



Geovisual analytics of historical morbidity and mortality data of Tuberculosis and Diphtheria in Zürich in the late 1920s and early 1930s

GEO 511 Master's Thesis

Author: Fabio Schilling, 19-704-048

Supervised by: Prof. Dr. Sara Irina Fabrikant, Prof. Dr. phil. Kaspar Staub
(kaspar.staub@iem.uzh.ch)

Faculty representative: Prof. Dr. Sara Irina Fabrikant

17.04.2025

Abstract

Tuberculosis and diphtheria are two infectious diseases that continue to be a major global health concern, particularly in developing countries. For instance, in 2023 alone, over one million lives were claimed by tuberculosis alone. Nevertheless, significant improvements have been made compared to the 19th and 20th centuries, when both diseases claimed a much higher death toll. This master's thesis aims to explore the distribution of tuberculosis and diphtheria cases in the city of Zurich in the late 1920s and early 1930s with a focus on geographical patterns and socioeconomic status. While there have been similar studies on the spread of other diseases, these have focused on other cities and diseases, most notably London for cholera, and more recently, Basel, (the canton of) Berne for the 1918-1920 influenza pandemic and emerging cities of developing countries. This thesis will explore the history of Zurich's fight against tuberculosis and diphtheria in relation to socioeconomic indicators, while aiming to provide new insights into its course, which could be central to understanding further developments in Zurich's health policy and geography of health. The dataset encompasses single case entries with attributes including registration dates, gender, age, occupation, place of residence, hospital treatment and mortality rates. Utilising these attributes, visualisations and clustering analyses were conducted. The results indicate a stronger socioeconomically underpinned clustering for tuberculosis than for diphtheria. Moreover, the mortality and case numbers in general are higher for tuberculosis. When contextualised within a historical framework, it is evident that the case numbers of both tuberculosis and diphtheria case numbers had declined from their respective peak values in the late 19th century, although tuberculosis remained a major public health problem. Consequently, during the time period under investigation, tuberculosis was identified as the most prevalent chronic disease. Conversely, diphtheria, often referred to as the "strangling angel" of children had already experienced a significant decline in its prevalence and mortality. The investigation revealed that both diseases exhibited clustering behaviour in space. However, during more thorough analyses conducted, age and particularly social class emerged as the only significant influences on tuberculosis mortality, while no significant influences on diphtheria mortality manifested themselves. The absence of statistical significance in the diphtheria data may be attributed to the limited number of fatalities, which rendered the sample size inadequate for conducting a meaningful analysis.

Keywords: Tuberculosis, Diphtheria, Historical Data, GIS, Cluster Analysis, Geographic Information Visualisation, Health Geography, Disease Mapping

Acknowledgements

I would like to sincerely thank all those who supported me throughout my academic life. Firstly, I would like to express my profound appreciation to my two supervisors, Prof. Dr. Sara Irina Fabrikant and Prof. Dr. phil. Kaspar Staub, who always supported me and were open to my frequent questions. Another person who deserves thanks, although I do sadly not know him personally, is Vithush Yogarasa, who transcribed the original, handwritten, hard-to-read data and without whom I would have had to do this laborious task all by myself. I would also like to express my sincere and heartfelt thanks to Takuya Takahashi and the members of the GIScience colloquium of the Department of Geography at the University of Zurich for their feedback and ideas.

Finally, I would also like to extend my gratitude to my family, especially my parents, who have supported me throughout my academic life. It is important for me to acknowledge the enthusiasm displayed by my maternal grandparents for my academic pursuits. Sadly, my grandfather did not have the opportunity to witness this milestone with me. Additionally, I would like to give a big thank you to my partner for putting up with me and always believing in me. And last but not least, I would like to thank all my friends, however long I have known them for, for their companionship and proofreading all my writings over the years. I would not have been able to complete this undertaking without the support you all provided me with.

Contents

Abstract.....	I
Acknowledgements.....	II
Contents	III
List of Figures.....	V
List of Tables.....	VII
1 Introduction	1
1.1 Motivation.....	1
1.2 Research gap	2
1.3 Research objectives and research questions	2
2 State of research	4
2.1 Disease mapping and health geography	4
2.2 GIS in health geography.....	5
2.3 Tuberculosis and diphtheria today	6
2.4 History of tuberculosis and diphtheria: a Swiss perspective.....	8
2.5 Zurich in the 1920s and 1930s	11
2.5.1 The 8 districts of Zurich	19
2.5.2 Socioeconomic characteristics of districts	21
3 Data	23
3.1 Disease datasets.....	23
3.2 Geodata.....	27
3.3 Statistical data	27
4 Methods	28
4.1 Geocoding residential addresses.....	28
4.2 Defining the 8 districts of Zurich	29
4.3 Socioeconomic data from statistical yearbooks and HISCO	32
4.4 Visualisation	32
4.4.1 Visualisation methods	32
4.4.2 Visual variables	33
4.5 Analysis	35

5	Results	37
5.1	Tuberculosis	38
5.2	Diphtheria.....	57
6	Discussion.....	74
6.1	Tuberculosis	74
6.2	Diphtheria.....	78
6.3	Synthesis	81
6.4	Limitations.....	84
7	Conclusions and future work	88
8	References	91
	Appendix	99
	Personal declaration	118

List of Figures

Aerial image of central sea-side Zurich (1925) (ETH-Bibliothek, 2025).....	11
Aerial image of Zurich's industrial outskirts (1929) (ETH-Bibliothek, 2025)	11
Zurich district map including Swiss reference map	12
Urban development of Zurich since 1850, adapted from (Brodbeck and Hermann, 2012) and (Behrens, Motschi and Schultheiss, 2015)	14
Relative age distribution pyramid of Zurich in 1930 (Stadt Zürich, 2018a)	15
Contemporary map of the city of Zurich including administrative districts (Statistik Stadt Zürich, 1931).....	19
Population density map	21
Building insurance value map.....	22
Rent map for 3 room apartments Figure 10: Rent map for 4 rooms apartments	22
First part of the original tuberculosis dataset.....	23
Second part of the original tuberculosis dataset	24
Original diphtheria dataset.....	25
Data manipulation steps (adapted by the author)	28
Border changes of district 3.....	30
Border changes of district 5.....	30
Border changes of district 6.....	31
Border changes of district 7.....	31
Population cartogram, coloured by population density.....	37
Tuberculosis age distribution	38
Tuberculosis social class distribution	39
Tuberculosis district distribution	40
Tuberculosis case map	41
Tuberculosis mortality map.....	42
Tuberculosis morbidity cartogram	43
HDBSCAN clusters of tuberculosis cases	44
HDBSCAN clusters of tuberculosis cases by social class	45
HDBSCAN clusters of tuberculosis fatalities by social class	46
Social class map of tuberculosis cases.....	47
NN-distances of tuberculosis cases	48
NN-distances of tuberculosis fatalities	49
G-function of tuberculosis cases.....	49
G-function of tuberculosis fatalities.....	50
Moran's I of tuberculosis cases Figure 37: Moran's I of tuberculosis fatalities.....	50
local G* cluster map of social class in tuberculosis data.....	51
local G* cluster map of social class in tuberculosis fatalities.....	52
Spatial distribution of age and social class influences on tuberculosis mortality	54
Tuberculosis case graph	55
Tuberculosis case histogram.....	55

Temporal tuberculosis incidence	56
Diphtheria age distribution	57
Diphtheria social class distribution	58
Diphtheria district distribution	59
Diphtheria case map	60
Diphtheria mortality map	61
Diphtheria morbidity cartogram Figure 50: Diphtheria mortality cartogram	62
HDBSCAN clusters of diphtheria cases	63
HDBSCAN clusters of diphtheria cases by social class	64
Social class map of diphtheria cases	65
Diphtheria incidences overlayed with school catchment areas	66
NN-distances of diphtheria cases	67
G-function of diphtheria cases	68
Moran's I of diphtheria cases Figure 58: Moran's I of diphtheria fatalities	68
local G* cluster map of social class in diphtheria data	69
Spatial distribution of age and social class influences on diphtheria mortality	71
Diphtheria case graph	72
Diphtheria case histogram	72
Temporal diphtheria incidence	73

List of Tables

Zurich's population subdivided into age cohorts (Statistik Stadt Zürich, 1931).....	16
Zurich's population subdivided by districts (Statistik Stadt Zürich, 1931)	16
Main characteristics of each district and defining social class	20
Transcribed attributes of tuberculosis data	24
Transcribed attributes of diphtheria data.....	26
Comparison of regression results of tuberculosis	53
Comparison of regression results for diphtheria	70
Comparative synthesis of tuberculosis and diphtheria data analysis.....	83

1 Introduction

1.1 Motivation

In the context of the recent global Covid-19 pandemic, the influence and geographic patterns of diseases have once again become a matter of significant importance. Notable diseases of the 20th century include tuberculosis, known as the “white death” or the “consumption”, and diphtheria, referred to as the “strangling angel”. Tuberculosis and diphtheria, which appeared last on a grand scale in Europe as epidemics during the Second World War, have persisted throughout the 20th century and have only been addressed through the introduction of vaccines since about the 1940s and 1950s (Holloway et al., 2013, 2014; Ritzmann, 2015; Müller et al., 2024).

Even today, tuberculosis and, to a lesser extent, diphtheria require attention from key global health players, including the United Nations and World Health Organisation (United Nations, 2024; WHO, 2024a, 2024b). In 2023, for instance, tuberculosis killed 1.25 million people out of 10.8 million cases, making it the deadliest infectious disease worldwide (WHO, 2024b). Recent diphtheria outbreaks have occurred in areas where vaccine coverage is insufficient. However, it should be noted that from 1980 to 2000, more than 90% of cases could have been prevented, indicating that the work of health organisations in this field is not yet complete (WHO, 2024a). Tuberculosis and diphtheria have long been associated with socioeconomic disadvantage, with lower living standards resulting in diminished resilience to these diseases (Holloway et al., 2014, 2013; Kistemann et al., 2002; Müller et al., 2024).

By approximately 1930, the global and Swiss case numbers and mortality rates of both tuberculosis and diphtheria had declined from their respective peaks in the late 19th century. When assessed in terms of case numbers and mortality, Switzerland occupied a median position in international comparisons, with Zurich demonstrating a lower incidence compared to other Swiss cities such as Basel, Berne and Geneva (Brunner and Senti, 1937; Kruker and Senti, 1932; Senti and Pfister, 1946). The city of Zurich meticulously documented notifiable cases of tuberculosis and diphtheria since at least the late 1920s and early 1930s, as mandated by epidemic laws (Bundesamt für Gesundheit, 2024a; Holloway et al., 2013, p. 81; Kruker and Senti, 1932; Senti and Pfister, 1946). The data encompasses attributes such as age, gender, occupation and address, thereby facilitating the analysis of both the spatial and socioeconomic characteristics of case numbers. The First Law of Geography, a well-known principle in the field, posits that “everything is related to everything else, but near things are more related than distant things.” (Tobler, 1970, p. 236). This phenomenon, termed TFL (Tobler’s first law of geography) (Miller, 2004; Waters, 2017) implies that, like everything else, poverty and disease are not randomly distributed in space, but tend to cluster. Consequently, it is logical to study diseases associated with poverty in space, a domain in which health geography and, by extension geographic information systems (GIS) shine. Health geography (Moon, 2020), which includes the

synonymously used terms medical geography (Tsatsaris et al., 2023) and spatial epidemiology (Meliker and Sloan, 2011; Wood et al., 2023) studies health outcomes in space. The content of these terms comprises the spatial distribution of disease, its relationship to environmental and socio-demographic factors, and studies of the population at risk (Lawson, 2006; Meliker and Sloan, 2011; Moon, 2020; Tsatsaris et al., 2023; Wood et al., 2023).

1.2 Research gap

The literature on the history of infectious diseases in Zurich itself is rather scarce, limited to the influenza pandemic of 1918-1920 and tuberculosis throughout the 19th and 20th centuries (Corti, 2012; Holloway et al., 2013; Ritzmann, 1998; Ziegler et al., 2024). Moreover, research on the influence of socioeconomic backgrounds on health outcomes is limited. This is of particular significance given the widely acknowledged fact, that a disadvantaged socioeconomic status is associated with an increased risk of infection and mortality. The question of the impact of tuberculosis, and indeed other infectious diseases, on specific occupational groups and social classes remains intriguing (Ritzmann, 1998, pp. 27–28). While other Swiss cities have been the focus of recent studies on infectious diseases (Birkhölzer, 2023; Leuch, 2021), there is a very limited number of studies that include tuberculosis or diphtheria. Zurich, as the largest city in Switzerland, is a logical location for this study. Despite its modest size, with a contemporary population of 450'000 and 250'000 inhabitants in 1930, it is a rational location for conducting research. The majority of previous research has focused on either larger millionaire cities (e.g., London, north-eastern American cities) or, more recently, rural areas in developing countries. Furthermore, Swiss cities did not possess the same levels of urban deprivation (i.e. slums) as larger European cities. Two datasets of tuberculosis and diphtheria cases in the late 1920s and early 1930s allow for the expansion of research on past health conditions in the City of Zurich and for the addition of data on infectious diseases throughout history, especially in connection with socioeconomic status. The analysis of individual case data of diseases, as are available for this thesis, can lead to new insights into the factors contributing to disease spreading, i.e. its morbidity and mortality.

1.3 Research objectives and research questions

Geographic information systems (GIS) have been extensively utilised for the purpose of disease mapping, with the employment of spatio-temporal analytic methods serving to enhance comprehension of local disease patterns (Gatrell and Löytönen, 1998; Kirby et al., 2017; Koch, 2017). Consequently, there is an opportunity to analyse this data in GIS. The subsequent analysis will involve the visualisation of the disease cases, thereby enabling a visual inspection of the distribution. In a further step, I would also like to analyse these distributions and look for any patterns, such as clusters of high or low disease incidence.

The tuberculosis and diphtheria data will be studied concerning at least two different aspects of health geography: first, links referring to social class or socioeconomic status

and secondly, time and space-time connections. The analysis will provide a visual description of the temporal progression of the spread of tuberculosis and diphtheria in the city of Zurich, and will reveal whether there are any striking connections between the social class or socioeconomic status of the infected people and their place of residence. The study will ascertain whether socioeconomically disadvantaged people, i.e. poor, are at a higher risk of infection due to various reasons, such as worse living conditions and/or workplace situation.

Drawing on the preceding paragraph, the research objectives of this master's thesis can now be specified as follows:

1. To geolocate the addresses of disease cases.
2. To visualise the spread of tuberculosis and diphtheria in space and time.
3. To describe how socioeconomic status / social class influences disease rates and leads to the emergence of hotspots.

These research objectives give rise to a plethora of research questions that require investigation. The focal point of this thesis is as follows:

How did socioeconomic factors influence the spatial patterns of tuberculosis and diphtheria cases in Zurich in the late 1920s and early 1930s, and where were the hotspots and coldspots of these two diseases, respectively?

With the research objectives and the main research question established, four supporting research questions can be formulated:

- How did tuberculosis and diphtheria cases spread spatially and temporally in Zurich during the late 1920s and early 1930s?
- Where are hotspots and coldspots of tuberculosis and diphtheria cases in Zurich, and how did disease prevalence vary across the city?
- What was the relationship between socioeconomic status and social class and the spatial distribution of tuberculosis and diphtheria cases?
- To what extent did socioeconomic disparities correlate with the spatial distribution of tuberculosis and diphtheria cases in Zurich, and how do disease patterns reflect socioeconomic inequalities in the city?

The selection of these questions is motivated by their capacity to facilitate a comprehensive overview of the two diseases in Zurich during the period of transition from the early to the mid-20th century. The incorporation of further inquiries into the socioeconomic status / social class of individuals and urban districts could facilitate an assessment of the influence of socioeconomic characteristics on the propagation of diseases and mortality rates.

2 State of research

2.1 Disease mapping and health geography

One of the earliest and most notable applications of disease mapping is exemplified by John Snow's written account of the 1854 cholera outbreak in London's Soho district (Pavia et al., 2019; Shiode et al., 2015; Shiode, 2012). Since the advent of computing technology, the field of digital health geography has significantly benefited from GIS, as they facilitate the mapping and the analysis of health-related data and its associated explanatory factors (Gatrell and Löytönen, 1998; Kirby et al., 2017; Koch, 2017).

The aforementioned umbrella term, health geography, includes methodologies such as disease mapping in addition to statistical calculations (Meliker and Sloan, 2011; Moon, 2020). Research in health geography involves the analysis of space-time data on diseases and the identification of patterns, integrating the disciplines of geography and health (Meliker and Sloan, 2011; Moon, 2020). The spatial dimension of disease mapping is a critical component of health geography, as it introduces an additional dimension, space, to the analysis. Important concepts within this discipline include morbidity and mortality. Morbidity is understood to describe the amount of disease within a population, while mortality is the number of deaths from a disease, either as an absolute or relative number. Consequently, both morbidity and mortality serve as indicators of the impact of a disease within a population. Morbidity is calculated by dividing the number of infected persons by the total population, or alternatively, by the number of new cases within a specified timespan. The calculation of mortality is derived from the number of deaths from a disease, expressed either as an absolute figure or as a rate, typically per 1000 persons. Normalisation to a rate enables comparison with other diseases. The evaluation of the impact of a disease on a population and the assessment of health outcomes are facilitated by both concepts (Hernandez and Kim, 2024).

Although cartography can be assumed to be as old as civilisation itself, the mapping of diseases is a relatively modern phenomenon. John Snow, who mapped the 1854 cholera outbreak in London's Soho neighbourhood, is perhaps the most well-known exponent of early disease mapping to geographers. However, it should be noted that he was not the only nor the first person to undertake such a task. The 19th century witnessed a proliferation of disease mapping. A multitude of influences and concerns would manifest themselves and propagate the technique of mapping diseases. The social question in conjunction with the rise of industrialisation and urbanisation, rapidly established a correlation between poverty and health (Koch, 2017, pp. 1–74).

The seminal work of John Snow (1813-1858) represented a pivotal moment in the evolution of disease mapping. Snow's efforts to map the source of an outbreak of cholera in London's Broad Street demonstrated the potential of this technique. His contributions elevated disease mapping to unprecedented heights (Koch, 2017, pp. 75–102; Meliker and Sloan, 2011). The advent of computing technology and GIS in the mid-20th century has

had a profound impact on disease mapping, generating a host of new possibilities. The integration of a vast array of data into GIS, and the expansion of analysis possibilities, were now suddenly possible (Koch, 2017, pp. 279–320; Meliker and Sloan, 2011).

2.2 GIS in health geography

A plethora of scientific papers have examined either the spread or the spatial distribution of cases of different diseases, such as tuberculosis (Chirenda et al., 2020; Gwitira et al., 2021; Kanturk, 2007; Kistemann et al., 2002; Sun et al., 2015; Tiwari et al., 2006), diphtheria (Setiawan et al., 2021) and cholera (Agbor, 2014; Bwire et al., 2017; Gaudart et al., 2013; Ngwa et al., 2021; Osei and Duker, 2008; Pezeshki et al., 2012; Ruiz-Moreno et al., 2010). However, the majority of these studies utilise more recent datasets, which are predominantly from developing countries in Africa and Asia. A significant proportion of research focuses on the well-known and well-documented 1854 cholera epidemic in London as well (Brody et al., 2000; Caplan et al., 2020; Koch and Denike, 2009; Shiode et al., 2015; Shiode, 2012; Walford, 2020).

A greater volume of research has been conducted on tuberculosis rather than on diphtheria, a discrepancy that may be indicative of its higher prevalence and greater perceived danger in the past and present. Research has demonstrated that hotspots of tuberculosis and diphtheria incidence are often associated with low socioeconomic status (Chirenda et al., 2020; Gubéran, 1980; Hermans et al., 2015; Kanturk, 2007; Liu et al., 2012; Setiawan et al., 2021; Shaweno et al., 2018; Sun et al., 2015; Tiwari et al., 2006; Vaughan, 2018). These factors result in lower living standards (i.e., overcrowding, malnutrition) including high population density, low education and poverty. Additionally, poor air quality was also mentioned as a contributing factor (Sun et al., 2015), which is not necessarily connected to low socioeconomic status, however a tendency can be observed. Chirenda *et al.* (2020) have found tuberculosis to be most prevalent among the economically active age group of 20- to 44-year-olds in Harare, Zimbabwe. A historical study of tuberculosis cases in Switzerland has also shown adolescents and young adults to be the most affected age groups (Gubéran, 1980). The transmission of tuberculosis, which often occurs within households, communities or workplaces gives rise to spatial heterogeneity in disease patterns (Shaweno et al., 2018). Although not explicitly mentioned, it can be assumed that the same pattern applies to diphtheria, due to the similarity of the transmission routes.

It is now widely acknowledged that both tuberculosis and diphtheria are particularly prevalent among population experiencing poverty (Coleman, 2018; Holloway et al., 2014, 2013; Kistemann et al., 2002; Müller et al., 2024). Furthermore, evidence suggest that co-infection and mutual reinforcement of tuberculosis and diphtheria occur, as there is a complex interdependence with iron deficiency (Coleman, 2018), which may also be traceable to poverty.

Research into historical health geography, focusing on the spatio-temporal analysis of health outcomes, has been conducted in a number of Swiss cities and cantons. For

example Birkhölzer (2023) investigated the 1855 Cholera epidemic in the City of Basel, while Leuch (2021), Staub *et al.*, (2021) and Bernhard *et al.*, (2023), examined the 1918 influenza pandemic in the Canton of Berne. Finally, Burkhard (2023) mapped the spatial distribution of birth weights in Basel and Lausanne at the beginning of the 20th century.

A commonality of all the Swiss examples under consideration is the mapping of the disease distribution and the spatio-temporal analysis, in conjunction with potentially contributing environmental and socio-demographic factors. Furthermore, they particularly investigate a potential link between social class or socioeconomic status and mapped infection rates, as people of lower socioeconomic status are more likely to fall victim to such types of diseases (Holloway *et al.*, 2014, 2013; Müller *et al.*, 2024). The potential causes of this phenomenon include the standard of living of the population at risk, the number of people who live in the same household, the living space available per household, and the hygiene situation of the households within a neighbourhood.

A significant number of studies to date have been based on residence, as is the case in this thesis. However, Shaweno *et al.* (2018) draw attention to workplaces and social gatherings which themselves pose a considerable risk of infection too and should thus not be dismissed. Regrettably, the workplace of the people in the dataset was not documented, only their occupation. This shortcoming must be kept in mind when interpreting the findings. It is also important to consider the limited mobility that was characteristic of this period, as motorised vehicles did not become widely available until after the Second World War, and public transport was significantly more limited than it is nowadays. Consequently, it can be hypothesised that most people were employed in close proximity to their place of residence. The validity of this claim is supported by the location of industries and factories in or close the districts known for working-class populations, such as the districts 3, 4, 5 and 6 where industrial plants and factories were situated along the Limmat river and on the western outskirts of the city of Zurich. However, a validation of this assumption would require further investigation.

2.3 Tuberculosis and diphtheria today

Tuberculosis and diphtheria are two infectious diseases caused by different bacteria. Tuberculosis is caused by the bacterium *Mycobacterium tuberculosis* and usually affects the lungs. However, it can also affect other parts of the body. The primary mode of transmission of tuberculosis is via airborne particles expelled by a cough or sneeze of a sick individual. The symptoms associated with tuberculosis include coughing, chest pain, fever, chills, weakness, fatigue and a loss of appetite. It is important to note that tuberculosis can be latent or inactive, and without treatment, approximately 10% of cases will become active. If left untreated, active tuberculosis can be fatal in about 50% of cases (CDC, 2024a). Diphtheria, another serious bacterial infection, is caused by toxins produced by the bacterium *Corynebacterium diphtheriae*. These toxins are capable of killing healthy tissue, most often in the nose, throat, or on the skin. Similar to tuberculosis, diphtheria is spread through airborne particles spread by coughing or sneezing individuals

who are already infected. The symptoms include fever, a pseudomembrane in the nose or throat, sore throat, swollen glands, weakness, pain, rash, redness and swelling of the skin. Respiratory diphtheria can lead to airway obstruction, kidney failure, heart muscle or nerve damage, and thus result in death in approximately 50% of untreated cases and 10% of treated cases (CDC, 2024b).

Research has concluded that there are gender differences in tuberculosis incidence rates and even deaths. A body of evidence has been accumulated showing that men are at higher risk of tuberculosis than women (Humayun et al., 2022; Miller et al., 2021). This phenomenon has been observed in countries across all income levels, as demonstrated by studies conducted by Horton *et al.* (2016) and Peer, Schwartz and Green (2023), who examined data from low- and middle-income as well as high-income countries. However, the reasons for this risk imbalance are not yet fully understood. Explanations include risk factors such as smoking, alcohol consumption and malnutrition, which are typically more strongly connected to the male gender (Humayun et al., 2022; Nhamoyebonde and Leslie, 2014; Peer et al., 2023). Two main hypotheses have been postulated: the behavioural and the physiological or biological. The behavioural hypothesis states that males are more susceptible to tuberculosis due to having a higher number of social contacts, working in high-risk occupations and the risky behaviours mentioned above (Horton et al., 2020; Humayun et al., 2022; Miller et al., 2021; Nhamoyebonde and Leslie, 2014; Peer et al., 2023). Conversely, the physiological or biological hypothesis searches for an explanation in biological differences in immune responses and higher genetic susceptibility (Humayun et al., 2022; Nhamoyebonde and Leslie, 2014). While this research is grounded in recent data, its findings can be extrapolated to a historical context, as many of these fundamental risk factors were evident a century earlier. A parallel observation of a higher tuberculosis prevalence among males was documented at the turn of the 20th century in New York (Nhamoyebonde and Leslie, 2014, p. 100).

In contrast to tuberculosis, diphtheria is not characterised by a gender disparity; rather, it is marked by a more pronounced age disparity. Historically, diphtheria was recognised as a major cause of childhood mortality (Byard, 2013). At the turn of the 20th century, it was a serious childhood disease, with 70% of those infected being younger than 15 years of age (Byard, 2013). However, in the 1930s, a shift in the age demographic of those affected was noted, with the incidence now affecting slightly older age groups, such as adolescents. Concurrently, the mortality rate simultaneously receded (Cheeseman et al., 1939; Dauer, 1950; Picken, 1937).

2.4 History of tuberculosis and diphtheria: a Swiss perspective

Throughout the 19th and the early 20th century, Switzerland experienced a prevalence of multiple diseases, among them tuberculosis and diphtheria. While diphtheria manifested in waves, tuberculosis was a chronic disease. Tuberculosis was identified as the foremost cause of mortality in Central Europe (Ritzmann, 1998, pp. 21–22). However, a decline in the incidence of both diseases was observed in Switzerland throughout the 19th century and into the 20th century. A decline in mortality has been observed since at least the 1880s (Kruker and Senti, 1932; Senti and Pfister, 1946). This decline has been attributed to improvements in living conditions, including better nutrition, sanitation, ventilation and hygiene (Gubéran, 1980; Holloway et al., 2014, 2013; Müller et al., 2024; Ritzmann, 1998, pp. 21–24). This suggests that the increased general immunity in the population resulting from these improved living conditions enables infected individuals' immune systems to adequately combat the disease, thereby reducing case outbreaks (Holloway et al., 2014, 2013; Müller et al., 2024).

