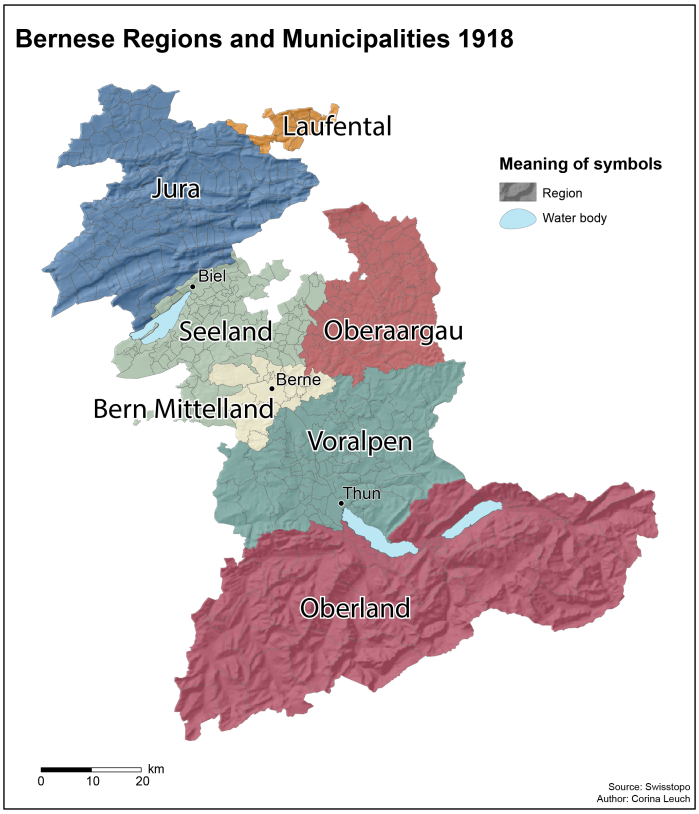

Figure 3.1: The greater Regions of the canton of Berne. Additionally, the relief shows that the Jura and the Oberland region are particularly mountainous.

In order to understand and correctly interpret the data, a few basic facts about the geography of the canton of Berne have to be established. In 1918, the canton of Berne was the second biggest canton of Switzerland in terms of surface (6798 km^2) and the one with the biggest population (around 675'000 inhab-

itants) (Bundesamt für Statistik, 1921). The three Swiss geographic regions can also be found in the canton of Berne: The Jura Mountains in the north, the Swiss Plateau in the central parts, and the Alps in the south of the canton. Furthermore, the canton of Berne consisted of 497 municipalities and 30 administrative districts. Figure 3.1 gives an overview of the municipalities, the regions, and the three biggest cities (Berne, Biel, Thun). The Jura region is part of French-speaking Switzerland while in the rest of the canton, the majority, is German-speaking. Figure 3.2 shows the population density on a municipality level for the entire study area. The map shows that the Swiss Plateau was heavily populated especially around the three biggest cities Berne, Biel, and Thun. The Jura region had some more heavily populated area between sparsely populated small municipalities. Finally, in the Bernese Oberland the municipalities around the lakes had a tendency to be a bit more heavily populated while the regions further in the south and the east were very sparsely populated mountain areas.

3.1.2 Sample

The outcome dataset is the same as used by Staub et al. and therefore the sample is the same. It contains raw case numbers of influenza-like illness based on doctors' reports. Berne was one of the first cantons to declare influenza a notable disease immediately after the outbreak in July 1918 (Regierungsrat des Kantons Bern, 1918). From then on, doctors had to report the number of influenza cases to the district authorities who were responsible to control that the doctors fulfilled their reporting duty (Staub et al.). These reports are now available in the cantonal archives of the canton of Berne (Sanitätsdirektion des Kantons Bern, 1918).

The received sample includes 143'389 disease reports, among which 131'725 were influenza reports and covers the time period from July 1918 to December 1919. An excerpt of these disease reports is shown in figure 3.3. The sample includes influenza cases from 472 municipalities (95% of all municipalities). Several of these reports had some inaccuracies: in 4723 cases (3.2%), there was no excact date/number pair available (reports like "10 cases in two weeks", "many cases"). For these cases, the middle date of the range was used as a reporting date or the number of cases from the immediately preceeding or following report was used. In another 2005 cases (1.4%), no exact municipality was reported but only the administrative district ("im ganzen Bezirk" – "in the entire district"). These cases were omitted for the analysis on a municipality level as no exact municipality could be assigned.

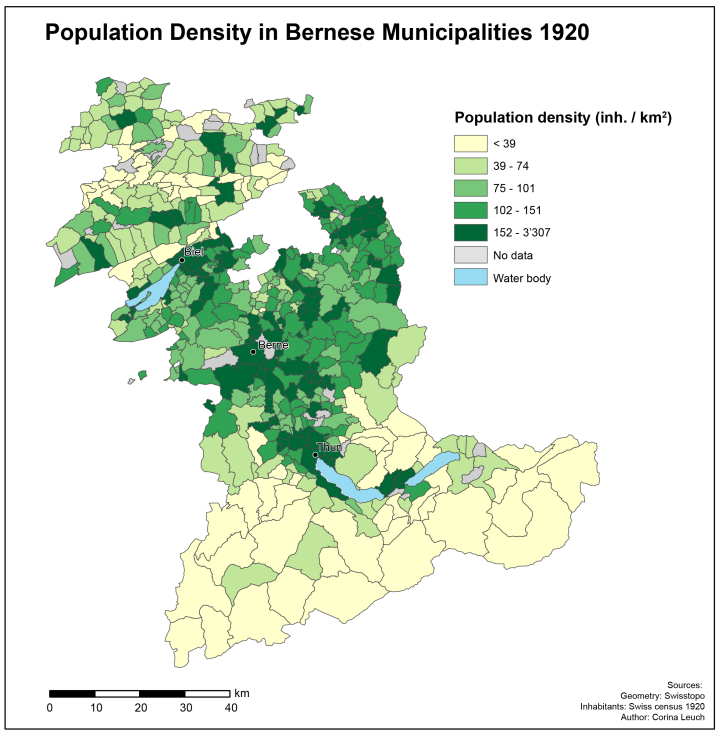

Figure 3.2: Population density per municipality in 1920. The darker an area, the higher the population per km^2. The Swiss Plateau (Seeland, Bern Mittelland, Oberaargau, and Voralpen) was more heavily populated than the mountainous regions (Jura/Laufental and Oberland).

3.1.3 Outcome variable

The outcome variable is the incidence rate of influenza in the canton of Berne. The raw reports of influenza-like illness had to be standardised by calculating incidence rates (cases per 100'000 inhabitants) for each municipality. As a standardisation size, the number of infected people per 100'000 inhabitants is chosen because it is a known ratio that is also often used nowadays in the presentation of Covid-19 cases. The population data needed for these calculations originates from the Swiss federal census of 1920 (Bundesamt für Statistik, 1921). This population dataset was collected 1 – 2 years after the flu took place, and therefore is influenced by the pandemic as well (i.e. places with high death tolls would have less inhabitants because of the pandemic). However, at the time a census only took place every ten years, therefore using the one from 1920 is the most accurate data available.

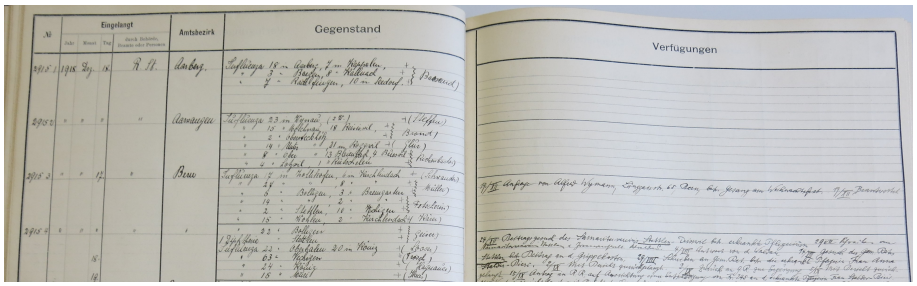

Figure 3.3: Excerpt from the original data as found in the cantonal Archives of Berne (Sanitätsdirektion des Kantons Bern, 1918). For this thesis, the data was already present in a digital form.

The temporal resolution of the data is 1 day meaning the reports contained an exact date. However, the analysis on a daily level does not make much sense in this context as daily disease reports fluctuate quite a bit. Therefore, the daily reports were temporally aggregated depending on the type of analysis. For the analysis of the spatial component (see figure 4.5), the were data temporally aggregated for the entire study period and per municipality, to gain an overview which areas were overall particularly affected by the pandemic. In the next step, the data were summed up by month for the months from July 1918 to June 1919, to bring in the temporal aspect and observe which places were affected when and how severely. For the epidemiological analysis, only the two principal waves were of interest. The available data covers influenza reports from July 1918 until the end of December 1919. Since this covers more than this time frame of interest, two subsets were selected:

1. Subset 1 consists of the reports that were dated for the months of June and July 1019. This was when the first spring/summer wave took place (see section 2.3)

2. Subset 2 consists of the reports that were dated for the months of October 1918, November 1918, December 1918, and January 1919. This was when Switzerland was struck by the second, more severe wave (see: section 2.3)

The smallest spatial resolution available is the municipality level. This level is used throughout all of the project because it allows for a fine-grained analysis of the pandemic with a potential to capture small-scale regional differences. For the logistic linear regression model, the data were split into two categories 0 and 1, where 0 is the bottom 80% of the incidences and 1 the top 20% of incidences for each of the two waves.

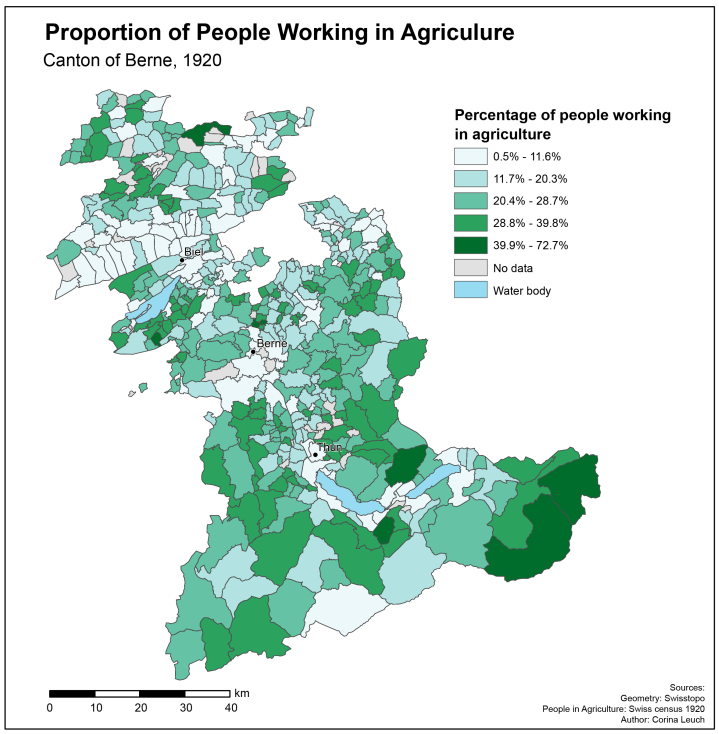

Figure 3.4: Proportion of people working in agriculture for each municipality. It is noteworthy that the southern Jura region had low proportions of people working in agriculture despite its seemingly rural setting.

3.1.4 Explanatory variables

Given their broad nature, the explanatory variables originate from a variety of sources. The socio-economic variables (proportion of people working in agriculture and number of inhabitants per municipality) originate from the Swiss national census of 1920 (Bundesamt für Statistik, 1921) and were already digitized. The rest of the variables comes from a variety of sources and had to be calculated first. A detailed overview of the distribution of the data and their correlation can also be found in the results section, specifically in figure 4.10 for the first wave/model 1 or figure 4.12 for the second wave/model 2.

Proportion of people working in agriculture

Bengtsson et al. (2018) conclude in their study that areas with a higher proportion of people working in agriculture were less affected than other areas. The idea was to test whether this also was true for the canton of Berne. The

number of people working in agriculture was standardised using the number of inhabitants in each municipality from the Swiss census 1920. This resulted in the percentage of people working in agriculture for each municipality. The distribution roughly followed the normal distribution, and it was not further classified.

The map in 3.4 shows the spatial distribution of the proportion of people working in agriculture. Generally speaking, the proportion of people working in agriculture was still quite high in 1920 particularly in the southern and eastern part of the Bernese Oberland, and south of Lake Biel. Noteworthy are the three cities Berne, Biel, and Thun that are clearly visible on this map with their low percentages of people working in agriculture. Furthermore, the southern part of the Jura region shows low values in the proportion of people working in agriculture. This can be attributed to the watch making industry which was resident in the southern Jura region (Fallet, 2020).

TB mortality

Mamelund (2011) finds that tuberculosis was an important explanatory factor for explaining both high morbidity and mortality of the 1918 influenza pandemic. The digital influenza dataset (the main dataset used for the outcome variable) also contained a few reports of tuberculosis. However, these reports only covered two municipalities and were not sufficient to be used in the analysis. Therefore, another solution had to be found: Every ten years, the canton of Berne published statistics that report on the number of tuberculosis deaths on a municipality level (Staatsarchiv Kanton Bern, 1910). For the purpose of this analysis, the most recent statistics from the years 1900 – 1910 was used (Staatsarchiv Kanton Bern, 1910). This gives an estimate of how high the prevalence of tuberculosis might have been in the different municipalities.

The map in figure 3.5 shows the spatial distribution of TB mortality in the canton of Berne. Compared to the rest of the canton, the Jura region had the highest TB mortality. In the Swiss Plateau, there were some areas that were slightly more affected, particularly around Lake Biel, north of the city of Berne, and in the region of Oberaargau. In the Bernese Oberland, the regions closer to the lake as well as in the very eastern part showed a higher TB mortality in the years prior to the 1918 influenza pandemic. The southern part of the Bernese Oberland did not seem to be much affected.

Just like the influenza data, the tuberculosis data was standardised by calculating the mortality per 100'000 inhabitants. Afterwards, the data was divided into two categories 0 and 1, where 0 represents the bottom 80% of the mortality

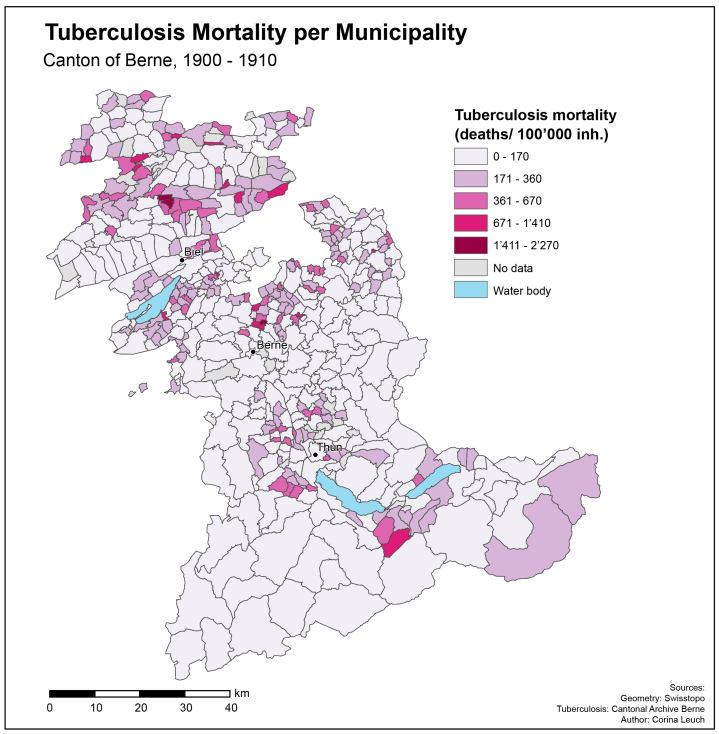

Figure 3.5: TB mortality (deaths per 100'000 people) in each municipality for the years 1900 – 1910. The map shows that the Jura region, areas around lake Brienz, and areas north of Berne were particularly affected by tuberculosis.

and 1 the top 20% with the highest mortality. One point that has to be noted is that the TB data from 1910 was standardised using the census from 1920 as it was the one available in a digital form. This introduces some potential inaccuracies. However, it remains open how using a census closer to 1910 would have influenced the result. Also, later on the data were classified into two categories, therefore small differences in mortality would have been "swallowed" by the classification anyway.

Urban vs. rural settings

The canton of Berne at the time of the 1918 influenza pandemic was still very rural with the exception of the three urban centres Berne, Biel, and Thun. There was a clear gap between these centres and the periphery with regard to development, but also in their number of inhabitants (Pfister, 2011). This classificaton is partially supported by the Swiss census of 1920 which lists Berne

and Biel as cities (Statistisches Bureau, 1921). Particularly, urban areas showed an earlier onset of the pandemic: several studies (Sonderegger, 1991; Chowell et al., 2008) found that urbanity was an explanatory determinant of spread of the 1918 influenza pandemic. This was considered for the analysis. The three urban areas Berne, Biel, and Thun were put into a single category and all other municipalities formed a second category. The idea was to test whether urban settings behaved differently than rural settings.

Access to the railway network

As stated in section 2, access to the railroad network was an important driver in the 1918 influenza pandemic in India (Reyes et al., 2018). The idea was to test if this also holds true for the canton of Berne on a smaller level. For this purpose, I obtained a dataset containing the Swiss railway stations in the year 1900 from Egli, Hans-Rudolf, Flury, Philipp, Frey, Thomas, Schiedt (2005) and the Swiss railway network from Büchel and Kyburz (2018). Several new lines were built between 1900 and 1918 (e.g. much of the Rhaetian Railway network (Rhätische Bahn AG, 2020)) and therefore had to be added into the network by hand. This was done by digitizing the misssing railway stations and lines using the online geoportal of the Swiss confederation and the tab "Zeitreise – Kartenwerke" (Bundesamt für Landestopographie (swisstopo), 2020). Furthermore, railway lines that had a purely touristic purpose were deleted for the purpose of this analysis (e.g. Rigi Railway, Brienz Rothorn Railway, etc.).

Afterwards, an undirected navigable railroad network was built using R version 4.0.2 and the packages igraph (Gsardi and Nepusz, 2006) and tidygraph (Pedersen, 2020). Unfortunately, no passenger data was available, therefore another proxy for the importance of a station within the network had to be found. This was achieved by calculating the node betweenness centrality which determines the importance of a node in a network by calculating the number of shortest paths that pass through a network (Geisberger, Robert & Sanders, Peter & Schultes, 2008). This number was then used as a proxy for how well a municipality is accessible through the railway network. Initially, the question was posed whether this analysis should be conducted for the entire country or just the canton of Berne (i.e. the area of interest). It was decided to use the entire country as an input because the railway network functions as a national network. Cutting out just the stations that are on the area of the canton of Berne would have led to several unconnected graphs which would have led to false results, especially in the area around the region of Oberaargau, where the railway lines cross the cantonal boundaries several times.

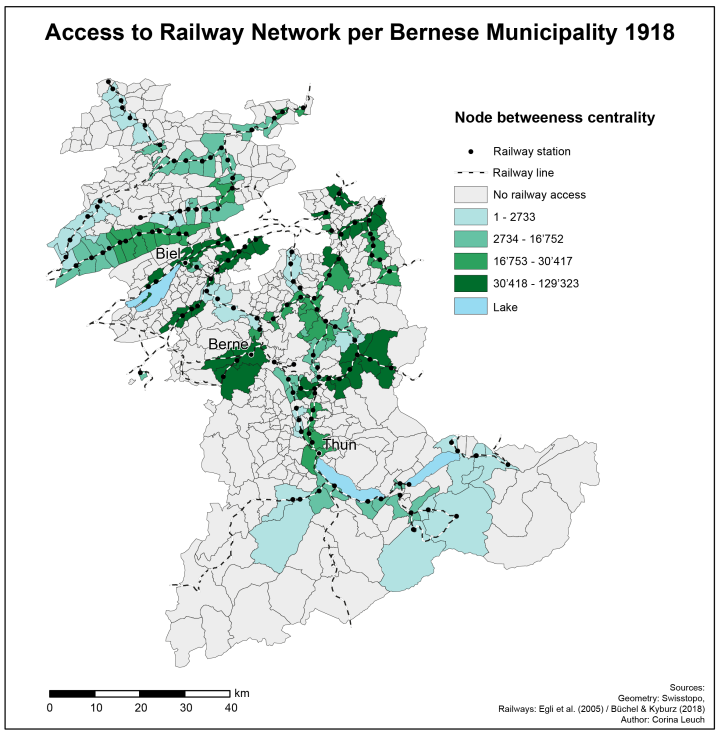

Figure 3.6: Node betweenness centrality per municipality in the canton of Berne. The node betweenness centrality is a measure of how central a node (in this context a railway station) is in a network. The map shows high values for the areas around Berne and Biel and particularly in the region of Oberaargau. Municipalities that have no access to the railway network are shown in light grey.

In a next step, the nodes were joined to the geometry of the municipalities. If one municipality had more than one railway station, the one with the largest value was selected to simulate the highest possible centrality. This way was chosen because in all cases where a municipality had more than one railway station they were all part of the same line, thus not making the place more accessible. After an evaluation of the node betweenness centrality, it was deemed necessary to split the data into groups, also to make the model easier to interpret. Splitting the data into quantiles was not possible as there was an excess rate of zeros in the data (256 municipalities had no railway access at all). Therefore, another way had to be found. Originally, a binary model with the categories “railway access” /“no railway access” was considered, but this would have meant an information loss because it would not have accounted for how central a station was within

a network (e.g. the muncipalities of Brienz and Berne would have been in the same category, even though Berne is much more accessible), hence an alternative way had to be found. All municipalities with no railway station were put into the same category. Afterwards, the remaining municipalities were split into four more categories, making a total of five categories. They were split according to their node betweenness centrality using an algorithm that minimizes the differences within categories and maximizes the differences between categories and that was originally developped for the use in cartography (Jenks and Caspal, 1971).

The map in figure 3.6 shows which areas in the canton of Berne were easily accessible by train. The map also shows that the accessibility is higher around the three cities Berne, Biel, and Thun. Particularly the region of Oberaargau contains areas with a high node betweenness centrality. This has to do with their relatively central location within Switzerland. Furthermore, the southern part of the Jura also has areas with high node betweenness centrality. In the Oberland, we see a difference between the low-lying regions around the Lake Brienz and Lake Thun that were accessible by rail and the harder-to-reach municipalities in the very south and east.

Precipitation

The weather data was obtained from the Swiss Federal Office of Meteorology and Climatology MeteoSwiss. The data included 29 stations spread out through the entire canton of Berne. The precipitation data covered daily measurements for the years of 1918 and 1919. This data was summed up for (1) the entire study period (July 1918 until December 1919 – see figure 3.7), (2) the first wave/ model 1 (July and August 1918), and (3) the second wave/model 2 (October 1918 until January 1919). Afterwards, a raster covering the entire extent of the canton of Berne was interpolated using the IDW (inverse distance weighted) method (Philip and Watson, 1982) available in the ArcGIS Spatial Analyst extension (ESRI, 2020a). Using Zonal Statistics (ESRI, 2020b), an average value for each municipality was calculated. This calculation was done for the three subsets (1 – 3) mentioned previously. Afterwards, the data was split into quintiles. Modelling quintiles instead of the raw data has several advantages: Firstly, precipitation data, by its nature, has a huge range of values, therefore modelling a one unit increase would not have yielded any meaningful results. Secondly, the data was interpolated and contained inaccuracies. Therefore, modelling the data in quintiles allowed to control for general tendencies, and reduces the risk of misinterpretation due to inaccurate interpolation results.

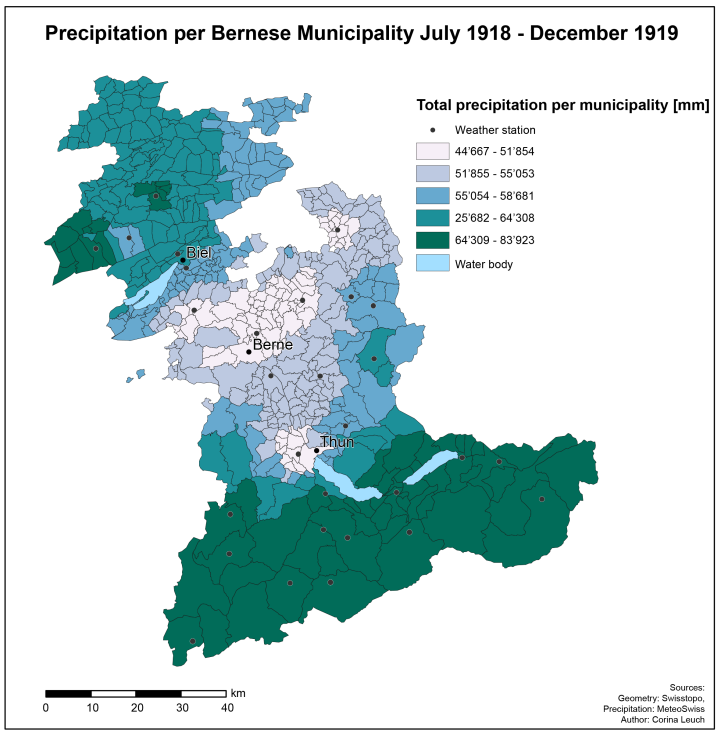

Figure 3.7: Total precipitation per Bernese municipality for the entire study period from July 1918 through December 1919. The mountain regions of the Alps and the Jura received more precipitation than the lower parts of the canton.

The map in figure 3.7 shows the amount of precipitation received in the canton of Berne per municipality for the entire study period (July 1918 through December 1919). The pattern follows the general precipitation pattern for Switzerland: the Oberland received the most precipitation, followed by the also mountainous area of the Jura region. The amount of precipitation decreases towards the low-lying areas and reaches its minimum in the area around the city of Berne. Finally, there are two areas in Thun and in the region of Oberaargau that also received little precipitation.

3.1.5 Geometry

Originally, most of the data was organized in two datasets. One dataset was a table containing all the municipalities, their socio-economic factors, a unique ID, as well as coordinates marking the centre of each municipality, which allows the dataset to be turned into a spatial dataset easily. The second dataset

contained the disease reports as well as their date and the same unique ID as the municipality dataset. For the purpose of a better visualization, it was seen as more meaningful to visualize the data as municipality polygons instead of a point representation. Unfortunately, no polygon dataset of the municipalities at the time of the pandemic was available and therefore one had to be created by hand. The basic dataset was created by the Swiss Federal Office of Topography and was kindly provided by Egli, Hans-Rudolf, Flury, Philipp, Frey, Thomas, Schiedt (2005). It contained the municipality boundaries as they were in the year of 1990. An older dataset was not available and therefore had to be manually created. The Swiss Federal Office of Topporgraphy provides old maps for Switzerland in their online map application (map.geo.admin.ch) under the tab "Zeitreise – Kartenwerke", as well as the possibility to import datasets into the online map application (Bundesamt für Landestopographie (swisstopo), 2020). The 1918 municipality polygon geometry was created by importing the available geometry data set from 1990 and adding missing boundaries by hand. Afterwards, these boundary drawings were downloaded and used to split the existing polygon geometry from 1990. Finally, the two datasets (the point data and the newly created polygon data) were matched using a spatial join. This way, the polygon dataset also contains the unique id of the point dataset which allows an easy matching of the influenza data with the geometry for the creation of better visualisations.

3.2 Methods

3.2.1 Research goal 1: Descriptive spatio-temporal analysis of the influenza data

The idea of this descriptive analysis was to gain an overview of the data, their potential and limitations. This was an important step for the next research goal because having a good overview of the data was essential for building a statistical regression model in a later stage. From an analytical point of view, the data can be split into three dimensions: (1) the incidence, (2) the temporal dimension, and (3) the spatial dimension. All three dimensions are related to each other. This theoretical division into these dimensions allowed to start the analysis simply by only looking at one dimension. Later on, the complexity could be increased by subsequently adding the other dimensions which should yield in more insight into the data and its patterns.

Incidence

Out of the previously mentioned three dimensions, the content-related dimension was arguably the most important one. It was also the only one that in this context could be analysed individually (assessing only the time or only the space would have made little sense in the context of this research). The analysis started simply by calculating conventional statistical key figures such as median, mean, standard deviation, etc. and presenting them in the form of boxplots in a first approach to describe the simple characteristics of the data. A second approach for an isolated analysis of the incidences was looking at the Lorenz curve. The Lorenz curve compares cumulative incidence numbers with cumulative population size among municipalities (ranked from the municipaliy with the lowest number of inhabitants to the one with the highest number of inhabitants) (Lee, 1997). In a Lorenz curve, the main diagonal means perfect equality, which in this case would have meant an equal distribution of case numbers according to their population size (i.e. the incidence is the same in each municipality). The further away the actual curve is from this main diagonal, the more inequality is present in the data.

Incidence and temporal dimension

The second step of the analysis was to determine the incidence over time. One first attempt to achieve this would have been by looking at new infections by time. This was already done by Staub et al. (see figure 2.3) and was already discussed in the state of the art section. Another aproach was to look at daily new disease reports compared to total disease reports. This graphic – although not as intuitive as looking at the daily (or weekly) new infections – provided further insights into the data.