The Industrial Revolution and the accompanying urbanisation provided ample breeding grounds for infectious diseases. These diseases proliferated in the unsanitary and overcrowded living quarters of rapidly expanding cities. The substandard and unhygienic working and living conditions experienced by factory workers were conducive to the rapid transmission of diseases. Since the 1880s tuberculosis had reached epidemic status, with the poor populations of highly urbanised regions being especially hard hit. By 1900, absolute tuberculosis numbers peaked, with over 9000 fatalities recorded in the entirety of Switzerland (Kruker and Senti, 1932). The disease burden did not spare the young, with the 15 to 35 age group facing the greatest risk due to their professional activities and social interactions (Corti, 2012; Gubéran, 1980). Indeed, young adults constituted the demographic with the highest number of tuberculosis cases (Ritzmann, 1998, pp. 21–24, 27–31).

During the 1930s, tuberculosis continued to exact a significant toll on lives worldwide, including Switzerland. While significant variations in incidence were observed among individual countries, Switzerland's position in international statistics was in the lower median range. Switzerland lagged behind wealthy industrialised nations such as Germany and the United Kingdom, but outperformed less industrialised countries such as Spain, Czechoslovakia and France. The mortality rate decreased to 124,5 out of 100'000 by 1930, though had reached 264,7 in 1901. In 1931, the number of nationwide tuberculosis fatalities was recorded as 4969 (Kruker and Senti, 1932; Senti and Pfister, 1946).

The treatment for tuberculosis that was available at the time was initially limited to stays in sanatoria located at higher altitudes. There, patients underwent various treatments, including air, lying and dietary treatments (Corti, 2012; Holloway et al., 2013; Ritzmann, 2010a, 1998, pp. 39–42; Rucker and Kearny, 1913; Silberschmidt, 1930). Prior to the advent of antibiotics, effective medical treatment was non-existent, with sanatoria being the sole option available. A specific cure would not be available until 1943 (Corti, 2012;

Holloway et al., 2014, 2013; Ritzmann, 2010a, 1998, p. 141). The earliest sanatoria, established from 1868 onwards, were private institutions, accessible only to the wealthy, such as the sanatoria in Davos, Arosa, Leysin and Crans-Montana, all located in the Swiss Alps (Peter, 2019; Ritzmann, 2017). In response to the widespread nature of the disease, public sanatoria were also established from the 1890s onwards by the cantons themselves and by charitable organisations, such as the sanatoria in Heiligenschwendi, Braunwald and Wald (Peter, 2019; Ritzmann, 2017, 2010a, 1998, pp. 42–44). The proliferation of sanatoria and the treatment of patients therein resulted in an increase in fatalities, with a decrease in the number of deaths in people's homes. By 1930, approximately 15 to 20% of tuberculosis fatalities occurred in sanatoria (Kruker and Senti, 1932; Senti and Pfister, 1946).

The history of medicinal tourism in Switzerland is long and distinguished, with the tradition of sanatoria dating back to the 19th century. During this period, foreigners would travel to the Swiss Alps with the hope of being cured by the allegedly beneficial climate. However, rich foreigners were seldom found in a densely populated city such as Zurich. Instead they favoured the highly prestigious and secluded locations in the Swiss Alps, where the sanatoria catering to such clientele were situated (Ritzmann, 2010a). Consequently, there is no need to be concerned about the potential influence of foreign entities.

Unfortunately, precise global or national numbers for diphtheria could not be found. However, an analysis of publications concerning the city of Zurich reveals that diphtheria has been a notifiable disease for a longer duration than tuberculosis. Furthermore, if the numbers of the city of Zurich are used as a reference, cases have dropped by two-thirds since 1900. From 1916 to 1925, Zurich recorded 2641 cases (128,2 per 100'000) and in the subsequent decade from 1926 to 1933, this figure decreased to an approximated 1405 cases (58,7 per 100'000). The number of fatalities was as low as 90 (4,4 per 100'000) from 1916 to 1925 and 36 from 1926 to 1933 (1,5 per 100'000). The majority of cases is observed in school-aged children. The temporal distribution of cases is also of interest, with the highest number of cases occurring during the winter months (Brunner and Senti, 1937).

Extensive vaccination efforts against diphtheria commenced in the 1940s, resulting in a long-term near-eradication of the disease (Müller et al., 2024; Ritzmann, 2015). In the late 19th century in Geneva, the use of serum as a therapeutic agent for diphtheria patients yielded favourable outcomes, with a significant reduction in mortality from 37% to 5% (Kaba, 2010). The dissemination of this serum appears to have occurred swiftly, and serotherapy was the only medical remedy until the advent of vaccines (Gubéran, 1980, p. 577). Furthermore, Switzerland initiated government controls of serum in 1926 (Kaba, 2010, p. 115), indicating a nationwide utilisation of this therapeutic approach.

In the contemporary era, the prevalence of these two diseases has been significantly reduced due to preventative measures and the advent of effective treatments. Since the advent of their respective cures, both diseases have been combated with great effectiveness, resulting in a significant decrease in their prevalence (Bundesamt für Gesundheit,

2024b, 2024c). However, 550 cases of tuberculosis are still registered annually, although many of these are among migrants from Africa and Asia, where tuberculosis remains endemic. Among the native population, a significant proportion of cases are observed in the elderly, who were infected during their childhood (Bundesamt für Gesundheit, 2024b). No data is available for diphtheria, although sporadic cases continue to be reported, particularly in areas where the vaccine protection is inadequate (Bundesamt für Gesundheit, 2024c).

2.5 Zurich in the 1920s and 1930s



Figure 1: Aerial image of central sea-side Zurich (1925) (ETH-Bibliothek, 2025)



Figure 2: Aerial image of Zurich's industrial outskirts (1929) (ETH-Bibliothek, 2025)

Figure 3 illustrates the 8 districts of Zurich from 1913 to 1933, with a map of Switzerland, indicating the location of Zurich with a small red square. The districts have been numbered in a clockwise direction, commencing with the centrally located district 1 in the medieval city centre, then moving south to district 2, and so forth.

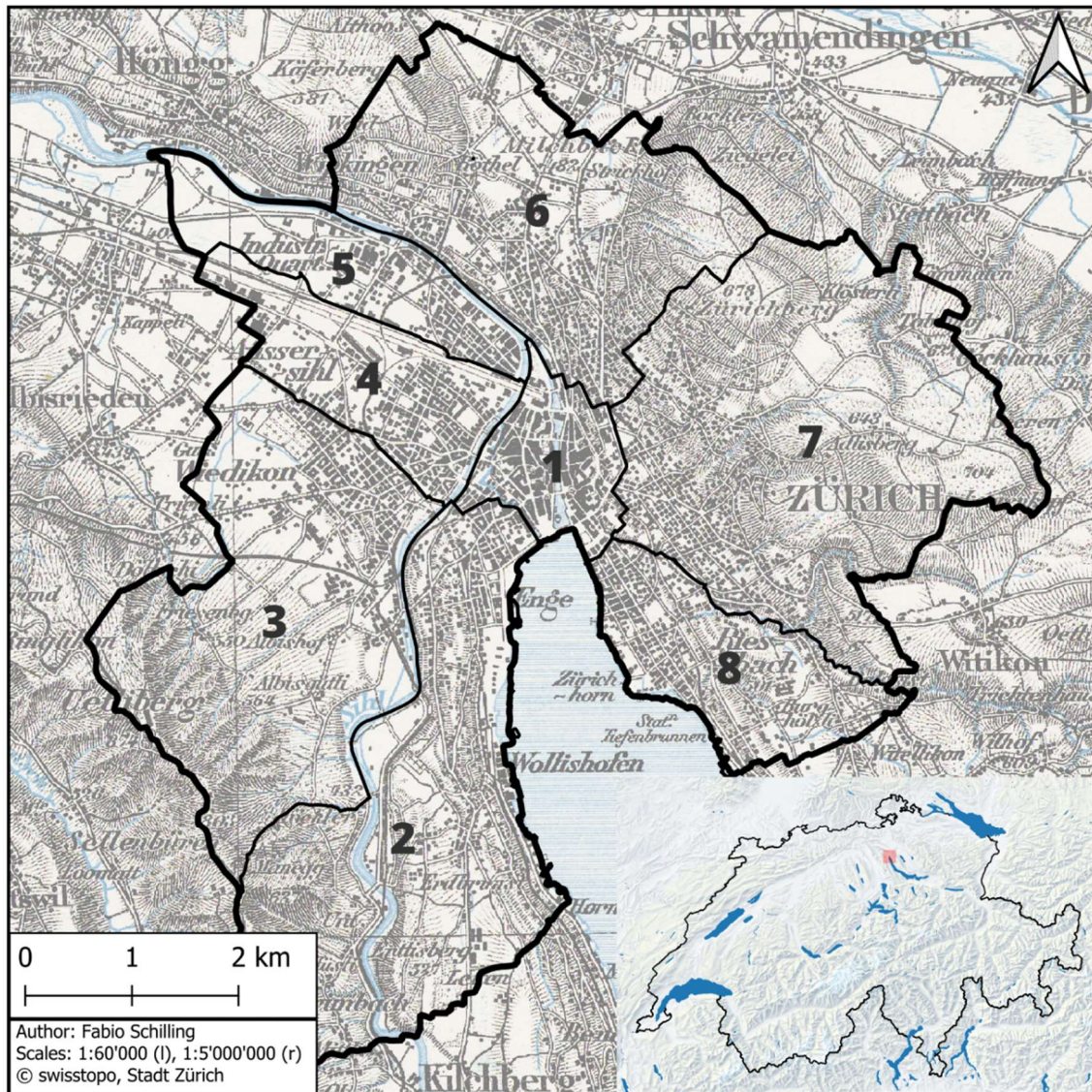


Figure 3: Zurich district map including Swiss reference map

Economic development

At the beginning of the 19th century, the textile industry (silk and cotton) along the Limmat river had dominated the urban economy. The arrival of the railway in 1847 led to Zurich transforming into an important industrial centre and transport hub. Concurrently, the banking and insurance industry was established to finance further developments. By the close of the 19th century, heavy industry had been relocated from the city centre to the outskirts of Aussersihl and Oerlikon, and the textile industry, which had previously dominated, was overtaken by the metal, engineering and electrical industries. Meanwhile, the city also became the financial centre of Switzerland (Behrens et al., 2015).

Population

In the late 19th century, a number of communities in the vicinity of Zurich were experiencing financial difficulties, which led to a desire to join the city. This was driven by the need to alleviate the administrative and financial burdens arising from rapid population growth. This eventuality was realised with the first city unification in 1893 (Behrens et al., 2015; Schulamt Stadt Zürich, 2016). Municipalities with a high proportion of socioeconomically disadvantaged residents, i.e. predominantly working class, such as Wiedikon and Aussersihl (later districts 3 and 4), vigorously advocated for this integration, while more affluent municipalities initially opposed it (Rebsamen et al., 1992, pp. 274–276; Stadt Zürich, 2018a).

Zurich's population grew exponentially during the 19th century, rising from approximately 10'000 inhabitants in the historic city centre in 1800 to around 150'000 in 1900 and 250'000 by 1930 (Statistik Stadt Zürich, 1931). The ongoing growth of the city by 1930 is primarily driven by population influx (Senti, 1928). During the period referred to as the “Goldene Zwanziger” (the German term for the Roaring Twenties), the city of Zurich underwent rapid growth. An extensive second expansion of the city was unsuccessful in 1929 due to the resistance of wealthy municipalities along the shores of Lake Zurich. Consequently, a compromise was reached with the negotiation of the second city unification in 1931, which was subsequently implemented in 1934 (Brodbeck and Hermann, 2012; Rebsamen et al., 1992; Schulamt Stadt Zürich, 2016; Stadt Zürich, 2018a).

The urban development map of Zurich in figure 4 the temporal evolution of the city with colour-coded initial development stages. For the purposes of this study, all phases up to the 1930s are of interest; however, as these are included in the 1910-1940 step, it can be assumed that all areas of Zurich up to the red, excluding the orange and yellow, had been developed by the time of the study. The most significant disparities are observed in western and northern Zurich, which, fortunately, are mostly outside the reference of the administrative boundaries of the city until 1934, with the exception of districts 3, 4 and 5.

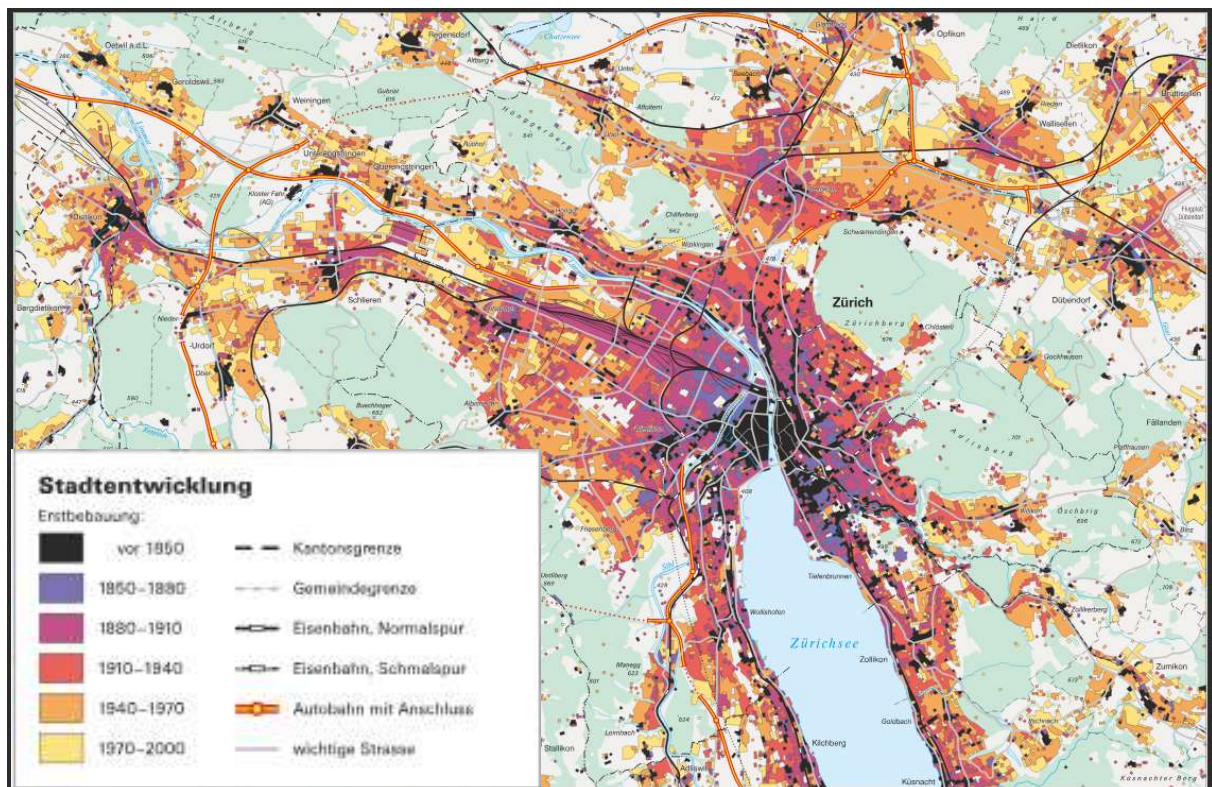


Figure 4: Urban development of Zurich since 1850, adapted from (Brodbeck and Hermann, 2012) and (Behrens, Motschi and Schultheiss, 2015)

Life expectancy in Switzerland at birth in 1930 was approximately 60 years, whereas in 2023 it is recorded as being over 82 years. However, this figure is arguably an underestimation due to the impact infant mortality. For instance, individuals born in 1917 could expect to reach 63 years of age and even 70 years of age for men and women, respectively. By 2017, this figure had increased to 91 and 94 years, respectively. Life expectancy at age 65 remained at 11 to 12 years in 1930, while it has increased to 20 to 22 years in 2023 (Bundesamt für Statistik, 2024)

The population numbers of the city of Zurich have been drawn from the statistical yearbook of the city of Zurich of 1930 (Statistik Stadt Zürich, 1931) and the city portrait (Stadt Zürich, 2018b). The age pyramid in figure 5 illustrates the relative population distribution of the city of Zurich in 1930 compared to today, symbolised by the dotted black line. The population in 1930 was characterised by a younger demographic, with women outnumbering men, beginning at approximately 15 years of age.

Alterspyramiden

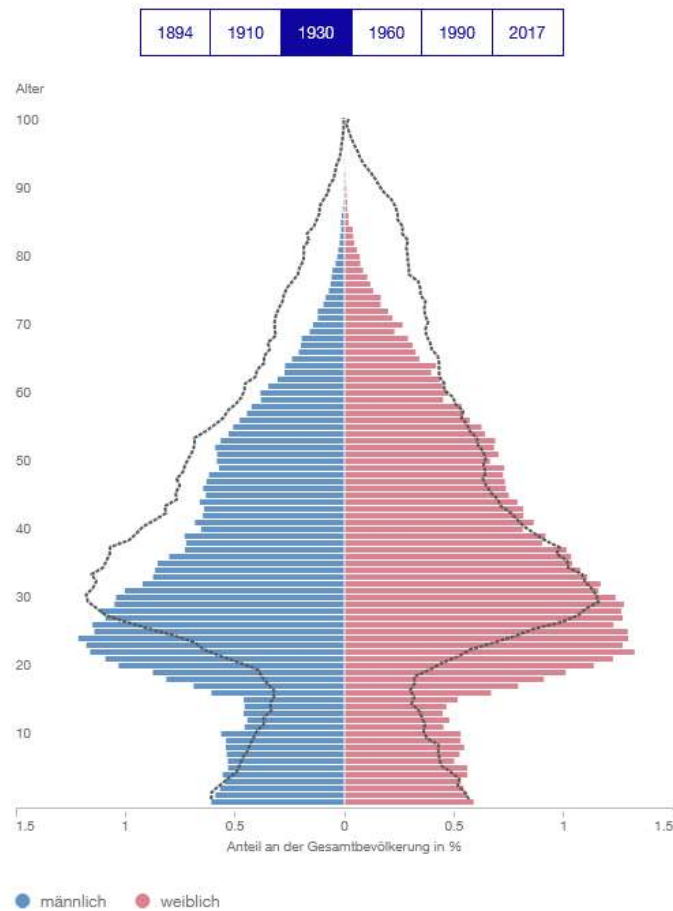


Figure 5: Relative age distribution pyramid of Zurich in 1930 (Stadt Zürich, 2018b)

The city's population is segmented into five-year age groups in the statistical yearbooks. For the purpose of clarity, these groups have been condensed into ten-year groups. Table 1 illustrates the population of the aforementioned age groups, ranging from 0-9 years of age to 60+ years of age. The largest single group is that of the 20-29-year-olds, constituting over 57'000 of a city population of almost 250'000 (23.6%). The second largest age group is the next older 30-39-year-olds with over 45'000 individuals (18.6%). The remaining age groups comprise between 22'000 and 30'000 individuals, with the exception of the 40-49 age group which includes 34'000 individuals (14%). The smallest age group is the elderly, here condensed into the 60+ age group, comprising 22'000 individuals (9%). A remarkable fact is that only in the age group 0-9 are more people of the male gender, while in all the other cohorts, the female gender dominates.

Table 1: Zurich's population subdivided into age cohorts (Statistik Stadt Zürich, 1931)

Age	Males	Females	Total	Share of total
0-9	14072	13884	27956	11.4%
10-19	14341	15456	29797	12.2%
20-29	27434	30222	57656	23.6%
30-39	20624	24849	45473	18.6%
40-49	15456	18799	34255	14%
50-59	12469	14994	27463	11.2%
60+	8697	13344	22041	9%
Total	113093	131548	244641	

When the population is subdivided into districts (see table 2), it becomes evident that the largest districts by population are districts 3, 4 and 6. Each of these districts houses over 15% of the city's population, totalling approximately 38'000 at least. In contrast, district 7, with a population of approximately 31'000 accounts for 12.5% of the city's total population. All other districts have a population of less than 25'000, thus constituting below 10% of the city's total population.

Table 2: Zurich's population subdivided by districts (Statistik Stadt Zürich, 1931)

District	Population	Share of population
1	21'824	8.7%
2	21'719	8.7%
3	38'716	15.5%
4	38'771	15.5%
5	17'613	7.1%
6	56'420	22.6%
7	31'137	12.5%
8	23'620	9.5%
Total	249'820	

Health

Statistics published by the city of Zurich demonstrate that a reduction of mortality from both tuberculosis and diphtheria has been ongoing since the 1880s. The canton of Zurich proudly claimed the rank of most effective tuberculosis controller by the 1930s, achieving the lowest tuberculosis mortality rate among all cantons (Ritzmann, 1998, p. 36). The City of Zurich meticulously collected district-based statistics on diseases. The districts in Zurich are particularly well-suited for statistical analysis as they align closely with the social conditions of their inhabitants, thereby facilitating insights into the correlation between poverty and disease (Ritzmann, 1998, pp. 36–37).

The city's statistics demonstrate that tuberculosis was the more deadly of the two diseases in question, affecting more than 100 per 100'000 people, in contrast with only 2 for diphtheria. Moreover, the data indicate that since the turn of the century, tuberculosis fatalities have fallen by about two-thirds and diphtheria fatalities by almost 90%. While the city experienced approximately 200 fatalities due to tuberculosis, diphtheria accounted for a mere 6 fatalities in 1928 (Kruker and Senti, 1932; Senti, 1928; Senti and Pfister, 1946; Statistik Stadt Zürich, 1931). Mortality by cause of death and per 100'000 is markedly higher for tuberculosis, at 84 for lung tuberculosis and 22 for tuberculosis of other organs, which adds up to 106. Conversely, diphtheria exhibits an exceptionally low mortality rate of 2. By 1930, the districts 3, 4 and 5, characterised by their industrialisation and substantial working class populations, experienced the highest tuberculosis mortality rates (Ritzmann, 1998, pp. 36–37; Statistik Stadt Zürich, 1931). The historical city centre of Zurich was also characterised by the same unfavourable living conditions, although there was no industrial presence here (Ritzmann, 1998, pp. 36–37).

In the 1930s, tuberculosis was among the most prevalent infectious diseases, with an annual incidence of approximately 500 to 600 cases in the city of Zurich alone. In comparison with other Swiss cities and with the world average, Zurich exhibits a tuberculosis mortality rate of 91,2 per 100'000 inhabitants (Kruker and Senti, 1932; Senti and Pfister, 1946). This figure represents the lowest recorded numbers in any major Swiss city, with Basel, Berne and Geneva all demonstrating higher rates of mortality. However, there are a number of cities with even lower mortality rates than Zurich, many of which are located in Germany and the USA (Brunner and Senti, 1937). The establishment of the public Zurich sanatorium had been opened in the community of Wald in the Zurich Oberland in 1898 was a significant development in the response to tuberculosis. The sanatorium was dedicated to providing free or affordable treatment to those, who were financially disadvantaged (Peter, 2019; Ritzmann, 2017). Significantly, the 1928 legislation mandating the reporting of tuberculosis cases also incorporated the stipulation that individuals diagnosed with tuberculosis should receive treatment regardless of their financial status (Holloway et al., 2013, p. 81). This suggests that, in principle the entire population of Zurich had equal access to treatment in sanatoria. The forest school at the Zürichberg was inaugurated in 1914. During the summer months, approximately 50 children at risk of tuberculosis were able to attend this school, which primarily took place outdoors. The time spent outside

and the location being free of pollution should have strengthened the accepted children, who were of fragile nature and descending from families in need, i.e. mostly working-class backgrounds (Matthijs, 2024). This suggests that at least some form of social awareness regarding tuberculosis was already present during this period, indicating the potential for well-directed interventions. The impact of equal treatment by legal mandate and well-directed interventions on this phenomenon is difficult to estimate, but it shows that there was some knowledge of the influence of different socioeconomic backgrounds and that steps were taken to counteract this.

By the 1930s, diphtheria appeared to have lost much of its former severity. In the 50 years prior, incidence and fatalities were drastically reduced (Brunner and Senti, 1937; Statistik Stadt Zürich, 1931). Nevertheless, the majority of cases continued to affect school-aged children, who, in addition to being more susceptible to infection, were also particularly exposed to the disease within the school environment. The decline in diphtheria cases was evident across the entire city of Zurich (Brunner and Senti, 1937). The protective efficacy of the diphtheria vaccination programme was documented as early as 1937, and it is plausibly that this initiative contributed to the observed decline in cases. Seasonal variation is evident, with higher numbers and mortality observed during the colder months (Brunner and Senti, 1937).

2.5.1 The 8 districts of Zurich

The city of Zurich is located at the outflow of lake Zurich where the Limmat river starts its course and merges with the Sihl river shortly after. The city is surrounded by several hills, namely the Uetliberg to its west, the Zürichberg to its east and the Käferberg and Höggerberg to the north. The city is located on the floodplain of the Sihl and along the banks of the Limmat rivers and the lakeshores.

The city of Zurich consisted of 8 districts (formerly 5 up to 1913) shown in the map in figure 6 together with the built-up areas. These had been formed from the villages and municipalities surrounding the city core of Zurich in 1893. Many of them were in financial trouble and their means were overstretched because of fast population growth. This is why the city was expanded to include several directly neighbouring municipalities in the first city unification. A cantonal vote was held, and overwhelming support was found for the merger. The support was especially strong in the affected, financially strained municipalities.



Figure 6: Contemporary map of the city of Zurich including administrative districts (Statistik Stadt Zürich, 1931)

The unification should ease the administrative burden and allow for better coordination of resources. Some wealthy municipalities such as Wollishofen and Enge opposed this move but were forced to abide by the federal court. The 8 districts would stay unchanged until the second city unification of 1934 and from then on there would be 11 districts. Eventually Schwamendingen would be split off from district 11 in 1971 to form the 12 districts known today (Rebsamen et al., 1992; Schulamt Stadt Zürich, 2016).

Table 3 is intended to provide a concise overview of the main characteristics of each district, as well as the social class that is its defining feature. The descriptions are based on individual district descriptions from Rebsamen, Bauer and Capol (1992, pp. 272–287).

Table 3: Main characteristics of each district and defining social class

District	Characteristics
1 (Altstadt)	<ul style="list-style-type: none"> - Medieval city centre - Mixed affluent and poor neighbourhoods
2 (Enge, Wollishofen, Leimbach)	<ul style="list-style-type: none"> - Historical residential buildings of middle- and upper-class - Among wealthiest districts of Zurich
3 (Wiedikon)	<ul style="list-style-type: none"> - Densely populated working-class district - Strong industrialisation
4 (Aussersihl)	<ul style="list-style-type: none"> - Rapidly developed to an industrial hub - Densely populated working-class district - Close to railways
5 (Industriequartier)	<ul style="list-style-type: none"> - Strong industrial presence - Close to Limmat river and railways - Working-class district
6 (Oberstrass, Unterstrass, Wipkingen)	<ul style="list-style-type: none"> - Suburban, working-class settlement structure and apartment buildings - Strong population growth since industrialisation
7 (Fluntern, Hottigen, Hirslanden)	<ul style="list-style-type: none"> - Best residential locations of Zurich - Undeveloped areas of forest for leisure - Representative, affluent suburban development
8 (Riesbach)	<ul style="list-style-type: none"> - Once middle-class country houses and natural idylls - Representative development along lakefront - Suburban, middle-class neighbourhood

2.5.2 Socioeconomic characteristics of districts

The districts of the city of Zurich can thus be classified into socioeconomic classes depending on the demographic composition of their inhabitants. The three classes are working- or lower-class, middle-class and upper-class. Working class districts are distinguished by high population density and a low-income structure. Conversely, upper-class districts are distinguished by lower population density and a higher income structure. Middle-class districts, as intermediate zones, exhibit median population densities and average income structures. The following districts can be characterised as working or lower class: districts 3, 4 and 5 and the Niederdörfli neighbourhood of district 1. Conversely, the districts 1, 2 and 7 are identified as the primary upper-class districts of the city of Zurich. The remaining districts of Zurich, namely 6 and 8, are assumed to be middle class, although their structure may vary, and they can be classified as either lower- or upper-class districts, depending on the specific neighbourhood.

The map below (figure 7) illustrates the population density of the city of Zurich in 1930. As is clearly evident, the northeastern districts (specifically 4, 5 and 6) are collectively with district 1 the most densely populated of the city. This observation lends further credence to their categorisation as districts of lower socioeconomic status. Conversely, the more affluent districts 2, 7 and 8, are distinguished by the comparatively lower population densities, attributable to the preponderance of wealth.

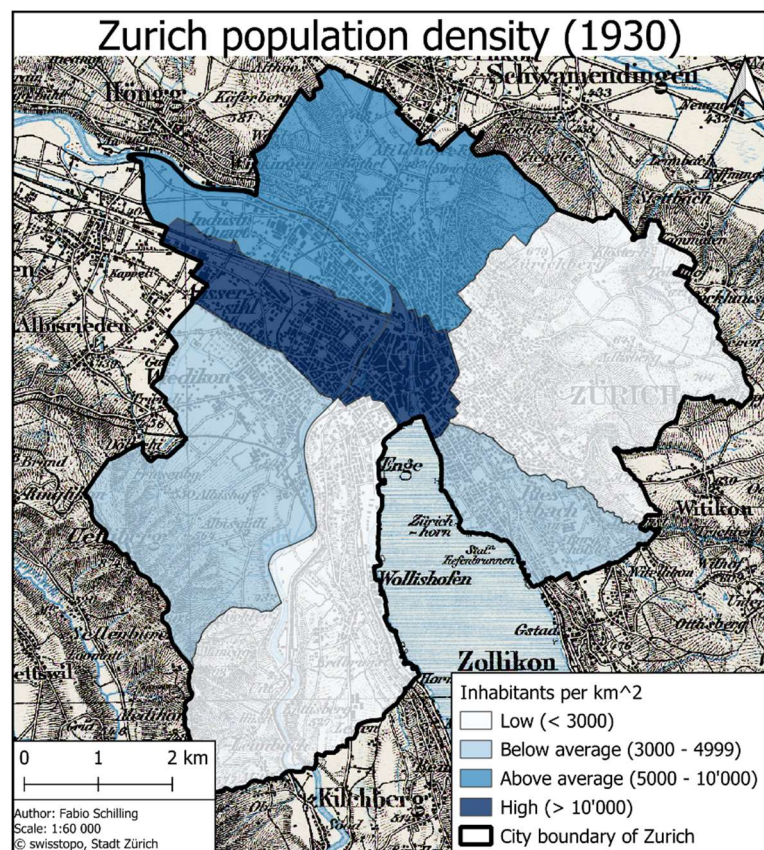


Figure 7: Population density map

The correlation between building insurance values (figure 8) and population density, as well as rental prices (figures 9 and 10), is indicative of a divergent relationship. Districts 1, 4 and 5 now display high values. The district with the highest values is district 1, which is characterised by historic buildings in the medieval old town and the splendour of highly prestigious streets, such as the Bahnhofstrasse and the main station. This distribution is likely attributable to the higher valuation of apartment buildings, predominantly present in working-class districts 4 and 5, in comparison to single-family residences and villas found in affluent districts. The presence of factories in the districts of 3, 4 and 5 may also

be a contributing factor, given that these factories can be quite valuable buildings. The middle to upper class districts 2 and 8 display average values, which can be attributed to their more expensive and urbanised building styles. The districts 3, 6 and 7 exhibit low values, with two of these (3 and 6) being intuitive as they contain a mixture of apartment and suburban buildings, resulting in lower values. District 3 has the lowest values.

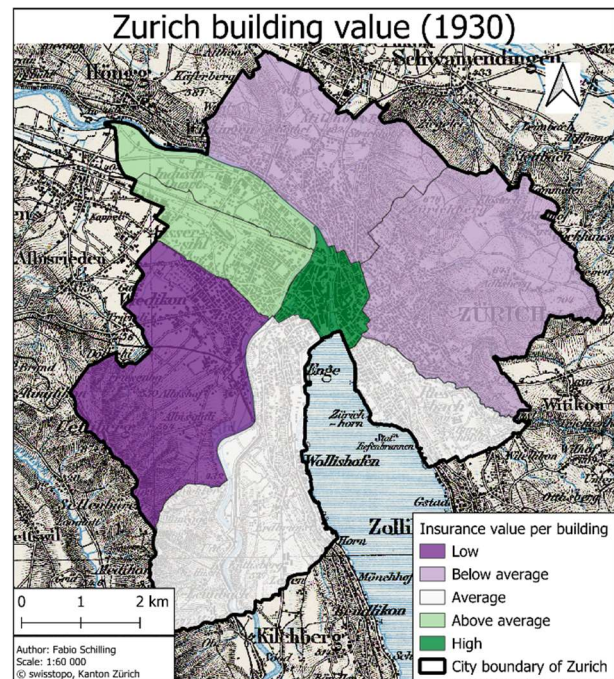


Figure 8: Building insurance value map

This phenomenon is illustrated in the subsequent maps depicting rental prices (figure 9 and 10). The working-class districts 3, 4 and 5 are the cheapest and most affordable districts in terms of rents. Conversely, the affluent districts 7 and 8 are among the most expensive residential districts in Zurich, along with district 1. Districts 2 and 6 are positioned in the median range of the rent price spectrum, aligning with their respective average population densities and suburban characteristic.

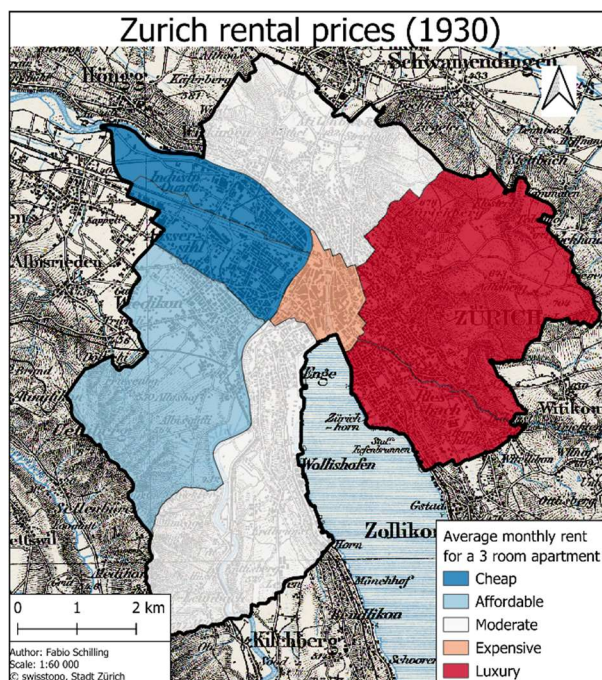


Figure 9: Rent map for 3 room apartments

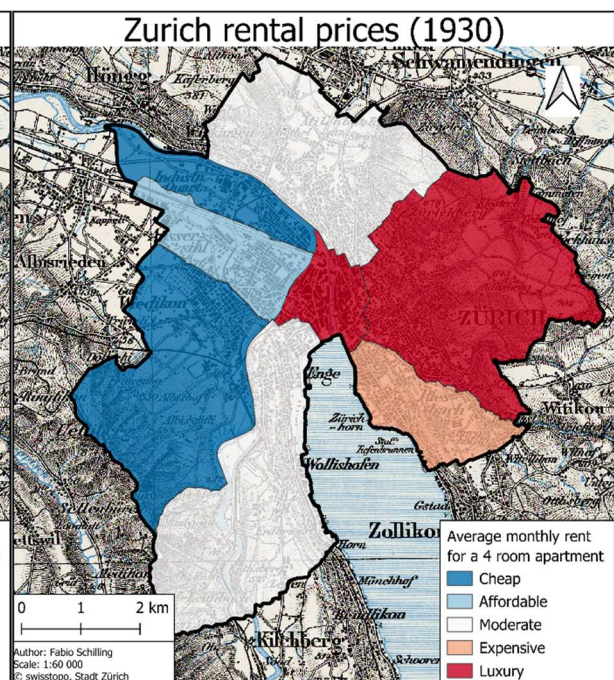


Figure 10: Rent map for 4 rooms apartments

3 Data

The data utilised in this thesis was reported by physicians in the city of Zurich based on legal obligations (Bundesamt für Gesundheit, 2024a). The obligation concerning tuberculosis was established in 1928 and came into effect in 1932 (Holloway et al., 2013, p. 81; Kruker and Senti, 1932; Senti and Pfister, 1946). Since 1876, the Swiss cause-of-death statistics have obliged physicians to register causes of death on a national scale since 1876 (Ritzmann, 2010b, 1998, pp. 24–27). The data was collected and registered by the health authorities of the City of Zurich, and was transcribed as part of an ongoing doctoral thesis in medicine at the Institute of Evolutionary Medicine of the University of Zurich (IEM, 2024). For the years 1927 to 1935, the original sources were photographed and digitized in the form of a Microsoft Excel spreadsheet. The data contains a wealth of information which will be presented in the next section.

3.1 Disease datasets

The images below (figures 11 and 12) illustrate the original data on tuberculosis cases. The information is documented in a notebook with prearranged pages, columns and rows, thus facilitating data capture. The tuberculosis dataset under consideration herein ranges from the start of 1932, when the reporting obligation came into effect, to the end of 1935, and includes 2046 entries. The following table 4 illustrates the attributes, their descriptions and an example. The original column names have been included for reference. It should be noted that not all columns were utilised nor required for the analysis. The columns marked “hospital admission date”, “date of death” and “tuberculosis type” were excluded from the analysis. The remaining columns will be utilised directly, with the exception of “district”, which will function as a control variable or reference during geocoding of addresses.

Nr	Zeit wann angeführt krank	Datum der Anzeige	Familienname	Vorname	Alter	Dist. bei Kindern der Eltern bzw. der Besorger	Kreis Quartier	Medizinischer Distrikt	Straße	Nr. des Haus- wirts	Spital- behandlung und Datum des Eintritts	Anzeige durch	Zugereist wahr?
✓	9/1/32	23. 6. 1932	Edermatt-Bernelli	Hans	56 J.	Tuberculum		4	Bäcker	199		Dr. Zingg	Ne. post. u. post.
✓	19/1/32	23. 6. 1932	Wossi	Arbano	47 J.	Kein		4	Bäcker	249	Wassgasse Kornhof	Dr. Zingg	Ne. post.
✓	15/1/32	28. 6. 1932	Allegel	Klara	33 J.	Kein		4	Rebasse	5	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	12/1/32	30. 6. 1932	Widmann	Emil	18 J.	Appelwein		4	Bäcker	219	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	26/1/32	3. 7.	Wider	Elise	17 J.	Kein		6	Widener	50	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	28/1/32	12. 7.	Lehli	Udo	30 J.	Kein		4	Hohl	355	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	29/1/32	13. 7.	Lehli	Frieda	23 J.	Kein		1	Niederdorf	67	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	28/1/32	12. 7.	Lehli	Emil	33 J.	Kein		1	Wite Gasse	3	Wand Dr. Zingg	Dr. Zingg	Ne. post.
✓	30/1/32	20. 7.	Wassgasse-Bolliger	Anna	19 J.	Kein		3	Tuberculum	149	Wand Dr. Zingg	Dr. Zingg	Ne. post.

Figure 11: First part of the original tuberculosis dataset

19 32

Kreismed. Bezirk	Straße	Nr. und Stad- werk	Spital- behandlung nach Datum des Eintritts	Anzeige durch	Zugereist wahr?	Mutmaßlicher Ansteckungsort	Sind nach andere Kinder verhanden?	Datum der ärztlichen Abmeldung	Datum der letzten Ab- schuppung, aus letztem Belag	Datum der Des- infektion	Aufhebung der Verfügung für die Gesunden	Bemerkungen
4	Bärber	199		Dr. Guggenberger	Ne. perlon gesten		ja			29.11.32	29.11.32	
4	Bärber b. Krenn	248	Kreuzgasse Häuslerhof	Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
4	Rebgasse	5	Wald b. Krenn	Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
4	Bärber	219	Wald b. Krenn	Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
6	Häuslerhof	50		Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
4	Holl	335		Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
1	Häuslerhof	62		Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
1	Wald Gasse	3		Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	
6	Häuslerhof	148		Dr. Krenn	Ne. perlon		ja			23.11.32	23.11.32	

Figure 12: Second part of the original tuberculosis dataset

Table 4: Transcribed attributes of tuberculosis data

Attribute name	Description	Example
Date of registration (Anzeigedatum)	Date	05.01.1932
Gender (Geschlecht)	Categorical	m
Age (Alter)	Number	30
Occupation (Beruf)	Text	Hilfsarbeiter
District (Kreis / Med. Distrikt)	Number	3
Street name (Strasse)	Text	Birmensdorfer
House number (Hausnummer)	Number	120
Hospital admission (Spital)	Categorical	1 [yes]
Hospital admission date (Spitaleintritt)	Date	11.01
Death (Gestorben)	Categorical	Yes
Date of death (Wann gestorben)	Date	27.11.1933
Tuberculosis type (Tuberkuloseart)	Text	Lungentuberkulose

Nr.	Zeit wann ausget. krank	Zeitraum der Krankh.	Familienname	Vorname	Alter	Bezt. bei Ausbr. der Epid. nach der Dauer	Erst krankh.	Wohnort Ort	Straße	Zeit mit Dau- erkr.
1	18. 10. 18	22. 10.	Eger	Peter	1. 7.	Köcher	1. 7.	1. 7.	Postplatz 11	11
2	18. 10. 18	22. 10.	Mayer	Eugen	6. 2.	Pacher	6. 2.	1. 7.	1. 7.	11
3	18. 10. 18	22. 10.	Böhm	Karl	1. 7.	Möller	1. 7.	1. 7.	1. 7.	11
4	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
5	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
6	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
7	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
8	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
9	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
10	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
11	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
12	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
13	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
14	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
15	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
16	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
17	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
18	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
19	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
20	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
21	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
22	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
23	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
24	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
25	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
26	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
27	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
28	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
29	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
30	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
31	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
32	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
33	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
34	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
35	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
36	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
37	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
38	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7.	1. 7.	11
39	18. 10. 18	22. 10.	Hoffmann	Karl	1. 7.	Böhm	1. 7.	1. 7		

25

Table 5: Transcribed attributes of diphtheria data

Attribute name	Description	Example
Disease start date (Krankheitsbeginn)	Date	16.03.
Date of registration (Anzeigedatum)	Date	20.03.1927
Gender (Geschlecht)	Categorical	w
Age (Alter)	Number	8
Occupation (Beruf)	Text	Käser
District / neighbourhood (Kreis/Quartier)	Number and Text	1 Z
Medical district (Med. Distrikt)	Number	1 16
Street name (Strasse)	Text	Peterhofstatt
House number (Hausnummer)	Number	11
Hospital admission (Spital)	Categorical	Ja
Hospital admission date (Spitaleintritt)	Date	20.03.
Outbreak centre (Ansteckungsherd)	Text	Wipkingen
Deregistration date (Abmeldungsdatum)	Date	30.04.
Death (Tod)	Categorical	nein
Date of death (Wann gestorben)	Date	27.02.1928
Diphtheria type (Diphtherieart)	Text	Mandeln

3.2 Geodata

The geodata utilised in this thesis encompasses the administrative districts and neighbourhoods of Zurich, derived from the city's geoportal (Stadt Zürich, 2025a, 2025b), in addition to Lake Zurich (Stadt Zürich, 2025c). Schools were selected as points of interest in conjunction with diphtheria case data, and the location of these is available for reference (Stadt Zürich, 2025d). Furthermore, a document exists which lists the construction year of each school, thereby allowing the extraction of those which were already built and in use at the time (Amt für Städtebau, 2008). All of the aforementioned geodata were provided and downloaded in ESRI shapefile format. To provide context on the Swiss national scale, the Swiss Map Vector dataset was used for the Borders and background map of Switzerland (swisstopo, 2024a). The historical background maps employed for the visualisations were obtained from the Federal Office of Topography and include the map sheets 158 Schlieren, 159 Schwamendingen, 160 Birmensdorf and 161 Zürich of the so-called "Siegfriedkarte" in the edition of 1930. The map is presented at a scale of 1:25'000, offering a high level of detail in smaller areas (swisstopo, 2024b). The map incorporates the map sheet 008 Aarau, Luzern, Zug, Zürich of the so-called "Dufourkarte". The scale of this map is 1:100'000, which provides less detail but is adequate for general overviews. These were provided as georeferenced tiff files (swisstopo, 2024c).

3.3 Statistical data

Data concerning individual districts and city-wide characteristics was drawn from different editions of the statistical yearbook of the city of Zurich (Stadt Zürich, 2024). The population numbers, district areas, number of buildings, insurance value of buildings, housing data and rental prices referring to the year 1930, found in the yearbooks of 1930 and 1935 (Statistik Stadt Zürich, 1936, 1931), were used. The data used for the HISCO classification was kindly provided by my supervisor Prof. Dr. phil. Kaspar Staub.

4 Methods

Regarding the supporting tools employed in this thesis, the AI-based ChatGPT and DeepL applications were incorporated to enhance the outcome. ChatGPT helped in writing code for the analysis as well as troubleshooting any errors encountered and interpreting the statistical outputs. As a non-native speaker, I relied on DeepL to improve my writing style and make it more concise. Although these technologies aided in the research, all decisions regarding interpretation and methodological choices were made autonomously.

The general data processing of this master's thesis is outlined in the adjacent figure 14. The data processing is comprised of four primary steps. Initially, the addresses were extracted from the original data through concatenation of street names and house numbers. Secondly, geocoding is employed to locate these addresses in space. This step is of paramount importance in ensuring the integrity of the geocoding process, which in turn, contributes to the overall data quality. Subsequently, the paths diverge, as the data can either be visualised most likely in the form of maps or processed for analysis. The visualisation was mainly conducted using the open-source software QGIS. The analysis was executed on various platforms, including GeoDa, R and Python.

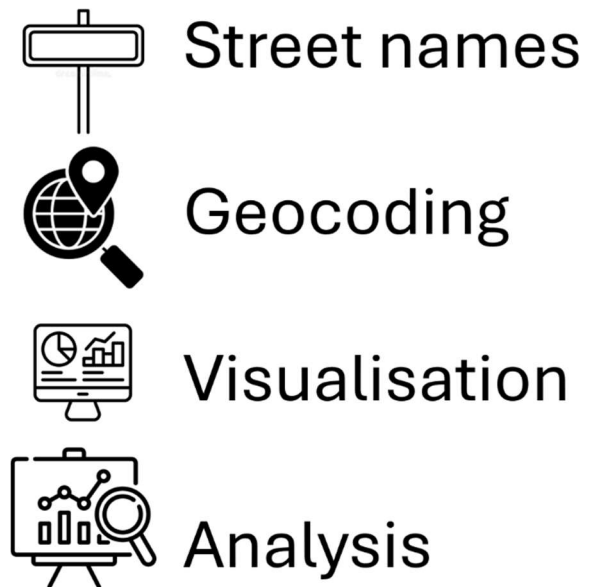


Figure 14: Data manipulation steps (adapted by the author)

4.1 Geocoding residential addresses

Geocoding, the process of transforming locational information such as addresses into geographic coordinates, is a widespread methodology to extract locational information from non-spatial data (Kirby et al., 2017).

The necessity to correct the display dates, by adding the year, was evident in the majority of cases where this information was absent. This correction was deemed essential for subsequent time-based analysis of the case chronology.

In order to geocode, i.e. extract a location, from the data at hand, three general steps were taken. Firstly, the columns marked “street name” and “house number” were linked, with the suffix “street” appended to the street name. The city (Zürich) was then added to the address to ensure the correct location. Secondly, general exceptions were defined in order to avoid adding the suffix “street”, if it was already present in the “street name”. Finally, the more fine-grained exceptions had to be identified in the data and subsequently implemented in the formula. Notably, in Zurich there exist addresses which are lacking the term

“street”, as found in squares, yards and alleys. The comprehensive description of the exact geocoding methodology can be found in the appendix.

Utilising this formula, the majority of addresses can be compiled for further utilisation in geolocation. The geocoding process was facilitated by an Excel tool provided by the Federal Office of Topography (swisstopo, 2023). This tool enables the entry of addresses, subsequently returning the corresponding coordinates in two distinct coordinate reference systems (CRS). The first is the official Swiss CRS, known as CH1903+ / LV95 (national survey of 1995), and the second is the well-known WGS84 (World Geodetic System 1984).

In the context of the tuberculosis data, this workflow of methods returns 2002 out of 2046 addresses, with the remaining 44 entries lacking addresses. However, approximately 500 addresses fail to yield any results, either due to non-localisable house numbers, incorrect street names or streets that have been renamed since. Manual inspection of the addresses that could not be geolocated reduced the number to 121. In addition, one address was removed because it was in Lausanne, resulting in 1922 geolocated addresses.

The data was then reduced to the extent of the city of Zurich before 1934 to guarantee the continuity of the reference frame. This process resulted in a reduction of the dataset from the original 2046 to 1852 entries, representing a loss of 194 entries. Additionally, the number of geographic locations decreased from 1922 to 1762, marking a loss of 160. Concurrently, the mortality data was reduced from 378 to 357 entries, representing a loss of 40 entries, and from 361 to 338 locations, representing a loss of 23 locations.

For the diphtheria data, an initial 670 entries yielded 468 addresses. Following the implementation of manual corrections and the deciphering of a few more addresses, 469 addresses were located within the city of Zurich out of a total of 477 addresses. However, it was noted that approximately 200 addresses were not present in the transcription of the original records and thus could not be incorporated into the subsequent visualisation and analysis.

4.2 Defining the 8 districts of Zurich

Firstly, the area of Lake Zurich needed to be subtracted from the shoreline districts. This was done, as the substantial lake area does not feature any habitation and is consequently, irrelevant for the analysis. In light of change to the districts, it was necessary to adapt them to correspond with the extents of the 1913 to 1933 period. The internal restructuring of the five districts (I through V) established in the first city unification of 1893 gave rise to eight districts (1 through 8) in 1913, through the reorganisation of neighbourhood or the replication of existing districts, which have remained predominantly unaltered to the present day. Districts 1, 2, 4 and 8 have retained their original boundaries. In contrast, the remaining four districts have undergone substantial changes since the aforementioned period. The district of Wiedikon (district 3) has undergone changes in its boundaries in the vicinity of the Triemli hospital, as illustrated in the accompanying map (figure 15). However, these boundary shifts have occurred in two distinct directions: the

district has expanded towards the Triemli hospital, while the border has been moved in the opposite direction, just north of the hospital.

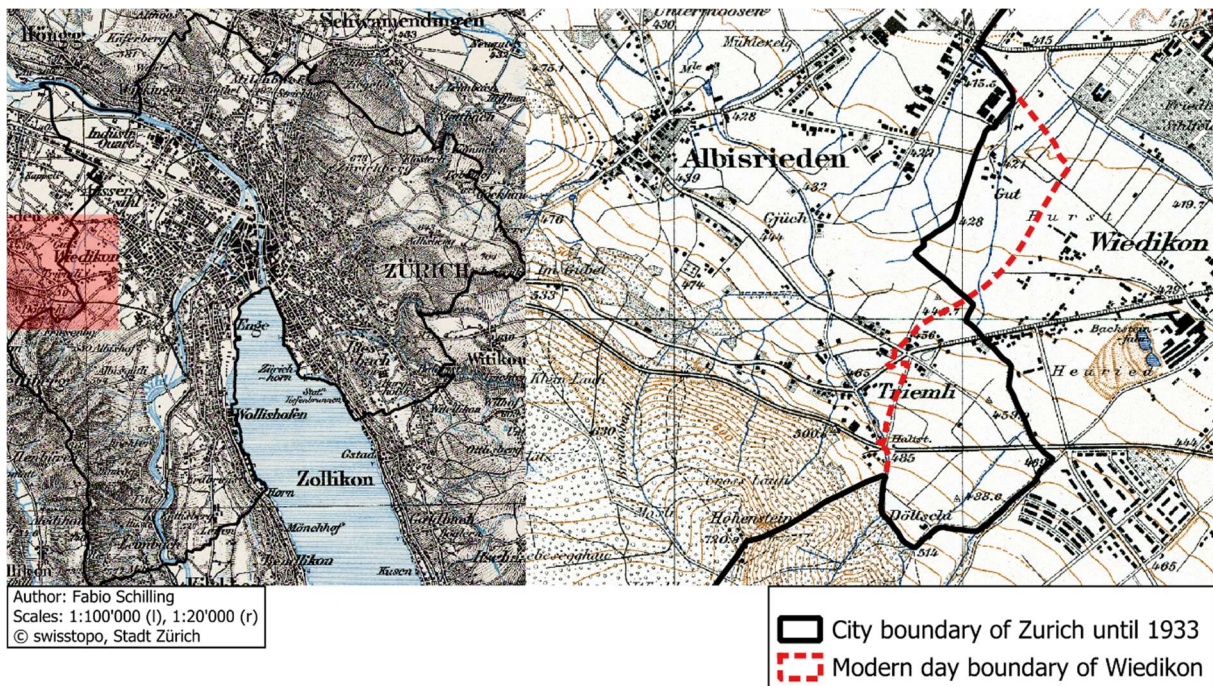


Figure 15: Border changes of district 3

The administrative boundaries of district 5 have been modified in the Hard area, situated in the extreme northwest of the district and city. Previously, the districts encompassed the area known as the “Hardhof” (refer to map label for details). The area in question today is part of district 9. The alteration in the district’s boundaries is illustrated in the accompanying figure 16.