Incidence and spatial dimension

One first approach to assess the spatial distribution of the cases was summing them up by municipality and visualizing them on a choropleth map (Slocum et al., 2008). This allowed a visual assessment of the data and helped to identify potential regional differences. Afterwards, the two dimensions were analyzed in a more formal way by applying spatial autocorrelation analyses. The global Moran's I is designed to reject the null hypothesis of spatial randomness in favor of the alternative hypothesis of clustering. This gave a first insight whether the incidences are randomly distributed in space or spatially dependent. The global Moran's I test does not give any indication as to the location of these clusters

(Moran, 1948). In order to gain insight into the location and the type of cluster (high rates or low rates), the local Moran's I was calculated (Anselin, 1995) using the software GeoDa (Anselin et al., 2003). The local Moran's I is an algorithm that compares the value of each observation with its neighbours and finds clusters of high-high or low-low rates. Furthermore, a p-value for each location was calculated which allowed to assess the statistical significance of each location. When combined with the location of each observation in space, this allowed a very powerful interpretation: a classification of the significant locations into clusters of high-high or low-low rates. Furthermore, outliers of the form high-low and low-high were calculated also using the software GeoDa (Anselin, 1995).

Incidence, spatial dimension and temporal dimension

In the last step of the discriptive analysis, all the three dimensions were combined. Firstly, the incidences were aggregated per month and municipality to produce maps that showed incidences per month per municipality. This again allowed to visualize how the pandemic spread across the canton of Berne. Therefore, this was a more profound descriptive analysis, that relied on disease data rather than obervations from historical sources. Secondly, similar maps were produced with incidence summed up per municipality for the entire duration of each of the two principal waves. Literature suggests that the different regions of the canton of Berne had different spatio-temporal characteristics (Sonderegger, 1991)(see section 2). Finally, the local Moran's I statistic was calculated for both waves using the software GeoDa (Anselin, 1995). These cluster maps then serve as a baseline to compare the model residuals to.

3.2.2 Research goal 2: Finding determinants of spread

Model design

For the statistical analysis, two linear logistic regression models were built. Model 1 included the incidence data from July and August 1918 and Model 2 included the incidence data from October 1918 until January 1919 (for a complete overview of the different variables and how exactly they were classified, please refer to section 3.1). These two time frames for the models were chosen based on the spatio-temporal analysis in research question 1 as well as suggestions from literature (Staub et al.). The spatial resolution was the municipality level, to observe small-scale local changes in the incidence rates. Together with the odds ratios, this allowed to make statements as to how the locally spe-

cific explanatory factors may have influenced the incidence rate in a particular municipality.

Choosing a logistic linear regression model is an easy and efficient way to "analyze the effect of a group of independent variables on a binary outcome by quantifying each independent variable's unique contribution" (Stoltzfus, 2011). Generally, a logistic regression model has a binary outcome variable, and it measures the probability of being in one outcome group versus the other. In the context of this thesis, the model measures the probability of a municipality being in the top 20% of influenza incidence versus being in the bottom 80% of influenza incidence, based on the characteristics of the municipalities. Furthermore, a logistic regression identifes the strongest linear combination of independent variables that increases the likelihood of detecting the observed outcome through an iterative process. This is known as maximum likelihood estimation (Stoltzfus, 2011).

As a prerequisite, all potential variables were tested for multicolinearity. Multicolinearity is present in the dependent variables if two dependent variables have a high correlation. It can be problematic because if multicolinearity is involved, it becomes hard to determine which variables are statistically significant and which are not (Mansfield and Helms, 1982). Multicolinearity was ruled out by assessing the pairwise correlation of all of the variables. If any problematic multicolinearity occured, it was adressed by removing one of the variables that had a high correlation with another variable.

In order to run a model, the appropriate explanatory variables had to be found. This was achieved through a data-driven approach: the n best models for each wave were found by performing an automatic model selection in R (version 4.0.2) using the glmulti package (Calcagno and de Mazancourt, 2010) and the AIC (Akaike Information Criterion) (Cavanaugh and Neath, 2019) as a criterion that defined a "good" model. The model with the lowest AIC is generally seen as the best model, however, Burnham and Anderson (2004) suggest that all models with an AIC value that is no more than 2 points higher than the smallest AIC in the group also have substantial support. Therefore, all the models from the set of best models were asessed and among them, a suitable one was selected which was then used in an attempt to explain local differences in incidences.

Another point that is briefly mentioned here, is how to interpret a logistic regression model. In this thesis, the argument is based on the odds ratios, which is why a brief explanation on how to interpret odds ratio follows here. An odds ratio is the ratio between the probability of an outcome 0 and the probability of an outcome 1 (Sperandei, 2014). Generally, odds ratios can have two meanings:

odds ratios greater than 1 point to a positive association between outcome and dependent variables, while odds ratios with values between 0 and 1 point to a negative association. For multinomial variables, the interpretation is slightly different. The interpretation of multinomial variables (variables that have more than two outcomes but are categorized – the majority of variables in these models) is slightly different but still relatively straightforward: the model takes the first category and compares it to each of the other categories (Sperandei, 2014). As an example: in the models of this thesis, different levels of railway access were modelled, where the first class is "no railway access". Therefore, the model compared the category "no railway access" to the different other levels of railway access (low, medium, high, etc.).

After a model for each wave was selected and run, a Breusch-Pagan test was run on the models to test for heteroscedasticity of the model residuals (Breusch and Pagan, 1980). This test was an attempt to assess model quality. If heteroscedasticity is present in a model this means that there is some variation in the outcome variable which cannot be explained by the model. Commonly, this is regarded as an indicator of at least one missing variable in a model.

Spatial autocorrelation

Another question that was of interest for this thesis, is to see how well the model performs spatially. This could be achieved by looking at the residuals of each municipality. These residuals were asessed using the same statistics as in research goal 1, the global and the local Moran's I. The difference between this part and research question 1 was that the correlation measures were based on the model residuals and not on the incidences themselves. This allowed to assess the quality of the outcome by comparing the model outcomes with the outcomes of the raw data. A perfect model would explain all the spatial variation in the incidences and therefore, no clusters would be present anymore.

In an attempt to visualize the spatially varying influence of the explanatory factors, bivariate choropleth maps were generated showing the correlation of influenza incidence with each of the explanatory factors separately. The advantage of a bivariate choropleth map over a univariate one is that the rates of two variables are shown in the map. This allows to not only observe the spatial distribution of a variable, but also the correlation between those two variables (Olson, 1981). This is a useful tool to gain further spatial insights into the pattern of the data. One inportant point to consider with bivariate choropleth maps is color. The color scheme had to be carefully chosen in order for the maps to be better understandable and accessible for people with color vision deficien-

cies (Robertson and O'Callaghan, 1986). The choropleth maps were generated using R version 4.0.2 and the package ggplot with the following considerations for the classification: for the bivariate choropleth maps, the maximum of possible categories is three. Anything with more than three categories would include too many colors (e.g. having four categories would mean 16 different colors, five categories 25 colors, etc.). Therefore, the data had to be categorized differently than in the model. Whenever possible, the data was simply split into terziles. For the variable "access to the railway system" this was not possible due to the many zero values. Therefore, all the zero values were put into the same category. The remaining data were split into two categories using Jenk's natural breaks algorighm (Jenks and Caspal, 1971). The maps are intended as an aid to correctly interpret the results of the models. Therefore, they are not included in the thesis itself but have been moved to appendix A.

3.2.3 Effective visualisation of results

One additional goal of this thesis that is not associated with a research question was the effective visualisation of results of my research. The current Sars-Cov-2 pandemic is a good reminder why this is of great importance. Researchers have since the outbreak called for "clear, honest and valid information" (Finset et al., 2020), to make the public understand the severity of the situation without causing panic. Furthermore, the February 2020 editorial of The Lancet, one of the most renowned and widely cited medical journals, expressed the need for easy and effective communication as a means for fighting conspiracy and inaccuracies that spread through social media and the internet, concluding: "There may be no way to prevent a COVID-19 pandemic in this globalised time, but verified information is the most effective prevention againt the disease of panic" (Amit Kumar Mandal , Paulami Dam , Octavio L. Franco , Hanen Sellami , Sukhendu Mandal , Gulten Can Sezgin , Kinkar Biswas , Partha Sarathi Nandi, 2020). This to some extent also holds true for this thesis as it has gathered growing interest since the start of the pandemic. For example: upon publishing a blog entry about my thesis on the University of Zurich Department of Geography's website (Leuch, 2020), I received several messages from interested people. This demonstrates a potentially greater interest in the topic since the outbreak of the Covid-19 pandemic.

One main output of the thesis are maps. One advantage of using maps is that many people are somewhat familiar with the use of maps. They allow the quick visual analysis of large datasets and the recognition of spatial patterns. This follows an interdisciplinary approach known as visual analytics (Andrienko et al.,

2010). There are well-established principles of how information can be conveyed in maps one of which is color (Bertin, 1983). The concept of color is described as being a combination of the three components color hue, saturation and lightness (Slocum et al., 2008). Brewer (1994) describes four color scemes: qualitative, binary, sequential, and diverging. Relevant for this paper were the latter two. When the data has quantitative steps from low to high, a sequential scheme was used, where low values were represented by light colors and high values by darker colors. Where the data was quantitative but had an obvious midpoint (e.g. for visualising residuals), a diverging scheme was used. It consists of two different sequential color hues that become lighther, the closer the values are to the midpoint. Finally, what had to be considered in making maps accessible was perception by people with color vision deficency. Luckily, there are numerous tools that allow a user to test a color palette for different kinds of color deficiency (Lu and Meeks, 2020; Brewer et al., 2002) which were used for the creation of maps in this thesis.

4 Results

4.1 Research Goal 1: Descriptive spatio-temporal analysis of the influenza data

4.1.1 Incidence

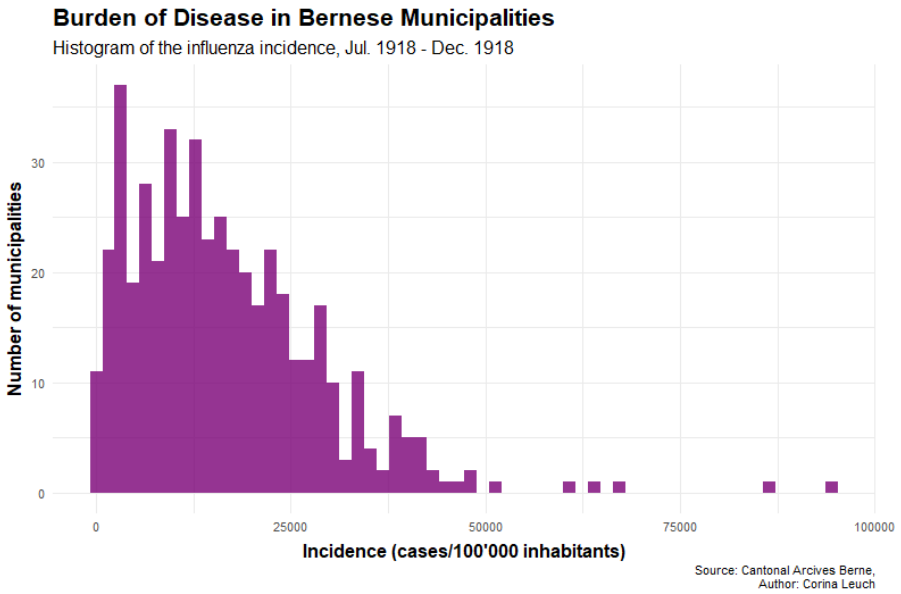

Figure 4.1: Distribution of incidences per municipality for the entire pandemic. The histogram is skewed to the right with some extreme outliers on the right side.

The histogram in figure 4.1 shows the distribution of the incidences, where the data was summed up for each municipality and the entire study period. Instead of raw case numbers, the incidences per 100'000 inhabitants is shown on the x-axis to make the data more comparable. The incidence numbers can easily be converted into morbidity in percent by dividing the number by 1000. The data does not seem to be normally distributed, but is a bit skewed to the right. Furthermore, the data contains some outliers on the higher end of the scale:

some municipalities reported more than 90'000 cases per 100'000 inhabitants, or more than 90% of the population was infected.

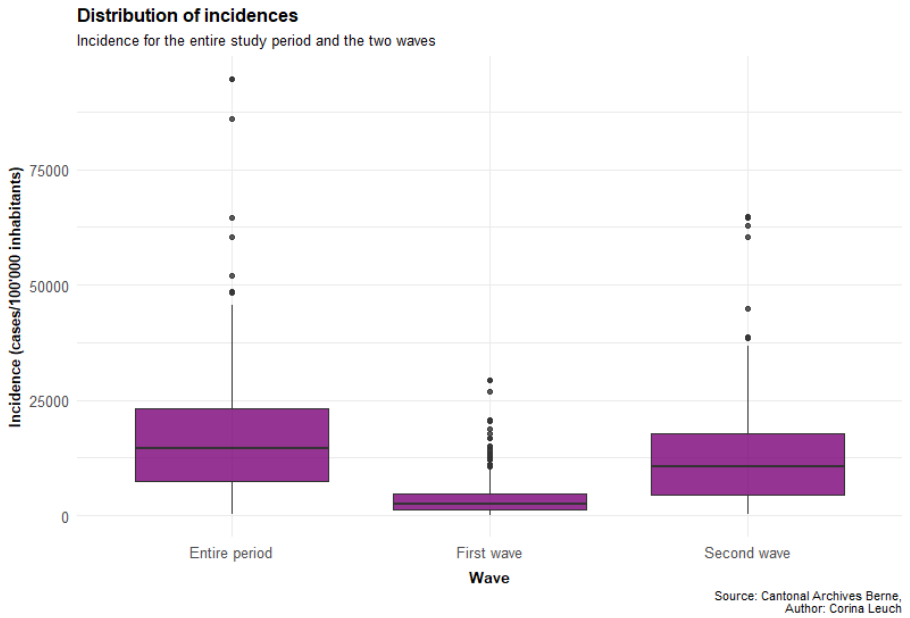

Figure 4.2: Boxplot of the incidences for the entire period (left plot), the first wave (July and August 1918; middle plot), and the second wave (October 1918 – January 1919; right plot). The plots show that both the variance and the mean were much smaller in the first wave but the first wave had more outliers.

The boxplot of the incidences (figure 4.2) underlines what the histogram already shows: the data contains some statistical outliers where the reported incidence was higher than statistically expected. There are three boxplots: one that includes the entire study period, and one for each of the two waves. This was a first step towards looking at the temporal dimension of the data. Generally speaking, incidence rates were higher during the second wave. The mean incidence for the entire study period was 16'505 cases per 100'0000 people (16.5% of the population). For the two waves the mean incidence was 3828 per 100'000 people (3.8% of the poulation, first wave) and 12'254 cases per 100'000 people (12.3% of the population, second wave). This means the burden of disease was higher in the second wave.

The variance was also higher in the second wave, while the first wave had the most outliers. However, it would not be wise to simply remove these outliers from the dataset. First, the spatial temporal dimensions have to be considered as well. It could well be that while these data points are outliers in the statisti-

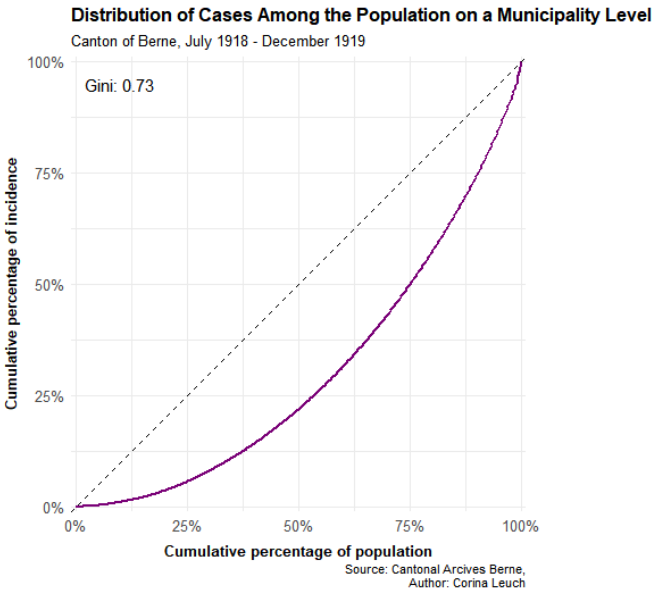

Figure 4.3: The Lorenz curve describes the cumulative incidence (in ascending order, x-axis) as a function of the cumulative population at municipality level. The grey dashed line signifies perfect equality, which in this case would mean that the incidence is equal in every municipality. The purple line represents the actual data and shows that there is moderate heterogeneity in the data. The curve is bent downwards, which means larger municipalities were more affected on average.

cal sense, there is a better explanation for their value and thus, removing them would draw an incomplete picture.

The Lorenz curve (figure 4.3) and the Gini coefficient confirm that there is some heterogenity in the data. This means that case numbers are not a linear function of population size. The graph shows that the 50% of the population from the smallest municipalities only accounted for about 25% of the total incidence while the 25% of the population from the largest municipalities was responsible for around half of the incidence. The Gini Coefficient is 0.73 which means that there is moderate heterogenity in the data. This confirms that the case numbers could not only be explained through the population size and that there had to be other factors that had an association with the case numbers.

4.1.2 Incidence and temporal dimension

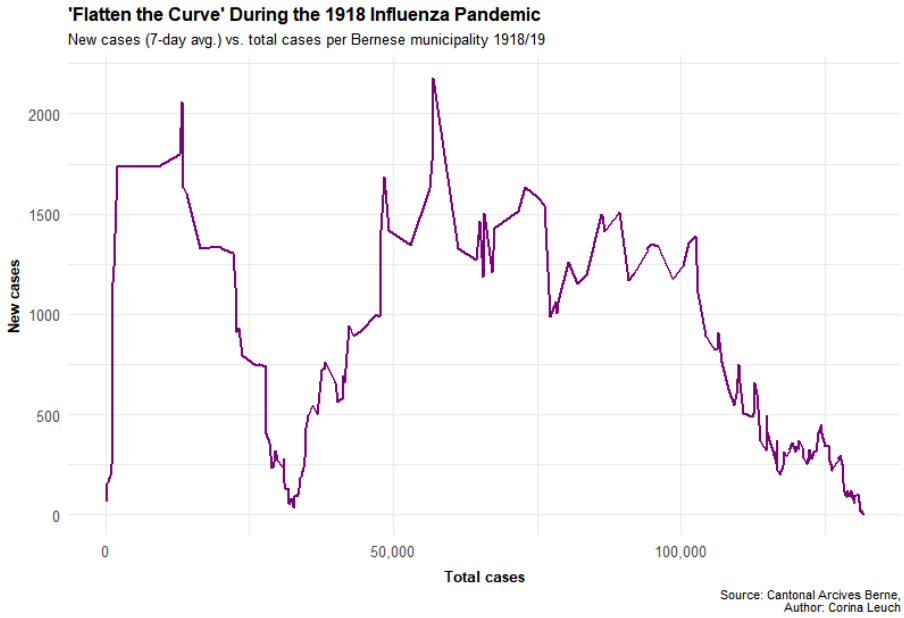

Figure 4.4: Daily new cases in relation to the total cases. If the curve has a positive slope, it means daily new infections were on average rising, and if the slope is negative, daily new infections were dropping. Note: This graph, even though it shows the pattern of the two waves, does not allow to make any statement about the temporal course as the x-axis is the total cases and not the time.

Figure 4.4 shows the the 7-day average of new cases in relation to total cases. Although the curve was smoothened through a 7-day rolling average, it still has some sharp turns and edges, indicating big differences in the daily reports. This curve allows to gain insights into how fast the pandemic was expanding and when daily new infections were decreasing, in other words, how the canton of Berne managed to "flatten the curve". Firstly, the curve shows the course of the pandemic with the two waves and allows to make comparisions as to their case numbers and how much they attributed to the entire pandemic. After a sharp increase in the beginning of the pandemic, the steep negative slope of the first wave indicates that infection rates went down relatively quickly. The curve does not allow to make any statement to how long (in terms of time) the case numbers were low in between the two waves. The course of the second wave indicates that case numbers rose slower than during the first wave, but stayed high for a longer period of time: the values on the y-axis were not necessarily

higher, but the total case number attributable to the second wave was higher, therefore indicating that the second wave must have lasted longer. Overall, the second wave shows several ups and downs, indicating that the case numbers kept rising and falling a bit. Finally, the case numbers slowly decended. The third smaller bump indicates a slight rise in infection numbers towards the end of the pandemic. The curve allows to tell roughly how many cases can be attributed to each of the two waves: during the first wave, around 30'000 infections were reported, while the second wave accounted for roughly 100'000 cases. This shows that the second wave had around three times as many infections as the first one.

4.1.3 Incidence and spatial dimension

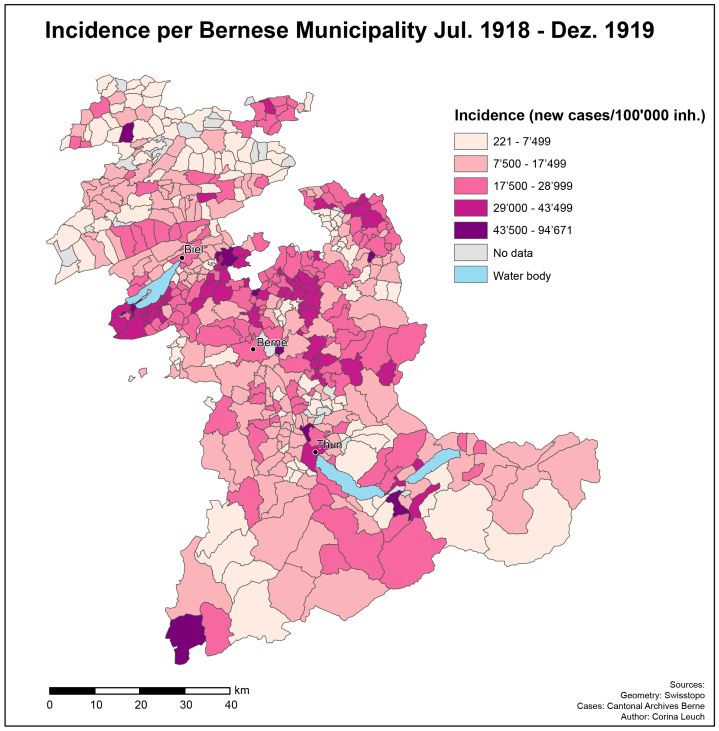

Figure 4.5: Incidence on a municipality level for the entire study period. The darker a municipality, the higher was the incidence. The map underlines the tendency that the Swiss Plateau was more heavily affected when considering the entire pandemic period.

The spatial distribution of incidence (cases per 100'000 inhabitants) over the entire study period (July 1918 – December 1919) does not provide a clear picture. The mountainous regions of the Jura and the Oberland were less affected than the lower-lying regions. Furthermore, municipalities with more inhabitants were generally more affected than smaller rural municipalities. This does however not hold true everywhere: there are also a few municipalities with only a few inhabitants that had a really high incidence. Interestingly enough, the Laufental (the northeastern part of the Jura) was particularly affected, contrary to the surrounding areas of the Jura. Finally, the municipality of Gsteig in the south-western corner of the canton was particularly affected.

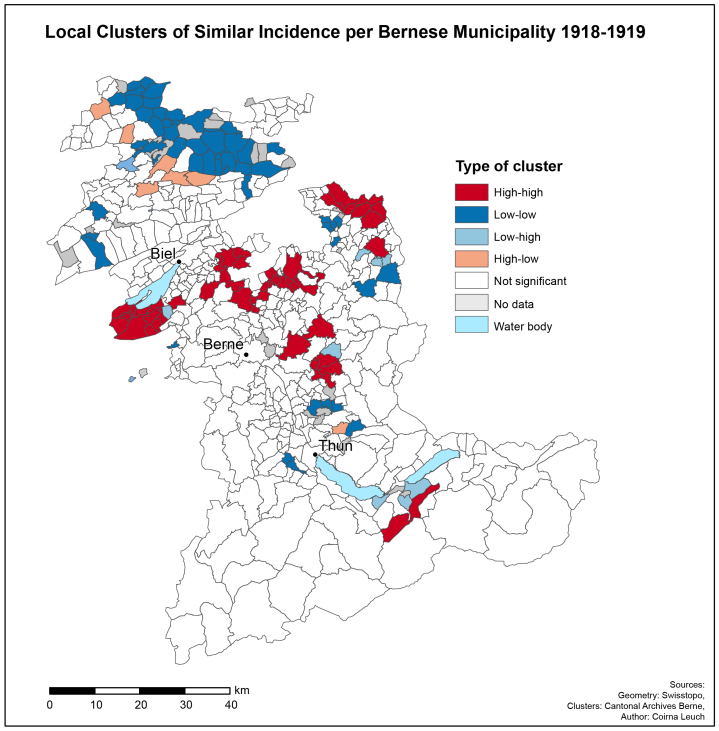

Figure 4.6: The local indicator of spatial association (LISA) shows local clusters of high-high (red) and low-low (blue) incidences, indicating that high rates were found next to high rates (or low rates next to low rates). The map further underlines that parts of the Jura region were significantly less affected, while the Swiss Plateau was significantly more affected, considering the entire pandemic period.

Looking at spatial autocorrelation provides further insight into the data. To test for spatial autocorrelation, a global Moran's I statistic was calculated which

suggests moderate positive autocorrelation (Moran's I = 0.33; p = <0.001). This means that the distribution of the incidences in space is not random and there are clusters of high and low incidences. Maps that show the local Moran's I are often called LISA (local indicators of spatial autocorrelation) maps. The map 4.6 shows this local spatial autocorrelation in clusters with statistically significant high-high rates and low-low rates and spatial outliers with high-low or low-high rates. The biggest cluster of the low-low rates can be found in the eastern part of the Jura region. Other than that, there are a few scattered clusters with significantly low rates. Sometimes, these "clusters" consist of only one municipality, indicating that the municipality and its neighbours both had high values. There are several high-high clusters spread out in the Swiss Plateau. One big cluster can be found south of Lake Biel and another one in the region of Oberaargau. Furthermore, there are a few areas with high-high rates north and east of the city of Berne. Finally, there are a few outliers. In the Jura region, some high-low outliers (high incidence but neighbours with low incidences) are present. In the Bernese Oberland, the opposite is the case; there are a few municipalities that have low incidences but their neighbours have high incidences between Lake Thun and Lake Brienz.

4.1.4 Incidence, temporal and spatial dimension

Looking at all the three dimensions (incidence, space and time) allows to see the course of the 1918 influenza pandemic in the canton of Berne (see figure 4.7). The disease first broke out in the Jura region which was particularly affected during the first wave in July and August 1918. The rest of the canton was not particularly affected, especially the Bernese Oberland. Afterwards, the infection rates went down in September 1918, before the second wave started in October 1918. This second wave affected the central regions first, before fully reaching the Oberland which was heavily affected in December 1918. In Janaury, the second wave was coming to an end and incidences in the municipalities were no longer as high. After that, there were a few more local outbreaks before the pandemic was eventually over, but these are not of interest for the scope of this thesis. Generally speaking, the pandemic lasted longer in the heavily populated Swiss Plateau than in the sparsely populated mountain areas. Finally, the data shows that the two waves behaved differently.