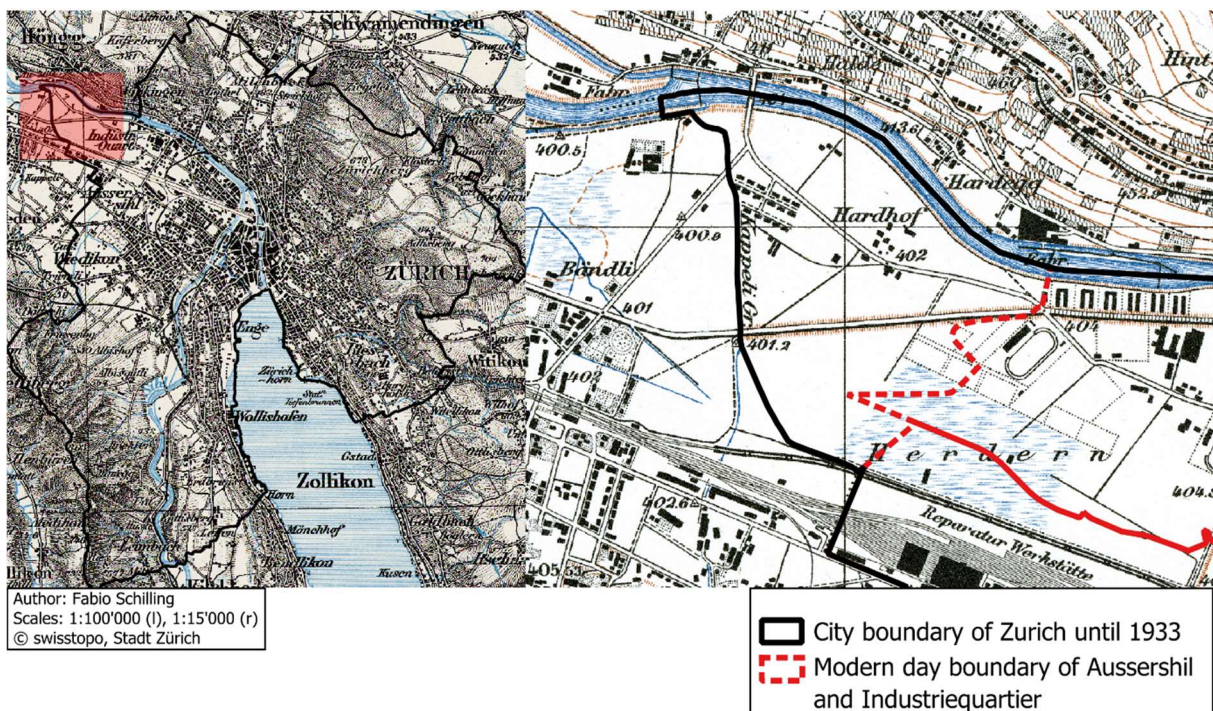


Figure 16: Border changes of district 5

The area of district 6 used to include the neighbourhood of Wipkingen represented by a dashed red line and accompanied by a corresponding map label (see figure 17). In 1934, Wipkingen was incorporated into district 10, along with its western neighbour, Höngg.

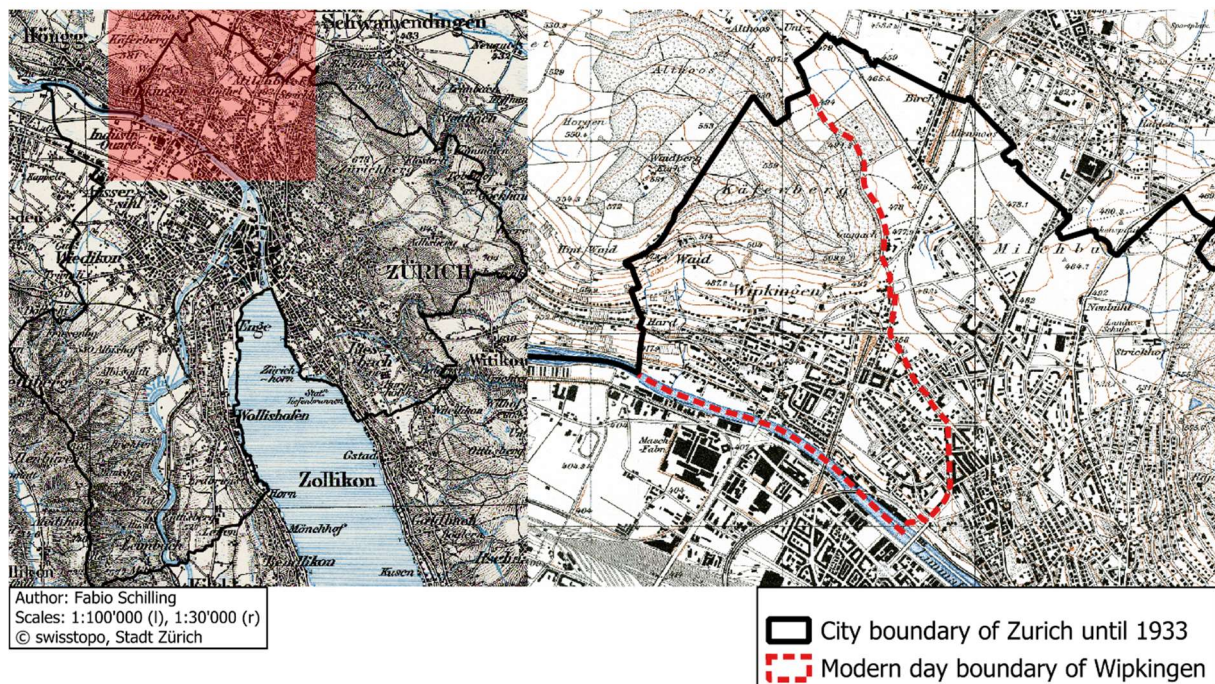


Figure 17: Border changes of district 6

A comparable situation is found in district 7, where the neighbourhood of Witikon was only added in 1934, but the hamlet of Eierbrecht (see map label), which is now part of the neighbourhood of Witikon, already belonged to the city of Zurich. The divergence is highlighted by a dashed red line in the image below (figure 18).

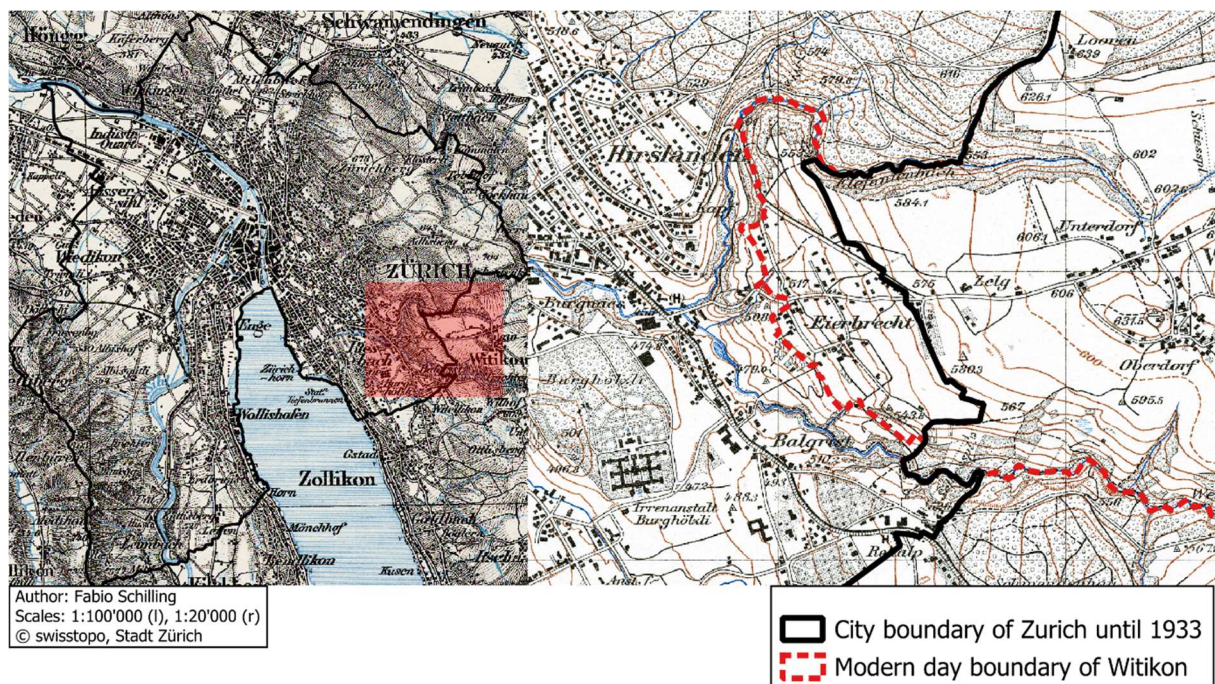


Figure 18: Border changes of district 7

The districts that necessitated adaptation were manipulated in QGIS. To this end, the polygons of the districts were adapted using the vertex tool available in the edit submenu. The borders were then adjusted in accordance with the dotted border lines present on the underlying and highly precise “Siegfriedkarte” background map. In the case of district 6, the three neighbourhoods of Oberstrass, Unterstrass and Wipkingen could simply be merged to form district 6.

4.3 Socioeconomic data from statistical yearbooks and HISCO

In order to undertake this research, socioeconomic data was required and found in two sources. The statistical yearbook(s) of the city of Zurich and the HISCO classification of occupations. The Historical International Standard of Classification of Occupations (HISCO) was defined as an information system on the history of work (van Leeuwen et al., 2002). The application of HISCO facilitates the allocation of occupations documented in each case entry to a designated social class and corresponding socioeconomic status. This approach enables an expansion of the analysis by incorporating this significant aspect of disease. The process involves a comparison of the registered occupations with lookup tables, where each occupation is assigned an individual code. This individual code is then compared with a reference table and assigned to a corresponding social class. The incorporation of individual- and district-level socioeconomic data this will significantly enhance the analytical value of the study.

District- and city-level socioeconomic data is available in the statistical yearbook of the City of Zurich (Stadt Zürich, 2024). As in 1930, a census was held and the year is situated amidst the studied time period, and importantly before the second city unification of 1934, this will be the primarily used yearbook (Statistik Stadt Zürich, 1931). This utilisation of this data facilitates the extraction of the socioeconomic character of individual city districts and enhance the understanding of the living conditions of their inhabitants. This will facilitate further insights into the potential association of disease occurrence with socio-demographic factors (i.e., poverty). The socioeconomic indicators selected for the analysis encompass population density (people/km²), insurance value per building (in 1000CHF) and rental prices (CHF). These indicators will be mapped by district. The utilisation of this comprehensive dataset facilitates the formulation of conclusions pertaining to the impact of sociodemographic factors at district level, thereby providing a more profound understanding of the information gained from the HISCO classification of occupations.

4.4 Visualisation

4.4.1 Visualisation methods

As proposed in the literature, I will create dot maps in order to visualise the distribution of individual disease cases in the city of Zurich (Shaweno et al., 2018, p. 4). This method bears a strong resemblance to John Snow’s seminal cholera map, in which each cholera case was represented by a dot denoting the respective residence (Osei, 2014). The merits

of this approach include its ease of interpretation and its widespread utilisation (Shaweno et al., 2018). Furthermore, the creating of density- or heatmaps, potentially kernel density estimations, is proposed to visualise disease densities and identify hotspots (Kirby et al., 2017, p. 5; Shiode et al., 2015). A further common visualisation method involves displaying case numbers or other statistical and attribute information, which has been aggregated to administrative units. This results in choropleth maps, which are frequently employed (Kirby et al., 2017, p. 6; Sasaki et al., 2008; Shaweno et al., 2018, p. 10). Another such visualisation method is cartograms, anamorphic maps, in which the geographic size of an area is adapted or distorted to be proportional to a specific attribute, such as population, while their colour remains free to display any other attribute, e.g. the population density, as in regular choropleth maps (Nusrat and Kobourov, 2016).

An objective of this study was to create animated chronological maps, for which the opportunity arose to create short animations of monthly disease incidence, visualised by heatmaps. Research has determined that approximately two-second intervals between frame changes are optimal for ensuring effective change recognition and interpretation, while minimising the risk of change blindness over the duration of exposure (Fabrikant et al., 2010; Griffin et al., 2006; Harrower and Fabrikant, 2008; Lowe, 2003; Sweller, 1994).

In order to enable the visualisation of the results, I used the layout creator of the free open-source GIS software QGIS. I also utilised the R programming language for statistical computing and data visualisation as well as the Python programming language. In both QGIS and Python, I tried to utilise the historical federal maps from the late 1920s and early 1930s to ensure a more coherent result, thereby facilitating the visualisation of the situation during that period. The general overview map, designated “Dufourkarte”, at a scale of 1:100'000, and the more detailed map entitled “Siegfriedkarte”, at a scale of 1:25'000, were utilised. Depending on scale and scope, I will utilise either of these maps as a background map. Using ESRI’s ArcGIS environment, a storymap was created to visualise the animated maps and to make them accessible to the broader public together with further findings. It can be found following this link: <https://arcg.is/0CzKDP>.

In order to project the data on a mapped space, it is necessary to select a CRS in which the data will be displayed. The CRS employed are enumerated below, along with their respective EPSG designations: 2056 (LV 95), the standard Swiss CRS, 21781 (LV 03), an older version of the standard Swiss CRS, 4326 (WGS 84), the Mercator projection and 3857, referred to as Web Mercator or Pseudo-Mercator. The focus of this study is the Swiss CRS (LV95), due to its reference to Switzerland and its metric coordinate system, which allows the best depiction of areas in Switzerland.

4.4.2 Visual variables

The optimal approach, colloquially known as best-practice, to cartographic principles is based on the foundational work of Bertin and Slocum (Bertin, 2011; Bertin and Berg, 1983; Slocum et al., 2022). These principles encompass a range of aspects, from fundamental guidelines such as ensuring high contrast, selecting appropriate sizes and shapes, to

more sophisticated concepts including the placement and arrangement of map elements. In regard to the selection of colours, including the choice of colour-blind safe colours, the ColorBrewer website (Brewer et al., 2013) was utilised. This website assists in the choice of compatible colours for maps. The classification scheme employed for the maps was quantiles, with a preference for quartiles and quintiles, as this allows classes to be presented in a quantifiable order, i.e. from highest to lowest values, while ensuring an equal number of members in each class (Slocum et al., 2022, pp. 87–88).

A square window was chosen for the layout of the maps, as this corresponds well with the dimensions of the 8-district city of Zurich, which measures approximately 10 by 10 kilometres. As suggested by best practice, a frame line was drawn around the map itself and individual map elements to ensure clear separation of content (Slocum et al., 2022, p. 267 ff). An appropriately sized title was placed above the map, but still within the frame line, to ensure clear communication of the map content. The legend, scale bar and source information were placed at the bottom of the map and arranged accordingly to ensure maximum visibility of the area of the city of Zurich itself. Finally, the maps were exported as PNG files, which ensure high data quality while maintaining ease of use.

The visual variables framework was employed, albeit predominantly one of its primary components, colour, was featured, with size arguably utilised to a similar extent. The visual variable colour, which consists of the three components hue, lightness and saturation, was employed and manipulated according to colour schemes provided by ColorBrewer2. Depending on the nature of data and number of classes, either sequential or diverging colour schemes with the appropriate number of grades were chosen. Therefore, for rental prices and building insurance values two distinct diverging colour schemes with 5 classes each, red-blue and purple-green, were chosen, respectively. The distinct colour schemes were chosen to ensure distinguishability between the two maps. The 5 classes were chosen, to allow for a median value to be established, which allows to depict which districts are below, above or on average. For sequential colour schemes, such as population density, morbidity and mortality, single hue colour schemes with 4 classes were chosen, in order to make full use of quartiles, which are a common method of data classification allowing representation of values above and below the median. The colours blue and red were chosen for morbidity and mortality respectively, as the colours are carrying cultural meaning, i.e. red signifying a negative event, such as death. Where deemed necessary to enhance contrast, the transparency of the basemap was adapted. This adjustment was implemented in maps that incorporated red dots, such as mortality dot maps, the local G* maps and the HDBSCAN cluster maps, as the relatively dark red colour of the dots compromised the visibility on the underlying dark basemap. In maps incorporating social class, the blue-red colour ramp was chosen. In this context, blue represents the working-class, with the colour being reminiscent of the term “blue-collar worker”, while red represents the upper-class. In the case of choropleth maps, the transparency of the choropleth layer was modified to ensure the preservation of orientation with respect to underlying basemap.

In order to capitalise on the enhanced visualisation capabilities of QGIS in comparison to GeoDa, the results were saved to the data in GeoDa and subsequently visualised in QGIS. This approach was adopted for the local G* and HDBSCAN cluster maps.

4.5 Analysis

As the work initially was conducted with point data, and subsequent analysis involved district-level data, it was necessary to aggregate the data. The point-in-polygon counting method of QGIS was utilised for the aggregation of the data to the reconstructed districts. The calculation of disease and mortality rates necessitated the disaggregation of the data collected over a period of four years to a yearly basis, thus facilitating meaningful comparisons. Furthermore, the data was normalised to a specific population, which was set at 1'000 individuals for morbidity, i.e. infection rates, as this is a frequently employed figure. However, for the mortality rate, 1'000 persons is inadequate in the context of the diseases in question, which is why 100'000 people were chosen as a reference frame. Consequently, the calculated district-level figures were then divided by the district's resident population and subsequently multiplied by 1'000 and 100'000, respectively. These values were then divided by the number of years, which is 4 in both cases, to represent yearly rates. The formula employed in the QGIS field calculator is as follows, where NUMPOINTS denotes the number of points of cases or fatalities, respectively, counted in a district:

$$(\text{NUMPOINTS} / \text{Population}) * 1000 / 4$$

The methods employed to achieve the initially stated objectives are informed by extant research, as previously referenced. The analysis is twofold: visual, as discussed above, and statistical, by the use of geographical data analysis methods and software. A comprehensive overview of the available methods for analysing epidemiological data can be found in the works of Osei (2014), Kirby, Delmelle and Eberth (2017), and Shaweno *et al.* (2018). Examples of these include nearest neighbour analysis, Ripley's alphabet of functions, Moran's I and local indicator of spatial association (LISA).

Cluster analysis is a subfield of spatial statistics and is well-suited to the research objectives of this thesis. The field of cluster analysis is broad and offers numerous possibilities, encompassing the investigation of non-random clustering of events in space. The utilisation of nearest neighbour analysis and the nearest neighbour index (NNI) as a fundamental metric for assessing the degree of clustering has been a pervasive practice across diverse studies (Bhatia, 2010; Kirby *et al.*, 2017; Peterson, 2009). The NNI can be viewed as a fundamental component of Ripley's alphabet of functions, with the G-function being a notable example. These functions have been employed not only to analyse the clustering behaviour of data but also to compare it to random distributions (Kirby *et al.*, 2017; Osei, 2014).

Finally, the method of density-based spatial clustering of applications with noise (DBSCAN) and its hierarchical twin (HDBSCAN) are more focused on densities and also have been utilised in disease research. They are capable of detecting clusters

independently of their size and shape, which is advantageous when dealing with real-world applications (Khan et al., 2014; Schubert et al., 2017).

Moran's I is another index that calculates the degree of clustering and is considered to be more sophisticated than the NNI. It is also frequently employed in health geography research. Moran's I is able to determine the spatial autocorrelation of variables (Agbor, 2014; Kirby et al., 2017; Osei and Duker, 2008). As Moran's I is a single value computed for the global distribution, it is expanded by local indicators of spatial association (LISA), which are another means to investigate the interaction between spatial units and the influence they take on each other on a localised level (Kirby et al., 2017; Osei, 2014). A related concept is that of local G / G* (Getis-Ord), which is capable of identifying clusters of high or low values, i.e. hotspots and coldspots, respectively (Chirenda et al., 2020; Getis and Ord, 1992). Local G* was chosen over LISA, as the identification of clusters is a research objective of this thesis.

Regression analysis constitutes yet another set of methods employed in the field of disease research. In essence, this analytical approach entails the establishment of a relationship between a dependent variable and one or more independent variables, with the objective of ascertaining the strength of the relationship (Agbor, 2014; Bernhard et al., 2023; Bingham et al., 2004; Bwire et al., 2017; Coleman, 2018; Horton et al., 2016; Humayun et al., 2022; Kirby et al., 2017; Kistemann et al., 2002; Pape et al., 2024; Peer et al., 2023; Pezeshki et al., 2012; Sun et al., 2015). Regression analysis can be enriched by incorporating space, as is demonstrated in geographically weighted regression (GWR), which accounts for the spatially varying influence of independent variables. GWR has been employed in research to explore the association between certain influencing factors and disease prevalence (Kirby et al., 2017; Sun et al., 2015).

The analysis was conducted using GeoDa, a geostatistical software tool that has been developed for the analysis of spatial data (Anselin et al., 2006). The following applicable methods were employed: DBScan, HDBScan, Moran's I, local G / G* and LISA. Notably, GeoDa was also employed by Burkhard (2023) in her Master's thesis. The implementation of the nearest neighbour analysis and associated methodologies, such as the G-function and regression, was conducted utilising the R programming language. The selection of R was motivated by its emphasis on statistical analysis and the availability of an extensive range of adaptable options in addition to dedicated packages.

Regarding the software applications employed for the analysis: R was utilised and run within the RStudio environment on version 4.3.1 (2023-06-16). The QGIS version employed is 3.34.12 "Prizren", GeoDa is run on the current version 1.22 and Python on version 3.11.8

5 Results

The results of the analysis are presented in a variety of forms. Firstly, there are the diagrams I have created, many of which are based on basic statistical calculations made in Excel. These include the calculation of the absolute number of cases and fatalities per age group or district, as well as the calculation and comparison of relative shares of the total amount in relation to population. Secondly, I have created maps to visualise the results of case geolocation and the social class attribute. The cartographic representations facilitate the visual examination of the distribution of cases and the formulation of preliminary observations regarding potential hotspots and influencing factors. Thirdly, I calculated statistical indices to demonstrate clusters or influences of certain factors on a more technical basis. These indices provide a reliable statistical foundation for cluster analysis. Finally, the temporal aspect of the data is presented in the form of histograms, graphs and sequentially arranged maps. This enables the analysis of temporal developments and changes over time in the intensity and distribution of the diseases. Consequently, the monitoring of hotspots facilitates the analysis of their persistence. The story-map with the animated maps, can be found here: <https://arcg.is/0CzKDP>.

For the purpose of reference, a general overview map of the city of Zurich and its population is provided here to support the results and allow for comparison (figure 19). Of particular note is the form chosen for the population distribution map by district, which is included in the form of a cartogram. This cartographic representation is intended to facilitate a more robust comparison with the disease's case numbers. The cartogram highlights the low population density and comparatively small population of districts 2 and 7. The below-average population density of district 3 can be attributed to its substantial size, remains underdeveloped. District 8, while exhibiting below-average population density, is notable for its substantial population size. Conversely, districts 1, 4, 5 and 6, which demonstrate above-average population density, are also characterised by substantial populations, as evidenced by their considerable sizes.

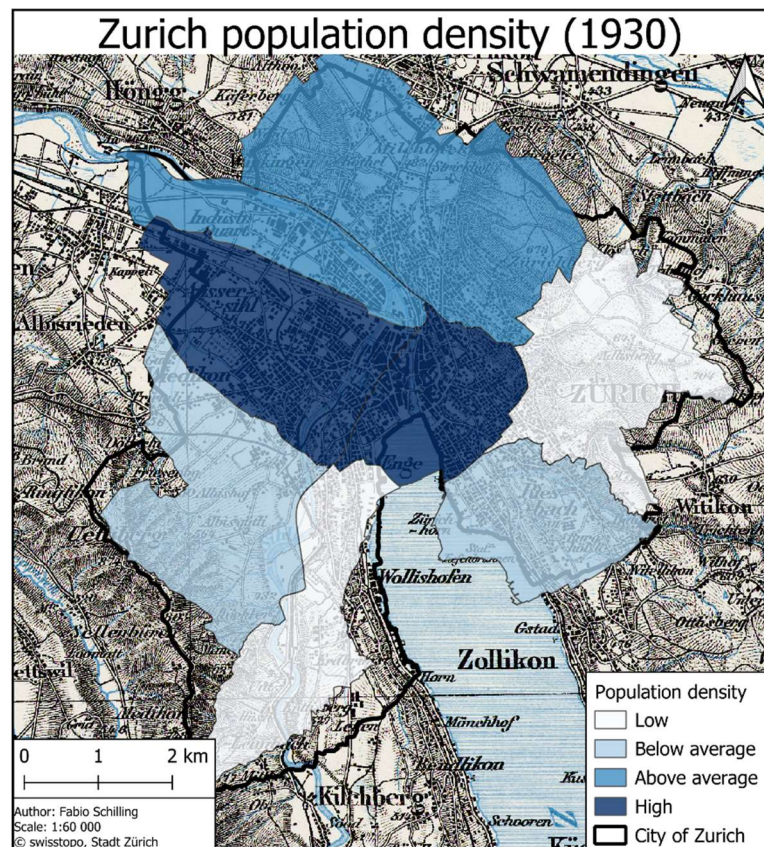


Figure 19: Population cartogram, coloured by population density

5.1 Tuberculosis

Regarding fundamental statistics, the data on tuberculosis reveals an average age of 31 years and a median age of 30 years. The distribution of cases across gender lines reveals a slightly higher prevalence of male cases, with 982 (53%) being male and 870 (47%) being female. This is an anomalous finding given the city of Zurich's demographic composition, which is characterised by a relatively higher proportion of women in the population. Furthermore, approximately three-quarters of all patients, amounting to 1426 cases (77%), have been referred to a hospital for treatment, while the mortality rate stands at 357 out of 1852 cases (approximately 20% of the total cases).

A more detailed analysis of the age in the following diagram (figure 20) shows that the 20–29-year-olds, who represented 23.6% of the population in 1930, suffered a morbidity higher than their proportion of the population, as 617 or 33.5% of the cases concerned this age group. The only other age group with more cases than their relative population size suggests is the 30–39-year-olds with 18.6% share of total population and 426 or 23.2% of cases. All other age groups exhibit case shares that are lower than their population shares. The youngest group (ages 0-9) exhibits the lowest case count of 102, constituting 5.5% of total cases, even though they constitute 11.4% of the population. The age group over 60 years old suffered 120 cases (6.5%), followed by the 10–19-year-olds, with 159 case (8.6%). Finally, the 50-59 age group accounts for 193 cases (10.5%), while the 40-49 age group accounts for 223 cases (12.1%).

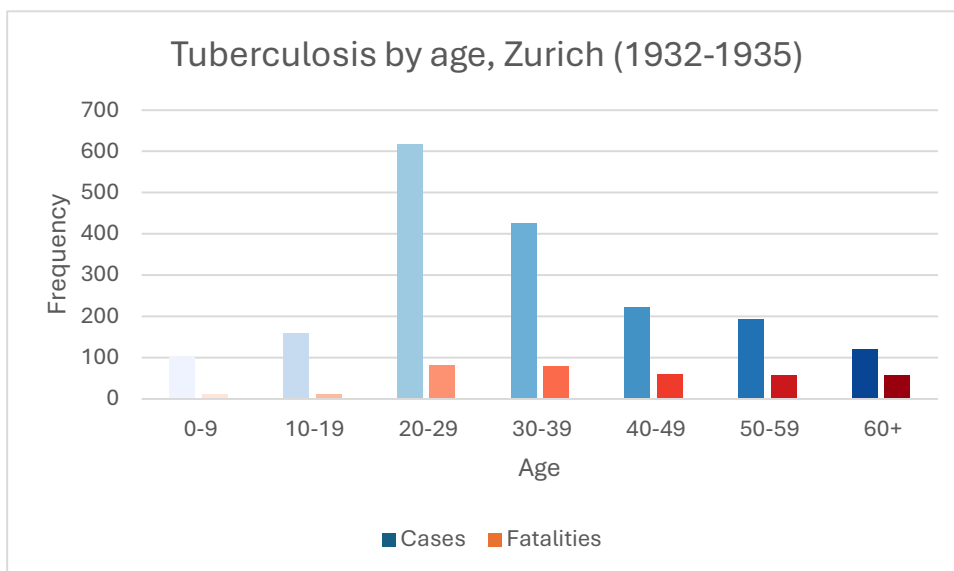


Figure 20: Tuberculosis age distribution

The distribution of fatalities is slightly more equal in relation to the population. The youngest age groups, ranging from 0 to 19 years, experience a minimal number of fatalities with only 11 cases (3.1%) in the 0-9 age group and 10 (2.8%) in the 10-19 age group. Notably, both age groups constitute more than 10% of the total population. The 20-29 age group exhibits a close correlation between its share of fatalities and its population share, with 81 deaths (22.9% of total fatalities) and 23.6% of the population, respectively. All older

age group exhibit a higher proportion of fatalities in comparison to their respective population shares. The 30-39 age group accounts for 80 deaths (22.7%), while representing 18.6% of the population. The remaining age groups all suffered 59 or 56 fatalities, corresponding to 16.7% and 15.9% of the total, with declining shares of population from 14% to 11.2% and 9%, respectively. The most pronounced disparity between the proportions of fatalities and population is observed in the over 60 age group, where the gap between the population share of 9% and the fatality share of 15.9% approaches a factor of two.

When comparing the proportion of the infected individuals in each age group who died from tuberculosis, the lowest values were observed in the youngest age groups. Conversely, the highest absolute values are observed among the working-age population between 20 and 40 years. However, the highest case-based mortality, defined as the proportion of fatalities occurring in relation to infections, is observed among the oldest age groups, with nearly half of the infected individuals succumbing to tuberculosis. Up to the age of 40, the share of fatalities from total cases is below the average of approximately 20%, i.e. less than 1 out of 5 infected persons dies. However, from the age of 40 onwards, the mortality rate increases to approximately 25% (26.5%) for the 40-49 age group, to 29% for those aged 50-59, and finally 46.7% for those aged 60 and older.

The subsequent diagram (figure 21) illustrates the frequency of cases and fatalities per social class. The most tuberculosis cases were registered in the middle-class with 578 cases. In contrast, the upper-class registered the lowest number of cases, with only 458 cases documented. The lower-class, with 532 cases, occupies the median position between the upper- and middle-classes. However, 409 cases could not be assigned to a social class. A similar distribution is observed in the fatality data, with the majority, 139 deaths, occurring among the middle-class. The lower-class has suffered the least number of deaths, with only 90, while the upper-class has suffered 101 deaths. Notably, 26 deaths remain unassigned to any social class. The mean of 20% mortality is observed across all three social classes, although the upper-class has experienced slightly higher mortality in proportion to cases.

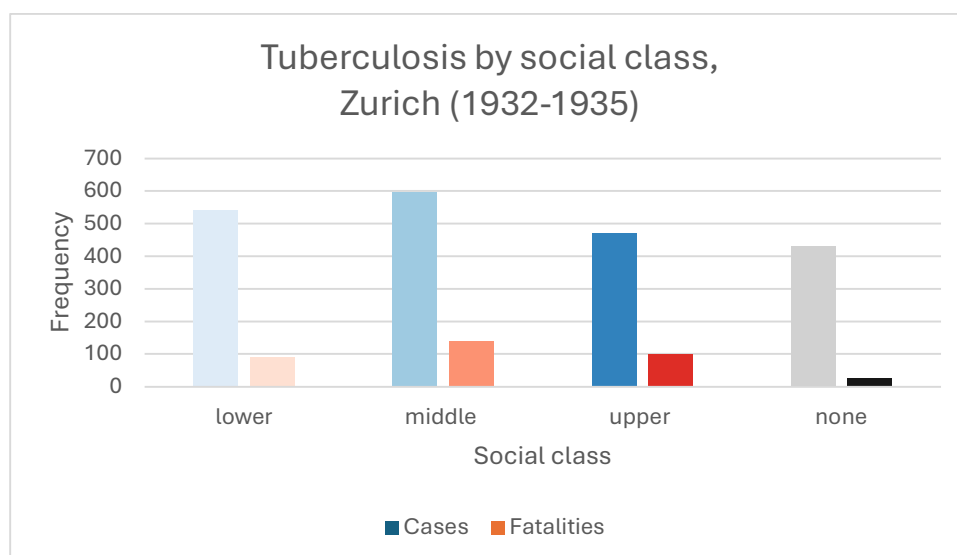


Figure 21: Tuberculosis social class distribution

The following diagram (figure 22) illustrates the allocation of tuberculosis cases and fatalities at district level, based on geolocation as determined by the address of the cases. Despite the districts having been registered separately, I considered the data extracted from the locations more reliable and conducted the primary analysis using this data. However, minor discrepancies were identified between the two datasets, potentially attributable to the reorganisation of the city's districts in 1934. The districts with the highest number of cases are districts 3 (327 cases), 4 (348 cases) and 6 (329 cases). The remaining districts have registered less than 200 cases, with district 1 having the most cases among them, almost reaching 200 cases. Districts 7 and 8 have been recorded with approximately 170 and 150 cases, respectively. Conversely, districts 2 and 5 reported the lowest number of cases, with slightly more than 100 cases each. Districts 3, 4 and 6 reported the highest numbers of fatalities, with 53, 58 and 73 cases, respectively. All other districts have recorded fewer than 40 fatalities, as illustrated in the accompanying diagram. The lowest number of deaths was recorded in districts 2 and 7, with approximately 25 cases each.

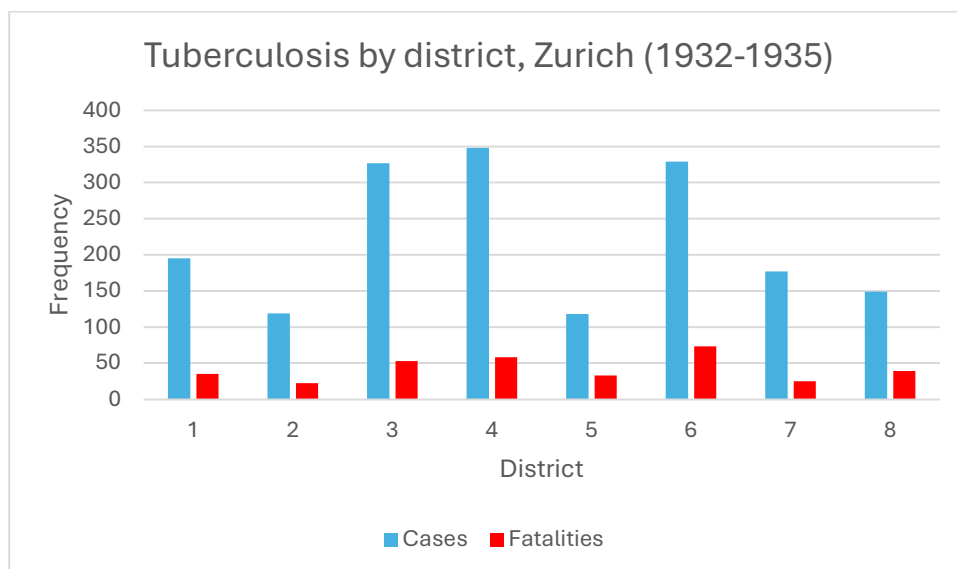


Figure 22: Tuberculosis district distribution

The distribution of the tuberculosis cases in the city of Zurich is illustrated in the map below, with each green dot representing a single case. The cases are distributed throughout the built-up inhabited areas of the city of Zurich. Three areas of higher case density can be discerned. The first of these is situated in the eastern half of district 1 (east of the Limmat river), where the historic old town is located, exhibits higher case density. Secondly, the working-class districts 3 and 4, located to the south of the railways display higher case density. Finally, district 5, situated north of the railways, also exhibits a higher case density. Conversely, district 2, situated along the western lakeshore, exhibits only few cases. A similar pattern can be seen in district 7, but more than district 2, which exhibits fewer cases than high-density areas. Districts 6 and 8 exhibit medium case densities.

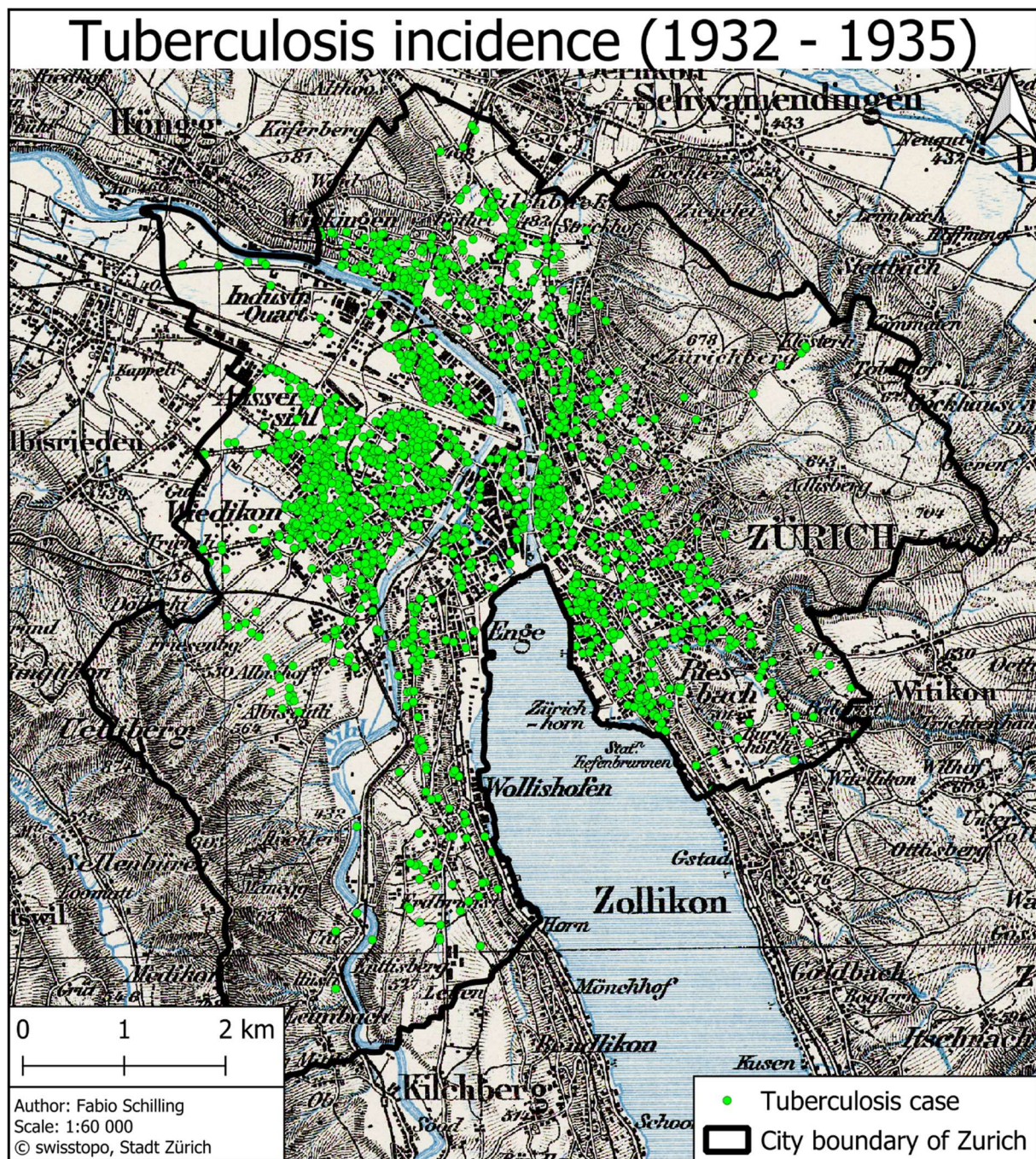


Figure 23: Tuberculosis case map

The distribution of tuberculosis fatalities in figure 24, symbolised by red dots, appears to closely mirror the tuberculosis case distribution, although it is significantly less dense, as it contains only a fifth of the data volume. The areas of higher density can again be described in a similar manner to the tuberculosis cases. The historic old town in district 1, located east of the Limmat river, exhibits the highest density of deaths, with the working-class districts 3, 4 and 5, situated to the south and north of the railways, also displaying a high density of fatalities, albeit to a lesser extent. Conversely, districts 2 and 7 exhibit a markedly lower incidence of fatalities. Districts 6 and 8, which are characterised by a medium density of fatalities, provide the middle ground.

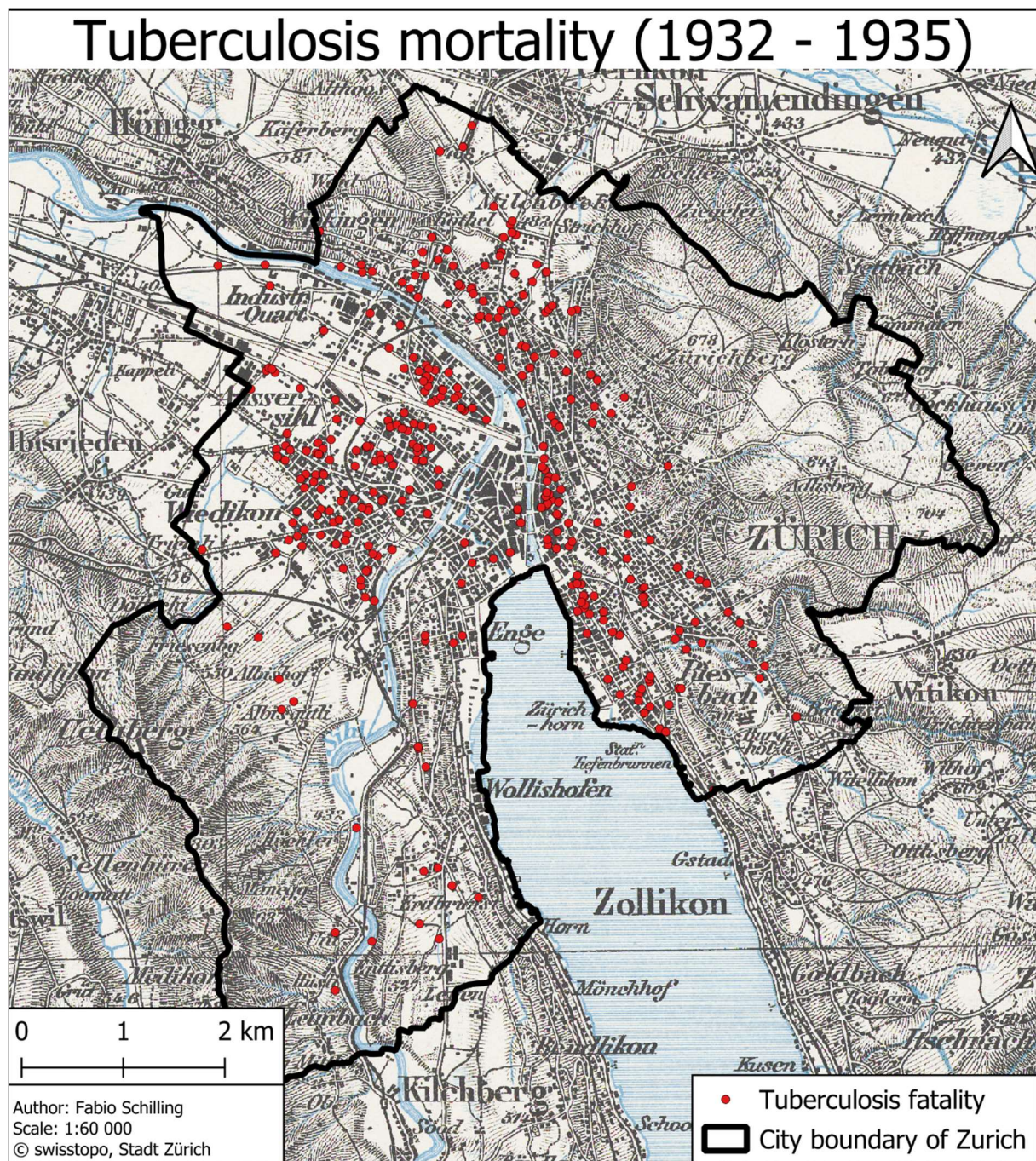


Figure 24: Tuberculosis mortality map

The following maps (figures 25 and 26) present data on the incidence of tuberculosis per 1'000 inhabitants and mortality per 100'000 inhabitants within the total population on a district level. The districts that have been most severely affected in terms of morbidity, or infection rate, are districts 4 and 6, followed by districts 1 and 3, as illustrated by the dark blue colour hues. Conversely, districts 7 and 8 exhibit comparatively lower rates, while districts 2 and 5 display the lowest rates, illustrated by light blue and white colour hues, respectively. The analysis of tuberculosis mortality rates reveals both similarities and differences. The districts with the highest mortality rates are districts 5 and 8, followed by districts 1 and 4, indicated by the dark red and red colour hues, respectively. The high mortality rates in districts 1 and 3 are reminiscent of their high morbidity rates, while the high mortality rates of districts 5 and 8 are quite the opposite of their low morbidity rates. Conversely, districts 3 and 6 exhibit comparatively lower mortality rates, while districts 2 and 7 display the lowest mortality rates, illustrated by their light red and white colour hues, respectively. The low mortality rates in districts 2 and 7 are once again reminiscent of their low morbidity rates. It is worth mentioning that district 6 has a relatively low mortality rate despite having the highest morbidity rate. It is important to note that the size of the districts is distorted according to the total number of cases and fatalities reported, respectively, in order to provide an indication of both the rates and the total numbers.

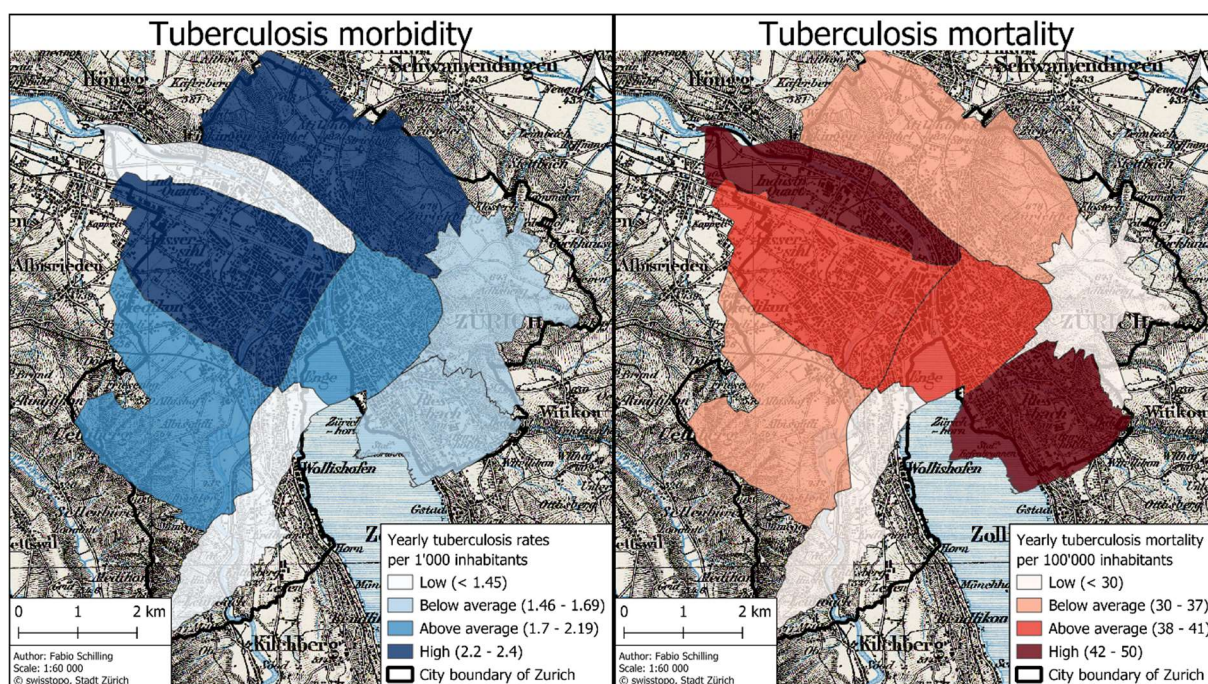


Figure 25: Tuberculosis morbidity cartogram

Figure 26: Tuberculosis mortality cartogram

The HDBSCAN clustering (figures 27 and 28) yielded eight distinct clusters of tuberculosis cases and six distinct clusters of tuberculosis fatalities, symbolised through distinct colours of the respective dots, while dots not assigned to any cluster remain white. The case clusters appear to align with the districts, although in districts 3 and 4 there is only a single, green cluster, while in district 8 there are two, brown and pink clusters. With these exceptions, it appears that every remaining cluster corresponds to a district. This is best illustrated by the blue cluster (district 5), the lavender-purple cluster (district 6) and the orange cluster (district 1). For the fatality clusters, the approximate locations of the clusters remain consistent, although their size and extent diminish. Notably, the two clusters in district 8 have merged into one, while the cluster in district 7 has disappeared, as has the cluster in district 2. Concurrently, the large cluster in districts 3 and 4 split up into two distinct green and blue clusters, roughly corresponding to districts 3 and 4, respectively. The orange, yellow and lavender-purple clusters correspond to districts 1, 5 and 6, respectively.

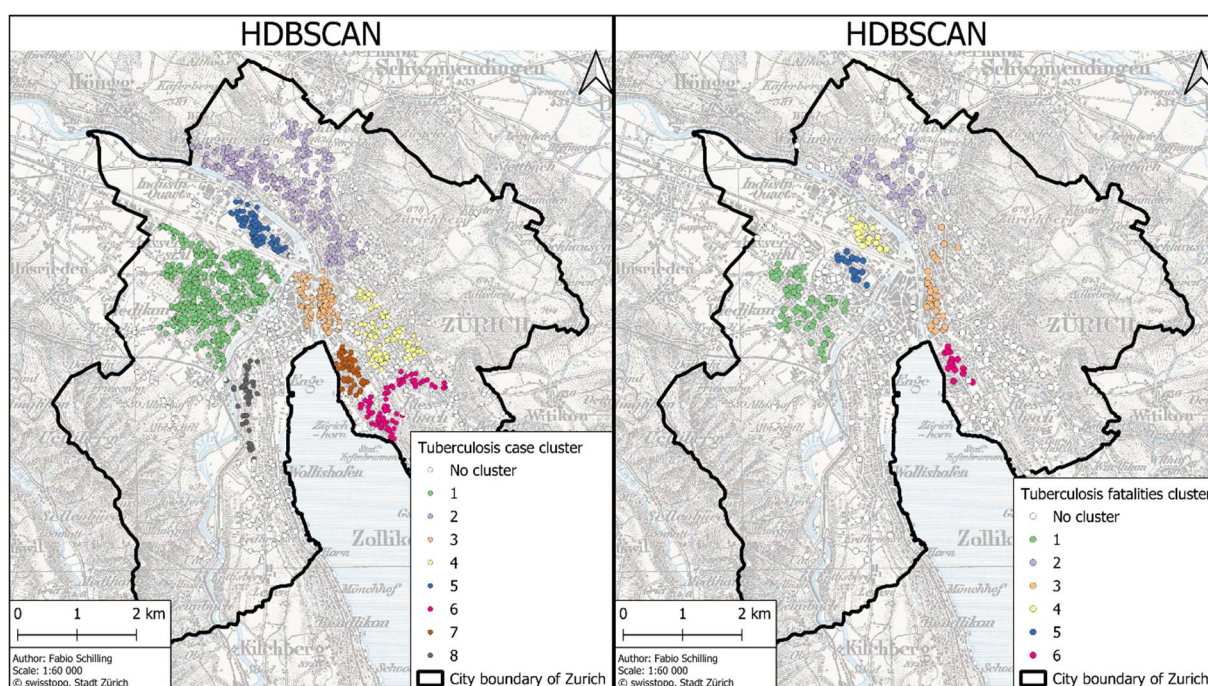


Figure 27: HDBSCAN clusters of tuberculosis cases

Figure 28: HDBSCAN clusters of tuberculosis fatalities

The results of the HDBSCAN clustering of tuberculosis cases by social class are displayed below (figure 29). It is evident to the observer that while the upper- and middle-classes each have a single cluster, designated by red and orange colours respectively, the lower-class is divided into nine such clusters. These clusters are dispersed throughout the city of Zurich, with notable variations in size. The largest cluster encompasses 74 points, while the smallest cluster includes 20 points. In contrast, the middle-class cluster counts 574 points, and the upper-class cluster 449 points. The clusters are distinguished by a range of blue colour hues and their locations are delineated as follows: The “Lower-class 1” cluster is located in district 1 and the southern part of district 6, “Lower-class 2” is situated in the west of district 4, “Lower-class 3” is in the remainder of district 6, “Lower-class 4” is in districts 7 and 8, “Lower-class 5” is in district 5, “Lower-class 6” is in district 3, “Lower-class 7” is in the northeast of district 4, “Lower-class 8” is in the southeast of district 4 and the northeast of district 3, and finally, “Lower-class 9” is in district 2. It was notably, some of these clusters were quite similar to the district boundaries, while others, mainly the clusters in district 3 and 4, were more fragmented. This is because there are four clusters in these two districts. The grey points represent non-clustered data points.