Based on these results, two additional maps are presented here for completeness, showing the incidences for each wave separately. This is a further step towards designing a model that includes locally specific factors that determine

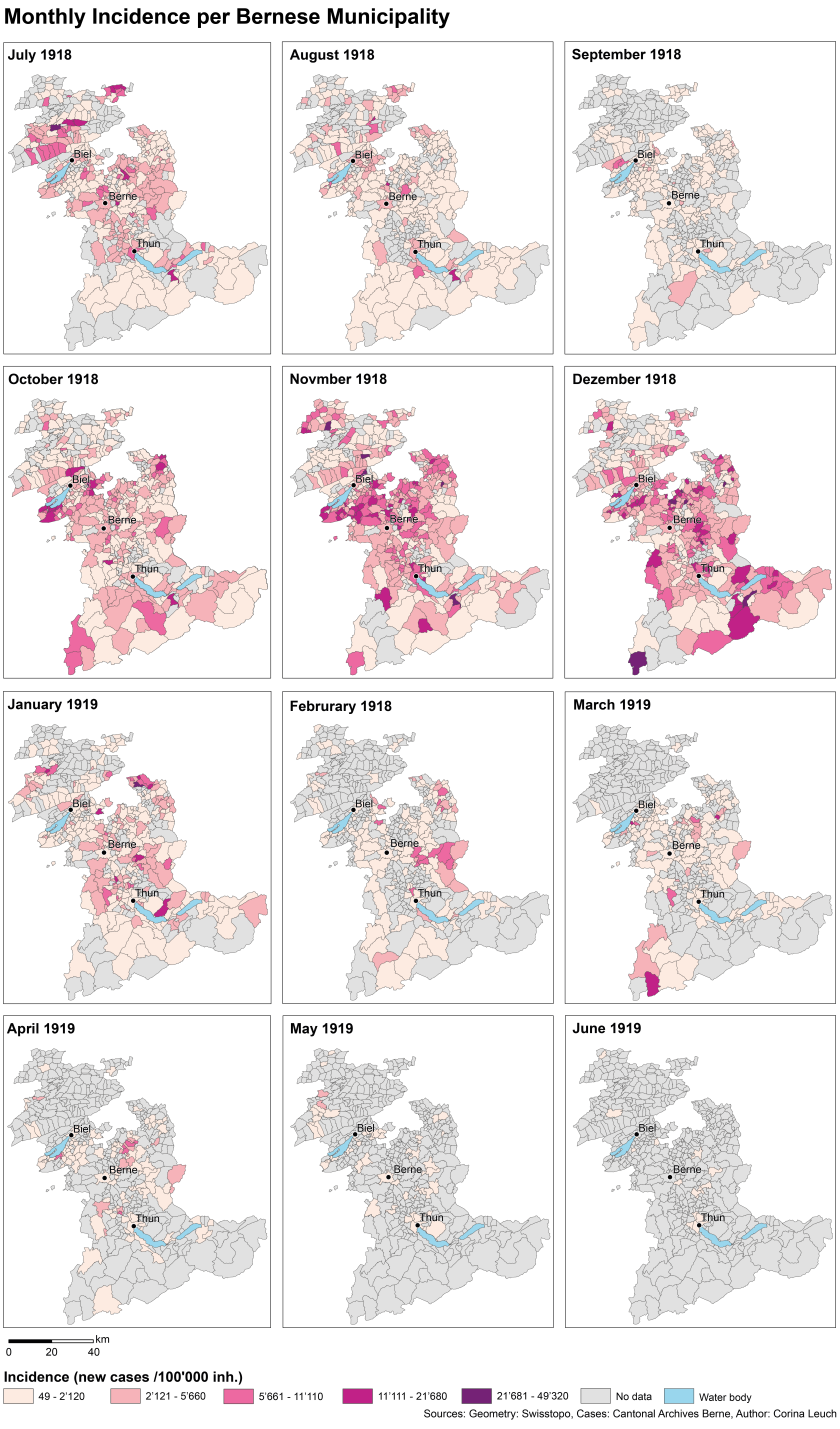

Figure 4.7: Monthly incidence rates on a communal level (July 1918 - June 1919). The legend is the same for all the maps, darker areas show high incidence rates. Note: these maps were originally created as part of an animation which can be found here: https://tinyurl.com/SpanishFluGIF.

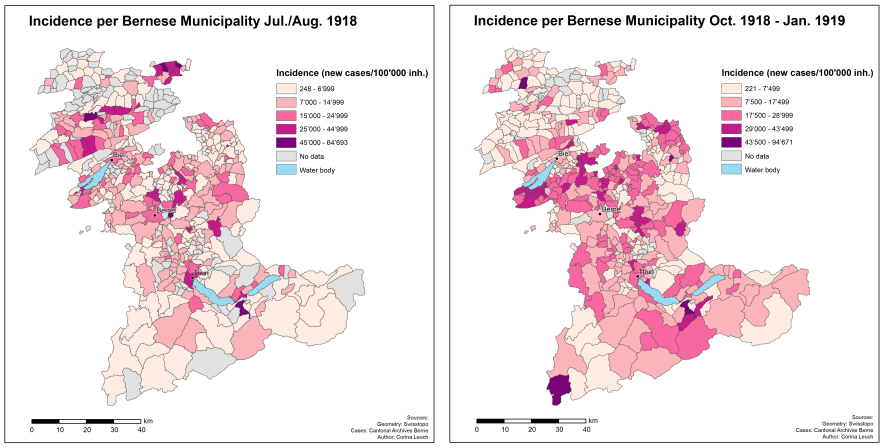

Figure 4.8: Incidence rates per municiality for each of the two waves. A different scale for each map was chosen because the second wave lasted longer and had much higher incidences. With the same scale, the spatial differences in the first wave would not have been visible. The map underlines that the Jura and especially the Laufental region were hit hardest by the first wave, while the Swiss Plateau and the Oberland were hit harder by the second wave.

the spread of the influenza virus (see figure 4.8). These two maps nicely show the spatial differences during the waves. During the first wave, the Jura mountains were heavily affected. However, this was not true for the entire region: while the Laufental region and the more southern part were heavily affected, many municipalities did not report any data during the first wave. In the rest of the canton, the Swiss Plateau was also affected while in the Alpine region only a few municipalities were somewhat more affected, while most municipalities were only mildly affected. During the second wave, it was the opposite: the Jura was not particularly affected, while the Swiss Plateau was heavily affected. In the Seeland region, south of Lake Biel, there seemed to be an entire cluster of municipalities that were heavily affected by the second wave. Furthermore, there was a second area east of the city of Biel that shows high incidences. In the Alpine region, the municipalities around the lakes were particularly affected. Again, the municipality of Gsteig stood out with its high incidence surrounded by municipalities with low incidences.

A Moran's I test further supports the observations of the raw incidences. The global Moran's I was statistically significant for both the first (Moran's I = 0.188; p = <0.001) and the second wave (Moran's I = 0.347; p = <0.001),

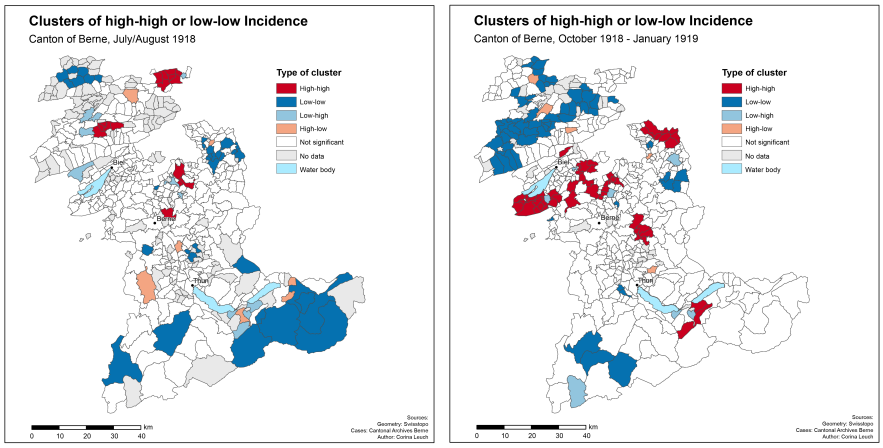

Figure 4.9: The local indicators of spatial association (LISA) maps show statistically significant clusters with high-high (red) and low-low (blue) incidences. The map shows areas where high rates were found next to high rates (or low rates next to low rates). This can be an indicator as to where locally specific factors may have had an influence. Furthermore, this map can be used as a baseline, to see how well the model performance is.

therefore rejecting the null hypothesis of spatial randomness. The spatial autocorrelation was mild for the first wave and moderate for the second wave. Additionally, local indicators of spatial autocorrelation in 4.9 show the local Moran's I in each municipality and the location of the clusters with high or low incidences. Again, differences between the first and the second wave become apparent. The first wave shows only a few clusters with high-high rates. The biggest one can be found in the Laufental, with a few other small ones in the Jura and the Mittelland. The Oberland shows an area with significant low-low clustering, illustrating that it was not particularly affected during the summer wave. During the second wave, the Swiss Plateau (Regions of Seeland, Mittelland and Oberaargau) show clusters with high-high incidence while the Jura region shows many low-low clusters. In both waves, a few outliers are visible.

4.2 Research goal 2: Finding determinants of spread

4.2.1 First wave: July 1918 – August 1918

Model prerequisites

Before fitting a model, the colinearity of the variables was assessed to rule out multicolinearity. Figure 4.10 shows the distribution of each variable, the correlation coefficient between the variables and some scatterplots. Its results indicate no problematic colinearity, therefore all of the explanatory variables can be kept for the model selection process.

Model selection process

The automatic model selection process returned a set of nine candidate models (models with an AIC no more than two points higher than the minimum AIC) shown in table 4.1. The variable urbanity appeared in all the models. The rest of the variables appeared in various combinations. From this large set of candidate models, the third model (incidence˜ 1 + urbanity + TB + railway + precipitation) was chosen as the final model 1.

Table 4.1: The nine candidate models returned from the model selection process for model 1. According to Burnham and Anderson (2004), all the models from this selection provide a good fit for the data.

	model	aic	weights
1	Top incidence ˜ 1 + urbanity + railway + precipitation	362.77	0.15
2	Top incidence ˜ 1 + urbanity + TB + agriculture	362.96	0.14
3	Top incidence ˜ 1 + urbanity + TB + railway + precipitation	363.38	0.11
4	Top incidence ˜ 1 + urbanity + TB + railway +precipitation + agriculture	363.84	0.09
5	Top incidence ˜ 1 + urbanity + TB + railway + agriculture	364.09	0.08
6	Top incidence ˜ 1 + urbanity + railway + precipitation + agriculture	364.13	0.08
7	Top incidence ˜ 1 + urbanity + precipitation + agriculture	364.26	0.07
8	Top incidence ˜ 1 + urbanity + TB + precipitation + agriculture	364.50	0.06
9	Top incidence ˜ 1 + urbanity + agriculture	364.52	0.06

Model results

Table 4.2 shows the results of the regression model for the model 1, which covers the first wave and the months of July and August 1918 (N = 376 municipalities, AIC = 383.48). The right column of the table contains the odds ratios and

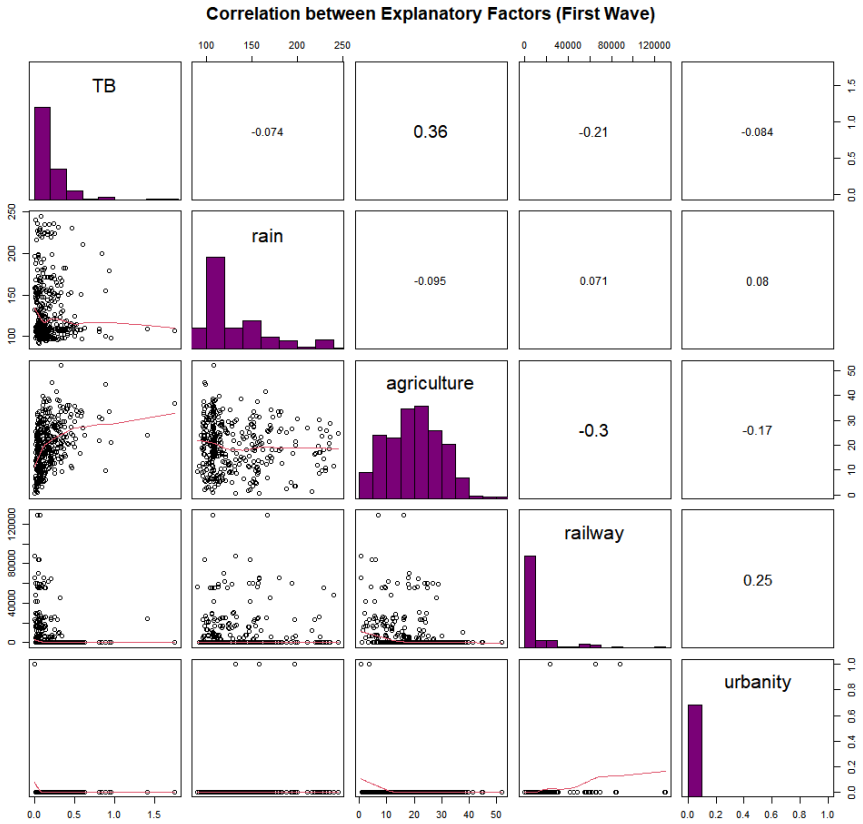

Figure 4.10: The pairwise correlation plot gives insights into the explanatory variables. In the upper right corner, the correlation coefficient between two variables is shown. The diagonal shows the distribution of the data and the left lower corner shows scatterplots between two variables with the red line as the correlation function. Note: the raw distributions of the variables are shown in this plot. Data source: Sanitätsdirektion des Kantons Bern (1918).

the confidence intervals. The table shows that most of the odds ratios were statistically significant at least on a 0.05 level with the exception of the highest two categories of the precipitation variable (indicating high precipitation). The odds ratio of urbanity was very high which means that the three cities Berne, Biel, and Thun were more likely to have an incidence rate in the highest quintile. However, the standard error for this variable was also quite high, and therefore the variable has to be interpreted with caution. The odds ratio for the TB variable was 1.5 (CI = (0.83, 2.17)), indicating that municipalities in the highest quintile of TB mortality were more likely to be in the highest quintile of influenza

Table 4.2: Model results of the logistic model for model 1. The left column shows the model coefficients (and the standard errors). The right column shows the odds ratios (and the confidence intervals). Asterisks behind the numbers indicate various levels of statistical significance.

	Dependent variable:	
	incidence	
	coefficients	odds ratio
	(1)	(2)
Top TB Quintile	0.41 (0.34)	1.50*** (0.83, 2.17)
urbanityCity	16.77 (832.95)	19,261,371.00*** (19,259,738.00, 19,263,003.00)
railway2	0.57 (0.40)	1.77*** (0.99, 2.56)
railway3	1.49*** (0.39)	4.42*** (3.65, 5.20)
railway4	0.40 (0.61)	1.50* (0.30, 2.69)
railway5	1.34 (1.45)	3.82** (0.98, 6.65)
rain2	0.19 (0.41)	1.21** (0.41, 2.00)
rain3	0.36 (0.40)	1.43*** (0.65, 2.21)
rain4	−1.06* (0.50)	0.35 (−0.63, 1.33)
rain5	−0.52 (0.46)	0.59 (−0.31, 1.49)
Observations	376	376
Log Likelihood	−170.74	−170.74
Akaike Inf. Crit.	363.48	363.48
Note:		*p<0.05; **p<0.01; ***p<0.001

incidence. The railway variable did not show a clear tendency, except that municipalities with railway access had a higher chance of having an incidence rate in the highest quintile. The railway group 5, with the best railway access, had the highest odds ratio of 3.82 (CI = (0.98, 6.65)), meaning municipalities with the highest railway access were almost four times more likely to be in the group with the high incidence. The precipitation finally showed a divided picture: while the second (OR = 1.21; CI = (0.41, 2.0)) and the third category (OR = 1.43; CI = (0.65, 2.21)) had odds ratios larger than 1 (indicating that municipalities in these categories were more likely in the top incidence group), this did not hold true for the two highest categories which had an odds ratio of

0.35 (CI = (-0.63, 1.33)) and 0.59 (CI = (-0.31, 1.49)) respectively. However, these two odds ratios were not statistically significant.

Finally, a Breusch Pagan test was run to test the model for heteroscedasticity as an attempt to assess the model fit. The Breusch pagan test (BP = 28.5; p = 0.001) indicates that the model suffers from heteroscedasticity.

Local spatial autocorrelation of the model results

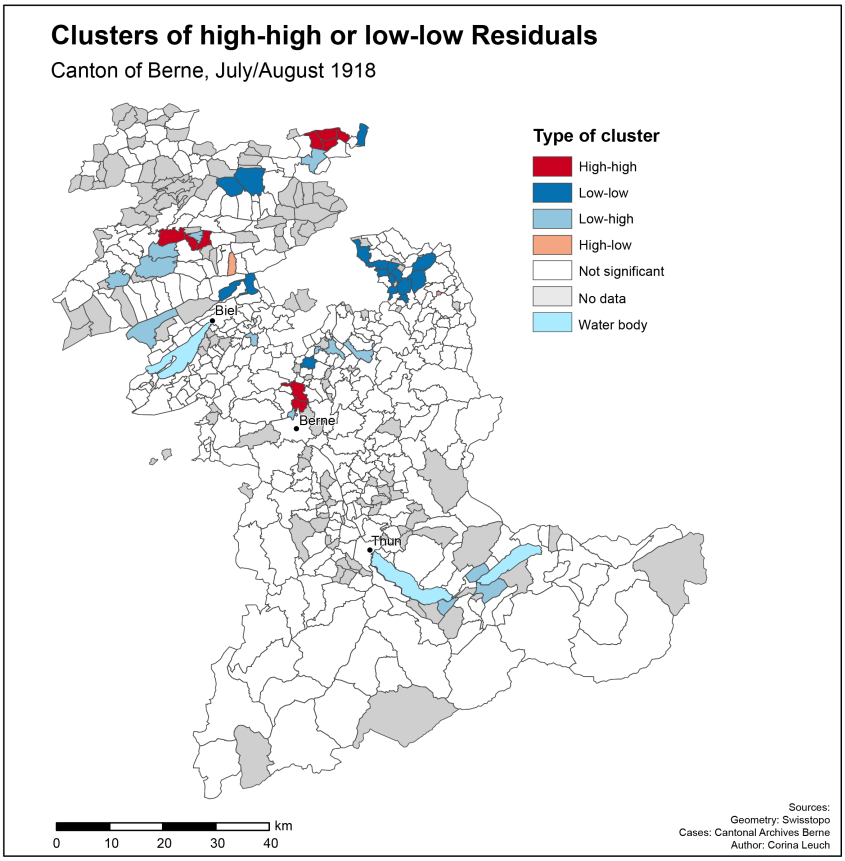

Figure 4.11: Local spatial autocorrelation of the residuals of the first wave model. The map shows clusters of high-high (and low-low) residuals, where high residuals indicate that the model underestimated the incidence rate and low residuals stand for an overestimation of the incidence rate. This means that areas with high residuals were found next to other values with high residuals (or low residuals next to low residuals). Some clusters still remain, despite the model's attempt to explain the spatial variation.

In an attempt to assess the model performance of model 1 in space, the spatial global and local spatial autocorreltation of the model residuals were calculated. In this context, a Moran's I test allows to test whether the residuals are randomly distributed in space. The global Moran's I test of the first wave (Moran's I = 0.057; p = 0.046) was statistically significant with the Moran's I statistic being close to zero. This indicates that there is practically no spatial autocorrelation present in the residuals which means that the distribution of the residuals in space is random and the model assumptions of the used regression model are not violated.

As already done for the descriptive analysis in research question 1, the local spatial autocorrelation for the municipalities was calculated. The map in figure 4.11 is shown for completeness, despite the fact that the global Moran's I indicated practically no global spatial autocorrelation. Other than in the previous section, this map shows the autocorrelation of the residuals and not of the incidences. Again, high-high clusters are areas where the residual value of a municipalitiy and its neighbours was high, therefore indicating areas where the model underestimated influenza incidence (and in contrast, low-low clusters show where the model overestimated the incidence). The map also shows that the model could explain some clusters (see figure 4.9). There were still two small high-high clusters, one in the Jura region and one just north of Berne. The model managed to explain the high-high cluster nort-north-east of Berne. When looking at low-low clusters, there are also fewer of them. The low-low cluster in the northern Jura region disappeared, while a new, smaller one appeared in the north-eastern part in between the two high-high clusters. Furthermore, in the Oberaargau region, there are still some unexplained low-low clusters but they are considerably smaller. The model worked particularly well in the Bernese Oberland, where all the clusters have disappeared with the exception of some low-high outliers between the two lakes. Throughout the rest of the canton, there are a few more scattered low-high outliers and two high-low outliers.

4.2.2 Second wave: October 1918 – January 1919

Model prerequisites

Before fitting a model, the colinearity of the variables was assessed to rule out multicolinearity which was done the same way as in model 1. Figure 4.12 shows the distribution of each variable, the correlation coefficient between the variables and the scatterplotts between all the explanatory variable pairs. The colinearities and distributions show a fairly similar picture as in the first wave. This has

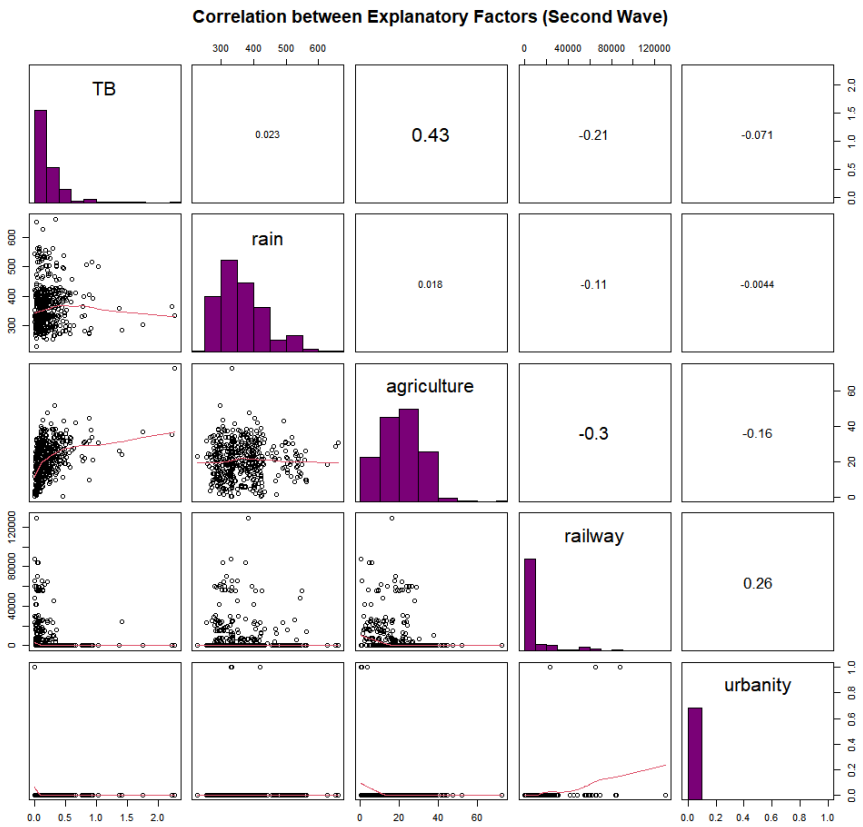

Figure 4.12: The pairwise correlation plot gives some insights into the explanatory variables. In the upper right half, the correlation coefficient between two variables is shown. The diagonal shows the distribution of the data and the left lower half shows scatterplots between two variables. Note: the graph shows the distribution of the raw data, classifications are not considered in this illustration. Data Source: Sanitätsdirektion des Kantons Bern (1918).

to do with the data sources of these dependent variables. With the exception of the precipitation variable, the same data was used in both waves, with minor changes due to slightly varying sets of municipalities (during the second wave, more municipalities reported data than during the first one, therefore more data entries were considered during the second wave). The precipitation variable contains different data covering the months from October 1918 to December 1918. There was no problematic colinearity between any of the variables, therefore all of the variables were kept for the following model selection process.

Model selection process

The next step in the modelling was an automated model selection process, just like for model 1. Again, a candidate model was defined as the model with an AIC two or less points bigger than the minimum AIC in the moedel set. For model 2 (second wave: October 1918 – January 1919), the automated model selection algorithm returned a candidate set of five models (the candidate models are shown in 4.3) that potentially provide a good fit for the data. From this candidate set, the fourth model (Top Incidence ~ 1 + urbanity + TB + railways + precipitation) was chosen, because it contained exactly the same explanatory variables as the model used for model 1. This is a convenient choice, because it allows to make comparisons between the two waves and potentially yields information whether the same factors had the same influence in both models.

Table 4.3: The set of candidate models for model 2. Burnham and Anderson (2004) suggest that all of these possible models have substantial support.

	model	aic	weights
1	Top incidence ~ 1 + railway + precipitation	458.82	0.19
2	Top incidence ~ 1 + TB + railway + precipitation	459.12	0.16
3	Top incidence ~ 1 + urbanity + railway + precipitation	459.47	0.14
4	Top incidence ~ 1 + urbanity + TB +railway + precipitation	459.77	0.12
5	Top incidence ~ 1 + railway + precipitation + agriculture	460.53	0.08

Model results

Table 4.4 shows the results of model 2 (N = 460; AIC = 461.66). From the odds ratios in the right column, half of them were statistically significant at least on the level 0.05. The urbanity did not have an influence in this second wave which is indicated by the odds ratio of 0 (CI = (-2309, 2309)). Again, this variable has an extremely high standard error which means that it has to be interpreted with caution. The variable TB had a positive association, meaning if the municipalitiy was in the top quintile of TB mortality, it was more likely to be in the top quintile for influenza incidence (OR = 1.48; CI = (0.9, 2.07)). For the variable access to the railway network there was no clear association. For this variable, different levels of railway access were compared to the category "no railway" access. As an example, between the first category (no railway access) and the third category (medium railway access) there was a negative association (OR = 0.75; CI = (-0.18, 1.68)), meaning municipalities with no railway access at all were more likely to be in the highest 20% of incidences than municipalities

in the third category with medium railway access. However, it has to be noted that this association was not statistically significant. Furthermore, the highest category of railway access had an odds ratio of 0 (CI = (-2116.21, 2116.31)) which means that there did not seem to be an association between having very high railway access and influenza incidence. For the variable precipitation, there was also no clear association. The second (indicating low precipitation) and the third category (indicating medium precipitation) had the highest odds ratios with 1.92 (CI = (1.23, 2.61)) and 1.3 (CI = 0.58, 2.01)) respectively. For the two higher categories (indicating even higher precipitation), the odds ratios showed a negative association which means municipalities in this category were less likely to be in the top incidence group.

As done for model 1, a Breusch-Pagan test was run in oder to assess model fit and test the model for heteroscedasticity. The Breusch-Pagan test (BP = 20.95, p = 0.021) indicates that the model indeed suffers from heteroscedasticity.

Spatial autocorrelation of the model residuals

In an attempt to assess model performance in space, the global Moran's I of the residuals was calculated (Moran's I = 0.323; $p < 0.001$). The test indicates that the model residuals are not randomly distributed in space, despite the fact that the Moran's I of the model residuals was lower than the moran's I of the raw incidences. This is a hint that there is another explanatory factor which is missing in model 2. The local spatial autocorrelation map of the residuals for the second wave supports the above-presented gloal Moran's I outcome (see 4.13). Just like in the model of the first wave, high residuals indicate that the model underestimated the incidences, while low values indicate areas where the model overestimated the incidence. In the Jura mountains there are no longer any clusters (with the exception of a single municipality in the Laufental region that has a statistically significant low-low result), which means that the model managed to explain all the spatial variation of incidences in those regions. In the Seeland region, there are still two larger high-high clusters, one south of Lake Biel and the other one just east of the city of Biel. In these areas, municipalities with high residuals have neighbours that also show a high residuals (i.e. the model underestimated the incidence rates). Furthermore, there is another high-high cluster in the northern part of the Oberaargau region but the model managed to explain the cluster just north of the city of Berne (see 4.9). Apart from those clusters, there are also some low-low clusters in the Swiss Plateau with a particularly large one located in between the cities of Berne and Thun. Furthermore, there are some more scattered low-low areas distributed around

Table 4.4: Model results of the logistic model for model 2. The left column shows the model coefficients (and the standard errors). The right column shows the odds ratios and the confidence intervals.

	Dependent variable:	
	Top_Inz	
	coefficients	odds ratio
	(1)	(2)
Top TB mortality	0.39 (0.30)	1.48*** (0.90, 2.07)
urbanityCity	−15.04 (1,178.45)	0.0000 (−2,309.72, 2,309.72)
railway2	0.63 (0.37)	1.87*** (1.14, 2.60)
railway3	−0.29 (0.47)	0.75 (−0.18, 1.68)
railway4	1.31** (0.45)	3.72*** (2.83, 4.61)
railway5	−14.68 (1,079.77)	0.0000 (−2,116.31, 2,116.31)
rain2	0.65 (0.35)	1.92*** (1.23, 2.61)
rain3	0.26 (0.37)	1.30*** (0.58, 2.01)
rain4	−0.71 (0.43)	0.49 (−0.35, 1.34)
rain5	−0.33 (0.40)	0.72 (−0.07, 1.50)
Observations	460	460
Log Likelihood	−219.83	−219.83
Akaike Inf. Crit.	461.66	461.66
Note:		*p<0.05; **p<0.01; ***p<0.001

the canton, with a few low-high outliers, particularly north of Berne. These outliers indicate areas with low model residuals surounded by areas with high model residuals. In the Bernese Oberland, a new high-high cluster appeared indicating an area where the incidences were higher than predicted by the model. Apart from that, there is an outlier cluster of low-high values just north of that, indicating an area with low residuals surounded by an area with high residuals. Finally, the low-low clusters that were previously present in the Bernese Oberland disappeared, indicating that the model managed to explain the spatial variation of the incidences in this area.

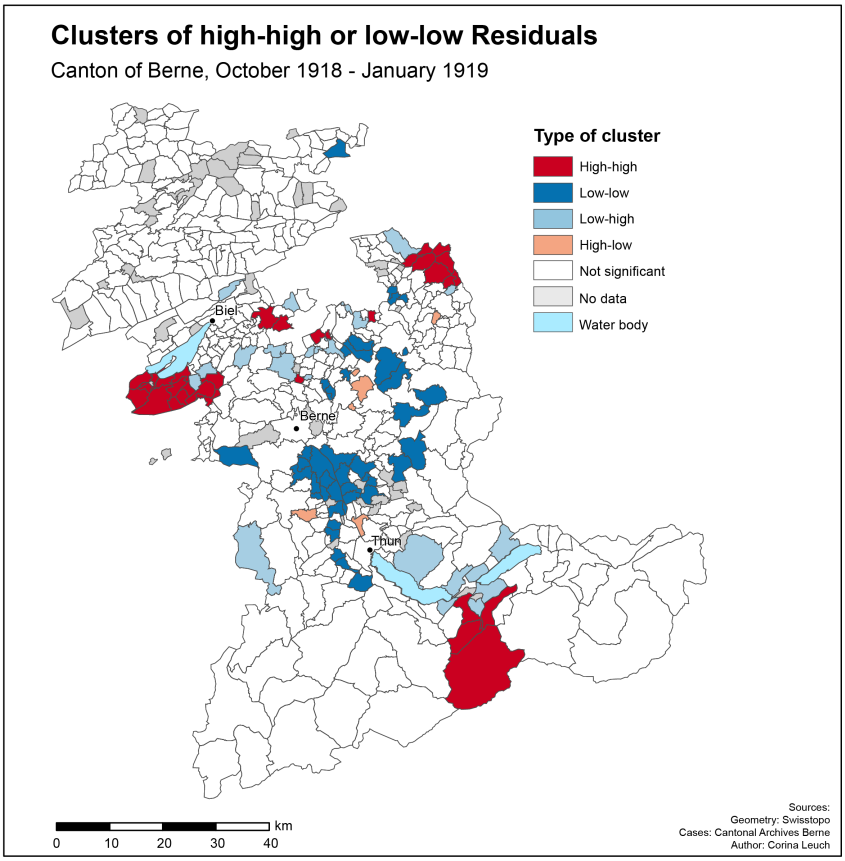

Figure 4.13: Local spatial autocorrelation of the model residuals of model 2. The map shows areas where high residuals were found next to high residuals (or low residuals next to low residuals). High residuals indicate that he model underestimated the incidence rate (i.e. it was in fact higher than predicted by the model) and low resdiuals indicate an underestimation of the incidences. The map has large high-high clusters in the regions of Seeland, Oberaargau and Oberland. Furthermore, there are some scattered low-low clusters in the regions of Mittelland, Voralpen and Oberaargau.

5 Discussion

The research presented in this thesis adds to the previous research provided in section 2 in various ways. First and foremost, it shows that, just like in most places, in the canton of Berne the pandemic struck in a mild summer wave (July/August 1918) followed by a more severe winter wave (October 1918 – January 1919). Furthermore, the analysis shows that there were locally specific factors that help explain why certain municipalities showed high incidences. Finally, the analysis shows the complexity of the topic and that innovative visualizations can help interpreting the results.

5.1 Research goal 1: Descriptive spatio-temporal analysis of the influenza data

The descriptive spatio-temporal analysis shows the complexity of the research field and that the outcome variable alone contains various layers and dimensions. Therefore, it can be very helpful to start with a basic analysis to get a first idea of the data, and later on increase the complexity of the data by adding more dimensions.

5.1.1 Incidence

Looking at the incidence of influenza-like illness in an isolated manner reveals that the data violates some statistical assumptions. First of all, the data is not normally distributed which has implications as to how it should be modelled in a regression model. Both the histogram (see figure 4.1) and the boxplots (see figure 4.2) show that the data has outliers. These outliers are all above the statistically expected values, with none below. This lies in the nature of these data; a municipality cannot have less than zero cases of an illness, and the municipalities that did not report any data were not considered in the analysis. Furthermore, the question remains whether these values are in fact outliers and should be treated as such, or if there is a better explanation why these values are so high, and they should be left within the data. The boxplot shows that particularly in the first wave, outliers occur. From literature, it is known that

the Jura region was more affected than the rest of the canton (Sonderegger, 1991). Therefore, it could well be that these values are in fact not outliers but the more heavily affected Jura region which means that removing them would draw an incomplete picture of the course of the "Spanish" flu pandemic. The same goes for the outliers visible in the histogram. The canton of Berne is an extremely regionally diverse canton whose municipalities had anything from 44 (municipality of Ballmoos) to 104'626 (city of Berne) inhabitants (Statistisches Bureau, 1921) in 1920. Small municipalities can have large incidences, because already a small number of cases as for example a single infected family can lead to a high incidence.

The other question that raw case numbers (or incidences) can answer, is how population size is related to incidence. In an ideal case, the incidence would be the same in all the municipalities. This would mean that the infection numbers are proportional to the poulation size, and therefore the explanation for the spread would be simply "more inhabitants equals more cases". However, the Lorenz curve (see figure 4.3) shows that this is not the case for the "Spanish" flu and that municipalities with more inhabitants disproportionally contributed to the overall incidence. This may be contrary to what Chowell et al. (2008) found in their study, as their findings showed a higher mortality in rural areas. However, these results have to be compared with caution, as their study used mortality rates while this project is based on incidence rates. A high mortality rate can point to a higher incidence rate but other locally specific factors may be involved as well.

5.1.2 Incidence and temporal dimension

To add a second layer of complexity, one can look at the temporal course of the pandemic. One fundamental way to do this is by looking at the new infections for a given time period in a bar chart. This was already done by Staub et al. with the exact same data (see figure 2.3), and was therefore not repeated in this thesis. As an alternative, the daily new infections were plotted against cumulative case numbers (see figure 4.4) which provides a nice alternative to the graph of the daily new infections. This image gives an overview of the structure of the pandemic and shows after how many cases the canton of Berne managed to "flatten the curve". Contrary to figure 2.3 which has time as a scale, figure 4.4 has the total case numbers as a scale. This allows to quantify the share of case numbers attributable to each wave. As already found by other studies, the 1918 influenza pandemic struck the canton of Berne in two waves, where the second wave was more severe by orders of magnitude, with more than

three times more cases of influenza-like illness than the first wave (Sonderegger, 1991; Sonderegger and Tscherrig, 2016). The sharp increase of case numbers at the beginning of the first wave suggests that the disease was already circulating in the population before the reporting mandate started in July 1918. However, without any data it is not possible to know when the disease was introduced in the canton of Berne.

5.1.3 Incidence and spatial dimension

One good starting point for the spatial analysis of the 1918 influenza pandemic is to look at the overall distribution of the incidences in the study area. Generally, the map of overall incidence rates (see figure 4.5) does not provide a very clear picture, but still, some things have to be put into a greater context in order to be understood. Firstly, the region of Laufental stands out with its very high incidence, despite its seemingly not very central location within the canton. However, a map that only shows the canton of Berne provides an incomplete picture: the region of Laufental is close to the city of Basel and was therefore more acesssible from there. Furthermore, the Laufental region was connected with Basel via the railway network and therefore much more accessible from Basel than from within the canton of Berne. Another similar example is the municipality of Gsteig. Again, the municipality does not seem very central but was in fact well-accessible through the canton of Vaud. These examples show the shortcomings of a model that only considers a region with boundaries that are arbitrary for the spread of a disease, such as state or regional boundaries: the cut leads to missing information that could lead to wrong interpretations. The map also shows that the more urbanized Swiss Plateau was more affected, which poses a contradiction to the findings of Chowell et al. (2008) who found that mortality was higher in rural areas but again, it is somewhat inaccurate to compare incidence and mortality rates. Literature suggests that on average in Switzerland, around half of the population was infected (Sonderegger, 1991). With the map alone, it is hard to tell how high the morbidity in the canton of Berne was. However, this can be achieved by looking at the raw case numbers: the data contain 143'389 disease reports (Sanitätsdirektion des Kantons Bern, 1918) and the population was around 666'000 (Statistisches Bureau, 1921). This leaves a morbidity rate of only around 21%. Furthermore, Staub et al. state that the reporting of case numbers was seen by many doctors as an unnecessary bureaucratic step. Therefore, not all cases might have been reported. Additionally, the canton of Berne accounted for a high number of deaths (almost 1 in 5 deaths were reported in the canton of Berne) which is a further hint, that not

all cases were reported. Finally, the fact that the canton of Berne was such a heavily affected region makes it an ideal study area for this thesis.

The map of the local spatial autocorrelation (see figure 4.6) provides a statistical estimation of the distribution of the incidences in space. This highlights potential areas of interest. The map shows several clusters on the border of the canton which have to be interpreted with caution. By its nature, the local Moran's I compares an entity with all of its spatial neighbours (Anselin et al., 2003). Therefore, if a municipality lies on the edge of the study area, it has fewer neighbours, and the individual neighbours have more weight. This potential bias could be overcome if the incidene of the neighbouring municipalities in other cantons (or possibly an average of that canton as an estimate) were available. These data were not available and therefore, potential edge effects have to be kept in mind when interpreting this map. The map statistically underlines what the incidence map already shows: the Swiss Plateau was overall more heavily affected than the more rural areas in the canton of Berne.

5.1.4 Incidence, spatial and temporal dimension

In order to gain a complete picture of the spatio-temporal characteristics of the pandemic, one has to combine all the three dimensions. The facet map in figure 4.7 provides a complete spatio-temporal picture of the 1918 influenza pandemic in the canton of Berne. The map confirms several findings noted by previous literature. First of all, it again confirms that there were two waves and additionally shows their spatial characteristics. Furthermore, it demonstrates how the French speaking part was hit harder by the first wave while the German speaking part was much more affected by the second wave. Therefore, these results map onto the findings of Sonderegger (1991) who notes that the pandemic in Switzerland generally moved from the west to the east. Furthermore, the findings in this study contribute to existing research by providing a quantification to Sonderegger (1991)'s more qualitative research. Additionally, Sonderegger and Tscherrig (2016) found that a thesis, according to which the pandemic was introduced via Basel, was not confirmed by mortality rates. However, the data finds some evidence that Basel might have played an important role in the spread of the virus in Switzerland: the Laufental region, which was relatively well-connected with Basel, was heavily affected early on in the pandemic. However, without any data covering Basel, it is not possible to definitely prove or disprove this hypothesis.

The facet map demonstrates that the pandemic first reached cities and other well-accessible places. These findings were already noted by Sonderegger (1991)

for the canton of Berne and by Chowell et al. (2008) in their study about Spain. The data show that the second wave lasted for approximately four months, which is longer than the second wave observed by Chowell et al. (2014) in Spain. Furthermore, this nicely pairs with the findings of Eggo et al. (2011), who found that cities were an important driver in the beginning of the pandemic, but with rising case numbers, between-city contacts where shut down and local spread became the main driver for the pandemic. Other studies (e.g. Olson et al. (2005)) found evidence of a spring wave as early as April 1918, which can neither be confirmed nor rejected, as the date only covers the time from July onward.

Based on the evidence of the facet map, it makes sense to provide the additional maps showing the impact of the two waves separately, since they were very different, as already noted by Sonderegger (1991). The map in figure 4.8 highlights the different characteristics of the two waves even better due to the different scales used for each of them. Using one single scale for the entire pandemic tends to "suffocate" finer characteristics of the first wave because the range of the incidences in the second wave is much bigger. This map also allows to draw conclusions, and contributes to previous research in different ways.

Earlier research suggests that the first wave was more virulent in the French speaking part (Sonderegger, 1991), however, this was not true for the entire region. While the southern and central part of the Jura, as well as the Laufental region, were heavily affected, large northern portions were mostly spared the virus. Reasons why the Laufental region might have been affected in such an early stage have already been discussed in the previous section (the proximity to Basel may have had an influence). One potential reason why the southern Jura was heavily affected at this early stage could have been the watch-making industry. Contrary to what one might think, the southern Jura region was not primarily a difficult-to-access rural area, but home to many watch factories that offered jobs to unskilled labourers form the surounding areas (Fallet, 2020). This is undermined by the map in figure 3.4 which shows indeed that the proportion of people working in agriculture was very low in the southern part of the Jura, indicating that the people there held other jobs. A further hint of this thesis is the map in figure 3.2 which shows that the area was not that heavily populated, which means that the population from surounding areas also found jobs in the watch factories. The watch industry might have contributed to the spread of the influenza virus for two possible reasons. Firstly, the working conditions in the watch factories may have been favourable for a spread of an airborne/droplet

infection: hundreds of unskilled workers worked in close proximity on production lines (Fallet, 2020). Secondly, the watch industry led to a certain exchange with other parts of the world, which may have favoured an early introduction of the disease into the area. In the context of literature, this maps onto the findings of Bengtsson et al. (2018), who found that people working in agriculture had a significantly lower risk of infection. This proves true for the Jura region in the sense that the northern parts where a higher percentage of people worked in agriculture were less affected. Furthermore, it is noteworthy that during the first wave, the incidence was high in the three biggest cities Berne, Biel, and Thun, as well as in their surounding areas. The fact that cities were an important factor in early spread was pointed out in different earlier studies (Sonderegger, 1991; Eggo et al., 2011; Chowell et al., 2008). This seems to manifest in this case study of the canton of Berne and can be seen as a contribution to Chowell et al. (2008)'s call for further geographically comprehensive studies.

During the second wave, the Swiss Plateau was heavily affected by influenza, while the Jura region was not particularly affected this time. Again, this was already described by Sonderegger (1991), and the data quantifies this. Contrary to the first wave, the three cities Berne, Biel, and Thun do not stand out with their high infection rates anymore. This again could be an indicator that local spread was the main driver of the second wave as noted by Eggo et al. (2011). Furthermore, areas of the Bernese Oberland were heavily affected, while others were largely spared the horrors of the pandemic. For one, areas around Lake Thun and Lake Brienz tended to be more affected. One possible explanation for this could be that they were simply easier to access which may have led to increased exchange with other places. In turn, this may have led to higher incidences by favouring an introduction into the municipality. Finally, many of the heavily affected areas were known tourist resorts such as Grindelwald and Lauterbrunnen that were more easily accessible (Dubler, 2009). Due to the ongoing World War I, it is questionnable whether tourism in fact played a role during the time of the 1918 influenza pandemic.

Again, the map of local spatial autocorrelation (see figure 4.9) provides insights into the dynamics of each of the waves. For the subsequent modelling process, the two principal waves were analysed, therefore these maps provide an important baseline to which the results of the models can be compared to. The map gives an overview into potential areas of interest. However, the map has two main shortcomings that are adressed at this point. The issue of polygons on the edge of the canton, where only a few neighbours account for the result,

has already been discussed in a previous section and it also prevails in these maps. Furthermore, municipalities with no incidence data turn out to be another shortcoming in the map of local spatial autocorrelation. As an example, the statistical analysis does not show any clustering for the greater region of Berne, despite the fact that the map of incidences (figure 4.8) shows high incidences in this area. This was the base for the previously discussed hypothesis that Berne may have been significantly affected at an early stage. However, the municipalities of Ittigen and Ostermundigen just east of Berne did not report any data, therefore interrupting a potential high-high cluster. This also may hold true for the cities of Biel and Thun, which both were not part of a cluster but bordered municipalities that did not report any influenza data. For the other areas discussed in the previous section, the map does show a significant high-high cluster for the region of Laufental, further supporting the thesis that its proximity to Basel may have been a factor for an early onset of the pandemic. For the southern Jura region, the map does not show a significant high-high cluster, instead producing one in the central part of the Jura region. Again, this has to be interpreted with caution. The local Moran's I considers all direct spatial neighbours and does not take topography into account. This proves as another shortcoming for a region that is heavily influenced by its topography such as the Jura region. As the relief in map 3.1 shows, the Jura consists of several steep valleys. Therefore, municipality A, which is not a spatial neighbour but within the same valley as municipality B, may actually have a bigger influence on municipality B's incidence than municipality C, which is a spatial neighbour but not within the same vallley, making the two not well-connected.

For the second wave, the map largely adds to the previously discussed findings. Generally, there are more high-high clusters, an interpretation for which may be that the second wave was simply more virulent than the first one. This interpretation maps onto the findings of various literature (Sonderegger, 1991; Sonderegger and Tscherrig, 2016). There are some scattered high-high clusters in the Swiss Plateau, indicating that it was indeed heavily affected. In the Bernese Oberland, the map further shows a high-high cluster around the previously discussed touristic areas, indicating that the incidence during the second wave was higher than during the first wave.

5.2 Research goal 2: Finding determinants of spread

The results of the descriptive spatio-temporal analysis yield two main findings. Firstly, that incidence is not only a function of population size, meaning that there must have been other factors that helped or hindered the spread of the virus. Secondly, that the two waves hit different areas with a different magnitude and therefore different locally specific factors may have played a role. As an additional help for interpreting the results, the bivariate choropleth maps from appendix A are considered. It is important to understand that interpretations based on associations only represent possible reasons why the disease spread the way it spread. These interpretations may have been true (to some degree) or completely false, as it is impossible to determine the real reason of the spread 100 years after the pandemic.

5.2.1 First wave: July 1918 – August 1918

Modelling process

One major advantage of using a logistic regression model is that it requires relatively few model prerequisites to be fulfilled (Stoltzfus, 2011). This makes it an easy application for the type of data available in this project, where various sources of data are combined into a single model. Another question that initially posed itself is how exactly to code the data into the two categories required for a logistic regression. During the modelling process, several ways of spliting the incidence data into categories were discussed. Finally, the decision was made to classify the quintile with the highest incidence vs. the lower four quintiles. This allows to study, which areas were particularly affected by the pandemic and to find potential reasons for why they were so heavily affected. The dependent variables were also grouped in different ways. Firstly, this makes the interpretation of them easier, and for some of them it makes the interpretation more meaningful (e.g. the node betweenness centrality is a network measure that not all readers might be familiar with, therefore grouping the results into variables with different levels of railway access makes it easier to interpret the data). The dependent variables and how each of them is classified is covered in section 3.1.

During the modelling process, a data-driven approach was followed. In the context of this project, this meant using an automated model selection to determine which model best fits the data. This approach seems suitable for the research question, as the literature suggests a variety of potential explanatory

factors (see section 2.5) and one of the objectives of the thesis was to find locally specific explanatory factors that contributed to the spread of the 1918 influenza pandemic in the case of the canton of Berne. One critical step consisted of selecting the "right" model from a set of models that the automatic model selection process returned (for all the candidate models, see table 4.1). One obvious approach would have been to simply select the model with the lowest AIC which was briefly considered as a choice. However, seeing that the difference in AIC was minimal led to the choice of the finally used model which contains a variety of different factors that cover various aspects: Firstly, the mortality from TB represents the health aspect, and shows how the population dealt with the exposure of a respiratory disease in the past. Secondly, the access to the railway system represents the geographic aspect by showing how central (or peripheral) a location is within the canton of Berne, which could be an indicator of how much exchange with outside locations took place. Furthermore, the weather covers the physical aspect of how the environment shaped the spread of the 1918 influenza pandemic. Finally, the variable urbanity serves as an attempt to take into account the different living conditions in rural and urban spaces.

Model results

The results of model 1 show that the association of some variances was of statistical significance (see table 4.2). However, the reason behind these associations is a matter of interpretation and cannot be definetely determined. One variable that model 1 controlled for was TB mortality in the years prior to the 1918 influenza pandemic. The model indicates a positive association between TB mortality and influenza incidence which means places that showed a high incidence were also more likely to be in the highest quintile of TB mortality. Therefore the initial hypothesis 2a) in section 2.6 is assumed to be true. In order to ease the interpretation, bivariate choropleth maps were created (see appendix A, specifically figure A.1). These maps add spatial information to the information gained by a traditional correlation analysis. Combined with the human ability of pattern recognition, this poses a powerful tool to assess the dynamics of the 1918 influenza pandemic. For the findings concerning the association between TB and influenza, several interpretations are possible. In literature, the correlation between pulmonary tuberculosis and influenza is already documented. A study found that in the city of Berne and Switzerland, the pulmonary tuberculosis deaths increased during the 1918 influenza pandemic (Zürcher et al., 2016). While the results of this thesis only show a positive association between TB in the years prior to the 1918 pandemic and influenza,

Zürcher et al. (2016) argue that influenza led to more TB deaths. Still, the question how the two illnesses influenced each other, remains unanswered. One possible interpretation of the results in this thesis could be that TB mortality functions as a proxy for TB incidence in the years before the 1918 influenza pandemic. This higher incidence is associated with widerspread lung damage due to earlier TB infections which increased the risk of falling severely ill from influenza and needing medical assistance. Secondly, the positive association between influenza and TB could simply mean that the overall conditions for the spread of an airborne virus were ideal but actually other factors were drivers (e.g. people working in a factory in close proximity). The map A.1 in appendix A shows that there are some areas, particularly in the Jura and Laufental regions, where both TB mortality and influenza incidence were high during the first wave. However, in many more municipalities, there was a high TB mortality in the years prior to the 1918 influenza pandemic but a low influenza incidence during the first wave. This makes the interpretation that the positive association is a sign for generally favorable conditions less likely. These are just two attempts to explain the weak positive association between TB mortality and influenza incidence. Finally, it is impossible to determine if any of the above or which one was actually the real reason. Therefore, the only statement that can definitely be made is that tuberculosis and influenza were somehow linked, which maps onto previous literature (Zürcher et al., 2016; Mamelund, 2011).

At this point it also has to be noted that the use of the TB mortailty statistics of 1900 – 1910 is not the most ideal solution, as it does not contain any information on the situation in the municipalities at the time of the 1918 influenza pandemic. Therefore, some uncertainty remains as to how the situation in the municipalities was in 1918. This could be improved by obtaining more accurate tuberculosis data. Furthermore TB data from 1910 were standardised using the population data from 1920. Using matching population data may also lead to a small improvement of these data.

The second variable that was included in model 1 was urbanity (i.e. Berne, Biel, Thun vs. the rest of the canton). The odds ratio returns very high values, indicating that the three cities were almost certainly in the highest quintile during the first wave, which may be an indicator that cities were an important factor in the early stages of the pandemic and indicating that hypothesis 2b) in section 2.6 also cannot be rejected for the first wave. This maps to findings of previous studies which showed that urban spaces showed an earlier onset of the pandemic than rural spaces (Chowell et al., 2008). Furthermore, it adds to the

findings of Sonderegger (1991) who observed that the panemic reached cities at an early stage before spreading to rural areas. When the pandemic reached the cities cannot be said with certainty, because the first data originate from July 1918, when the case numbers were already quite high. Therefore, from these data only, it is not possible to determine the onset of the pandemic in a specific municipality. These results support the findings of Eggo et al. (2011) who found that long-distance transmission was an important factor in the beginning stages of the pandemic, because cities tended to be better connected with places that are further away. Again, the model does not allow to make any statements about the exact transmission ways of the pandemic, but the fact that all the major cities were so heavily affected by the first wave is an indicator that they played an important role in the introduction of the virus into the canton of Berne.

The third variable that model 1 controlled for was railway access. Contrary to the first two variables, this variable does not show a clear tendency. While generally there seems to be a positive association between railway access and influenza incidence, the association does not become stronger with increasing railway access. This makes it hard to accept or reject hypothesis 2c) in section 2.6. The hypothesis cannot be seen as fully satisfied, but it is also not entirely wrong. However, the findings of this thesis map to the findings of Reyes et al. (2018) who found that railway lines were an important explanatory factor for the spread of the pandemic. For the interpretation of this factor in particular, figure A.3 proofs as insightful, as it helps to better understand the model results by adding spatial information. The map shows that indeed, there is some degree of correlation in the southern Jura region, and in the bigger cities. Furthermore, the northern part of the Jura region and the southern part of the Oberland region show areas that had no railway access and also showed a low incidence during the first wave of the 1918 influenza pandemic. Particularly in the Swiss Plateau, however, there are areas where the the correlation does not hold true. Therefore, it seems that the association between influenza and railway was stronger in hard-to-reach areas.

The data that describes the variable access to the railway system contains several shortcomings that have to be adressed. Firstly, calculating the node betweenness centrality for entire Swizerland and using it as a proxy for how well accessible a place was leads to several inaccuracies within the canton. Some regions, particularly in the region of Oberaargau and in the eastern part of the Voralpen region, contain places that have a high node betweenness centrality

because of their central location within Switzerland which results in many shortest paths leading through those railway lines. However, when only the canton of Berne is considered, they are not very central at all. Furthermore, the node betweenness centrality only calculates the shortest paths (in terms of horizontal distance) and does not consider topography which is an important factor in a mountainous country such as Switzerland. Another shortcoming of using the node betweenness centrality is that it does not consider passenger or train information: the algorithm does not take into account a) how many trains pass at a station per day, b) whether they stop, and c) how many passengers board and exit the train at a given station. All these factors potentially make a station seem more important than it actually is. Additionally, only using the railway system as a proxy for accessability ignores the fact that there are other means of transport. While cars may not have been very common, stagecoaches were an important means of transportation, particularly (but not exclusively) in the Alpine regions. Finally, the variable does not consider forms of local movement such as a person commuting to the next village for work, etc.

Even considering that all the factors discussed above introduced uncertainty into the results, some broad tendencies remain. Similar to the findings of Reyes et al. (2018), the railway system could have made an impact in the dissemination of the disease in the sense that it helped introduce the virus to a municipality. However, with the current data and methods, it remains difficult to link the railway system to the dissemination of the 1918 influenza pandemic.

The fourth and final variable that model 1 controlled for is precipitation. Again, model tendencies do not show a clear association. While the second and third group (indicating low to medium precipitation) show a positive association between precipitation and influenza incidence, the relationship is reversed for the highest two groups (which indicate high precipitation). Initially, I hypothesized that there would be a negative association between precipitation and influenza incidence (see 2.6, hypothesis 2d)). While the model shows a negative association for areas that received high amounts of precipitation, hypothesis 2d) from section 2.6 does not prove to be true for areas that receive a medium amount of precipitation. Again, the bivariate choropleth map A.4 in appendix A provides further insights into the spatial relation of influenza incidence and precipitation. The map shows a broad tendency where the Bernese Oberland and the northern Jura received a higher amount of precipitation but had a lower influenza incidence, and in the more low-lying areas (and parts of the Jura region) the pattern was reversed. The map also shows that there are outliers in each cate-

gory, highlighting that in this map only one explanatory factor is considered in an isolated manner, which is not capable of explaining all of the spatial variation of incidences. These findings prove contrary to the findings of Reyes et al. (2018) who found that rainfall helped explain the spatial variation of excess deaths in India. However, the findings of this thesis link to Roussel et al. (2016) who found no statistically significant association between the weather and the epidemic but conclude that the climate may have an impact on the spread of influenza on an intra-annual scale. Furthermore, they conclude that it remains unclear which climatic factor exactly contributed to the spread.

The precipitation data used for this thesis also contain some shortcomings. First of all, the data are not very accurate. While the measurements at the individual stations may have a high accuracy, the interpolation process is unable to accurately display local variations. Having data at a higher resolution would help to better capture local patterns. Furthermore, it remains questionnable whether "precipitation" is the best variable to picture weather and if available variables such as temperature or humidity may yield better results (see Roussel et al. (2016)).

Finally, it remains hard to say whether precipitation had an influence on influenza incidence during this first summer wave of the 1918 influenza pandemic. Using more precise data might yield better results but as Roussel et al. (2016) pointed out, it remains difficult to identify climatic factors that explain spatial variation of influenza incidences.

Local spatial autocorrelation

The local spatial autocorrelation was also used in the first research question to determine whether clusters of high (or low) incidence are present in the data (see figure 4.1.4). After creating a model, the same technique was used on the model residuals to gain an understanding of how well model 1 performs in space (i.e. how well model 1 manages to explain spatial variations in incidence). With an ideal model, the residuals would show no more clusters and no spatial autocorrelation because the model would manage to explain all the spatial variation in incidence. On a global scale, the Moran's I index decreased to almost zero, which means that there is practically no spatial autocorrelation in the model residuals. Indeed the map of the local spatial autocorrelation of the model residuals for the first wave shows very few clusters. Comparing the left map of figure 4.9 with figure 4.11 allows to determine in wich areas the model managed to explain spatial variations in incidence. The maps show that the model managed to explain most of the clusters in the Bernese Oberland, where many low-low

clusters were present in the initial data. Hovever, some clusters do remain, particularly in the Jura region and north of Berne. This means that the model did not manage to explain the spatial variation of influenza incidence in the case of high-high clusters and the model was better at explaining why the incidence was low in certain areas than why it was high in other areas. This is a hint that there must be at least one more locally specific factor, or that there are spatial interaction effects (i.e. incidence or explanatory factors of neighbouring municipalities have an additional effect on the incidence). Model 1 suffered from heteroscedasticity. This supports the thesis that the model was lacking at least one important explanatory factor but does not question the validity of the regression model.

5.2.2 Second wave: October 1918 – January 1919

Modelling process

For the modelling of the second wave, the same methods were used. They are extensively discussed in section 5.2.1 and are not repeated here. Noteworthy at this point is the model selection for the second wave. Among the candidate models that the model selection returned, one of the models was the same one as model 1, which included the explanatory variables TB mortality, urbanity, railway access and precipitation. Therefore this model was chosen as model 2, as the reasons that led to the selection of model 1 also hold true for model 2. Since the two models are the same for both waves, this leads to further possibilities to explore and compare the two waves (e.g. did the same factors play the same role in the spread during both waves?; are some tendencies even reversed?)

Model results

Model 2 also returned some statistically significant results. Around half of the variables have statistically significant odds ratios. The first variable the model controlled for was TB mortality. The result of model 2 (second wave) returns similar results as model 1 (first wave). Therefore, the initial hypothesis 2a) 2.6 also cannot be rejected for the second wave. As already discussed in the first wave, this maps to the findings of Zürcher et al. (2016) who found that TB and influenza seem to be related. As in the first wave, one possible explanation is that municipalities that had a higher TB mortality also had a higher influenza incidence. This then led to more severe influenza cases in the following influenza epidemic due to the lung damage caused by TB. Another possible explanation could be that the positive association between TB mortality and influenza inci-

dence is just a general sign that conditions for the spread of an airborne virus were favorable in certain municipalities. The map in appendix A again shows the spatial correlation between TB mortality and influenza incidence which can be of help for the interpretation. The map shows that areas with both high TB mortality and high influenza incidence include regions north of Berne, south of Lake Biel and in between Lake Brienz and Lake Thun, just to name a few. The Jura region, despite its high TB mortality in the years of 1900 – 1910, was not much affected by the second wave of the 1918 influenza pandemic. However, the study does not allow to make a final statement about the actual reason why TB mortality and influenza incidence are linked.