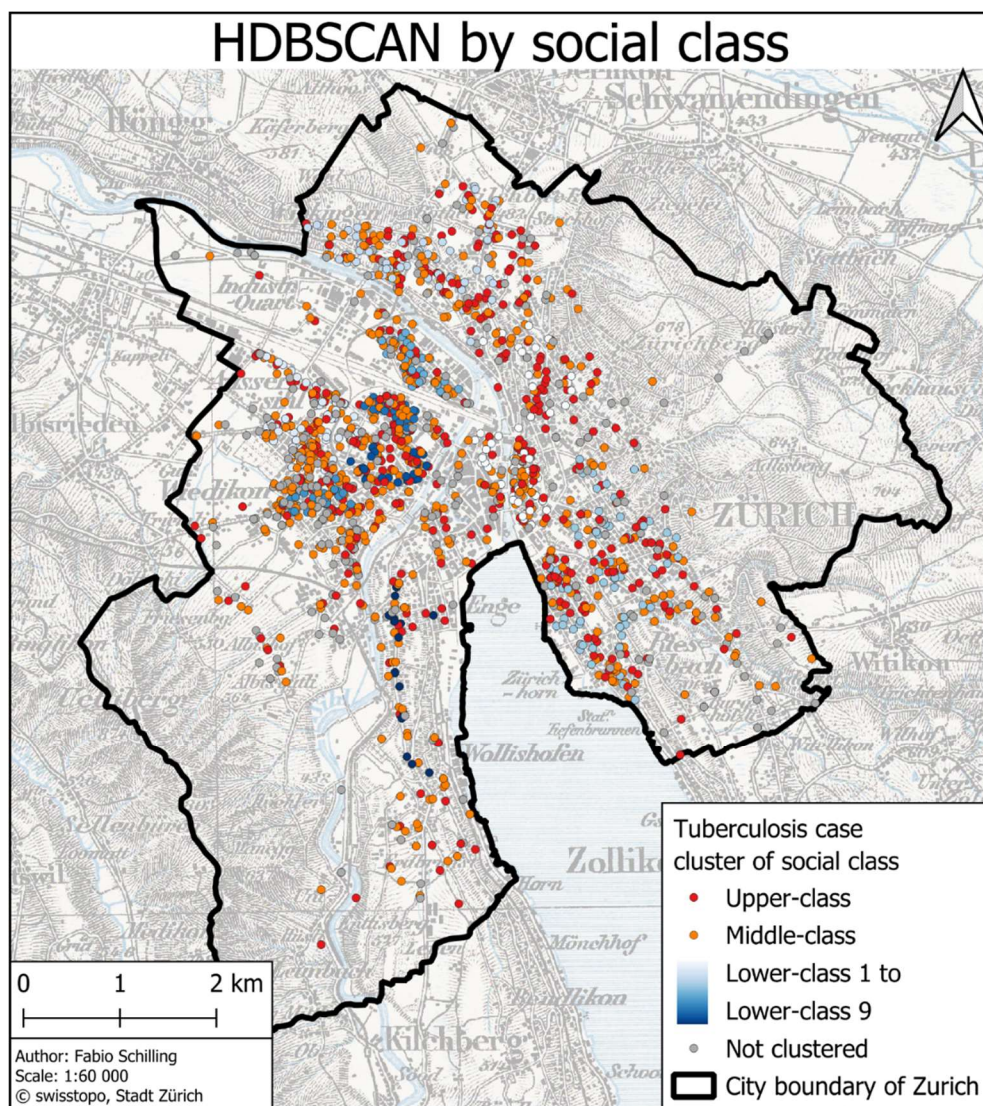


Figure 29: HDBSCAN clusters of tuberculosis cases by social class

The results of the HDBSCAN clustering of tuberculosis fatalities are presented below (figure 30). The upper- and middle-class demographics are each represented by a single cluster, coloured red and orange, respectively. The lower-class was subdivided into four clusters, covering most of the city and being coloured in hues of blue. The points that were not part of any cluster are coloured grey. The lower-class clusters encompass between 12 and 25 points each, while the middle- and upper-class clusters encompass 132 and 93 points, respectively. The distribution of the lower-class clusters is as follows: “Lower-class 1” is located in districts 1, 7 and 8, “Lower-class 2” in districts 3 and 4, “Lower-class 3” in districts 4 and 5, and “Lower-class 4” in district 6.

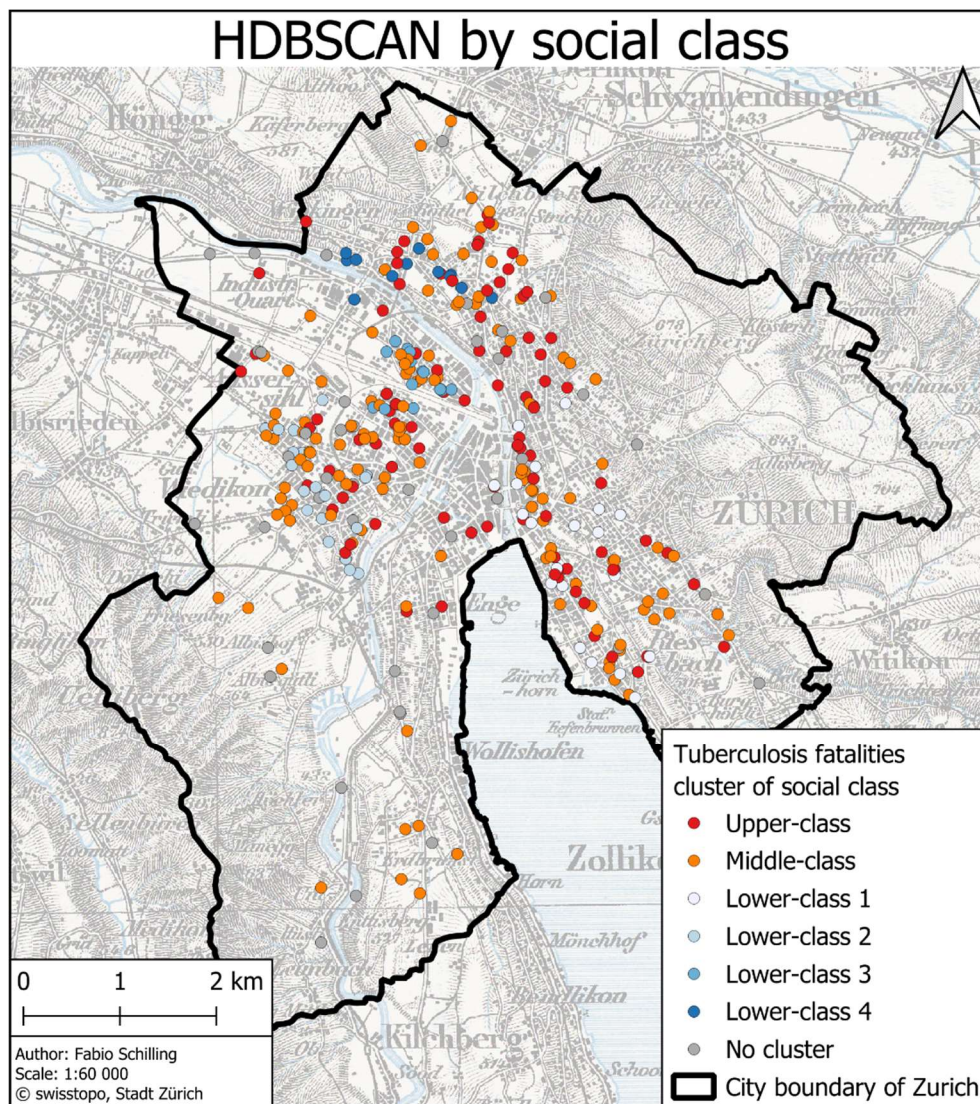


Figure 30: HDBSCAN clusters of tuberculosis fatalities by social class

The social class distribution of tuberculosis cases (figure 31) exhibits high heterogeneity, as no district is clearly dominated by any specific social class. This is illustrated by the intermingling of blue, white, and red dots on the map. Nevertheless, certain trends can be discerned. Specifically, a higher prevalence of blue dots, indicative of the lower-class, is observed in districts 3, 4 and 5. Conversely, district 2 exhibits a predominance of red, indicate of the upper-class, over blue dots. White dots, representing the middle-class, are visible across all districts. The map does not provide any indication of a region or district that is clearly dominated by a certain social class or that is devoid of another.

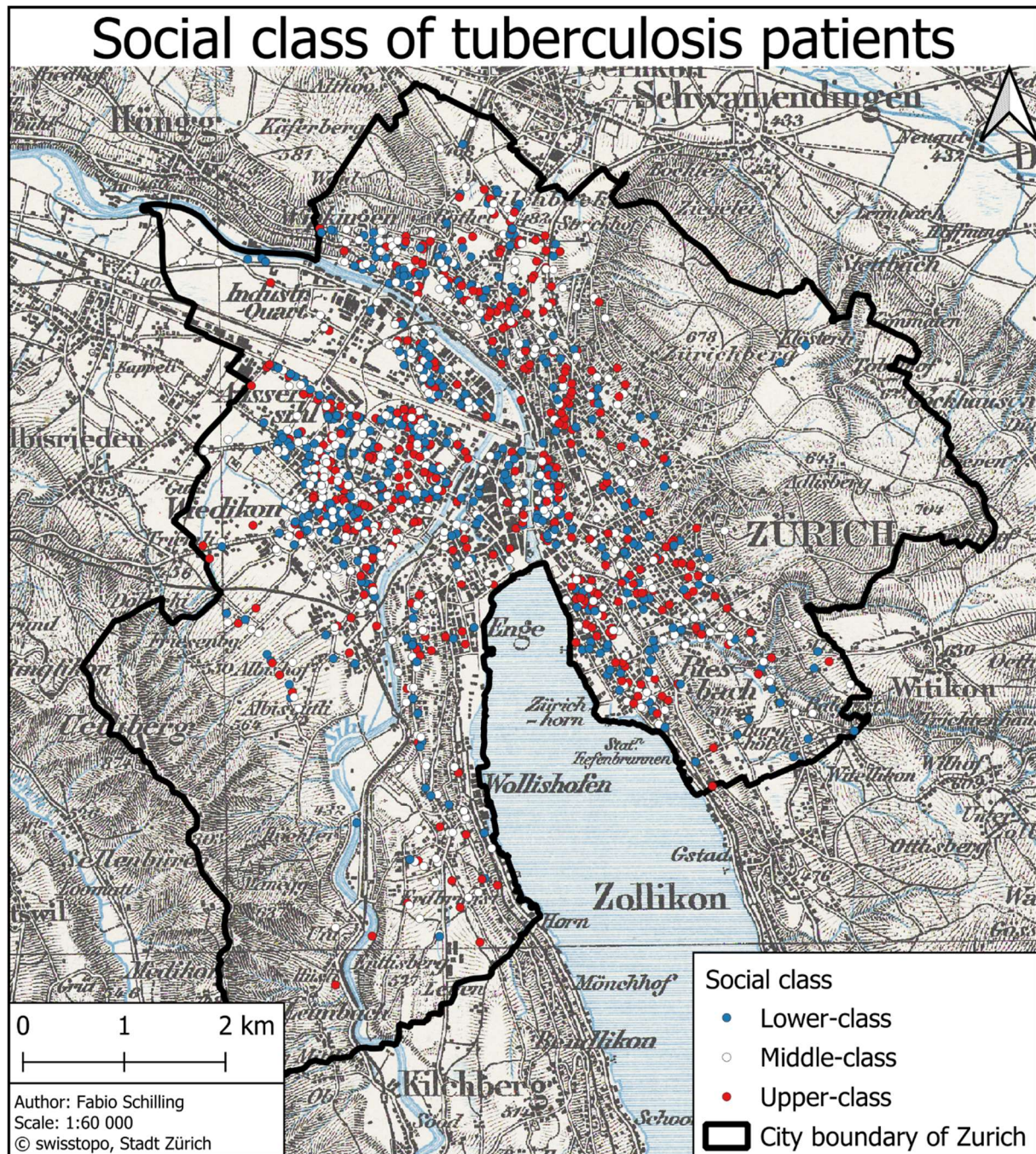


Figure 31: Social class map of tuberculosis cases

The NNI for tuberculosis cases is 0.518. The corresponding z-score is -38.72, with a p value extremely close to 0. The NNI for tuberculosis fatalities is 0.735. The corresponding z-score is -8.91, with a p value of approximately 0. The range of nearest neighbour distances for tuberculosis cases range from 0 to 615 metres. The median nearest neighbour distance is 29m, the mean distance is 38m. For tuberculosis fatalities, the median distance is 78m, the mean distance is 110m, and the distance range extends from 0 to 789m. The distribution of nearest neighbour distances for both tuberculosis cases and fatalities is presented below (figures 32 and 33). The histogram of tuberculosis cases displays an exponential decline. The majority of cases are located within 100m of their nearest neighbour. In contrast, the histogram of tuberculosis fatalities exhibits a less pronounced decline, with the majority of nearest neighbours situated within approximately 300m. This is reasonable, as there are fewer fatalities, and therefore distances can be expected to increase.

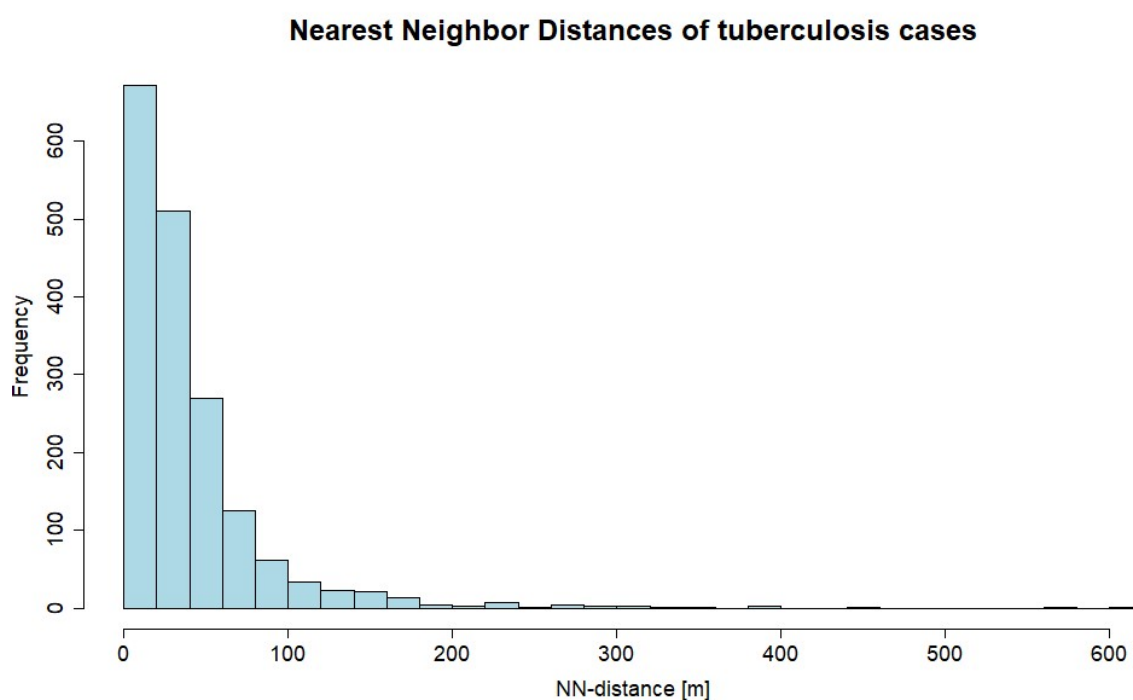


Figure 32: NN-distances of tuberculosis cases

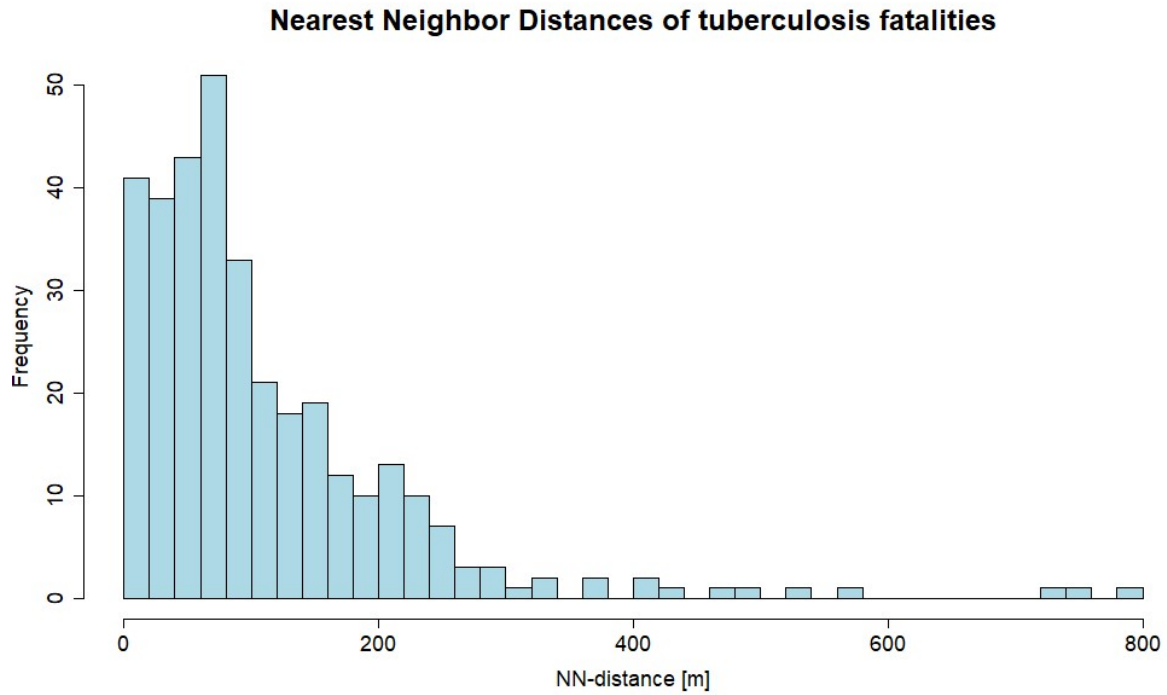


Figure 33: NN-distances of tuberculosis fatalities

The G-function, or cumulative distance plot, for tuberculosis cases (figure 34) reveals a steep rise to 20% within the first meters of distance. Subsequently, approximately half of the nearest neighbours are within 30m and 80% within 60m of each other.

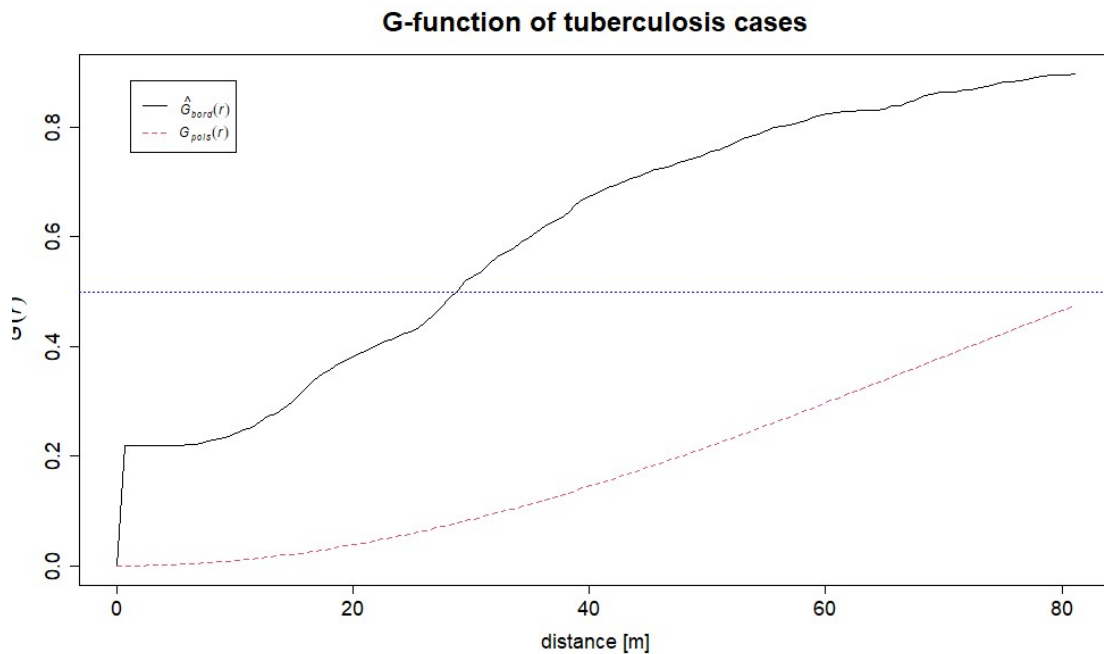


Figure 34: G-function of tuberculosis cases

The same plot for the tuberculosis fatalities (figure 35) reveals a steep, albeit comparatively reduced, incline within the first metres. 20% of nearest neighbours are reached within approximately 25m of distance. Furthermore, 50% are reached within approximately 75m and 80% within 150m. The black line representing the observed data are consistently situated above the red reference line of a theoretical distribution in both plots.

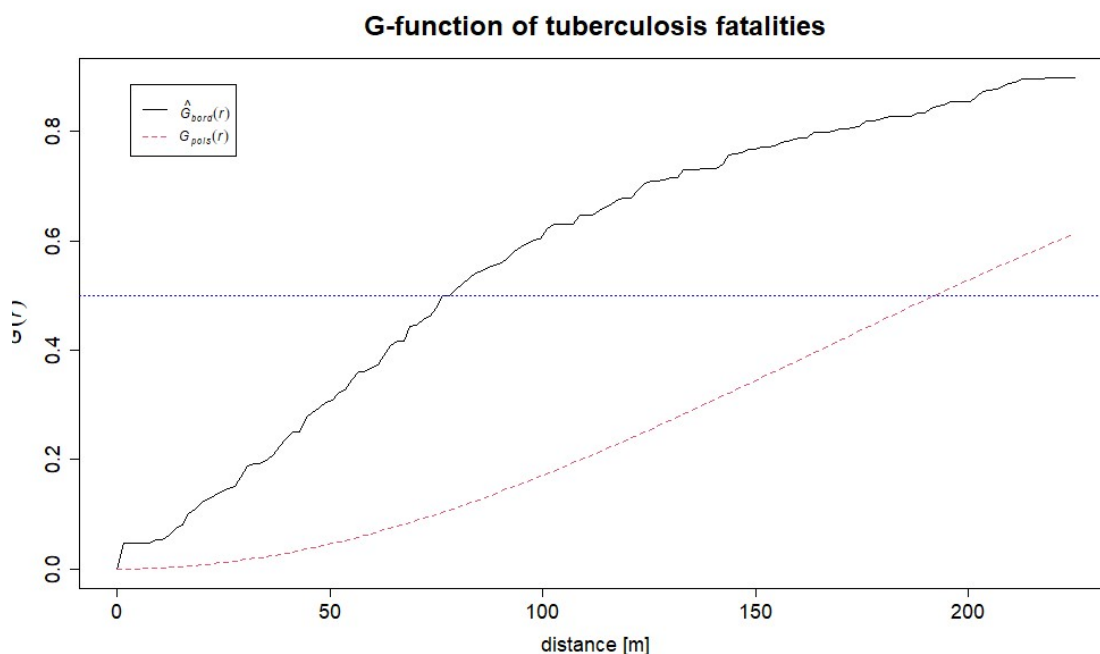


Figure 35: G-function of tuberculosis fatalities

Moran's I for both the tuberculosis cases and fatalities at district level is negative, as indicated by the negative slopes of the fitted lines (figures 36 and 37). The value of Moran's I for tuberculosis cases is -0.289, and for tuberculosis fatalities it is -0.366.

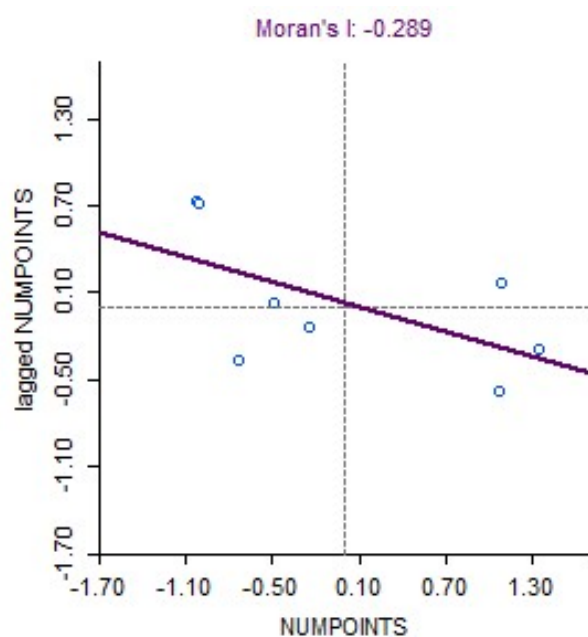


Figure 36: Moran's I of tuberculosis cases

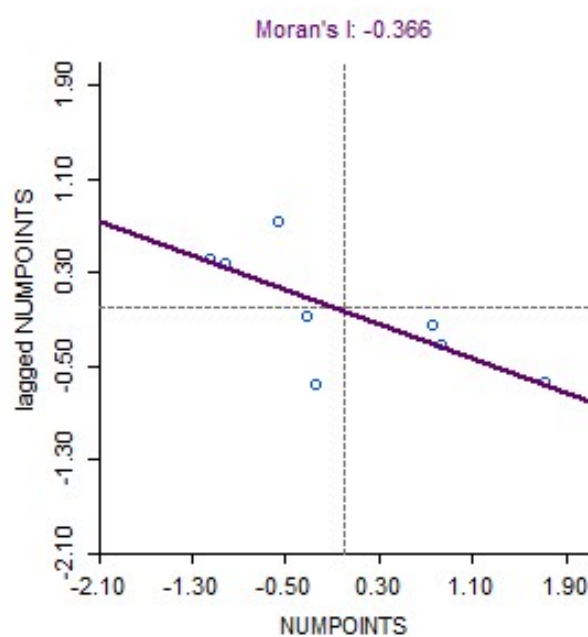


Figure 37: Moran's I of tuberculosis fatalities

The results of local G^* of the social class of tuberculosis cases (figure 38) demonstrate a cluster of blue points, representing the lower-class, in the west of Zurich, corresponding to the working-class districts 3 and 4, which are known for the low socioeconomic status of the inhabitants. Conversely, two distinct clusters of red points, representing the upper-class, are observed in the south-east and north-east of the city, corresponding to parts of the districts 6, 7 and 8. The southern cluster, located in the districts 7 and 8, is likely to encompass affluent neighbourhoods. However, the northern cluster, situated in the Oberstrass and Unterstrass neighbourhoods of district 6, is less likely to encompass such affluent neighbourhoods. Additionally, there are some less clustered red points found in parts of districts 1 and 2, which are also regarded as affluent. The white points are not significant and are not part of any cluster.