The model results of model 2 show that the odds ratio of the urbanity were not statistically significant during the second wave (October 1918 – January 1919). Furthermore, their odds ratio was 0, indicating that the variable "urbanity" (i.e. Berne, Biel, Thun vs. the rest of the canton) did not have an influence on the dissemination of the influenza virus. Therefore, hypothesis 2b) from section 2.6 has to be rejected, as it cannot be said that urban spaces had higher incidences during the second wave. In this respect, the first and the second wave were different: In the first wave model, urbanity had an extremely high odds ratio, indicating that urbanity was an important explanatory factor for the spread. At this point, it again has to be mentioned, that the standard error for the variable urbanity was high in both models which means that the variable did not provide a good fit for the data. However, together with the descriptive analysis and the findings of Sonderegger (1991) it is still safe to say that cities had an influence, particularly during the first wave.

The odds ratios of the variable railway access also do not draw a clear picture, neither in respect to statistical significance nor in the direction of association. This means that the hypothesis formulated in section 2.6 has to be rejected, as it cannot be definitely said that railway access helps explain the spread of the influenza virus. In this respect, the results are also different for the first and the second wave. While the results of model 1 did not show a strictly positive association between railway access and influenza incidence, at least there was a clear difference between the groups that had railway access and the group that did not. This is not the case in the second wave, as the results from model 2 indicate that group 3 (which translates to medium railway access) had a negative association. Therefore, hypothesis 2c) in section 2.6 cannot be accepted. The map in figure A.7 shows indeed that there is a less clear picture than for the

first wave. One reason for the negative association in group 3 could be the Jura region, which was practically not affected but relatively well-connected to the railway network. On the other hand, areas in the Bernese Oberland that did not have railway access were heavily affected. While this does not agree with the findings of Reyes et al. (2018) who found that long-distance railway access had an influence on the spatial variation of the 1918 influenza pandemic, it does provide a quantification for the results of Sonderegger (1991), who concluded that the German speaking parts were more heavily affected during the second wave. One potential explanation as to why the correlation between railway access and influenza incidence does not provide a clear picture during the second wave, could be Eggo et al. (2011)'s conclusion that in the initial phase, long-distance transmission was of importance while in the later stages local spread became dominant. In the context of this case study, this would mean that during the first wave, long-distance transmission along the railway lines was an important driver of spread. In the second wave, local spread became important, therefore railway access did not play such a big role anymore, or it could simlply be the case, that the second wave was so virulent and widespread, that the railway lines did not matter anymore. Of course this is only a hypothesis that cannot be definitely proven. The data is the same as used for model 1 and therefore has the same shortcomings (for an extensive discussions of them, please refer to 5.2.1). A more accurate modelling of general accessability could prove insightful in further studies.

The last variable, precipitation, also failed to provide a clear picture. While groups 2 and 3 (low and medium precipitation) showed a positive association, groups 4 and 5 (high and very high precipitation) had a negative association, however the odds ratio was not statistically significant. Therefore, hypothesis 2d) (see section 2.6) has to be rejected, as the evidence does not support it. Furthermore, the results of model 2 also don't support those of Reyes et al. (2018) who found that precipitation was an explanatory factor for the 1918 influenza pandemic in India. However, the findings in this study are similar to the ones in Roussel et al. (2016) who found no statistically relevant association between climatic variables and incidence on an epidemic level. The map in figure A.8 shows that the precipitation variable does show the common pattern associated with precipitation in Switzerland (more precipitation in the mountain regions than in the flat lands), but no clear pattern correlation with influenza incidence. Therefore, precipitation is not such a good variable for explaining the spread of the influenza virus during the second wave. This is consistent with the results

of model 1 where precipitation also failed to provide a clear association with influenza incidence. Hence, the seasons also did not seem to have an impact when it comes to precipitation. Potentially better results could be achieved with different climatic data such as temperature or humidity, which may play a role, as suggested by Roussel et al. (2016).

Local spatial autocorrelation

Again, the map in figure 4.13 shows the local spatial autocorrelation of the model residuals of model 2. Compared to the residual map of model 1 (see figure 4.11), the map shows considerably more and larger clusters. This indicates that the model performance was worse during the second wave. This is also undermined by the global Moran's I which was considerably higher than for model 1. The global Moran's I for model 2 was considerably higher than for model 1.

Again, comparing the LISA (local indicators of spatial autocorrelation) map of the model residuals to the LISA map of the incidences (see figure 4.9) allows to assess how well the model explained the spatial variation of the incidence. On a global level, the Moran's I of the model residuals is lower than the Moran's I of the incidence rates, which indicates that the model managed to explain some of the spatial variation of the incidences. Just like model 1, model 2 managed to explain a large portion of the low-low clusters in the regions that were not very much affected (e.g. the Oberland region for model 1 and the Jura region for model 2). Despite the fact that the model managed to explain some of the high-high clusters spread out across the Swiss Plateau, substantial clusters of both high-high and low-low type remain. Furthermore, the residual map shows new low-low clusters, particularly in the area between Berne and Thun. This indicates that in this area, the model overestimated the influenza incidence. The reason why this was the case remains unclear. Possibly, the model omitted an important explanatory factor or spatial interaction effects had an influence (i.e. the incidence or explanatory factors of neighbouring municipalities influenced the incidence). Furthermore, a new high-high cluster appeared in the Bernese Oberland, south of Lake Brienz. One potential explanation for this new cluster may be the tourism industry. The municipalities that belong to this high-high cluster were well-known tourism resorts such as Wengen or Grindelwald, which is why they were relatively well-connected to to other places, despite the fact that they seem very isolated mountain regions, particularly since the model does not consider influences such as tourism. However, due to the fact that World War I was just ending at the time, it remains questionnable whether tourism in fact played a role. Finally, it has to be noted that model 2 suffered

from heteroscedasticity, just like model 1. This means that some variation in the incidence is not well-explained in the model, and hints that at least one important explanatory variable is omitted. These findings go well with the LISA map which still shows substantial clustering, indicating that there is still missing information that may explain the local variation of the spread more accurately.

5.2.3 Differences between the two waves

Sonderegger (1991); Eggo et al. (2011); Smallman-Raynor (2004); Patterson and Pyle (1991) note that the second wave of the 1918 influenza pandemic was more severe by orders of magnitude, which is something which these data also show. Furthermore, the French-speaking part of the canton of Berne was hit harder by the first wave and largely spared by the second wave, while the pattern is reversed for large areas in the German-speaking part of the canton. This provides a quantification to the findings of Sonderegger (1991) and therefore, hypothesis 1 from section 2.6 can be accepted: the pandemic, broadly speaking, spread from west to east. At the model level, it is noteworthy that the model performance was better in model 1 with an AIC of 363.46, compared to model 2 which had an AIC of 461.66, meaning the model fit was better for model 1. Furthermore, the results of the explanatory variables produced outcomes that were clearer and more statisticallly sound than in model 1. One possible way of explaining this fact is that due tothe incidences being generally higher during the second wave, the disease was also more widespread than during the first wave. This makes it harder to find meaningful explanatory variables. This is just a possible explanation as to why the model performance was worse in model 2. The real reason remains unknown, just like it remains unknown why the second wave was so much more virulent. Finally, both models suffred from heteroscedasticity, which indicates that some information is missing that would help to explain the spread even further but does not question the validity of the two models.

5.2.4 Implications for future pandemics

These findings provide a broad idea as to what might have contributed to the spread of the 1918 influenza pandemic. The question remains what these results can teach us and how they can help to be better prepared for a future pandemic. The current Sars-Cov-2 pandemic teaches the world that preparedness is of essence, especially in a world as globalized as our world is today. The current

pandemic shows that in our globalized world, a virus has the potential to spread much faster both on a global scale, where air travel makes it possible to reach even the most remote places within hours. On a local scale, increased local exchange (e.g. through commuting) helps to accelerate the spread. Covid-19 teaches us that when a new disease develops, the available means to fight it are no different from the means available during the 1918 influenza pandemic, at least in the beginning stages until effective medication or a vaccine are invented. This means, governments rely on nonpharmazeutical interventions at least in the beginning stages. Therefore, it is of great importance to learn from past pandemics to be prepared for the next one. The results from this thesis show two associations that could have implications for future pandemics. Firstly, cities were affected at an early stage of the pandemic, therefore showing the possibility that long-distance connections were an important driver in the initial stages of the pandemic. Secondly, the access to the railway network showed some association, therefore giving hints that traffic ways are an important driver of spread, which for a current pandemic might extend to air travel as well. These two findings have possibly even bigger implications today due to the before-mentioned possibility of an accelerated spread. Therefore, governments should be prepared to act quickly in order to contain or at least delay a possible outbreak.

5.3 Limitations

Like every research project, this one has its limitations. For the sake of structure, limitations are split into three categories: (1) limitations of the data which show how better or different data could lead to better results, (2) limitations in the analysis that show how a different modelling process coluld lead to even more accurate results, and (3) limitations in the analysis that show how more advanced modelling techniques could produce more accurate results.

5.3.1 Limitations of the data

The different data variables all have their limitations. Firstly, the urbanity variable is a simple binary classification that compares the three biggest cities to the rest of the canton which resulted in a high standard error in the data. To obtain a better result, a more advanced method could be developed that possibly includes other factors such as tourism or industry. This would allow to capture the structure of the settlement more accurately. Secondly, the tuberculosis variable contains the TB mortality from the years before the actual pandemic.

This may lead to an incomplete picture, as it does not capture more recent outbreaks. Furthermore, the TB mortality from 1910 does not make any statements about the real incidence other than the presumption that a higher TB mortality points to higher influenza case numbers. Therefore, obtaining more accurate TB data (if available) could lead to improved results. Additionally, the access to the railway system is modelled with a simple network measure. This does not capture a complete picture as it makes some train stations more central than they seem in the model compared to reality, and it does not allow to make a statement as to how frequently a line is used. A more advanced approach could lead to better results concerning access to the railway network. The weather data consisted of a simple interpolation of precipitation measurements from the Federal Office of Meteorology and Climatology. Recently, I was made aware of more precise historical weather data, originally intended to study climate change. The data includes high-resolution temperature and precipitation data (Pfister, 2019). Using these data may yield more accurate results, particularly the use of temperature data seems an interesting opportunity, as for the current models no temperature data was available.

Finally, some limitations in the outcome data have to be discussed. As stated in the previous paragraph, two facts are important to consider when working with the outcome data. Firstly, in 1918 the only way to diagnose influenza was by clinical methods – assessing the symptoms a patient showed. Having no test available to definitely say whether a patient had a common cold or was in fact suffering from influenza introduced inaccuracies and can lead to a reporting bias. This means a patient was diagnosed with influenza when in fact they were not suffering from influenza or vice versa. Secondly, even though the reporting of influenza cases was mandatory for doctors, they often saw it as an unnecessary bureaucratic step (Staub et al.). Therefore, this also introduces some inaccuracies, as doctors might not have reported all the data or reported data inaccurately.

5.3.2 Missing data

The issue of missing data has two dimensions. The first dimension is relatively simple to discuss and concerns the fact that not all municipalities reported data, particularly during the first wave. Given the locations of some of them (e.g. Ostermundigen and Ittigen just east of Berne) which were surrounded by municipalities with high incidences, it seems very unlikely that they did not have any cases. This is a limitation that is almost impossible to overcome but it has to be kept in mind when interpreting the model outcomes.

Secondly, the model outcomes point to some missing variables. Mamelund (2011) and Jester et al. (2018) suggest that young adults were more likely to die from influenza. It remains questionnable whether the age structure has an influence on the incidence rate. However, it is possible that young adults, due to their tendency to have more severe courses, led to more medical consultations which might manifest in a higher reported incidence.

Additionally, the data has shown that accessability played a role in the spread of the 1918 influenza pandemic. However, part of how well accessible a municipality was, was completely omitted by not considering roads and stagecoach lines. Including information on how well connected to the road and stagecoach network a municipality was could yield better results, particularly in the Bernese Oberland. Also, the current modelling of accessability does not consider local spread, which previous studies identified as an important driver in the later stages of the pandemic (Eggo et al., 2011).

Finally, it remains questionnable whether there is data for all phenomena that may have contributed to the spread of the influenza virus. It is now known that the general strike was associated with increasing case numbers during the peak of the second wave (Staub et al.). It is highly questionnable whether there is any data that could be included in any model to help explain the variations of the spread afterwards. Furthermore, other factors such as co-morbidities or background immunity may also have played a role as suggested by Chowell et al. (2014). It is questionnable whether good data to capture these factors can be found.

5.3.3 Limitations of the analysis

Last but not least, the way the analysis was conducted contains some inaccuracies. The models consist of simple logistic regression models. Their advantage is that they have relatively few model assumptions and therefore, they are relatively easy to use and interpret. However, they are somewhat limited in the statements they allow to make. Therefore, using a more advanced modelling technique such as general additive models might provide further and more accurate insights.

The model also has another particular shortcoming. The standard error and the odds ratio of the urbanity variable is extremely high, which means the data does not provide a good fit for the model. One possible explanation for this lies within the way the data was modelled. The variable contains three points in one class (Biel, Berne, and Thun) and all the rest of the datapoints belonged to another class. Naturally, this classification does not provide a good fit as also

other areas had similar incidence rates as the three cities. Therefore, the model can be improved by coming up with a better classification scheme that captures urbanness, or the settlement structure in genereal, better. However, modelling this variable more accurately would have been a considerable effort and would have been out of scope for this thesis. Finally, the TB variable used data from 1900 – 1910 and was standardised using the census data from 1920 which could lead to inaccurate results due to population movements (e.g. because of World War I). Using population data from the 1910 census may lead to more accurate results.

5.4 Further research

The results of this thesis represent a first step towards a better understanding of the 1918 influenza pandemic and its determinants of spread. This thesis is by no means conclusive and further research should be conducted in order to understand the "Spanish" flu even better. After writing this thesis, I see further research potential in three areas.

The first area concerns how the current approach could be improved by using more accurate data and building more advanced models. Firstly, by including additional variables such as the age structure or general indicators of population health, better results may be achieved simply due to the fact that these data are missing in the current analysis. Secondly, most explanatory variables are modelled relatively simply as putting more effort into modelling them more precisely would have been out of scope for this thesis. Putting more effort in modelling the explanatory variables as precisely as possible could yield better results (e.g. a better classification for urbanness and railway access). Therefore, this could be adressed in a further research project. Additionally, the models used for this thesis were relatively simple and building more advanced models may yield more insights.

Another big area that could be interesting for further research is the temporal resolution. The current approach incorporates the temporal dimension by comparing the two waves to each other. Using a finer temporal resolution (i.e. weeks or months) could lead to further insights and help to capture the spread even better.

Finally, the third area where I see potential is in the effective communication of the research through appealing visualizations. Earlier in this thesis, the point was made that an effective communication of research is important in the case of pandemics in order to allow access to information without causing panic.

6 Conclusion

The research in this thesis contributes to our understanding of the 1918 influenza pandemic, a pandemic that is sometimes described as the greatest demographic shock in human history. The canton of Berne represents a unique case study with a regionally very diverse setting, stretching from the northern border with France to remote places in the Swiss Alps and high case numbers. This paper yields two main findings: First of all, it adds to current literature that the Swiss canton of Berne, like many places in the Northern Hemisphere, was struck in two waves. The first wave was a relatively mild summer wave in July/August 1918 and it was followed by a much more virulent second wave that lasted from around October 1918 to January 1919. Furthermore, this thesis agrees with previous research that the Jura region was affected more by the first wave, while the rest of the canton was struck harder by the second wave. Secondly, the thesis identified locally specific factors that may have had an influence on the spread of the virus. In this case study, especially traffic routes and urbanness seemed to have had an influence. The model results show, that particularly for the second wave, the model performence could be improved. Therefore, more effort should be put in finding locally specific factors that help explain the spread of the much more virulent winter wave as well as more advanced modelling techniques.

Finally, the thesis argues that the study of past pandemics can be a great help in developing an emercency plan for future pandemics. The current Sars-Cov-2 pandemic is a great reminder that, despite greatly improved possibilities through modern medicine – if a new virus arises, in the beginning, the available measures are no different than the ones available 100 years ago, before the development of modern medication. Therefore, identifying possible determinants of spread and coping mechanisms is one of the most valuable tools available to prepare for a future pandemic. Because one thing is certain: "The world will face another influenza pandemc – the only thing we don't know is when it will hit and how severe it will be" (World Health Organization, 2019).

Bibliography

Ismail Ocsoy Amit Kumar Mandal , Paulami Dam , Octavio L. Franco , Hanen Sellami , Sukhendu Mandal , Gulten Can Sezgin , Kinkar Biswas , Partha Sarathi Nandi. COVID-19: fighting panic with information. *Lancet*, (January):19–21, 2020.

C. E. Ammon. Spanish Flu epidemic in 1918 in Geneva, Switzerland. *Euro Surveillance*, 7(12):5–9, 2002. doi: Ammon,C.E.(2020) .SpanishFluuepidemicin1918inGeneva,Switzerland.EuroSurveillance,7(12) ,5âĂŞ9.https://www.eurosurveillance.org/content/10.2807/esm.07.12.00391-en. URL `https://www.eurosurveillance.org/content/10.2807/esm.07.12.00391-en`.

Gennady Andrienko, Natalia Andrienko, Urska Demsar, Doris Dransch, and Jason Dykes. Space , time and visual analytics. *International Journal of Geographical Information Science*, 24(10):1577–1600, 2010. doi: 10.1080/13658816.2010.508043.

Luc Anselin. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2):93–115, apr 1995. ISSN 0016-7363. doi: https://doi.org/10.1111/j.1538-4632.1995.tb00338.x. URL `https://doi.org/10.1111/j.1538-4632.1995.tb00338.x`.

Luc Anselin, Julia Koschinsky, Xun Li, Dylan Halpern, Qinyun Lin, Susan Paykin, Marynia Kolak, and Kevin Credit. GeoDa, 2003.

Tommy Bengtsson, Martin Dribe, and Björn Eriksson. Social class and excess mortality in Sweden during the 1918 influenza pandemic. *American Journal of Epidemiology*, 187(12):2568–2576, 2018. ISSN 14766256. doi: 10.1093/aje/kwy151.

Jacques Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983. ISBN 0299090604.

W. Besorger. Als das Fieber nach Zug kam: Die "Spanische" Grippe von 1918/19. *Tugium*, 34:193–211, 2018.

Martin C.J. Bootsma and Neil M. Ferguson. The effect of public health measures on the 1918 influenza pandemic in U.S. cities. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18):7588–7593, 2007. ISSN 00278424. doi: 10.1073/pnas.0611071104.

Gabrielle Brankston, Leah Gitterman, Zahir Hirji, Camille Lemieux, and Michael Gardam. Transmission of influenza A in human beings. *Lancet Infectious Diseases*, 7(4):257–265, 2007. ISSN 14733099. doi: 10.1016/S1473-3099(07)70029-4.

T. S. Breusch and A. R. Pagan. The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The Review of Economic Studies*, 47(1):239, 1980. ISSN 00346527. doi: 10.2307/2297111.

Cynthia A. Brewer. Chapter 7 - Color Use Guidelines for Mapping and Visualization. In ALAN M MACEACHREN and D R FRASER TAYLOR, editors, *Visualization in Modern Cartography*, volume 2 of *Modern Cartography Series*, pages 123–147. Academic Press, 1994. doi: https://doi.org/10.1016/B978-0-08-042415-6.50014-4. URL `http://www.sciencedirect.com/science/article/pii/B9780080424156500144`.

Cynthia A. Brewer, Mark Harrower, Ben Sheeshley, Andy Woodruff, and David Heyman. Color Brewer 2.0 - Color Advice for Cartography, 2002. URL `https://colorbrewer2.org/{#}type=sequential{&}scheme=BuGn{&}n=3`.

Konstantin Büchel and Stephan Kyburz. Fast track to growth? Railway access, population growth and local displacement in 19th century Switzerland. *Journal of Economic Geography*, 20(1):155–195, 2018. ISSN 14682710. doi: 10.1093/jeg/lby046.

Bundesamt für Landestopographie (swisstopo). Karten der Schweiz, 2020. URL `https://map.geo.admin.ch/?lang=de{&}topic=ech{&}bgLayer=ch.swisstopo.pixelkarte-farbe{&}layers=ch.swisstopo.zeitreihen,ch.bfs.gebaeude{_}wohnungs{_}register,ch.bav.haltestellen-oev,ch.swisstopo.swisstlm3d-wanderwege{&}layers{_}opacity=1,1,1,0.8{&}layers{_}visibility=false,`.

Bundesamt für Statistik. *Statistisches Jahrbuch der Schweiz 1920*. Bern, 1921.

Kenneth P Burnham and David R Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, nov 2004. ISSN 0049-1241. doi: 10.1177/0049124104268644. URL `https://doi.org/10.1177/0049124104268644`.

Vincent Calcagno and Claire de Mazancourt. glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. *Journal of Statistical Software*, 34(i12), 2010. doi: DOI:http://hdl.handle.net/10.,. URL `https://ideas.repec.org/a/jss/jstsof/v034i12.html`.

Joseph E Cavanaugh and Andrew A Neath. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3):e1460, may 2019. ISSN 1939-5108. doi: https://doi.org/10.1002/wics.1460. URL `https://doi.org/10.1002/wics.1460`.

Siddharth Chandra and Eva Kassens-Noor. The evolution of pandemic influenza: Evidence from India, 1918-19. *BMC Infectious Diseases*, 14(1):1–10, 2014. ISSN 14712334. doi: 10.1186/1471-2334-14-510.

G. Chowell, C. E. Ammon, N. W. Hengartner, and J. M. Hyman. Estimation of the reproductive number of the Spanish flu epidemic in Geneva, Switzerland. *Vaccine*, 24(44-46):6747–6750, 2006. ISSN 0264410X. doi: 10.1016/j.vaccine.2006.05.055.

Gerardo Chowell, Luís M.A. Bettencourt, Niall Johnson, Wladimir J. Alonso, and Cécile Viboud. The 1918-1919 influenza pandemic in England and Wales: Spatial patterns in transmissibility and mortality impact. *Proceedings of the Royal Society B: Biological Sciences*, 275(1634):501–509, 2008. ISSN 14712970. doi: 10.1098/rspb.2007.1477.

Gerardo Chowell, Anton Erkoreka, Cécile Viboud, and Beatriz Echeverri-Dávila. Spatial-temporal excess mortality patterns of the 1918-1919 influenza pandemic in Spain. *BMC Infectious Diseases*, 14(1):1–12, 2014. ISSN 14712334. doi: 10.1186/1471-2334-14-371.

Laura Cilek, Gerardo Chowell, and Diego Ramiro Fariñas. Age-specific excess mortality patterns during the 1918–1920 influenza pandemic in Madrid, Spain. *American Journal of Epidemiology*, 187(12):2511–2523, 2018. ISSN 14766256. doi: 10.1093/aje/kwy171.

Bernard Degen. "Landesstreik", 2012. URL `https://hls-dhs-dss.ch/de/articles/016533/2012-08-09/`.

Sharon N. DeWitte. Mortality risk and survival in the aftermath of the medieval Black Death. *PLoS ONE*, 9(5), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0096513.

Anne-Marie Dubler. "Berner Oberland", 2009. URL https://hls-dhs-dss.ch/de/articles/010296/2009-05-20/.

Rosalind M. Eggo, Simon Cauchemez, and Neil M. Ferguson. Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *Journal of the Royal Society Interface*, 8(55):233–243, 2011. ISSN 17425662. doi: 10.1098/rsif.2010.0216.

Hans-Ulrich Egli, Hans-Rudolf, Flury, Philipp, Frey, Thomas, Schiedt. «GIS-Dufour» – Verkehrs- und Raumanalyse auf historischer Grundlage. *Geomatik Schweiz/Geomatique Suisse*, (5):5246–249, 2005.

ESRI. What is the ArcGIS Spatial Analyst extension?, 2020a. URL https://desktop.arcgis.com/en/arcmap/latest/extensions/spatial-analyst/what-is-the-spatial-analyst-extension.htm.

ESRI. How Zonal Statistics works, 2020b. URL https://desktop.arcgis.com/de/arcmap/10.3/tools/spatial-analyst-toolbox/h-how-zonal-statistics-works.htm.

Estelle Fallet. "Uhrenindustrie", 2020. URL https://hls-dhs-dss.ch/de/articles/013976/2020-08-11/.

H J Field and E De Clercq. Antiviral drugs - a short history of their discovery and development. *Microbiology Today*, 31:58–61, 2004.

Arnstein Finset, Hayden Bosworth, Phyllis Butow, Pål Gulbrandsen, Robert L. Hulsman, Arwen H. Pieterse, Richard Street, Robin Tschoetschel, and Julia van Weert. Effective health communication – a key factor in fighting the COVID-19 pandemic. *Patient Education and Counseling*, 103(5):873–876, 2020. ISSN 18735134. doi: 10.1016/j.pec.2020.03.027.

A. Fleming. On the Antibacterial action of cultures of a penicillinum, with special reference to their use in the isolation of B. influenza. *Bull World Health Organ*, 79(8):780–790, 1929.

Dominik Geisberger, Robert & Sanders, Peter & Schultes. Better Approximation of Betweenness Centrality. In *Proceedings of the 10th Workshop on Algorithm Engineering and Experiments and the 5th Workshop on Analytic Algorithmics and Combinatorics*, pages 90–100, 2008. doi: 10.1137/1.9781611972887.9.

Julia R. Gog, Sébastien Ballesteros, Cécile Viboud, Lone Simonsen, Ottar N. Bjornstad, Jeffrey Shaman, Dennis L. Chao, Farid Khan, and Bryan T. Grenfell. Spatial Transmission of 2009 Pandemic Influenza in the US. *PLoS Computational Biology*, 10(6), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003635.

Michael Greenberger. Better prepare than react: Reordering public health priorities 100 years after the Spanish flu epidemic. *American Journal of Public Health*, 108(11):1465–1468, 2018. ISSN 15410048. doi: 10.2105/AJPH.2018.304682.

Gabor Gsardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Sy:1965, 2006.

Trevor Hoppe. " Spanish Flu " : When Infectious Disease Names Blur Origins and Stigmatize Those Infected. 108(11):1462–1464, 2018. doi: 10.2105/AJPH.2018.304645.

George Jenks and Fred Caspal. Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*, 62:217–244, 1971. doi: https://doi.org/10.1111/j.1467-8306.1971.tb00779.x.

Barbara J. Jester, Timothy M. Uyeki, Anita Patel, Lisa Koonin, and Daniel B. Jernigan. 100 Years of Medical Countermeasures and Pandemic Influenza Preparedness. *American Journal of Public Health*, 108(11):1469–1472, 2018. ISSN 15410048. doi: 10.2105/AJPH.2018.304586.

Niall P.A.S. Johnson and Juergen Mueller. Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the history of medicine*, 76(1):105–115, 2002. ISSN 00075140. doi: 10.1353/bhm.2002.0022.

Daniel Koch. Meldepflichtige übertragbare Krankheiten und Erreger - Leitfaden zur Meldepflicht. Technical report, Bundesamt für Gesundheit BAG, Bern, 2020.

Raymond Kohli. Todesfälle in der Schweiz auf Rekordniveau - Die Spanische Grippe von 1918. *BFS Aktuell*, (November 2018):1–8, 2018. URL `https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/geburten-todesfaelle.assetdetail.6467464.html`.

Wen-Chung Lee. Characterizing Exposure-Disease Association in Human Populations using the Lorenz Curve and Gini Index. *Statistics in Medicine*, 16(7):729–739, apr 1997. ISSN 0277-6715. doi:

https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7<729::AID-SIM491>3.0.CO;2-A. URL https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7{%}3C729::AID-SIM491{%}3E3.0.COhttp://2-a.

Corina Leuch. Researching a Pandemic - During a Pandemic, 2020. URL https://www.geo.uzh.ch/en/department/125/blog/pandemic.html.

Susie Lu and Elijah Meeks. Viz Palette, 2020. URL https://projects.susielu.com/viz-palette.

Svenn Erik Mamelund. Geography May Explain Adult Mortality from the 1918-20 Influenza Pandemic. *Epidemics*, 3(1):46–60, 2011. ISSN 17554365. doi: 10.1016/j.epidem.2011.02.001. URL http://dx.doi.org/10.1016/j.epidem.2011.02.001.

Edward R. Mansfield and Billy P. Helms. Detecting Multicollinearity. *The American Statistician*, 36(3a):158–160, 1982. ISSN 0003-1305. doi: 10.1080/00031305.1982.10482818.

L. Marino. La Grippe Espagnole en Valais (1918-1919), 2014.

Christina E. Mills, James M. Robins, and Marc Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004. ISSN 00280836. doi: 10.1038/nature03063.