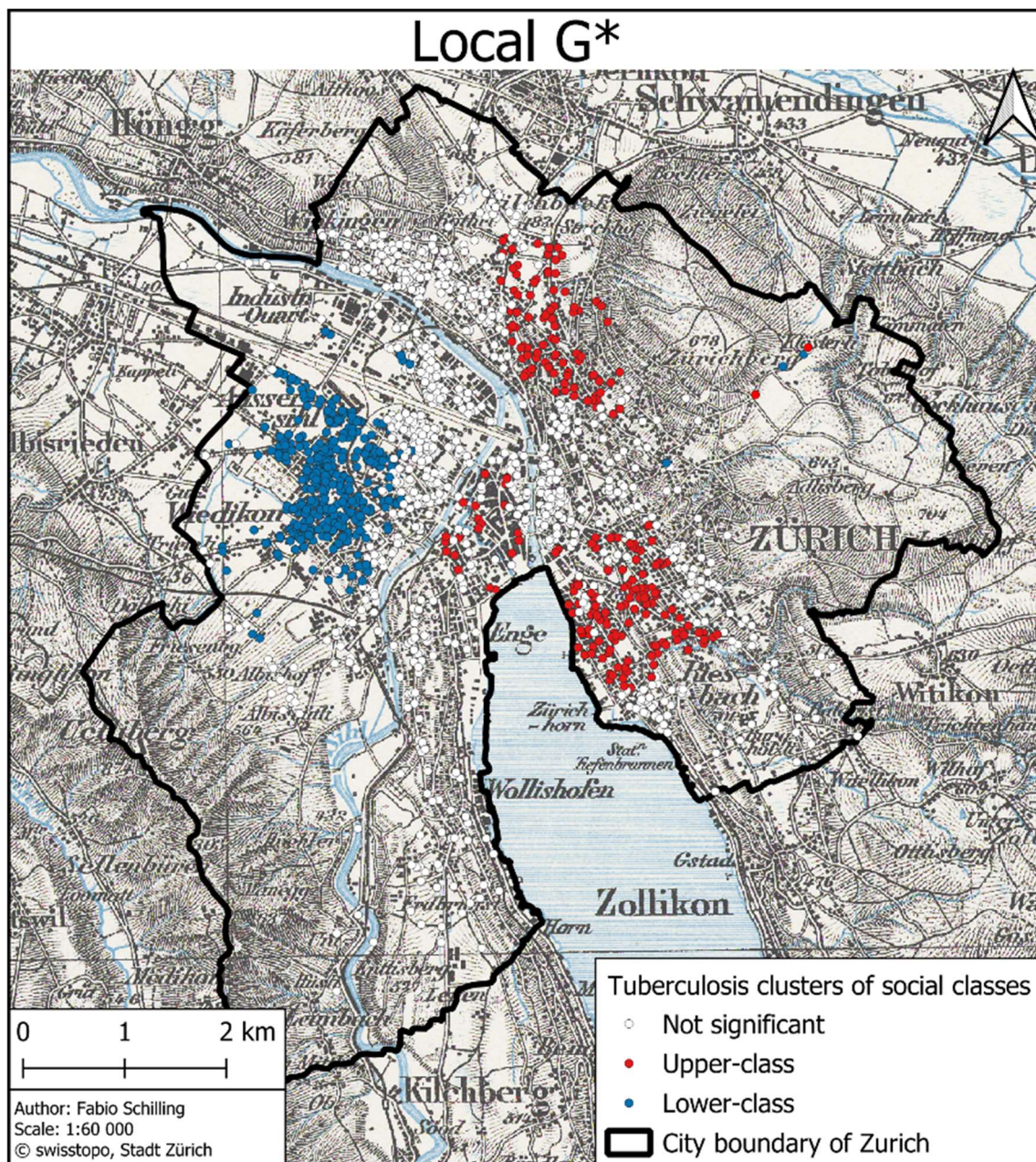


Figure 38: local G^* cluster map of social class in tuberculosis data

Regarding the tuberculosis fatalities (figure 39), the clusters which were detected in the cases appear to be relatively constant, albeit on a smaller scale, as they still occupy roughly the same locations. The blue dots, indicating the lower-class, is still present in parts of districts 3 and 4. The two red clusters in the southeast and northeast have considerably shrunk, and now only count a handful of fatalities. The cluster in districts 1 and 2 has completely vanished.

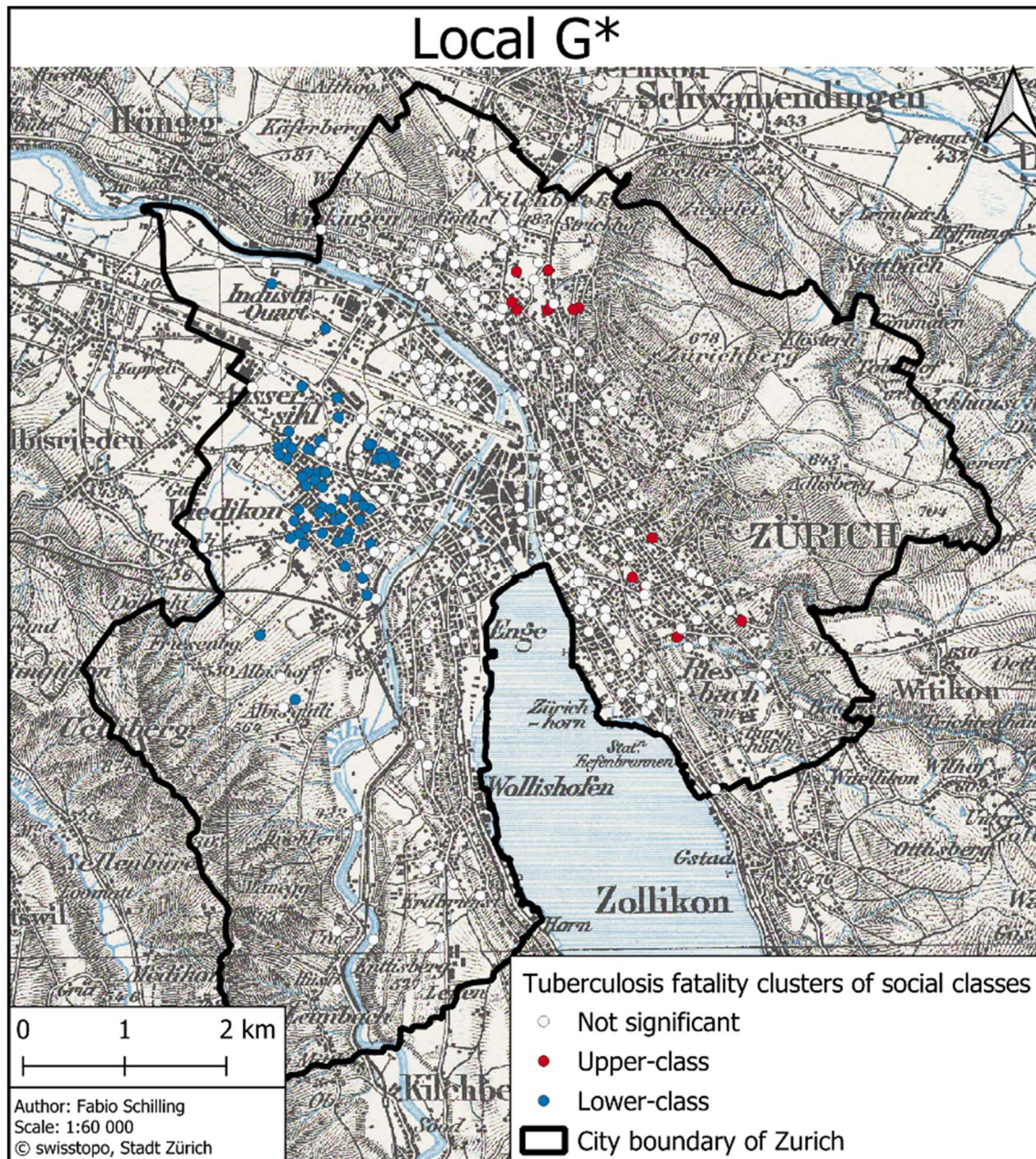


Figure 39: local G* cluster map of social class in tuberculosis fatalities

Regression analysis of all tuberculosis cases, with death as the dependent variable and gender, age, hospital admission, and social class as independent variables has revealed that all but gender to have positive values. Gender was the only negative value identified, while age, hospital admission and affiliation to higher social classes all had positive values. The R-squared value is 0.053, and the adjusted R-squared is 0.05 with a p-value of $< 2.2e^{-16}$.

The geographically weighted regression is executed analogously, with death designated as the dependent variable and gender, age, hospital admission and social class as the independent variables. GWR attests spatially varying influences on all four investigated attributes (see table 6). The only negative median concerns the hospital admission variable, while all other variables had positive medians. The R-squared value increases from 0.063 to 0.358, while the adjusted R-squared is 0.061 in both cases. The residual sum of squares (RSS) decreases from 255 to 175 when compared to global regression. The Akaike information criterion (AIC) also decreases from 1610 to 1368.

Table 6: Comparison of regression results of tuberculosis

	Global regression	Geographically weighted regression
Variable	Coefficient	Coefficient (median)
<i>Gender</i>	-0.015	0.002
<i>Hospital</i>	-0.014	-0.032
<i>Age</i>	0.006***	0.007
<i>Social class</i>	0.021*	0.015
Observations	1747	1747
Multiple R ²	0.063	0.358
Adjusted R ²	0.061	0.061
RSS	255	175
AIC	1610	1368
F-statistic	29.35 (4 & 1742 DF)***	
Note:	* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$	

The precise results are documented in the appendix, within the R code section.

The maps below (figure 40) depict the spatial variance of the influence of the age and social class attributes in the tuberculosis case data. While the influence of age on the outcome is positive, as indicated by the almost homogeneously spread red values in most areas, the influence of social class is much more heterogeneously spread and subjected to more variance over shorter distances. Here, red indicates higher risk in higher social classes, while blue indicates higher risk in lower social classes. Nonetheless, some similarity with the local G^* clusters can be observed, such as the northern cluster in district 6, which repeats itself here, as well as the southern cluster in districts 7 and 8, which repeats itself too. A certain resemblance of the blue cluster in districts 3 and 4 is also evident.

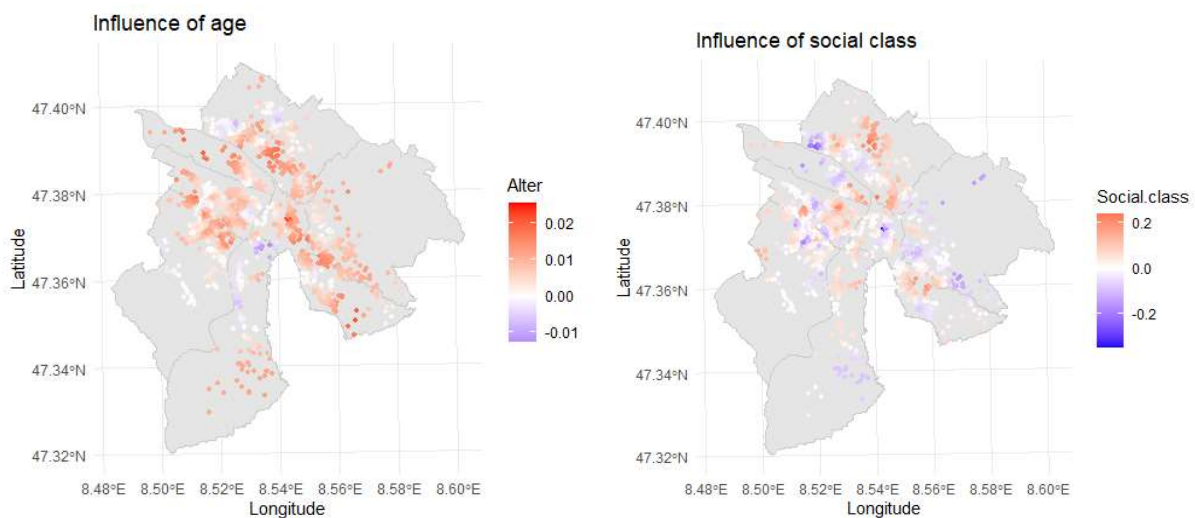


Figure 40: Spatial distribution of age and social class influences on tuberculosis mortality

The majority of tuberculosis cases each year have been documented during the late winter months, specifically in February and March, with a maximum of approximately monthly 50 cases. The minima are observed in late 1932 and 1933 with as few as 20 cases. The data reveal a wave-like pattern, with fluctuations occurring each year, reaching a maximum in late winter to early spring, thereby indicating seasonality.

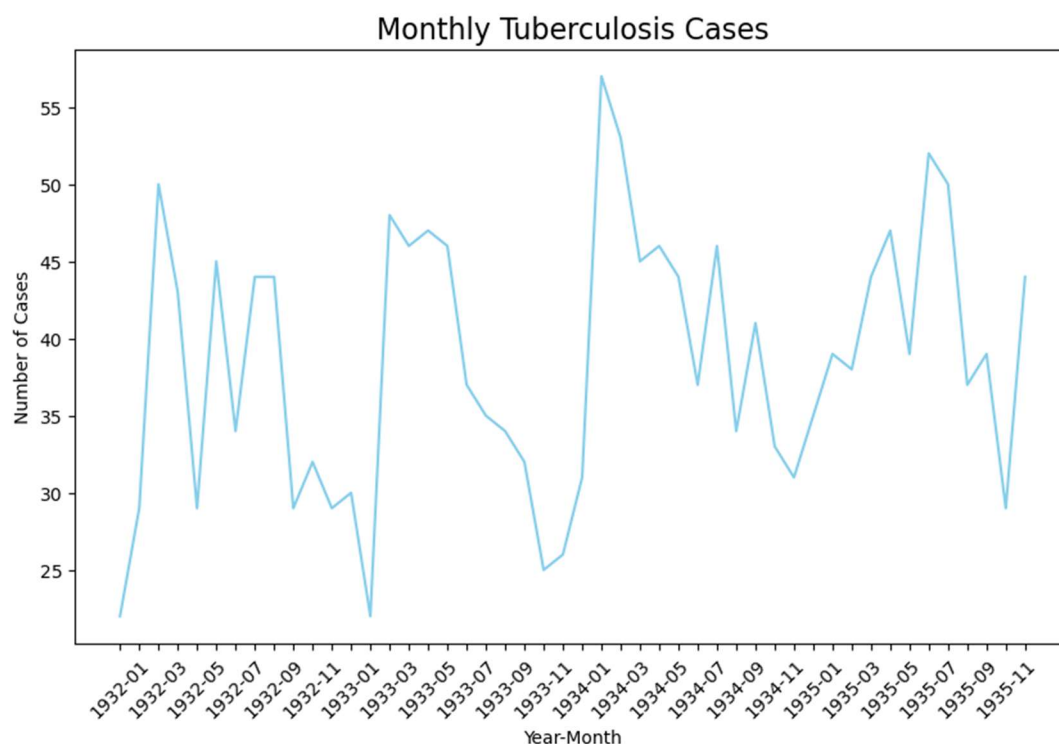


Figure 41: Tuberculosis case graph

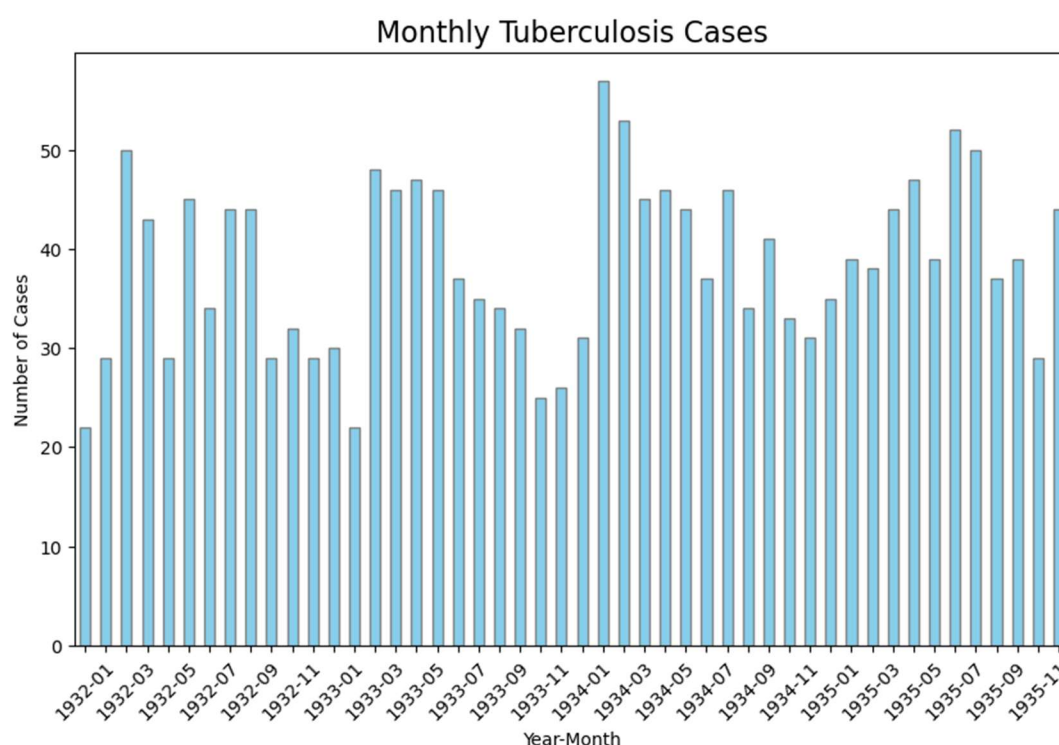


Figure 42: Tuberculosis case histogram

The tuberculosis case heatmaps for each year shown below (figure 43) exhibit minimal interannual variation. A substantial and consistent hotspot is observable in the working-class districts 3 and 4 across all four years, even extending into district 5. Meanwhile, a smaller, but similarly dense hotspot is evident in district 1 from 1932 to 1934, with a subsequent weakening observed in 1935. Two additional areas of interest are evident in districts 6 and 8, where medium density values were consistently recorded throughout the observed period. Conversely, districts 2 and 7 appear to have been relatively unimpacted by tuberculosis during the investigated period. A similar picture is shown by the animated map and the GIF created from monthly heatmaps. The clusters in district 1 (Niederdörfli) and districts 3 and 4 are remarkably stable throughout the four years, being present in most months of each year.

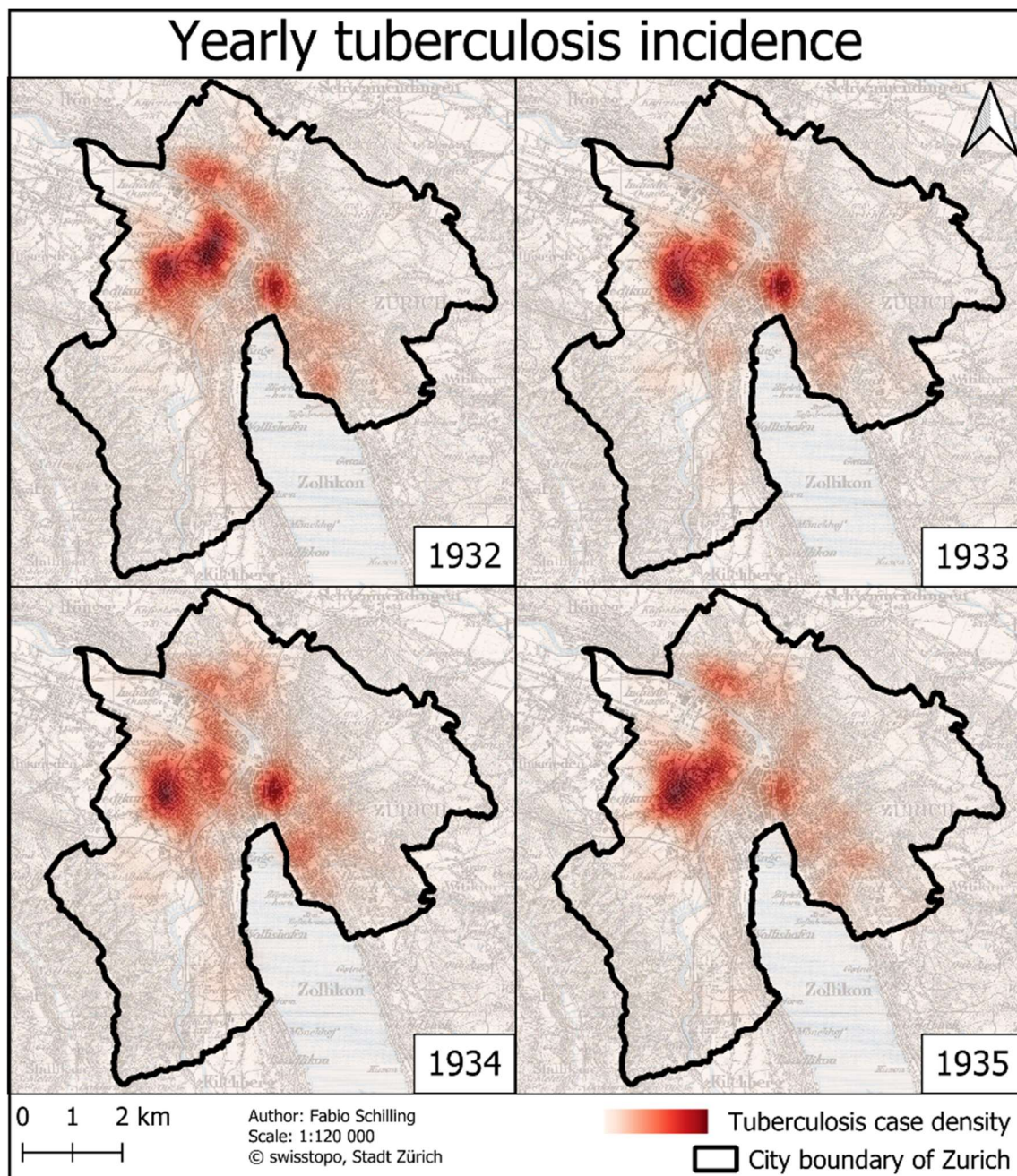


Figure 43: Temporal tuberculosis incidence

5.2 Diphtheria

The following diagram (figure 44) illustrates the age groups, and their respective number of diphtheria cases. The highest number of cases by far is observed in the youngest age group from 0 to 9 years, with 373 cases, constituting over half (58.5%) of the total registered cases. The next-older age group of 10-19-year-olds demonstrates a significant number of cases as well, with 138 cases (21.6%). Older age groups exhibit a decline in the number of cases, with the 60+ age group having no recorded cases. The total number of diphtheria fatalities is 19, which corresponds to a mortality rate of just 2.8%. Of the 19 fatalities, 12 (70.6%) were in the youngest age group. The three age groups up to 39 years of age have each one or two fatalities, while the age groups above 40 years of age have no fatalities.

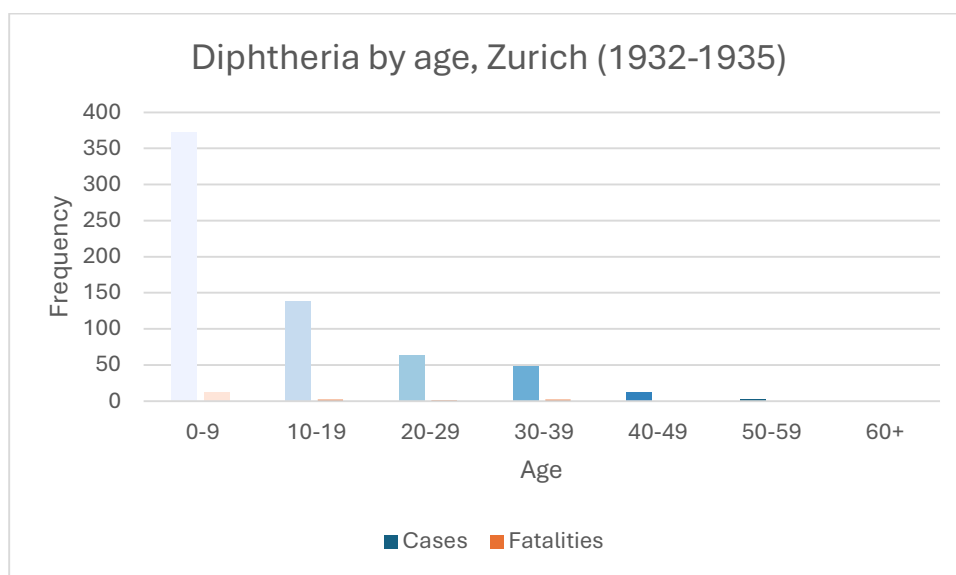


Figure 44: Diphtheria age distribution

The distribution of cases among the three social classes (figure 45) is an intriguing phenomenon, with the upper-class reporting the highest number of cases at 247. This figure is considerably higher than the relatively similar values observed for the lower- and middle-class, which registered 192 cases and 183 cases, respectively. Notably, 48 cases could not be allocated to any specific social class. The upper-class also exhibited the highest number of fatalities, with 8 deaths, closely followed by the middle-class with 7 deaths. The lower-class demonstrated the lowest number of fatalities, with a mere 4.

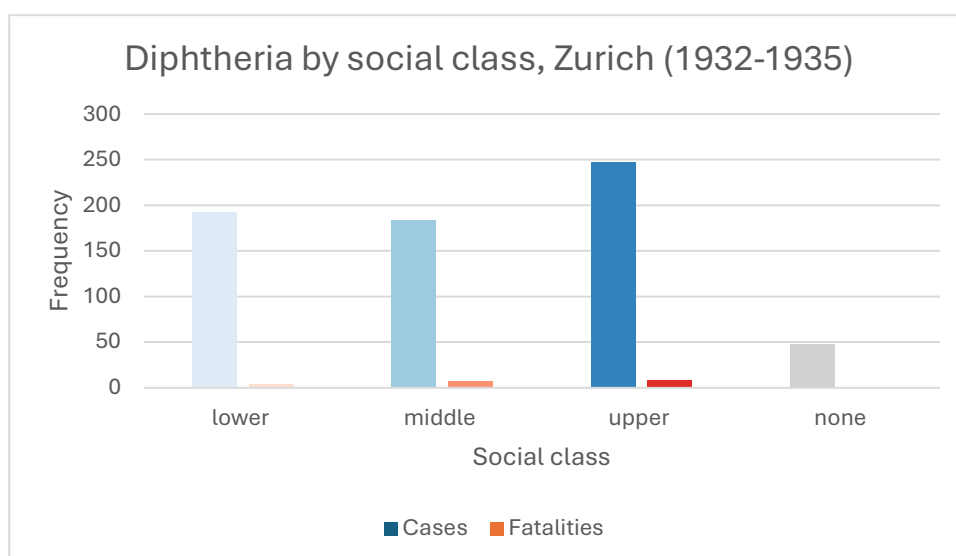


Figure 45: Diphtheria social class distribution

The following diagram (figure 46) presents the distribution of diphtheria cases and fatalities by district. As with the data on tuberculosis, the location of the mapped cases is the principal source of information used. However, due to the unavailability of addresses for approximately 200 cases, the missing entries will also be considered, following the description of the cases extracted from the geolocated addresses. The district with the highest number of cases is district 6, with 90 cases (13.5%), followed by districts 7 and 8, with 69 cases each (10.3%). District 5 has 64 registered cases. The remaining districts (1 through 4) all have fewer than 60 cases, with the lowest number found in district 2 at just 28 cases (4.2%). Of the 19 fatalities, 5 (or approximately one-third) occurred in district 8, 2 in district 4, and 1 in every other district, except district 3, where no fatalities were recorded. The distribution of cases across districts, as indicated in the district column of the original data, is broadly reflected in the current data. However, it is important to note that the distribution of fatalities is not consistent with the distribution of cases. In district 8, for example, 6 deaths were reported, while in district 4, 4 deaths were reported. In districts 6 and 3, 3 and 2 fatalities were reported, respectively, and a single fatality in every remaining district.