P A P Moran. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society: Series B (Methodological)*, 10(2):243–251, jul 1948. ISSN 0035-9246. doi: https://doi.org/10.1111/j.2517-6161.1948.tb00012.x. URL https://doi.org/10.1111/j.2517-6161.1948.tb00012.x.

David M. Morens and Jeffery K. Taubenberger. The mother of all pandemics is 100 years old (and going strong)! *American Journal of Public Health*, 108 (11):1449–1454, 2018. ISSN 15410048. doi: 10.2105/AJPH.2018.304631.

Donald R. Olson, Lone Simonsen, Paul J. Edelson, and Stephen S. Morse. Epidemiological evidence of an early wave of the 1918 influenza pandemic in New York City. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11059–11063, 2005. ISSN 00278424. doi: 10.1073/pnas.0408290102.

Judy M. Olson. Spectrally Encoded Two Variable Maps. *Annals of the Association of American Geographers*, 71(2):259–276, jun 1981. ISSN 0004-5608. doi: 10.1111/j.1467-8306.1981.tb01352.x. URL https://doi.org/10.1111/j.1467-8306.1981.tb01352.x.

Our World in Data. World population since 10,000 BCE (OurWorldInData series), 2020.

John Paget, Peter Spreeuwenberg, Vivek Charu, Robert J. Taylor, A. Danielle Iuliano, Joseph Bresee, Lone Simonsen, and Cecile Viboud. Global mortality associated with seasonal influenza epidemics: New burden estimates and predictors from the GLaMOR Project. *Journal of Global Health*, 9(2):1–12, 2019. ISSN 20472986. doi: 10.7189/jogh.09.020421.

Wendy E. Parmet and Mark A. Rothstein. The 1918 Influenza Pandemic: Lessons Learned and Not-Introduction to the Special Section. *American journal of public health*, 108(11):1435–1436, 2018. ISSN 15410048. doi: 10.2105/AJPH.2018.304695.

David Patterson and Gerald Pyle. The Geography and Mortality of the 1918 Influenza Pandemic. *Bulletin of the History of Medicine*, 65(1):4–21, 1991.

Thomas Pedersen. A Tidy API for Graph Manipulation, 2020. URL `https://www.rdocumentation.org/packages/tidygraph`.

C Pfister. Im Strom der Modernisierung. Bevölkerung, Wirtschaft und Umwelt 1700-1927. *Bern: hier+jetzt*, 2011.

Lucas Pfister. Statistical Reconstruction of Daily Precipitation and Temperature Fields in Switzerland back to 1864. *Climate of the Past*, 16(2):663–678, oct 2019. doi: 10.1594/PANGAEA.907579. URL `https://doi.org/10.1594/PANGAEA.907579`.

G M Philip and D F Watson. A precise Method for Determining Contoured Surfaces. *The APPEA Journal*, 22(1):205–212, 1982. URL `https://doi.org/10.1071/AJ81016`.

Regierungsrat des Kantons Bern. Regierungsratsbeschluss 3779: Influenza-Epidemie Massnahmen, 1918.

Olivia Reyes, Elizabeth C. Lee, Pratha Sah, Cécile Viboud, Siddharth Chandra, and Shweta Bansal. Spatiotemporal Patterns and Diffusion of the 1918 Influenza Pandemic in British India. *American journal of epidemiology*, 187 (12):2550–2560, 2018. ISSN 14766256. doi: 10.1093/aje/kwy209.

Rhätische Bahn AG. Geschichte: Schon seit 1889 faszinierend anders unterwegs, 2020. URL `https://www.rhb.ch/de/unternehmen/portraet/geschichte`.

Philip K. Robertson and John F. O'Callaghan. The Generation of Color Sequences for Univariate and Bivariate Mapping. *IEEE Computer Graphics and Applications*, 6(2):24–32, 1986. ISSN 02721716. doi: 10.1109/MCG.1986.276688.

Max Roser. The Spanish flu (1918-20): The global impact of the largest influenza pandemic in history, 2020. URL `https://ourworldindata.org/spanish-flu-largest-influenza-pandemic-in-history{#}:{~}:text=Estimatessuggestthattheworld,{%}25)oftheworldpopulation.`

Marion Roussel, Dominique Pontier, Jean Marie Cohen, Bruno Lina, and David Fouchet. Quantifying the role of weather on seasonal influenza. *BMC Public Health*, 16(1):1–14, 2016. ISSN 14712458. doi: 10.1186/s12889-016-3114-x. URL `http://dx.doi.org/10.1186/s12889-016-3114-x`.

John Graham Royde-Smith and Thomas A. Hughes. World War II, 2020. URL `https://www.britannica.com/event/World-War-II/Costs-of-the-war`.

Sanitätsdirektion des Kantons Bern. Verwaltungsberichte Der Sanitätsdirektion Für Die Jahre 1918 & 1919, 1918.

Terry A Slocum, Robert M McMaster, Fritz C Kessler, Hugh H Howard, and Robert B Mc Master. *Thematic Cartography and Geovisualization: Pearson New International Edition*. Prentice Hall, Upper Saddle River, New Jersey, 3 edition, 2008. ISBN 129204067X.

Matthew Smallman-Raynor. Winter and Seasonal Ailments. In *World Atlas of Epidemic Diseases*, chapter 6., pages 84–97. 2004. doi: 10.1201/b13526-7.

Matthew Smallman-Raynor, Niall Johnson, Andrew D Cliff, Matthew Smallman-raynor, and Niall Johnsont. The Spatial Anatomy of an Epidemic : Influenza in London and the County Boroughs of Published by : Wiley on behalf of The Royal Geographical Society (with the Institute of British Geographers) Stable URL : http://www.jstor.org/stable/3804472 The spatial. *Transactions of the Institute of British Geographers*, 27(4):452–470, 2017.

Christian Sonderegger. *Die Grippewelle 1918/19 in der Schweiz*. Lizenziatsarbeit, Universität Bern, 1991.

Christian Sonderegger and Andreas Tscherrig. Die Grippepandemie 1918-1919 in der Schweiz. *«Woche für Woche neue Preisaufschläge» Nahrungsmittel-, Energie- und Ressourcenkonflikte in der Schweiz des Ersten Weltkrieges*, pages 259–284, 2016.

Sandro Sperandei. Lessons in biostatistics Understanding logistic regression analysis. *Lessons in Biostatistics*, pages 12–18, 2014.

Staatsarchiv Kanton Bern. Mortalitätsstatistik infolge von Tuberkulose in den Gemeinden des Kantons Bern, 1910. URL `https://www.query.sta.be.ch/detail.aspx?ID=156549`.

Eidgenössisches Statistisches Bureau. Bevölkerungsbewegung, 1921.

Kaspar Staub, Peter Jüni, Martin Urner, Katarina Matthes, Corina Leuch, Gina Gemperle, Nicole Bender, Sara Irina Fabrikant, Milo Puhan, Frank Rühli, Oliver Grübner, and Joël Floris. Public Health Interventions, Epidemic Growth, and Regional Variation of the 1918 Influenza Pandemic Outbreak in a Swiss Canton and Its Greater Regions. *Annals of Internal Medicine*.

Jill C. Stoltzfus. Logistic regression: A brief primer. *Academic Emergency Medicine*, 18(10):1099–1104, 2011. ISSN 10696563. doi: 10.1111/j.1553-2712.2011.01185.x.

Andreas Tscherrig. Krankenbesuche Verboten! Die Spanische Grippe 1918/1919 Und die kantonalen Sanitätsbehörden in Basel-Landschaft und Basel-Stadt, 2016.

Andreas Tscherrig. "Die Totenglocken wollten nicht mehr verstummen", Die Tragödie der Grippepandemie von 1918/1919 in Nidwalden. In D Krämer and K Schleifer, editors, *Nidwalden im Ersten Weltkrieg*, pages 116–135. 2018.

Heather L. Van Epps. Influenza: Exposing the true killer. *Journal of Experimental Medicine*, 203(4):803, 2006. ISSN 15409538. doi: 10.1084/jem.2034fta.

World Health Organization. World Health Organization Best Practices for the Naming of New Human Infectious Diseases, 2015. URL `http://www.who.int/classifications/icd/revision/Content{_}Model{_}Reference{_}Guide.January{_}2011.pdf?ua=1,`.

World Health Organization. Ten threats to global health in 2019, 2019. URL `https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019`.

Wan Yang, Elisaveta Petkova, and Jeffrey Shaman. The 1918 influenza pandemic in New York City: Age-specific timing, mortality, and transmission dynamics. *Influenza and other Respiratory Viruses*, 8(2):177–188, 2014. ISSN 17502640. doi: 10.1111/irv.12217.

Kathrin Zürcher, Marcel Zwahlen, Marie Ballif, Hans L. Rieder, Matthias Egger, and Lukas Fenner. Influenza pandemics and tuberculosis mortality in 1889 and 1918: Analysis of historical data from Switzerland. *PLoS ONE*, 11(10): 1–11, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0162575.

A Bivariate choropleth maps

As an aid for the the interpretation of the model outputs, bivariate choropleth maps were created to show the correlation in space between influenza incidence and each variable.

A.1 First wave: July/August 1918

A.1.1 TB mortality vs. influenza incidence

The choropleth map in figure A.1 shows the correlation between tuberculosis mortality and influenza incidencce during the first wave of the pandemic and adds spatial information to it. The map is a bit difficult to understand at first, which is why a few reading examples are given here. The values that are probably most interesting for the correlation of these two variables are the dark grey to blue values on the main diagonal of the legend. Those values indicate a positive correlation between tuberculosis mortality and influenza incidence, meaning a municipality had either a low influenza incidence and low tuberculosis mortality or high values in both of the indicators. There are no clear tendencies, but for example in the Bernese Oberland, there were quite a few municipalities that had low-low values. Furthermore, in the Jura region and the Laufental, some dark blue municipalities can be found which indicates high-high values. Green values mean a high tuberculosis mortality but a low influenza incidence. This is the case for some municipalities in the Mittelland between Berne and Thun as well as in the northern part of the Jura. Pink values, which make up the majority of the municipalities, are the ones that showed a high influenza incidence but a low TB mortality. Dark grey municipalities did not report any influenza data and therefore no correlation could be calculated.

A.1.2 Population density vs. influenza incidence

Figure A.2 shows the correlation between population density and influenza incidence. This is a bit inconsistent with the models, where urbanity and not population density was modelled. However, a similar map with the urbanity variable would not have produced meaningful results which is why the pop-

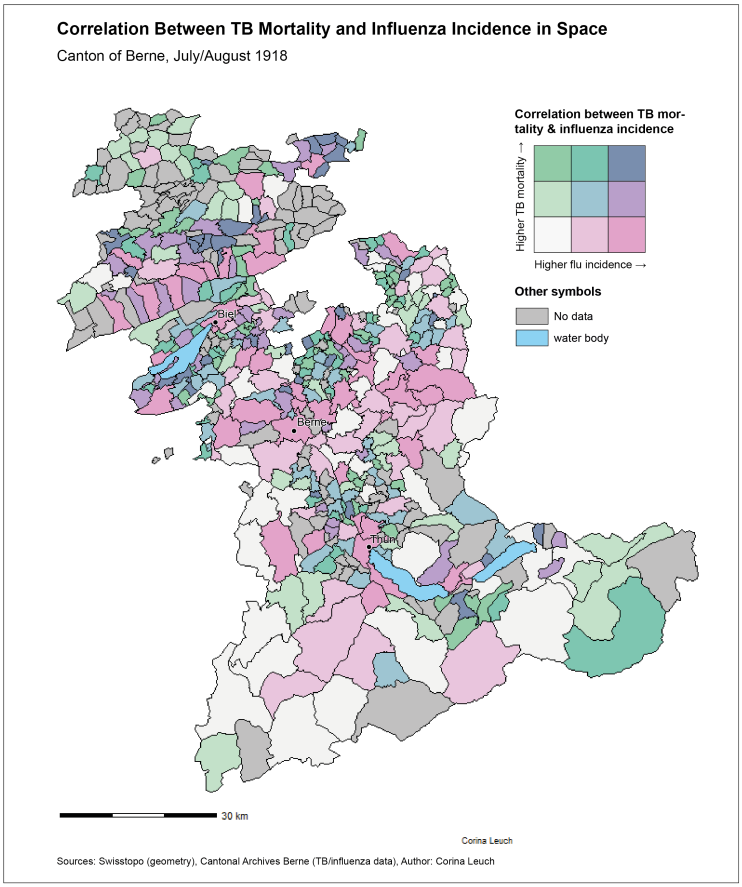

Figure A.1: Bivariate choropleth map of the variables TB mortality and influenza incidence during the first wave which gives insight into spatial dynamics of the correlation. Grey-blue values indicate a positive correlation. Green values indicate a low influenza incidence but a high TB mortality, and pink values indicate a high influenza incidence and low TB mortality.

ulation density was modelled instead. The Jura region is somewhat divided: the southern part has areas that are both densely populated and have a high incidence. However, they are next to municipalities with a low population density, therefore there is no clear pattern. The northern region also shows some isolated areas with a medium-high population density but a low influenza incidence. The Laufental region was hit hard by this first wave, no matter the population density of its municipalities. In the Swiss Plateau, there is no clear pattern visible. What stands out at first glance are the cities of Biel and Berne and their surrounding areas that both had a high population density and a high influenza incidence. This also holds true for the city of Thun. Apart from that,

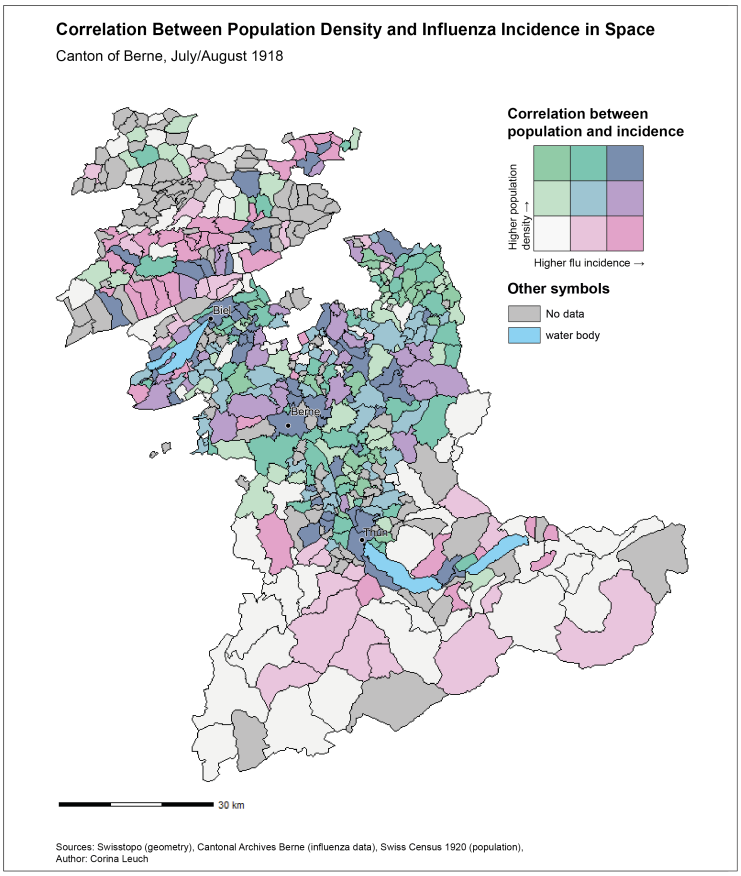

Figure A.2: Bivariate choropleth map that shows the correlation between population density and influenza incidence during the first wave. It gives insight into spatial dynamics of the correlation. Grey-blue values indicate a positive correlation. Green values indicate a low influenza incidence but a high population density, and pink values indicate a high influenza incidence and low population density.

there are some scattered municipalities that were both densely populated and had a high incidence. In the Oberaargau region, there seems to be an entire cluster of municipalities that had a medium-high population density but were not particularly affected by the first wave of the 1918 influenza pandemic. In the Bernese Oberland, there is a difference visible between the more low-lying municipalities around the lakes and the more mountainous regions: the areas around the lakes were both densely populated and had a high incidence. Finally, in the southern and eastern part of the Oberland region, there are also many municipalities with both a low incidence and a low mortality.

A.1.3 Access to railway network vs. influenza incidence

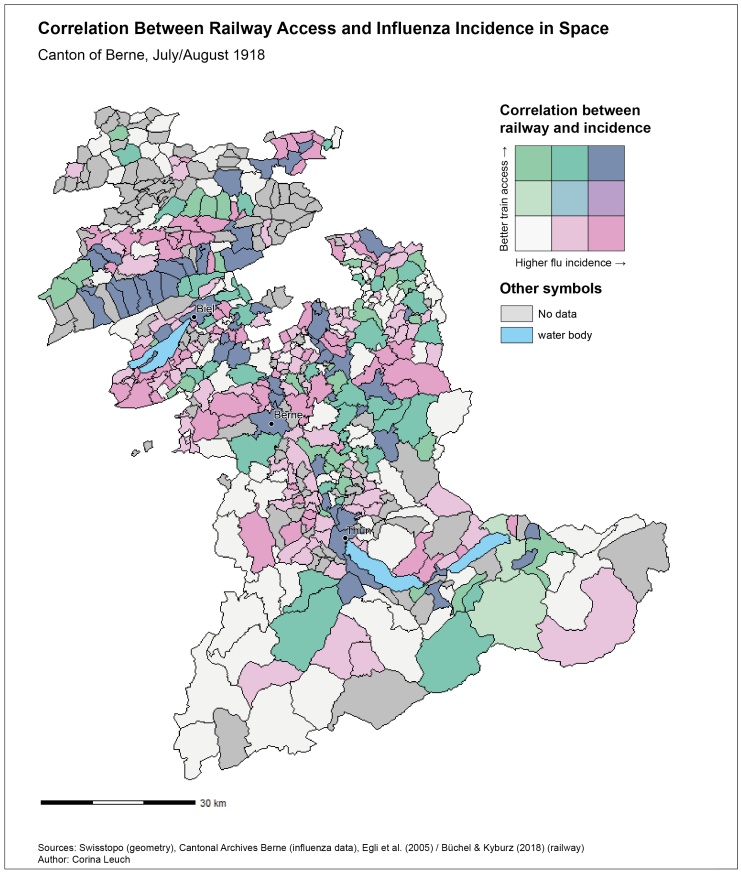

Figure A.3: Bivariate choropleth map showing the correlation between railway access and influenza incidence during the first wave which gives insight into spatial dynamics of the correlation. Grey-blue values indicate a positive correlation. Green values indicate a low influenza incidence but good access to the railway network, and pink values indicate a high influenza incidence and no access to the railway network.

The correlation between the access to the railway system and influenza incidence during the first wave in figure A.3 reveals that the railway access might have played an important part in some areas of the canton. Particularly the south of the Jura region had both a high incidence and good access to the railway system. In the northern neighbouring valley, there was no railway access, but incidences were still high. Again north of that, there were municipalities that did have railway access but low influenza incidence. Furthermore, in the Seeland region,

the city of Biel stands out with a high incidence and high railway access but also the region south of Lake Biel. This region showed medium-high incidences despite having no access to the railway system. In the Oberaargau region, there are municipalities with high influenza incidence but no railway access and areas with good raiway access but fewer cases, therefore showing negative correlation between railway access and influenza incidence in this particular area. Finally in the Oberland, the city of Thun and its surrounding areas strike out with good railway access and high incidences. The eastern part of the Oberland was also less affected for some reason, even though it had access to the railway network. The southern part of the Bernese Oberland towards the Valais finally, had no access to the railway network but was also not particularly affected.

A.1.4 Precipitation vs. influenza incidence

The correlation between precipitation and influenza incidence during the first wave (see A.4) shows clearer tendencies than the two previous maps. The Jura region had little to medium precipitation and the incidences show the well-known pattern, that the southern part and the Laufental region were more affected than the northern part. Around the Seeland region, the pattern was farely similar with a medium-high flu incidence and little precipitation. In the eastern part of the canton, in the Oberaargau region, there is a cluster with both high incidences and high precipitation, whereas in the northern part of the Oberaargau region, the municipalities received less precipitation and sho<wed lower incidences. Towards the Alpine region in the south, the precipitation gradually increases. Generally speaking, the incidences in that region were low, with the exception of Thun and its surrounding municipalities as well as some more municipalities around the two lakes, while the precipitation was high.

A.2 Second wave: October 1918 – January 1919

A.2.1 TB mortality vs. influenza incidence

Figure A.5 shows a bivariate choropleth map which shows the correlation between incidence and tuberculosis mortality during the second wave of the 1918 influenza pandemic. It adds to what previous findings in this thesis have already shown: Firstly, much of the Jura region was only mildly affected by the second wave but there are areas that had a high mortality from tuberculosis. Again, the southernmost valley of the Jura region was affected most while having a lower TB mortality than the rest of the region. Secondly, the Laufental also seems

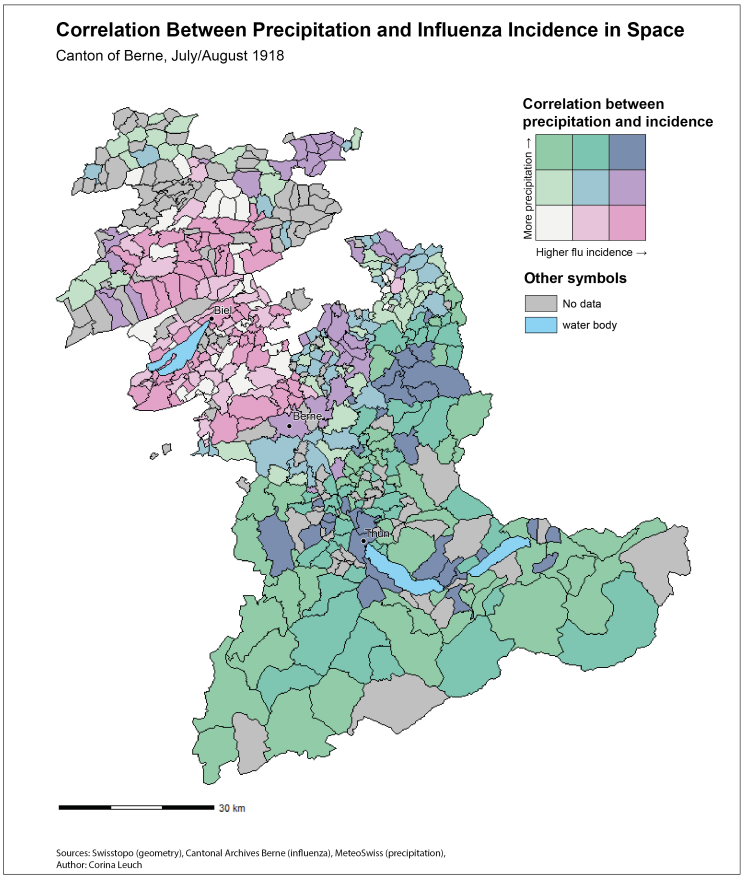

Figure A.4: Bivariate choropleth map showing the correlation of precipitation and influenza incidence during the first wave which gives insights into the spatial dynamics of the correlation. Grey-blue values indicate a positive correlation. Green values indicate a low influenza incidence but a high precipitation, and pink values indicate a high influenza incidence and low precipitation.

to have been more affected. In the Seeland region, there is a cluster south of Lake Biel, where both the incidence and tuberculosis mortality were high. The Mittelland region shows a north-south gradient when it comes to tuberculosis mortality, while incidences were high in most municipalities of the regions. In the Oberaargau region, there almost seems to be a negative correlation between tuberculosis and influenza: the eastern part seems to have suffered more from influenza while the eastern part had medium-high TB mortality. The Oberland region does not show a clear picture. There is an area of both high incidence and high TB mortality inbetween Lake Thun and Lake Brienz. Furthermore, the eastern part also shows medium-high tuberculosis mortalty, with low-medium

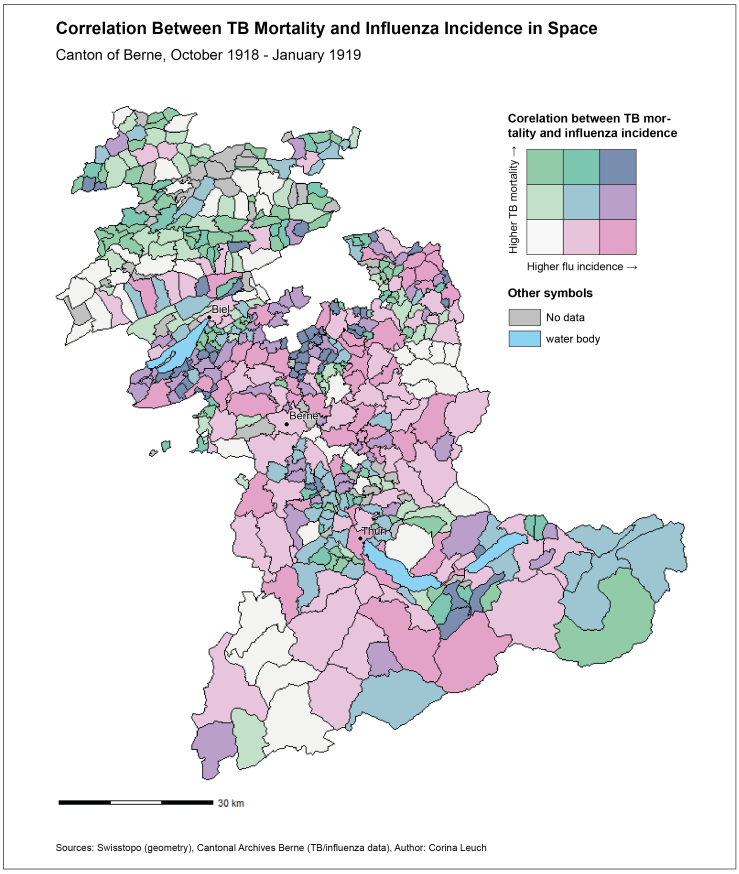

Figure A.5: Bivariate choropleth map showing the correlation of TB mortality and influenza incidence during the second wave and how the correlation is distributed in space. Grey-blue values signify a positive correlation. Green values mean a negative correlation, with high TB mortality and low influenza incidence. Finally, pink values also mean a negative correlation where the influenza incidence is high and the TB mortality is low.

incidences. Towwards the southern tip of the canton of Berne, many municipalities were not particularly affected by tuberculosis but did have varying degrees of influenza incidence.

A.2.2 Population density vs. influenza incidence

Figure A.6 shows the correlation between population density and influenza incidence during the second wave of the 1918 influenza pandemic. Again, the use of the population density rather than the urbanity is inconsistent with the models but yields more meaningful results. In the Jura region, many municipalities had

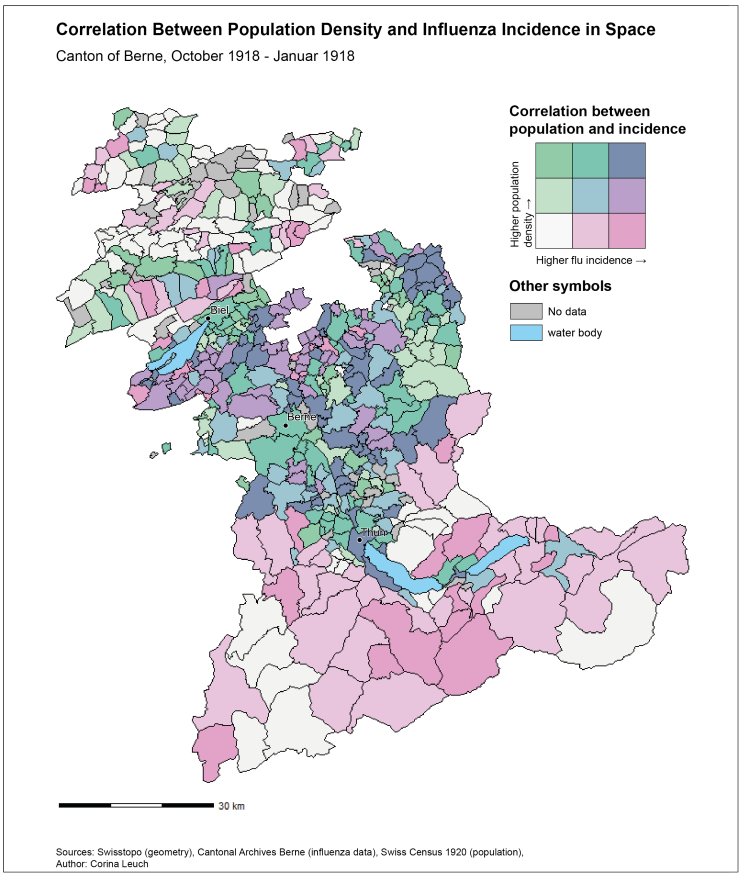

Figure A.6: Bivariate choropleth map showing the correlation between population density and influenza incidence and how the correlation is distributed in space. Grey-blue values signify a positive correlation. Green values mean a negative correlation, with high population density and low influenza incidence. Finally, pink values also mean a negative correlation where the influenza incidence is high and the population density is low.

both a low population density and a low incidence. Again, the southern part of the Jura region stands out with its higher population density and incidence compared to the rest of the Jura region. Particularly affected by the second wave of the 1918 influenza pandemic was the region on the southern border of Lake Biel, despite the fact that the majority of the municipalites only had a medium population density. Contrary to that, the northern part had a high population density but only low-medium incidences of influenza. Interestingly, the cities of Biel and Berne, though heavily populated, were not in the highest category for incidence. Furthermore, in the Oberaargau region there is no clear tendency in

incidence, even though it was generally more densely populated. In the Bernese Oberland, many municipalities had medium-high incidences, even though they were not very densely populated. Generally, there is a tendency in the Oberland, that the municipalities that were closer to the lakes were more heavily populated than the ones far away and many of those more densely populated municipalities were heavily affected by the second wave.

A.2.3 Railway access vs. influenza incidence

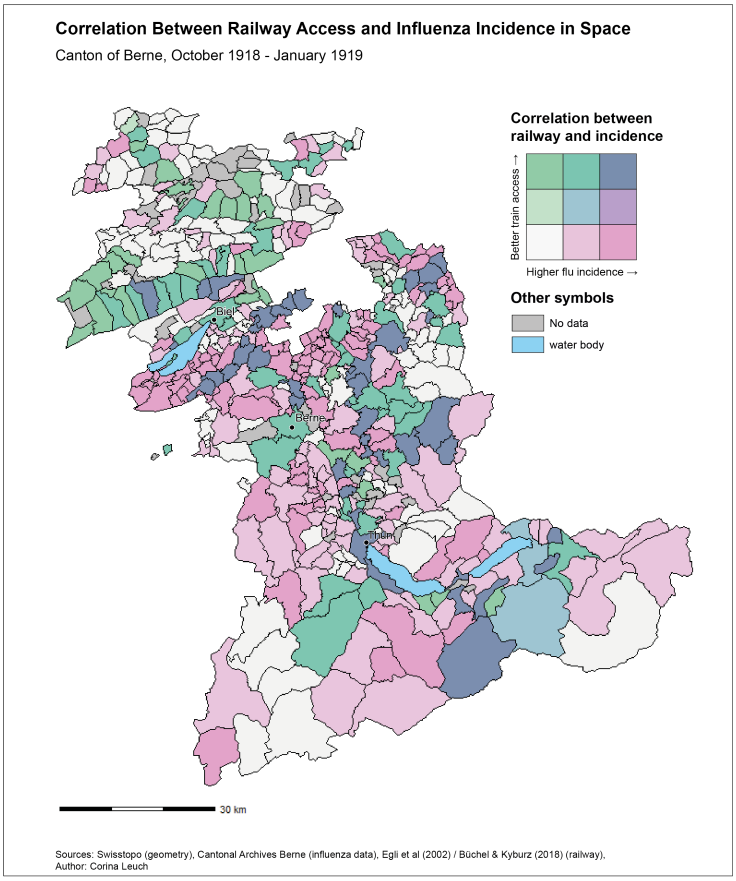

Figure A.7: Bivariate choropleth map showing the correlation between railway access and influenza incidence during the second wave and how the correlation is distributed in space. Grey-blue values signify a positive correlation. Green values mean a negative correlation, with high railway access and low influenza incidence. Finally, pink values also mean a negative correlation where the influenza incidence is high and the railway access is low.

Figure A.7 shows the correlation between railway access and influenza incidence during the second wave. The pattern in the bivariate choropleth map is less clear than during the first wave. Along the southern Jura railway line, there is no clear pattern, there are a few municipalities with a medium or high incidence. In the rest of the Jura region, many municipalities did not have railway access but were not particularly affected by the second wave. In the Seeland region, there is a cluster with municipalities that all had a high incidence but no railway access on the southern bank of Lake Biel, while the northern part was less affected despite the fact that the access to the railway system was better. One area where railway access (or rather having no railway access) may have played a role is in the southern Oberaargau region in the east, where there is a cluster with no railway access and low incidences. In the northern part of the Oberaargau region, there are some areas that had both good railway access and high incidences. In the Bernese Oberland, there is no clear correlation between railway access and influenza incidence. Many municipalities were affected no matter their railway access.

A.2.4 Precipitation vs. influenza incidence

The comparison between the two variables precipitation and incidence (A.8) looks clearer than the other bivariate choropleth maps at first sight. This has to do with the nature of the precipitation variable, that has a smaller degree of local variation than the rest of the variables. The Jura region mostly received high precipitation but low influenza incidence. The Laufental region was also more affected by influenza while receiving less precipitation than the neighbouring Jura region. The Seeland region as a whole received a medium amount of precipitation. Again, the difference in incidence between the north and the south is visible. In the Mittelland region, there are differing degrees of influenza incidences with little precipitation. The Oberaargau region in the east received a medium amount of precipitation with differing degrees of influenza incidence in the medium-high categories in the northern parts and a low incidence in the southern part. The Alpine region received a high amount of precipitation. There is a cluster of high-high values south of Lake Brienz and Lake Thun. The rest of the Oberland showed low-medium incidences.

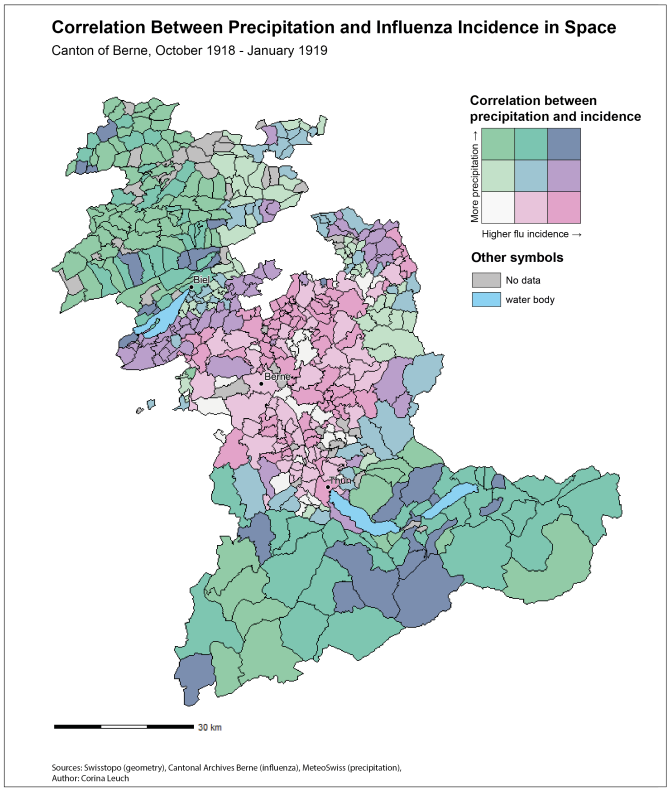

Figure A.8: Bivariate choropleth map showing the correlation between precipitation and influenza incidence during the second wave and giving insight into the spatial dynamics of the correlation. Grey-blue values indicate a positive correlation. Green values indicate a low influenza incidence but high precipitation, and pink values indicate a high influenza incidence and low precipitation.

B R Code used for the analysis

Note: the data source files used in this code are not provided with this thesis. On its own, the code is not executable. The code itself is commented throughout to facilitate its understanding. The code is also available in the following GitHub repository: (https://github.com/coleuc/SpanishFlu).

Descriptive visualisations

Histogram

```
# make histogram: this code prepares the data and afterwards
    creates a nice looking histogram
# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

# import packages using function
pkgTest("tidyverse")
pkgTest("here")
pkgTest("sf")
pkgTest("lubridate")


```

```r
# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

# set up coordinate system projection strings
WGS84 <- "+init=epsg:4326"
LV03 <- "+init=epsg:21781"

# read the data -> this part is hard-coded. Change if code needs to be executed
gemeinden <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_Gemeinden.csv'), sep= ';', encoding  = "UTF-8") %>% rename(GEM_ID= X.U.FEFF.GEM_ID) # have to rename the id number to match it
grippe <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_Faelle.csv'), sep= ';', encoding = "UTF-8")

# get only the influenza cases and sum them up by district/year
grippe <- grippe %>% filter(CatDisease == 1) %>% select(District, NumbCasesAdjust2, GEM_ID, Year, Month, Day) # %>% group_by(GEM_ID) %>% summarise(totalCases = sum(NumbCasesAdjust2)) %>% ungroup()

# joining the gemeinden data to the disease data
grippe_gemeinden <- left_join(grippe, gemeinden, by = 'GEM_ID')

# check for N/As
na <- grippe_gemeinden %>% filter(is.na(N)) %>% as_tibble() # data contains some entries that could not be matched with a district. These data are omitted for now
# nas seem to be things that are aggregated by districts --> good that they are na, so that they aren't accidentally included in the analysis

# grippe_gemeinden <- drop_na(grippe_gemeinden, c(N,E)) %>% select(Gemeinde_Name, totalCases, Wohnb, E, N, GEM_ID, BfS_GEM_ID)
# if not summarized by district, use following code
grippe_gemeinden <- drop_na(grippe_gemeinden, c(N,E)) %>% select(Gemeinde_Name, NumbCasesAdjust2, Wohnb, E, N, GEM_ID, BfS_GEM_ID, Year, Month, Day) %>%
  unite(Date, c(Year, Month, Day), sep = "/")

# calculate morbitity
grippe_gemeinden <- grippe_gemeinden %>% group_by(GEM_ID) %>% mutate(inccidence = sum(NumbCasesAdjust2)/Wohnb*100000) %>%
  select(GEM_ID, inccidence, Gemeinde_Name, Wohnb) %>% unique()# calculate cases per 1000 inhabitants

```

```r
# make a histogram --> one value for each municipality
ggplot(data = grippe_gemeinden, aes(grippe_gemeinden$inccidence)) +
  geom_histogram(fill="#7a0177", alpha=0.8, bins = 60) +
  expand_limits(x = 10000) + # make sure axis labels arent being cut off
  ggplot2::labs(title = "Burden of Disease in Bernese Municipalities",
                subtitle = "Histogram of the influenza incidence Jul. 1918 - Dec. 1918",
                y = "Number of municipalities",
                x = "Incidence (cases/100'000 inhabitants)",
                caption = "Source: Cantonal Arcives Berne,
                           Author: Corina Leuch") +
  theme_minimal() +
  theme(axis.title.x=element_text(vjust=-0.5, size = 12, face="bold"),
        axis.title.y = element_text(vjust=2, size=12,face= "bold"),
        axis.text = element_text(size = 8),
        plot.title = element_text(size=16, face = "bold"),
        plot.caption = element_text(size =8),
        plot.subtitle = element_text(size = 12)) +
  theme(legend.position = "none", plot.margin=unit(c(0.5,0.5,0.5,0.5),"cm"))
```

Boxplots

```r
# boxplots ----
# write csv

grippe_levels <- read.csv(grippe_levels, path = paste0(outputFolder, "/Grippe_Final.csv"))
first_levels <- read.csv(first_final, path = paste0(outputFolder, "/First_Wave_Final.csv"))
second_levels <- read.csv(second_final, path = paste0(outputFolder, "/Second_Wave_Final.csv"))

# the incidence data seems to have some outliers. Create boxplot
colnames(first_levels)[3] <- "Inz"
first_levels$wave <- "First wave"


colnames(second_levels)[3] <- "Inz"
second_levels$wave <- "Second wave"

grippe_levels$wave <- "Entire period"

dat <- bind_rows(grippe_levels, first_levels,second_levels)
```

```
dat$wave <- factor(dat$wave, levels = c("Entire period", 'First
    wave', 'Second wave'))

ggplot(dat, aes(x=wave, y=Inz), group_by(wave)) +
  geom_boxplot(fill="#7a0177", alpha=0.8) +
  theme_minimal() +
  scale_x_discrete(name = "Wave") +
  ggplot2::labs(title = "Distribution of incidences",
                subtitle = "Incidence for the entire study period
    and the two waves",
                y = "Incidence (cases/100'000 inhabitants)",
                x = "Wave",
                caption = "Source: Cantonal Archives Berne,
                            Author: Corina Leuch") +
  theme_minimal() +
  theme(axis.title.x=element_text(vjust=-0.5, size = 10, face="bold
    "),
        axis.title.y = element_text(vjust=2, size=10,face= "bold"),
        axis.text = element_text(size = 10),
        plot.title = element_text(size=12, face = "bold"),
        plot.caption = element_text(size =8),
        plot.subtitle = element_text(size = 10))
```

Lorenz curve

```
# this script draws a lorenz curve to show the inequality in
    incidence rates

# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

```

```r
# import packages using function
pkgTest("tidyverse")
pkgTest("here")
pkgTest("sf")
pkgTest("lubridate")
pkgTest("zoo")
pkgTest("scales")
pkgTest("gglorenz")

# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

# set up coordinate system projection strings
WGS84 <- "+init=epsg:4326"
LV03 <- "+init=epsg:21781"

grippe <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_Faelle.csv'), sep= ';', encoding = "UTF-8")%>%
  filter(CatDisease == 1) %>% select(NumbCasesAdjust2, GEM_ID) %>% drop_na() %>% group_by(GEM_ID) %>%
  mutate(Cases = sum(NumbCasesAdjust2)) %>% select(GEM_ID, Cases) %>% unique()

gemeinden <- read.csv(file= paste0(dataFolder, '/SpanischeGrippe_Gemeinden.csv'), sep= ';', encoding = "UTF-8") %>%
  rename(GEM_ID=X.U.FEFF.GEM_ID)

grippe <- left_join(grippe, gemeinden, by= "GEM_ID")
grippe <- grippe %>%select(Wohnb, Cases) %>% rename(Population=Wohnb) %>% ungroup() %>% select(Cases, Population) %>% mutate(Inz = Cases/Population*100000)

ggplot(grippe) +
  stat_lorenz(aes(Inz),color="#7a0177", size=1)+
  coord_fixed()+
  geom_abline(linetype = "dashed") +
  theme_minimal() +
  theme(axis.title.x=element_text(vjust=-0.5, size = 10, face="bold"),
        axis.title.y = element_text(vjust=2, size=10,face= "bold"),
        axis.text = element_text(size = 10),
        plot.title = element_text(size=12, face = "bold"),
        plot.subtitle = element_text(size = 10),
        plot.caption = element_text(size = 8))+
  hrbrthemes::scale_x_percent() +
  hrbrthemes::scale_y_percent() +
```

```
  labs(x = "Cumulative percentage of population",
       y = "Cumulative percentage of incidence",
       title = "Distribution of Cases Among the Population on a
    Municipality Level",
       subtitle = "Canton of Berne, July 1918 - December 1919",
       caption = "Source: Cantonal Arcives Berne,
                              Author: Corina Leuch") +
  annotate_ineq(grippe$Cases) +
  theme(legend.position = "none", plot.margin=unit(c
    (0.5,0.5,0.5,0.5),"cm"))
```

"Flatten the curve"

```
# this script creates a line graph with the 7-day rolling average
    on the y axis and the total cases on the x axis. Therefore, it
    creates a graph that shows the 'flattening of the curve'
# in the Swiss canton of Berne.

# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package
pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

# import packages using function
pkgTest("tidyverse")
pkgTest("here")
pkgTest("sf")
pkgTest("lubridate")
pkgTest("zoo")
pkgTest("scales")



# set data folder and R folder
```

```
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

# set up coordinate system projection strings
WGS84 <- "+init=epsg:4326"
LV03 <- "+init=epsg:21781"

grippe <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_
    Faelle.csv'), sep= ';', encoding = "UTF-8")

# get only the influenza cases and sum them up by district/year
grippe <- grippe %>% filter(CatDisease == 1) %>% select(District,
    NumbCasesAdjust2, GEM_ID, Year, Month, Day)
grippe$Month <- sprintf("%02d", as.numeric(grippe$Month))
grippe$Day <- sprintf("%02d", as.numeric(grippe$Day))

# if not summarized by district, use following code
grippe <- drop_na(grippe, NumbCasesAdjust2) %>% select(
    NumbCasesAdjust2, Year, Month, Day) %>%
  unite(Date, c(Day, Month, Year), sep = ".")

grippe <- grippe %>% mutate(Date = as.Date(grippe$Date, "%d.%m.%Y")
    ) %>% select(NumbCasesAdjust2, Date) %>% group_by(Date) %>%
  mutate(Cases = sum(NumbCasesAdjust2)) %>% select(Date, Cases) %>%
     unique() %>% ungroup()#  %>%
# mutate(Date = format(Date, "%d.%m.%Y"))

grippe <- grippe %>% arrange(Date) %>%
  dplyr::mutate(cases07 = zoo::rollmean(Cases, k =7, fill = NA))
    %>%
  dplyr::mutate(totalDay = cumsum(Cases))

# need to drop N/A s or it screws up ggplot later
grippe_narm <- grippe %>% drop_na()

# make the actual graph
ggplot(data=grippe_narm) + geom_line(aes(x=totalDay, y=cases07),
    size=1, color="#7a0177") +
  theme_bw() +
  labs(title = "'Flatten the Curve' During the 1918 Influenza
    Pandemic",
       subtitle = "New cases (7-day avg.) vs. total cases per
    Bernese municipality 1918/19",
       y = "New cases",
       x = "Total cases",
       caption = "Source: Cantonal Arcives Berne,
                             Author: Corina Leuch")  +
```

```
  scale_x_continuous(labels = comma) +
  theme_minimal() +
  theme(axis.title.x=element_text(vjust=-0.5, size = 10, face="bold
    "),
        axis.title.y = element_text(vjust=2, size=10,face= "bold"),
        axis.text = element_text(size = 10),
        plot.title = element_text(size=12, face = "bold"),
        plot.caption = element_text(size =8),
        plot.subtitle = element_text(size = 10))
```

Incidence maps and GIF

```
# this script works as a baseline for various maps showing the
    incidence per month. The maps themselves were later created in
# QGIS and illustrator using the shapefiles generated in this code.

# set up part
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

# import packages using function
pkgTest("tidyverse")
pkgTest("here")
pkgTest("sf")
pkgTest("tmap")
pkgTest("rgdal")


# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

```

```
#projection strings
WGS84 <- "+init=epsg:4326"
LV03 <- "+init=epsg:21781"

# prepare the data
dat <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_Faelle.csv'), sep= ';', encoding = "UTF-8")
gemeinden <- read.csv(file = paste0(dataFolder, '/SpanischeGrippe_Gemeinden.csv'), sep= ';', encoding  = "UTF-8") %>% rename(GEM_ID= X.U.FEFF.GEM_ID) # have to rename the id number to match it

# clean up dataset ----
# join together disease data and district data in order to get the number of inhabitants. Furthermore: some rows have no Gemeinde associated with it (things like
# "Im Amt", "im ganzen Bezirk"). These can not be used for this analysis and are removed
dat <- left_join(dat, gemeinden, by = 'GEM_ID') %>% drop_na(GEM_ID)
dat$Month <- sprintf("%02d", as.numeric(dat$Month)) # convert 1 to 01 in month --> easier sorting


# group by month, calculate incidence per district and month
dat_grouped <- dat %>% filter(CatDisease == 1) %>% select(Gemeinde_Name, NumbCasesAdjust2, Wohnb, GEM_ID, Year, Month, BfS_GEM_ID, E, N) %>% group_by(GEM_ID, Gemeinde_Name, Year, Month, Wohnb, BfS_GEM_ID, E, N) %>%
  summarise(montlyCases = sum(NumbCasesAdjust2)) %>% unite(Month, c(Year, Month), sep = "_") %>% mutate(Inz = montlyCases/Wohnb*100000) %>% ungroup()

write_csv(dat_grouped, paste0(RFolder, "Monthly_Cases.csv"))

# check for N/A
na <- dat_grouped %>% filter(is.na(GEM_ID)) %>% as_tibble()

geometry <- st_read(dsn = paste0(dataFolder, "/Bern_Punkt_id.shp")) %>% `st_crs<-`(LV03)

dat_grouped_sf <- left_join(geometry, dat_grouped, "GEM_ID") %>% na.omit(dat_grouped_sf$Month)
# st_write(dat_grouped_sf, paste0(outputFolder,'/Monthly_Incidences.shp'), driver = 'ESRI Shapefile') # issue here: doesn't write date; seems to be a known problem
months <- read.csv(file=paste0(dataFolder, "/months.csv"), sep = ";", encoding="UTF-8") %>% rename(Month= X.U.FEFF.Month)

```

```
66 dat_grouped_sf <- left_join(dat_grouped_sf, months, by="Month") #
       could be used to make title pretty, but doesn't work for now
67
68 month_list <- dat_grouped %>% distinct(Month) %>% pull()
69
70 # this function creates a shapefile with the incidence per
       municipality for each month
71 createMonthlyData <- function(month, data=dat_grouped_sf){
72   monthly <- data %>% filter(Month==month)
73   st_write(monthly,paste0(outputFolder, "/07_monthly_incidence",
       month, ".shp"))
74 }
75
76 lapply(month_list, createMonthlyData)
77 # use same breaks as in gif
78
79 lapply(month_list, createStaticMaps, field="Inz", name="01_Monthly_
       Incidence_")
80
81
82 #### if all the maps are prepared and exported to pngs, this code
       creates a nice looking gif
83 files <- list.files(path = enFOlder, pattern = "*.png", full.names=
       TRUE)
84 all_im <- image_read(files)
85 scaled <- image_scale(all_im,"1000!")
86
87 animation <- image_animate(all_im, fps = 1, dispose = "previous")
88 image_write(animation, paste0(outputFolder, "/01_Cases_Month_EN.gif
       "))
89
90
91 # the code beow creates a shapefile for each wave which can then be
        turned into a map using QGIS/Ilu
92
93 # select data of first wave and bring to form we want
94 first_wave <- dat %>%
95   filter(CatDisease == 1) %>%
96   select(Gemeinde_Name, NumbCasesAdjust2, Year, Month, Wohnb,GEM_ID
       ) %>%
97   group_by(GEM_ID, Gemeinde_Name, Wohnb, Year, Month) %>%
98   unite(Month, c(Year, Month), sep = "_") %>%
99   filter(Month == "1918_07"| Month == "1918_08") %>% ungroup() %>%
100  group_by(GEM_ID, Wohnb, Gemeinde_Name) %>%
101  summarise(overallCases = sum(NumbCasesAdjust2)) %>%
102  mutate(Inz = overallCases/Wohnb*100000) %>% ungroup()
103
104 first_wave_sf <- left_join(geometry, first_wave, by="GEM_ID")
```

```
st_write(first_wave_sf, paste0(outputFolder,'/Incidences_1st_wave.shp'), driver = 'ESRI Shapefile') # issue here: doesn't write date; seems to be a known problem

# select data for second wave and bring to form we want
second_wave <- dat %>%
  filter(CatDisease == 1) %>%
  select(Gemeinde_Name, NumbCasesAdjust2, Year, Month, Wohnb,GEM_ID) %>%
  group_by(GEM_ID, Gemeinde_Name, Wohnb, Year, Month) %>%
  unite(Month, c(Year, Month), sep = "_") %>%
  filter(Month == "1918_10"| Month == "1918_11"| Month == "1918_12" | Month == "1919_01") %>% ungroup() %>%
  group_by(GEM_ID, Wohnb, Gemeinde_Name) %>%
  summarise(overallCases = sum(NumbCasesAdjust2)) %>%
  mutate(Inz = overallCases/Wohnb*100000) %>% ungroup()

second_wave_sf <- left_join(geometry, second_wave, by="GEM_ID")
st_write(second_wave_sf, paste0(outputFolder,'/Incidences_2nd_wave.shp'), driver = 'ESRI Shapefile')
```

Bivariate choropleth maps

Parts of this script were copied from Timo Grossenbacher's blog (Link).

```
# this codes draws a nice bivariate choropleth map that can later be exported as a picture or pdf (for further editing).
# This is only an example showing how to create a bivariate choropleth map using the variables influenza incidence and
# Tb mortality for the first wave. Other maps can be created using other variables

# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
```

```
  }
}

# load packages
pkgTest("tidyverse")
pkgTest("sf")
pkgTest('cowplot')
pkgTest('ggspatial')

# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output/")

# read datasets
dat <- read.csv(file = paste0(dataFolder, '/First_Wave_Final.csv'))
geometry <- st_read(dsn = paste0(dataFolder, '/Bern_Punkt_id.shp'))
geometry$area <- st_area(geometry)
lakes <- st_read(dsn = paste0(dataFolder, '/SEEN_1990.shp'), ) %>%
    filter(ID_See == 9040| ID_See == 9060 | ID_See == 9070) %>%
  st_set_crs(21781)

dat <- right_join(geometry, dat, "GEM_ID") %>% mutate(pop_dens =
    Wohnb.x*1000000/area)

# calculate terziles for the flu
quantiles_flu <- dat %>%
  pull(Inz_first) %>%
  quantile(probs = seq(0, 1, length.out = 4))

quantiles_tb <- dat %>%
  pull(TBratio) %>%
  quantile(probs = seq(0, 1, length.out = 4), na.rm = T)

bivariate_color_scale <- tibble(
  "3 - 3" = "#7B8EAF", # high influ, high other
  "2 - 3" = "#7FC6B1",
  "1 - 3" = "#8AE1AE", # low influ, high other
  "3 - 2" = "#BB9FCE",
  "2 - 2" = "#9EC5D3", # medium influ, medium other
  "1 - 2" = "#C2F0CE",
  "3 - 1" = "#E6A3D0", # high influ, low other
  "2 - 1" = "#EAC5DD",
  "1 - 1" = "#F3F3F3" # low influ, low other
) %>%
  gather("group", "fill")

# assign class according to quintile
```

```
flu_tb <- dat %>%
  mutate(
    quantiles_flu = cut(
      Inz_first,
      breaks = quantiles_flu,
      include.lowest = TRUE
    ),
    quantiles_tb = cut(
      TBratio,
      breaks = quantiles_tb,
      include.lowest = TRUE
    ),
    # by pasting the factors together as numbers we match the
    groups defined
    # in the tibble bivariate_color_scale
    group = paste(
      as.numeric(quantiles_flu), "-",
      as.numeric(quantiles_tb)
    )) %>%
  # we now join the actual hex values per "group"
  # so each municipality knows its hex value based on the his gini
    and avg
  # income value
  left_join(bivariate_color_scale, by = "group")

# define map theme
theme_map <- function(...) {
  theme_minimal() +
    theme(
      text = element_text(family = "Ubuntu Regular", color = "
    #22211d"),
      # remove all axes
      axis.line = element_blank(),
      axis.text.x = element_blank(),
      axis.text.y = element_blank(),
      axis.ticks = element_blank(),
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      # remove grid
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      plot.margin = unit(c(.5, .5, .2, .5), "cm"),
      panel.border = element_blank(),
      panel.spacing = unit(c(-.1, 0.2, .2, 0.2), "cm"),
      legend.background = element_blank(),
      plot.title = element_text(size = 12, color = "black"),
      plot.subtitle = element_text(size = 8, color = "black",
                                   margin = margin(b = -0.1,
```

```
                                                    t = -0.1,
                                                    l = 0,
                                                    unit = "cm"),
                                    debug = F),
      plot.caption = element_text(size = 7,
                                  hjust = 0,
                                  margin = margin(t = 0.2,
                                                  b = 0,
                                                  unit = "cm"),
                                  color = "black"),
      ...
    )
}

map <- ggplot(
  # use the same dataset as before
  data = flu_tb
) +
  # color municipalities according to their flu / tb combination
  geom_sf(data = geometry,
    aes(fill = "#c0c0c0"
    ),
    # use thin white stroke for municipalities
    color = "black",
    size = 0.1
  )+
  # color municipalities according to their flu / tb combination
  geom_sf(
    aes(
      fill = fill
    ),
    # use thin white stroke for municipalities
    color = "black",
    size = 0.1
  ) +
  geom_sf(data = lakes,
          aes(fill = "#8CD1F2"
          ),
          # use thin white stroke for municipalities
          color = "black",
          size = 0.1
  ) +
  # add titles
  labs(x = NULL,
       y = NULL,
       title = "Correlation between access to the railway network
    and inflenza inccidence",
       subtitle = "Canton of Berne, July/August 1918",
```

```
160       caption = 'Sources: Swisstopo (geometry), state archive of
      the Canton of Berne (influenza data), Schiedt (railway) Author:
       Corina Leuch') +
161
162   # as the sf object municipality_prod_geo has a column with name "
      fill" that
163   # contains the literal color as hex code for each municipality,
      we can use
164   # scale_fill_identity here
165   scale_fill_identity() +
166   # add the theme
167   theme_map() +
168   annotation_scale(line_width = 0.1, height=unit(0.1, "cm"), text_
      cex = 0.7,
169                    pad_x=unit(1, "cm"))
170
171
172 bivariate_color_scale <-  bivariate_color_scale %>%
173   separate(group, into = c("quantiles_flu", "TB_Terzile"), sep = "
      - ") %>%
174   mutate(quantiles_flu = as.integer(quantiles_flu),
175          TB_Terzile = as.integer(TB_Terzile))
176
177
178 legend <- ggplot() +
179   geom_tile(
180     data = bivariate_color_scale,
181     mapping = aes(
182       x = quantiles_flu,
183       y = TB_Terzile,
184       fill = fill)
185   ) +
186   scale_fill_identity() +
187   labs(x = "Higher flu inccidence ???",
188        y = "Better railway access ???",
189        title = 'Bivariate choropleth map') +
190   # make font small enough
191   theme(
192     axis.title = element_text(size = 7, hjust = 0),
193     axis.text = element_blank(),
194     axis.ticks = element_blank(),
195     panel.background = element_blank(),
196     plot.title = element_text(size = 10, color = "black", hjust =
      0),
197
198     ) +
199   # quadratic tiles
200   coord_fixed()
```

```
201
202 # make two separate maps to get legends for placement in final map
203 # build a manual map to get the legend
204 map2 <- ggplot() +
205   geom_sf(data = geometry,
206           aes(fill = '#F3F3F3'
207           ),
208           # use thin white stroke for municipalities
209           color = "black",
210           size = 0.1
211   ) +
212   scale_fill_identity(name=element_blank(),labels = c('no data'),
213                       guide = guide_legend(title = "Other symbols",
214                                            title.theme = element_
     text(size = 10, color = "black"),
215                         keyheight = unit(3, units = "mm"),
216                         keywidth = unit(6, units = "mm"),
217                         label.position = "right", label.theme =
     element_text(size = 7)
218                       )) + theme(
219                         legend.background = element_rect(fill = NA)
220                       )
221
222
223 map3 <- ggplot() + geom_sf(data = lakes,
224                            aes(fill = "#8CD1F2"),
225                            # use thin white stroke for
     municipalities
226                            color = "black",
227                            size = 0.1) +
228   scale_fill_identity(name=element_blank(),labels = c('water'),
229                       guide = guide_legend(title = "Other symbols",
230                       title.theme = element_text(size = 10, color =
      NA),
231                       keyheight = unit(3, units = "mm"),
232                         keywidth = unit(6, units = "mm"),
233                         label.position = "right", label.theme =
     element_text(size = 7)
234                       )) + theme(
235                         legend.background = element_rect(fill = NA)
236                       )
237
238 legend2 <- get_legend(map2)
239 legend3 <- get_legend(map3)
240
241 ggdraw() +
242   draw_plot(map, 0, 0, 1, 1, hjust=0) +
243   draw_plot(legend, 0.7, 0.6, 0.2, 0.2, hjust = 0, vjust = 0)+
```

```
  draw_plot(legend2, 0.68, 0.45, 0.2, 0.2, hjust = 0, vjust = 0)+
  draw_plot(legend3, 0.68, 0.425,0.2,.2, hjust=0, vjust = 0)
```

Preparation of variables and modelling

Prepare variables

Precipitation

```
# this script rearranges the weather data and joins the actual data to the staions. Using the output of this script,
# a surface with the precipitation for each wave and the overall data can be interpolated using ArcGIS.

local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

# import packages using function
pkgTest("tidyverse")
pkgTest("here")
pkgTest("sf")
pkgTest("tmap")


# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

#projection strings
WGS84 <- "+init=epsg:4326"
```

```r
LV03 <- "+init=epsg:21781"

# read datasets
stationen <- read.csv(file = paste0(dataFolder, '/stationen.csv'), sep= ';', encoding = "UTF-8")
stationen <- na.omit(stationen)

wetter <- read.csv(file = paste0(dataFolder, '/wetter_neu.csv'), sep= ';', encoding = "UTF-8") %>% as_tibble()
wetter$month <- sprintf("%02d", as.numeric(wetter$month)) # convert 1 to 01 in month --> easier sorting

# entire dataset----
# summarise the amount of rain for every station --> later used for interpolation, filter out the timespan where we don't have
# any influenza data
wetter_grpd <- wetter %>% unite(month, c(year, month), sep = "_") %>% group_by(month,stn) %>%
  mutate(rain=sum(as.numeric(rre150d0))) %>% select(stn, month, rain) %>% filter(month != "1918_01") %>%
  filter(month != "1918_02") %>% filter(month != "1918_03")%>% filter(month != "1918_04")%>% filter(month != "1918_05")%>%
  filter(month != "1918_06") %>% group_by(stn) %>% mutate(rain = sum(as.numeric(rain))) %>% select(stn, rain) %>% unique()

first_wave <- wetter %>% unite(month, c(year, month), sep = "_") %>% group_by(stn) %>%
  filter(month== "1918_08"| month == "1918_07") %>%
  mutate(rain_first = sum(rre150d0)) %>% select(stn, rain_first) %>% unique() %>% ungroup()

second_wave <- wetter %>% unite(month, c(year, month), sep = "_") %>% group_by(stn) %>%
  filter(month== "1918_10"| month == "1918_11"| month == "1918_12"| month=="1919_01") %>%
  mutate(rain_second = sum(rre150d0)) %>% select(stn, rain_second) %>% unique() %>% ungroup()

# now the dataset is in the form that we want it, we can join it to the stations
station_weather <- left_join(stationen, wetter_grpd, by="stn")
station_weather <- left_join(station_weather, first_wave, by="stn")
station_weather <- left_join(station_weather, second_wave, by="stn")

# convert to sf
stations_sf <- st_as_sf(station_weather, coords = c("E", "N"), crs= LV03)
```

```
68
69 st_write(stations_sf, paste0(outputFolder,"/stations.shp"), driver=
       "ESRI Shapefile")
70 write_csv(station_weather, paste0(outputFolder, "/stations_weather.
       csv"))
71 write_csv(first_wave, paste0(outputFolder, "/stations_weather_first
       _wave.csv"))
72 write_csv(second_wave, paste0(outputFolder, "/stations_weather_
       second_wave.csv"))
```

Railway access

```
1  # this code creates a navigable railway network using the prepared
       railway stations.
2  # in order for this code to work, the dataset has to be in the
       following form:
3  # nodes: point data with one point at each station
4  # edges: line data, that is cut at every node. Each line data has
       to start at a point and end at a point.
5
6  # for the scope of this thesis the data was prepared by hand and
       using basic python.
7
8
9  local({r <- getOption("repos")
10 r["CRAN"] <- "http://cran.r-project.org"
11 options(repos=r)
12 })
13
14 # function: checks if a package is installed or not
15 # if installed --> loads package
16 # else --> installs package
17
18 pkgTest <- function(x)
19 {
20   if (!require(x,character.only = TRUE))
21   {
22     install.packages(x,dep=TRUE)
23     if(!require(x,character.only = TRUE)) stop("Package not found")
24   }
25 }
26
27 # import packages using function
28 pkgTest("tidyverse")
29 pkgTest("here")
30 pkgTest("sf")
31 pkgTest("tmap")
32 pkgTest("tidygraph")
33 pkgTest("igraph")
```

```
pkgTest("leaflet")


# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output")

#projection strings
WGS84 <- "+init=epsg:4326"
LV03 <- "+init=epsg:21781"


# import the edes and nodes
nodes <- st_read(dsn=paste0(dataFolder, "/nodes_new.shp")) %>%
    select(-1)
edges <- st_read(dsn=paste0(dataFolder, "/edges_new.shp")) %>% st_
    set_geometry(NULL) %>% select(9, 10, 11) # drop geometry - will
     be added again later

node_coords <- do.call(rbind, st_geometry(nodes)) %>%
  as_tibble() %>% setNames(c("y","x"))

nodes <- bind_cols(nodes, node_coords) %>% st_set_geometry(NULL) #
    drop geometry, so it doesn't get mixed up later (will be added
    again)

# We add the short names and coordinates for "from station" and "to
     station"
edges <- edges %>%
  # Match "from stations"
  inner_join(select(nodes, c(from_y = "y",
                             from_x = "x",
                             from_id = "stop_id")),
             by = c("From_Node" = "from_id")) %>%
  # Match "to stations""
  inner_join(select(nodes, c(to_y = "y",
                             to_x = "x",
                             to_id = "stop_id")),
             by = c("To_Node" = 'to_id'))

nodes <- nodes %>%
  st_as_sf(coords = c("y", "x"), crs=21781)

# transform to wgs84 (for leaflet)
nodes <- st_transform(nodes, WGS84)

edges_sf <- edges %>%
```

```
  mutate(from = paste(from_y, from_x, sep = " ")) %>%
  mutate(to = paste(to_y, to_x, sep = " ")) %>%
  mutate(coords = paste0("LINESTRING(",from, ", ", to,")")) %>%
  select(-one_of(c("from", "to",
                   "to_y", "to_x",
                   "from_y", "from_x"))) %>%
  st_as_sf(wkt="coords", crs=21781)  %>% st_transform(WGS84)

network <- igraph::graph_from_data_frame(as_tibble(edges_sf),
    vertices = nodes) %>% as_tbl_graph()



# One more thing and we are done: remove loops(edges that start and
    end at the same vertex)
network <- as_tbl_graph(igraph::simplify(network,
                                         remove.multiple=F,
    remove.loops=T))

# calculate node betweeness centrality
network <- network %>%   activate(nodes) %>%
  mutate(btw = centrality_betweenness(weights=edges_sf$Shape_Leng,
    directed = FALSE))


# Plot the leaflet map
pal <- colorNumeric(
  palette = "YlGnBu",
  domain = network %>%
    activate (nodes) %>%
    as_tibble() %>%
    select("btw") %>%
    pull())


leaflet() %>%
  addProviderTiles("Esri.WorldTopoMap", group = "Terrain")  %>%

  # Add edges
  addPolylines(data = network %>%
                 activate (edges) %>%
                 as_tibble() %>%
                 st_as_sf()) %>%
  # Add point marker
  addCircleMarkers(data=network %>%
                   activate(vertices) %>%
                   as_tibble() %>%
                   st_as_sf(),
```

```
                    color =~pal(btw),
                    stroke = FALSE,
                    radius = 1.5,
                    fillOpacity = 0.8)

# now we can write the nodes back into a shapefile which we can
    then use in the gis
nodes_btw <- network %>% activate(nodes) %>% as_tibble() %>% st_as_
    sf(crs=WGS84) %>% st_transform(LV03)

btw_tibble <- network %>% activate(nodes) %>% as_tibble()

st_write(nodes_btw, paste0(RFolder, '/nodes_btw.shp'))

geometry <- st_read(paste0(dataFolder, "/Bern_Punkt_id.shp"),
    encoding="UTF-8")
railways <- st_read(paste0(dataFolder, "/railway_stations_bern.shp"
    ))

gemeinden <- st_join(geometry, railways) %>% select(Gemeinde, GEM_
    ID, btw) %>% st_transform(WGS84)
st_write(gemeinden,paste0(RFolder, "/Gemeinden_Erreichbarkeit.shp")
    )
```

Gather variables for modelling

```
# this script gathers all the variables in one dataset and creates
    the correlation plots

# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

```

```
23
24 # load packages
25 pkgTest("MASS")
26 pkgTest("tidyverse")
27 pkgTest("classInt")  # needed for Jenks
28 pkgTest("graphics")
29 pkgTest("sf")
30 pkgTest("lattice")
31 pkgTest("RColorBrewer")
32
33 # set data folder and R folder
34 dataFolder   <- here::here("data")
35 RFolder      <- here::here()
36 outputFolder <- here::here("output")
37
38
39 # load oliver's helper function
40 source("helper.r")
41
42 # read data  and calculate overall incidence ----
43 grippe <- read_csv(paste0(dataFolder, '/Monthly_Cases.csv')) %>%
       group_by(GEM_ID) %>%
44   mutate(cases=sum(montlyCases)) %>% dplyr::select(GEM_ID, Wohnb,
       cases) %>%
45   unique()  %>% mutate(Inz = cases/Wohnb*100000) %>% na.omit(grippe
       )
46
47 gemeinden <- read_delim(paste0(dataFolder, "/gemeinden.csv"), delim
       =";") %>%
48   select(GEM_ID, Gemeinde_Name, LWS, TB, Haush, Hoehe, Wohnb, E, N)
       %>% mutate(HauGr =Wohnb/Haush) %>%
49   mutate(AntLWS = LWS/Wohnb*100) %>% mutate(TBratio = TB/TB/Wohnb*
       100) %>% select(-Wohnb, -TB, -Haush, -LWS) %>%
50   na.omit()
51
52 # calculate first wave incidence ----
53 first <- read_csv(paste0(dataFolder, '/First_Wave.csv')) %>% group_
       by(GEM_ID) %>%
54   mutate(cases=sum(overallCases)) %>% dplyr::select(GEM_ID, Wohnb,
       cases) %>%
55   unique() %>% mutate(Inz_first = cases/Wohnb*100000) %>% na.omit()
56
57 second <- read_csv(paste0(dataFolder, '/Second_Wave.csv')) %>%
       group_by(GEM_ID) %>%
58   mutate(cases=sum(overallCases)) %>% dplyr::select(GEM_ID, Wohnb,
       cases) %>%
59   unique() %>% mutate(Inz_second = cases/Wohnb*100000) %>% na.omit
       ()
```

```
60
61
62 # join all the datasets together
63 first <- inner_join(first, gemeinden, "GEM_ID")
64 second <- inner_join(second, gemeinden, "GEM_ID")
65 grippe <- inner_join(grippe, gemeinden, 'GEM_ID')
66
67 # now I need to calculate quintiles for every variable I want. This
       is TB and inccidence for now
68
69 inz_total_quint <- grippe %>%
70   pull(Inz) %>%
71   quantile(probs = seq(0, 1, length.out = 6))
72
73 TB_total_quint <- grippe %>%
74   pull(TBratio) %>%
75   quantile(probs=seq(0,1, length.out = 6))
76
77 grippe <- grippe %>%
78   mutate(
79     Inz_Quintiles = cut(
80       Inz,
81       breaks = inz_total_quint,
82       include.lowest = TRUE
83     ),
84     TB_Quintiles = cut(
85       TBratio,
86       breaks = TB_total_quint,
87       include.lowest = TRUE
88     )) %>%
89   select(GEM_ID, Gemeinde_Name, Inz, Inz_Quintiles, TBratio, TB_
       Quintiles, AntLWS, HauGr, Wohnb)
90
91 # now I have to do the same thing for the first and second wave
       respectively
92 inz_first_quint <- first %>%
93   pull(Inz_first) %>%
94   quantile(probs = seq(0, 1, length.out = 6))
95
96 TB_first_quint <- first %>%
97   pull(TBratio) %>%
98   quantile(probs=seq(0,1, length.out = 6))
99
100 first <- first %>%
101   mutate(
102     Inz_Quintiles = cut(
103       Inz_first,
104       breaks = inz_first_quint,
```

```
      include.lowest = TRUE
    ),
    TB_Quintiles = cut(
      TBratio,
      breaks = TB_first_quint,
      include.lowest = TRUE
    )) %>%
  select(GEM_ID, Gemeinde_Name, Inz_first, Inz_Quintiles, TBratio,
    TB_Quintiles, AntLWS, HauGr, Wohnb)

# second wave
inz_second_quint <- second %>%
  pull(Inz_second) %>%
  quantile(probs = seq(0, 1, length.out = 6))

TB_second_quint <- second %>%
  pull(TBratio) %>%
  quantile(probs=seq(0,1, length.out = 6))

second <- second %>%
  mutate(
    Inz_Quintiles = cut(
      Inz_second,
      breaks = inz_second_quint,
      include.lowest = TRUE
    ),
    TB_Quintiles = cut(
      TBratio,
      breaks = TB_second_quint,
      include.lowest = TRUE
    )) %>%
  select(GEM_ID, Gemeinde_Name, Inz_second, Inz_Quintiles, TBratio,
     TB_Quintiles, AntLWS, HauGr, Wohnb)

# some variables are still missing: urbanity, weather and railways
# urbaniity: everything that's smaller than 15'000 in one class

grippe <- grippe %>% mutate(
  urbanity = ifelse(Wohnb < 10000, 0,1)
)

first <- first%>% mutate(
  urbanity = ifelse(Wohnb < 10000, 0,1)
)

second <- second %>% mutate(
  urbanity = ifelse(Wohnb < 10000, 0,1)
)
```

```r

# weather data ----
wetter <- read_delim(paste0(dataFolder, "/rain_gemeinden.csv"),
    delim =";") %>%
  mutate(first_wave = prec_1,second_wav = perc_2,Total_rain = prec_
    all) %>%
  select(first_wave, second_wav, Total_rain, GEM_ID)

geometry <- st_read(dsn = paste0(dataFolder, '/Bern_Punkt_id.shp'))

grippe <- left_join(grippe, wetter, "GEM_ID") %>% select(-first_
    wave, -second_wav)
first <- left_join(first, wetter, "GEM_ID") %>% select( -Total_rain
    , -second_wav)
second <- left_join(second, wetter, "GEM_ID") %>% select(-Total_
    rain, -first_wave)


wetter_total_quint <- grippe %>%
  pull(Total_rain) %>%
  quantile(probs = seq(0, 1, length.out = 6))

grippe <- grippe %>%
  mutate(
    Rain_Quintile = cut(
      Total_rain,
      breaks = wetter_total_quint,
      include.lowest = TRUE
    ))

wetter_first_quint <- first %>%
  pull(first_wave) %>%
  quantile(probs = seq(0, 1, length.out = 6))

first <- first %>%
  mutate(
    Rain_Quintile = cut(
      first_wave,
      breaks = wetter_first_quint,
      include.lowest = TRUE
    ))

wetter_second_quint <- second %>%
  pull(second_wav) %>%
  quantile(probs = seq(0, 1, length.out = 6))

second <- second %>%
  mutate(
```

```
    Rain_Quintile = cut(
      second_wav,
      breaks = wetter_second_quint,
      include.lowest = TRUE
    ))

# finally: the railway stations
railway_sf <- st_read(paste0(dataFolder, "/Gemeinden_Erreichbarkeit
    _final.shp"))
railway <- st_set_geometry(railway_sf, NULL) %>% mutate(btw =
    replace_na(btw, 0))

freq_table <- railway_sf %>%
  dplyr::count(GEM_ID) %>%
  group_by(GEM_ID) %>%          # now required with changes to
    dplyr::count()
  mutate(prop = prop.table(n))

# some municipalities have two train stations and this needs to be
    adressed by looking at the geometry (i.e selecting the relevant
     one)
# furthermore the ones where GEM_ID is zero are adressed
# read new dataset (don't have GEM_ID yet) -> this is done by hand
railway_new_sf <- st_read(paste0(dataFolder, "/Gemeinden_
    Erreichbarkeit_final.shp")) %>%
  select("Gemeinde", "btw", "GEM_ID")

freq_table <- railway_new_sf %>%
  dplyr::count(GEM_ID) %>%
  group_by(GEM_ID) %>%          # now required with changes to
    dplyr::count()
  mutate(prop = prop.table(n))

# select only the GEM ID and btw, delete Geometry (not required)
railway_gemeinden <- railway_new_sf %>% st_set_geometry(NULL) %>%
  select(btw, GEM_ID)

grippe_final <- left_join(grippe, railway_gemeinden, by="GEM_ID")
first_final <- left_join(first, railway_gemeinden, by='GEM_ID')
second_final <- left_join(second, railway_gemeinden, by='GEM_ID')

# classify btw
btw_not_zero <- filter(grippe_final, btw>0)
first_not_zero <- filter(first_final, btw>0)
second_not_zero <- filter(second_final, btw>0)


classIntervals(btw_not_zero$btw, n=4, style = "jenks")
```

```r
classIntervals(first_not_zero$btw, n=4, style = "jenks")
classIntervals(second_not_zero$btw, n=4, style = "jenks")

grippe_final <- grippe_final %>% mutate(btw_classes =
                                          ifelse(btw == 0, 1,
                                          ifelse(btw <= 13896, 2,
                                          ifelse(btw <= 30624, 3,
                                          ifelse(btw <= 87852,4,5))))
    )

first_final <- first_final %>% mutate(btw_classes =
                                          ifelse(btw == 0, 1,
                                                 ifelse(btw <= 15167,
     2,
                                                        ifelse(btw <=
     42005, 3,
                                                               ifelse
    (btw <= 87852,4,5)))))
second_final <- second_final %>% mutate(btw_classes =
                                            ifelse(btw == 0, 1,
                                                   ifelse(btw <=
    13896, 2,
                                                          ifelse(btw
    <= 30624, 3,

    ifelse(btw <= 69866,4,5)))))


# pairs plot first wave, absolute
# rename the names to make pairs plot more clear
first_pairs <- first_final %>%
  rename(
    TB = TBratio,
    rain = first_wave,
    agriculture = AntLWS,
    railway = btw

  )

variables <- c("TB", "rain", "agriculture", 'railway', "urbanity")
pairs(first_pairs[variables], lower.panel=panel.smooth, upper.panel
    =panel.cor, diag.panel=panel.hist, cex.labels=2,
      main = "Correlation between Explanatory Factors (First Wave)"
    )


# pairs Plot second wave, absolute
second_pairs <- second_final %>%
```

```
  rename(
    TB = TBratio,
    rain = second_wav,
    agriculture = AntLWS,
    railway = btw
  )

pairs(second_pairs[variables], lower.panel=panel.smooth, upper.panel=panel.cor, diag.panel=panel.hist, cex.labels=2,
      main = "Correlation between Explanatory Factors (Second Wave)", sub='test')
# write csv
write_csv(grippe_levels, path = paste0(outputFolder, "/Grippe_Final.csv"))
write_csv(first_final, path = paste0(outputFolder, "/First_Wave_Final.csv"))
write_csv(second_final, path = paste0(outputFolder, "/Second_Wave_Final.csv"))
```

Modelling

```
# set up part ----
# define standard repo for rpackages
local({r <- getOption("repos")
r["CRAN"] <- "http://cran.r-project.org"
options(repos=r)
})

# function: checks if a package is installed or not
# if installed --> loads package
# else --> installs package

pkgTest <- function(x)
{
  if (!require(x,character.only = TRUE))
  {
    install.packages(x,dep=TRUE)
    if(!require(x,character.only = TRUE)) stop("Package not found")
  }
}

# load packages
pkgTest("tidyverse")
pkgTest("glmulti")
pkgTest("stargazer")
pkgTest("xtable")
pkgTest("arm")

```

```r

# set data folder and R folder
dataFolder   <- here::here("data")
RFolder      <- here::here()
outputFolder <- here::here("output/")

# first wave ----
# first have to read data set and add factors

dat_first  <- read_csv(paste0(outputFolder, "/First_Wave_Final.csv"))

dat_first <- within(dat_first, {
  urbanity <- factor(urbanity, levels = 0:1, labels=c("Dorf", "City"))
  GEM_ID <- factor(GEM_ID)
  Rain_Quintile <- factor(Rain_Quintile, levels = c("[90.8,104]", "(104,110]" , "(110,125]" , "(125,157]", "(157,250]" ), labels = c(1,2,3,4,5))
  Inz_Quintiles <- factor(Inz_Quintiles, levels = c("[71,934]", "(934,2.02e+03]", "(2.02e+03,3.23e+03]", "(3.23e+03,5.62e+03]", "(5.62e+03,2.95e+04]")  , labels = c(1,2,3,4,5))
  TB_Quintiles <- factor(TB_Quintiles, levels = c("[0.000956,0.0501]", "(0.0501,0.0969]", "(0.0969,0.159]", "(0.159,0.286]", "(0.286,1.75]") , labels = c(1,2,3,4,5))
  btw_classes <- factor(btw_classes)
})

dat_first <- dat_first %>%
  mutate(Top_Inz = ifelse(Inz_Quintiles == 5, 1, 0)) %>%
  mutate(Top_TB = ifelse(TB_Quintiles == 5, 1, 0)) %>%
  mutate(Bottom_TB = ifelse(TB_Quintiles == 1, 1, 0))

# need to calculate terzile for TB
tb_terziles <- dat_first %>%
  pull(TBratio) %>%
  quantile(probs = seq(0, 1, length.out = 4))

dat_first <- dat_first %>%
  mutate(
    TB_Terziles = cut(
      TBratio,
      breaks = tb_terziles,
      include.lowest = TRUE
    )) %>%
  mutate(
    TB_Terziles = factor(TB_Terziles, labels = c(1,2,3))
  )
```

```
68
69 dat_first <-  dat_first %>%
70   mutate(Top_Inz = factor(Top_Inz, levels=0:1)) %>%
71   mutate(Bottom_TB = factor(Bottom_TB, levels = 0:1)) %>%
72   mutate(Top_TB = factor(Top_TB, levels = 0:1))
73
74 # model selection
75 g1 <- glmulti(Top_Inz~urbanity+AntLWS+Top_TB+btw_classes+Rain_
       Quintile, data = dat_first,
76               level = 1,
77               crit = "aic",
78               plotty =TRUE, report = TRUE,
79               family = "binomial",
80               method = "h",
81               confsetsize = 250)
82
83 top <- weightable(g1)
84 top <- top[top$aic <= min(top$aic) + 2,]
85 xtable(top)
86
87 # model
88 summary(m1 <- glm(Top_Inz ~ 1 + urbanity + Top_TB + btw_classes +
       Rain_Quintile, data = dat_first, family = 'binomial'))
89
90
91 # breusch-pagan test
92 lmtest::bptest(m1)
93
94 # calculate odds ratio and CI
95 odds1 <- exp(cbind(OR = coef(m1), confint(m1)))
96
97 # prepare results for reporting in table
98 m1.OR <- m1 # small workarounds, exponantiate coefficients of model
        to generate a second fake model which has the OR
99 m1.OR$coefficients <- exp(m1$coefficients)
100 stargazer(m1, m1.OR, ci=c(F,T), column.labels = c('coefficients', '
       odds ratio'),
101           single.row = TRUE, star.cutoffs = c(0.05,0.01,0.001),
102           digits = 2, column.sep.width = "0.5pt", no.space = TRUE,
103           covariate.labels = c("urbanityCity", "Top TB Quintile", "
       railway2", "railway3", "railway4", 'railway5', 'rain2', 'rain3'
       , 'rain4', 'rain4'))
104
105 dat_first$residuals <- residuals(m1)
106 dat_first$predicted <- predict(m1)   # Save the predicted values
107
108 # write model data (used for Moran's I in GeoDA)
```

```r
write.csv(dat_first, file = paste0(outputFolder, '/FirstWaveModel.csv'))

# do this exact thing for the second wave ----
dat_second  <- read_csv(paste0(outputFolder, "/Second_Wave_Final.csv"))

# prepare the data (assign factors)
dat_second <- within(dat_second, {
  urbanity <- factor(urbanity, levels = 0:1, labels=c("Dorf", "City"))
  GEM_ID <- factor(GEM_ID)
  Rain_Quintile <- factor(Rain_Quintile, levels = c("[228,304]","(304,334]","(334,374]", "(374,410]", "(410,661]"), labels = c(1,2,3,4,5))
  Inz_Quintiles <- factor(Inz_Quintiles, levels = c( "[284,3.63e+03]", "(3.63e+03,8.09e+03]", "(8.09e+03,1.28e+04]", "(1.28e+04,1.9e+04]",  "(1.9e+04,6.47e+04]")  , labels = c(1,2,3,4,5))
  TB_Quintiles <- factor(TB_Quintiles, levels = c("[0.000956,0.0624]", "(0.0624,0.117]", "(0.117,0.189]", "(0.189,0.337]", "(0.337,2.27]") , labels = c(1,2,3,4,5))
  btw_classes <- factor(btw_classes)
})

dat_second <- dat_second %>%
  mutate(Top_Inz = ifelse(Inz_Quintiles == 5, 1, 0)) %>%
  mutate(Top_TB = ifelse(TB_Quintiles == 5, 1, 0)) %>%
  mutate(Bottom_TB = ifelse(TB_Quintiles == 1, 1, 0))


dat_second <-  dat_second %>%
  mutate(Top_Inz = factor(Top_Inz, levels=0:1)) %>%
  mutate(Bottom_TB = factor(Bottom_TB, levels = 0:1)) %>%
  mutate(Top_TB = factor(Top_TB, levels = 0:1))

# second model selection
g2 <- glmulti(Top_Inz~urbanity+AntLWS+Top_TB+btw_classes+Rain_Quintile, data = dat_second,
              level = 1,
              crit = "aic",
              plotty =TRUE, report = TRUE,
              family = "binomial",
              method = "h",
              confsetsize = 250)

# create table with candidate models (R)
top2 <- weightable(g2)
top2 <- top2[top2$aic <= min(top2$aic) + 2,]
```

```
147
148 # create table of candidate models for latex
149 xtable(top2)
150
151 # actual model
152 summary(m2 <- glm(Top_Inz ~ 1 + urbanity + Top_TB + btw_classes +
        Rain_Quintile, data = dat_second, family = 'binomial'))
153
154 # breusch pagan test
155 lmtest::bptest(m2)
156
157 # latex table
158 m2.OR <- m2 # small workarounds, exponantiate coefficients of model
         to generate a second fake model which has the OR
159 m2.OR$coefficients <- exp(m2$coefficients)
160 stargazer(m2, m2.OR, ci=c(F,T), column.labels = c('coefficients', '
        odds ratio'),
161           single.row = TRUE, star.cutoffs = c(0.05,0.01,0.001),
162           digits = 2, column.sep.width = "0.5pt", no.space = TRUE,
163           covariate.labels = c("urbanityCity", "Top TB inccidence",
         "railway2", "railway3", "railway4", 'railway5', 'rain2', "
        rain3", "rain4", "rain5"))
164
165 # calculate KI and odds ratio
166 # calculate odds ratio and 95% KI
167 odds2 <- exp(cbind(OR = coef(m2), confint(m2)))
168
169 # save residuals and predicted
170 dat_second$residuals <- residuals(m2)
171 dat_second$predicted <- predict(m2)
172
173 write.csv(dat_second, file = paste0(outputFolder, '/SecondWaveModel
        .csv'))
```

Personal declaration

I hereby declare that the submitted Thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.