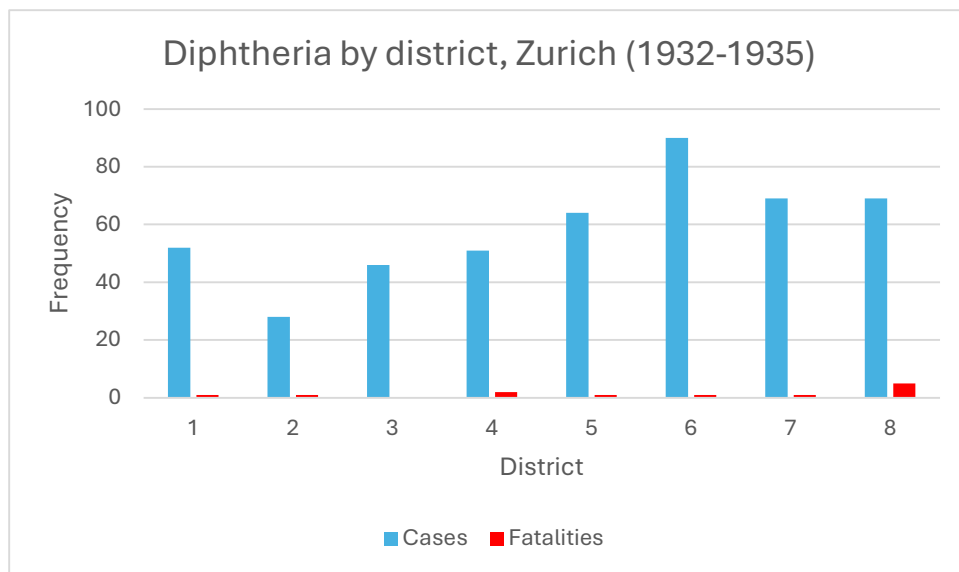


Figure 46: Diphtheria district distribution

The distribution of diphtheria cases in figure 47, represented as green dots, appears to demonstrate an even distribution at first glance. However, upon closer inspection, it becomes evident that the number of cases in district 2 is minimal. In addition, a relatively high density of cases has been observed in several locations, including district 8. This higher density is also present in the Niederdörfli neighbourhood of district 1 on the east side of the Limmat river, as well as in district 5 between the railyards to the south and the Limmat river to the north. Medium case densities are found in district 6, district 7, and districts 3 and 4 west of the Sihl river.

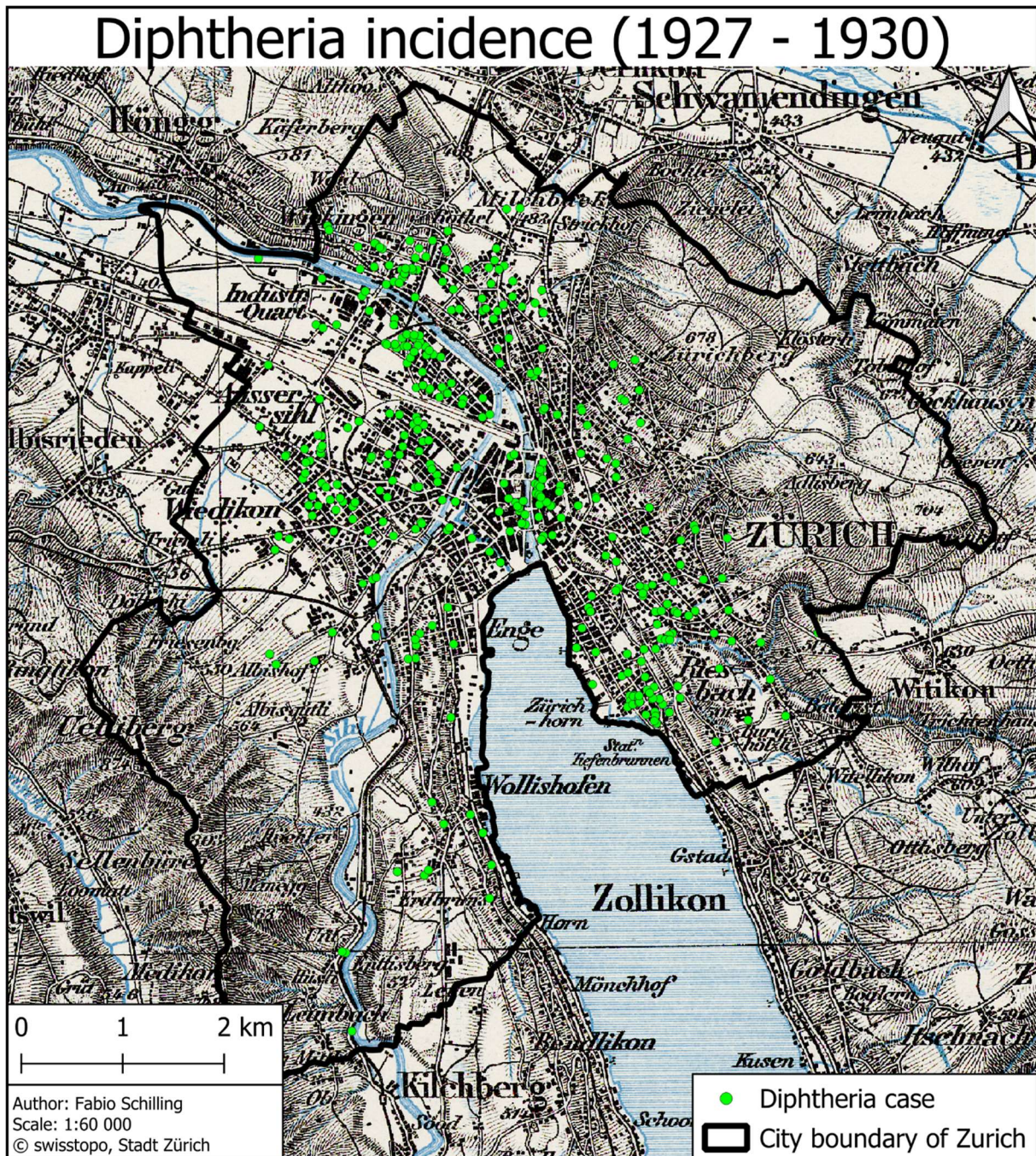


Figure 47: Diphtheria case map

The distribution of diphtheria fatalities (figure 48), symbolised as red dots, demonstrates a minimal number of fatalities in all districts, as previously mentioned. The only exception to this is district 8 on the eastern lakeshore, where approximately half of all fatalities are mapped. The remaining fatalities are dispersed throughout the city, extending from district 6 in the north to district 2 in the south and from district 4 in the west to district 7 in the east.

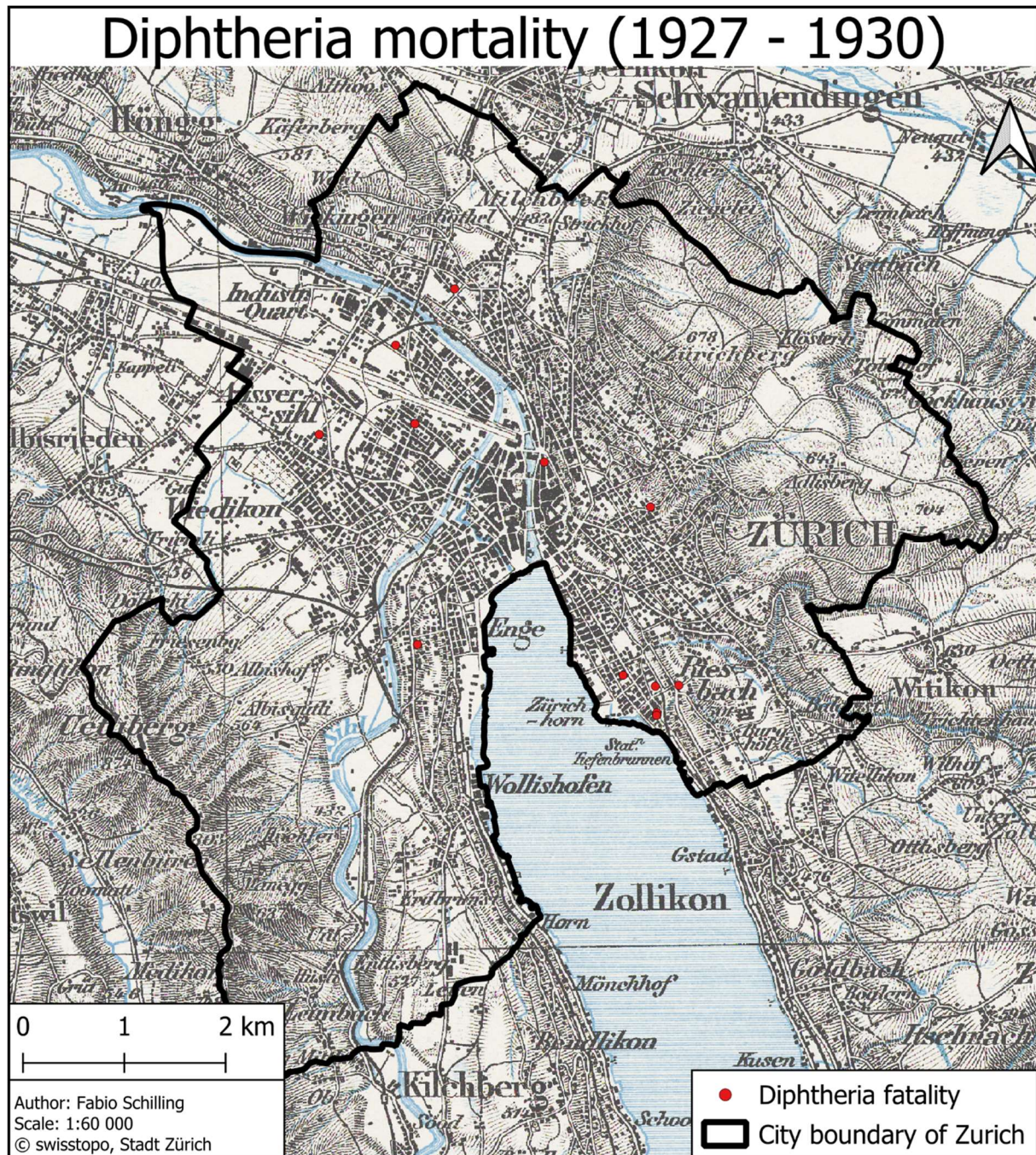


Figure 48: Diphtheria mortality map

The cartograms below (figures 49 and 50) offer a visual representation of the district-level diphtheria morbidity and mortality rates per 1'000 and 100'000 inhabitants, respectively. The districts with the highest morbidity rates are districts 5 and 8, followed by districts 1 and 7 with above-average morbidity rates, which are symbolised by the dark blue and blue colour hues, respectively. District 6, shown in light blue, has a below-average morbidity rate. Finally, districts 2, 3 and 4 exhibit the lowest morbidity rates, represented by a white colour hue. Diphtheria morbidity rates appear similar to the diphtheria morbidity rates. The districts with the highest mortality rates are districts 5 and 8, symbolised by their dark red colour hue, consistent with the morbidity rates. These are followed by districts 1 and 4, which exhibit above-average mortality rates and are represented by red colour hues. The districts 2 and 7 exhibit below-average mortality rates, indicated by the light red colour hues. Finally, districts 3 and 6 exhibit the lowest mortality rates, symbolised through white colour hues. However, it should be noted that due to the very low number of deaths, the cartogram may not be significant. Both maps are distorted by the number of cases and fatalities, respectively, and the mortality cartogram is exaggerated.

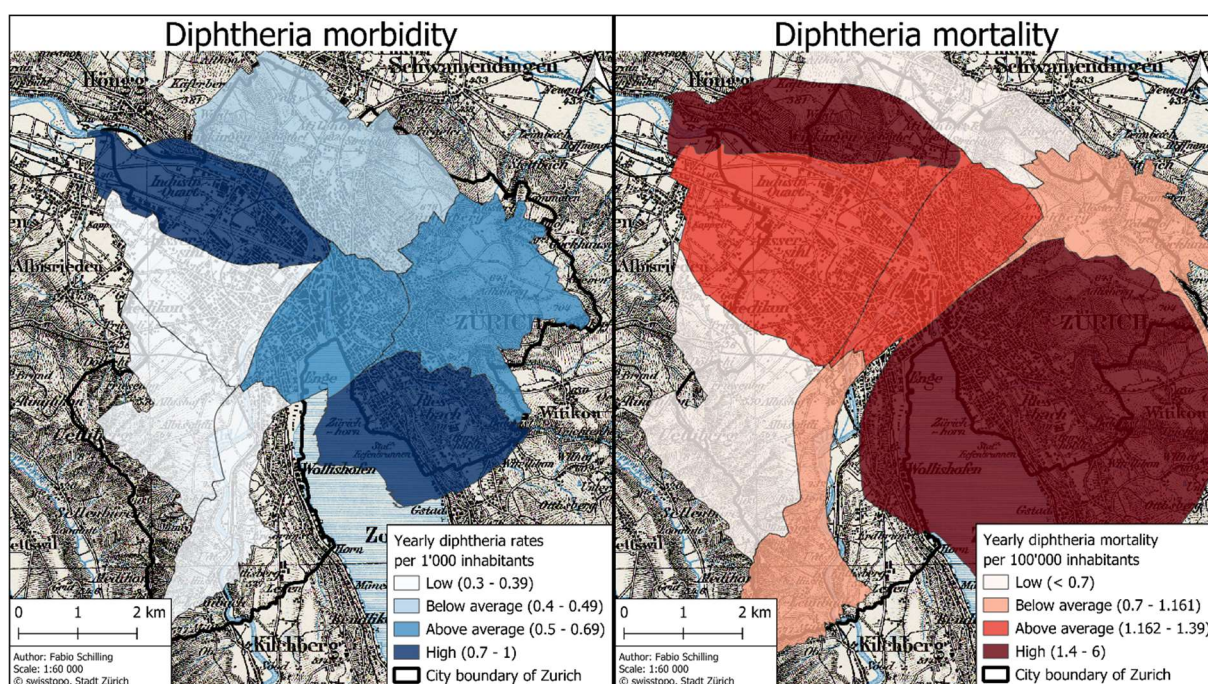


Figure 49: Diphtheria morbidity cartogram

Figure 50: Diphtheria mortality cartogram

The results of the HDBSCAN clustering for diphtheria cases are presented below (figure 51). The points have been coloured according to their assigned cluster, and those points assigned to no cluster remain white. As was the case for tuberculosis, the districts are roughly delineated by the clusters, although there are only six total clusters here. The green cluster encompasses the majority of districts 7 and 8. The orange cluster covers district 1 and the rest of district 7. The lavender-purple cluster primarily corresponds to district 6. The yellow, pink and blue clusters, in turn, approximately cover districts 5, 4, and 3, respectively. Meanwhile, no cluster is present in district 2.

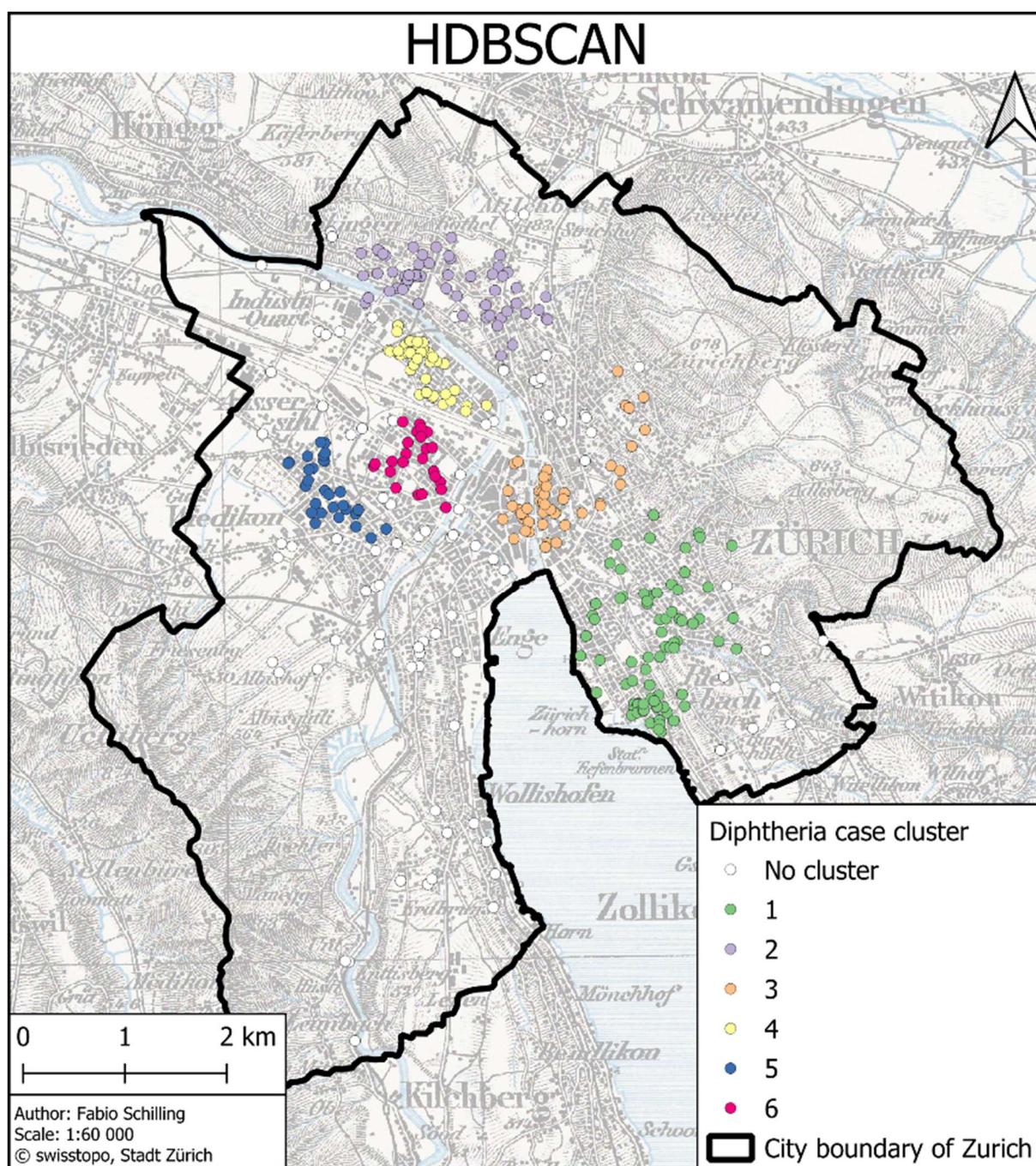


Figure 51: HDBSCAN clusters of diphtheria cases

When the HDBSCAN clustering is adapted to include the social class, a new pattern emerges (see figure 52). The cases belonging to the upper- and middle-classes coalesce into a single cluster per class, which is designated as red or orange, respectively. Meanwhile, the lower-class clusters are subdivided into three distinct clusters, designated dark blue, blue and light blue colour, respectively. The dark blue cluster, labelled “Lower-class 3”, contains points in districts 3, 4, 5 and 6, i.e. the north and north-west of Zurich. The light blue cluster, labelled “Lower-class 1”, encompasses cases in the southern districts 7 and 8, while the blue “Lower-class 2” cluster contains cases in district 1. Cases that could not be assigned to a cluster, are marked in grey and labelled “Not clustered”.

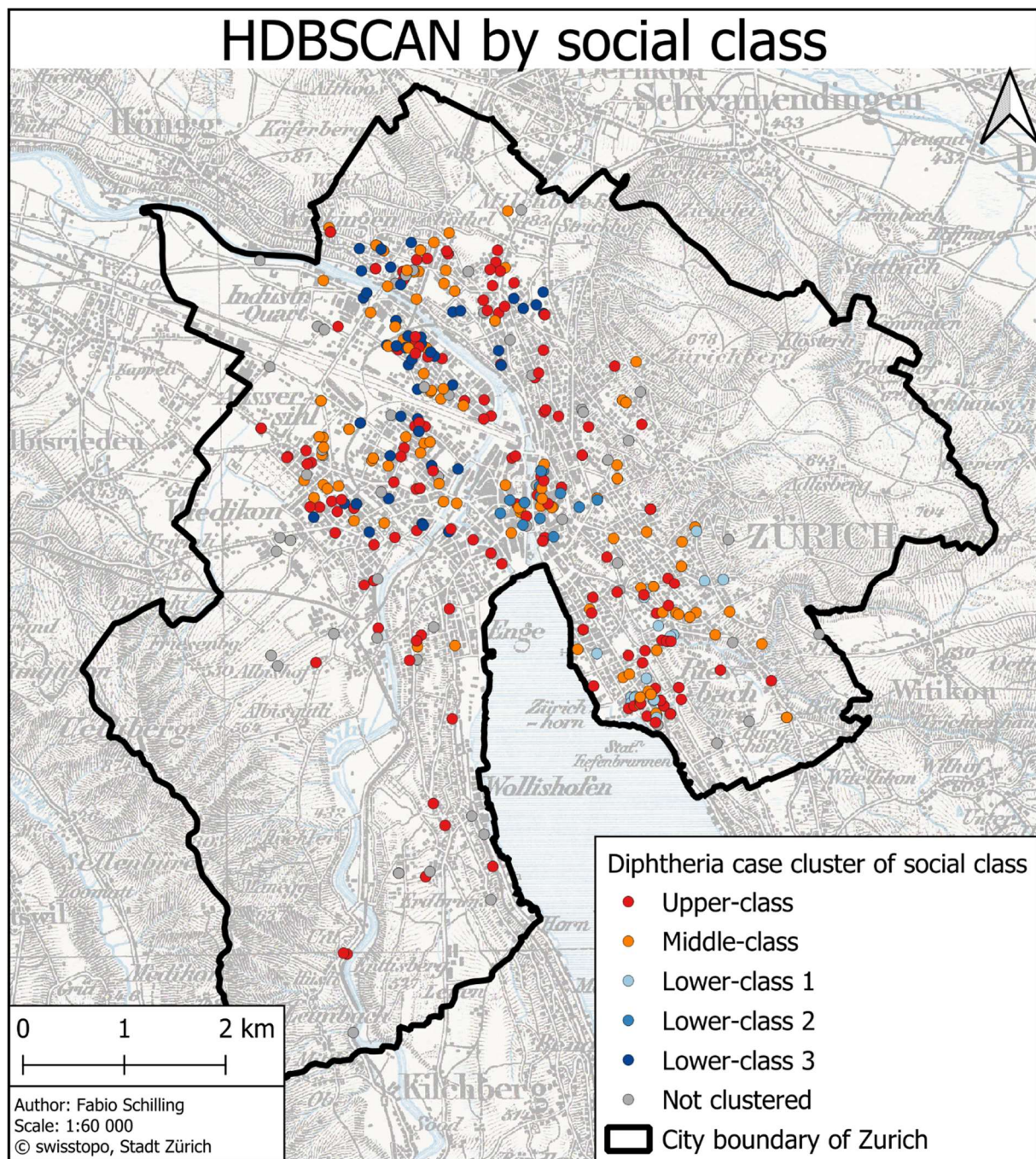


Figure 52: HDBSCAN clusters of diphtheria cases by social class

The social class distribution of diphtheria cases (figure 53) exhibits a complex pattern of social classes diversity throughout the city, with certain social classes seemingly pre-dominant in specific districts. Nevertheless, high heterogeneity is the predominant feature, with red points, denoting the upper-class, being the most prominent. District 2 is mostly dominated by those red dots, indicating the upper-class, as are districts 7 and 8, all of which are characteristically affluent areas. For the remaining districts, the heterogeneity is so high that any statements are challenging to make.

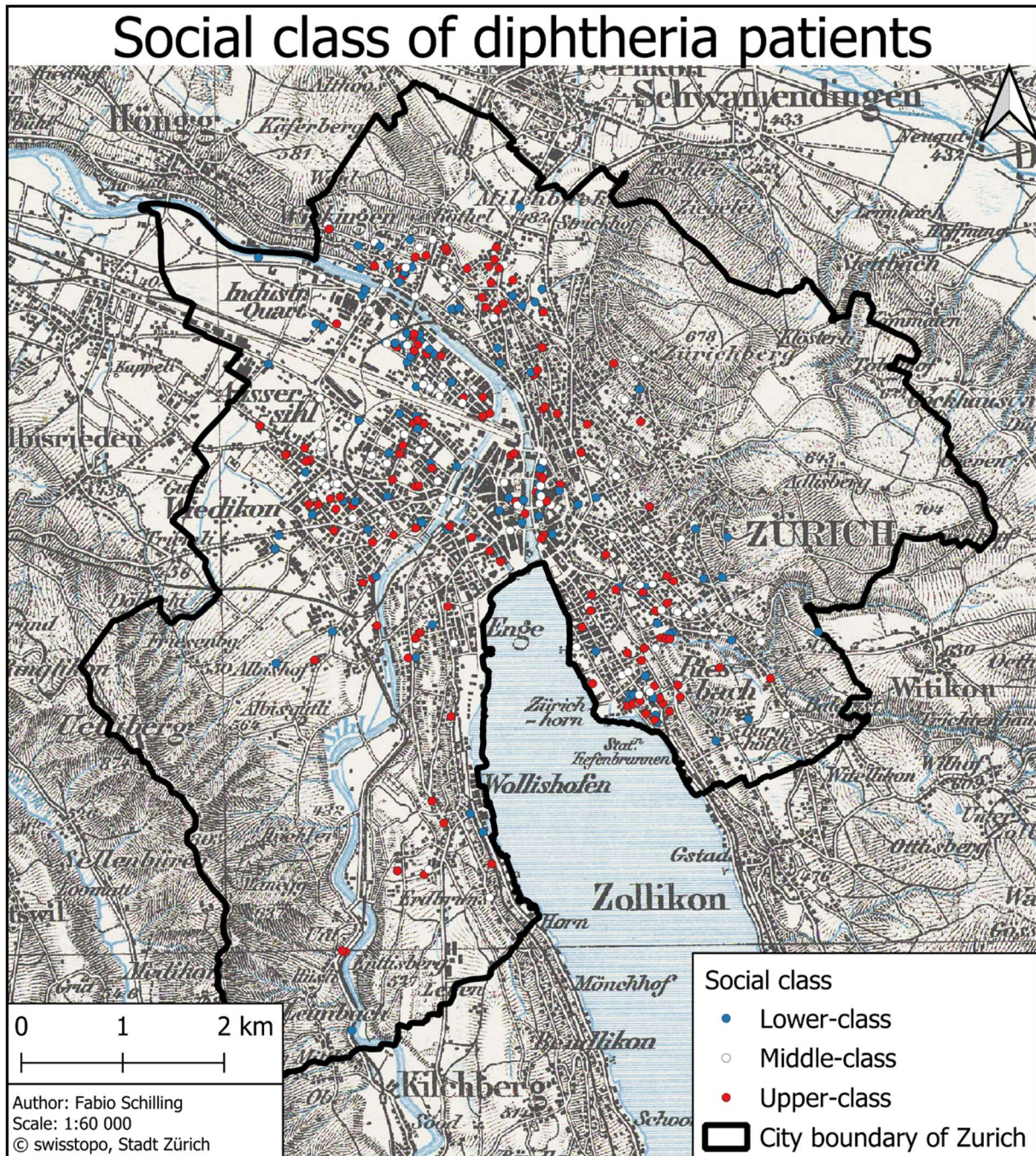


Figure 53: Social class map of diphtheria cases

The subsequent map (figure 54) illustrates the schools, along with their respective catchment areas, which have been calculated as Voronoi polygons, superimposed on the density distribution. It appears that the catchment areas of some schools in the districts 6, 5, 1 and 8 correlate with the major hotspots of diphtheria cases. Additionally, smaller yet potentially significant hotspots, which also correlate with school catchment areas, have been identified in districts 3, 4 and 7.

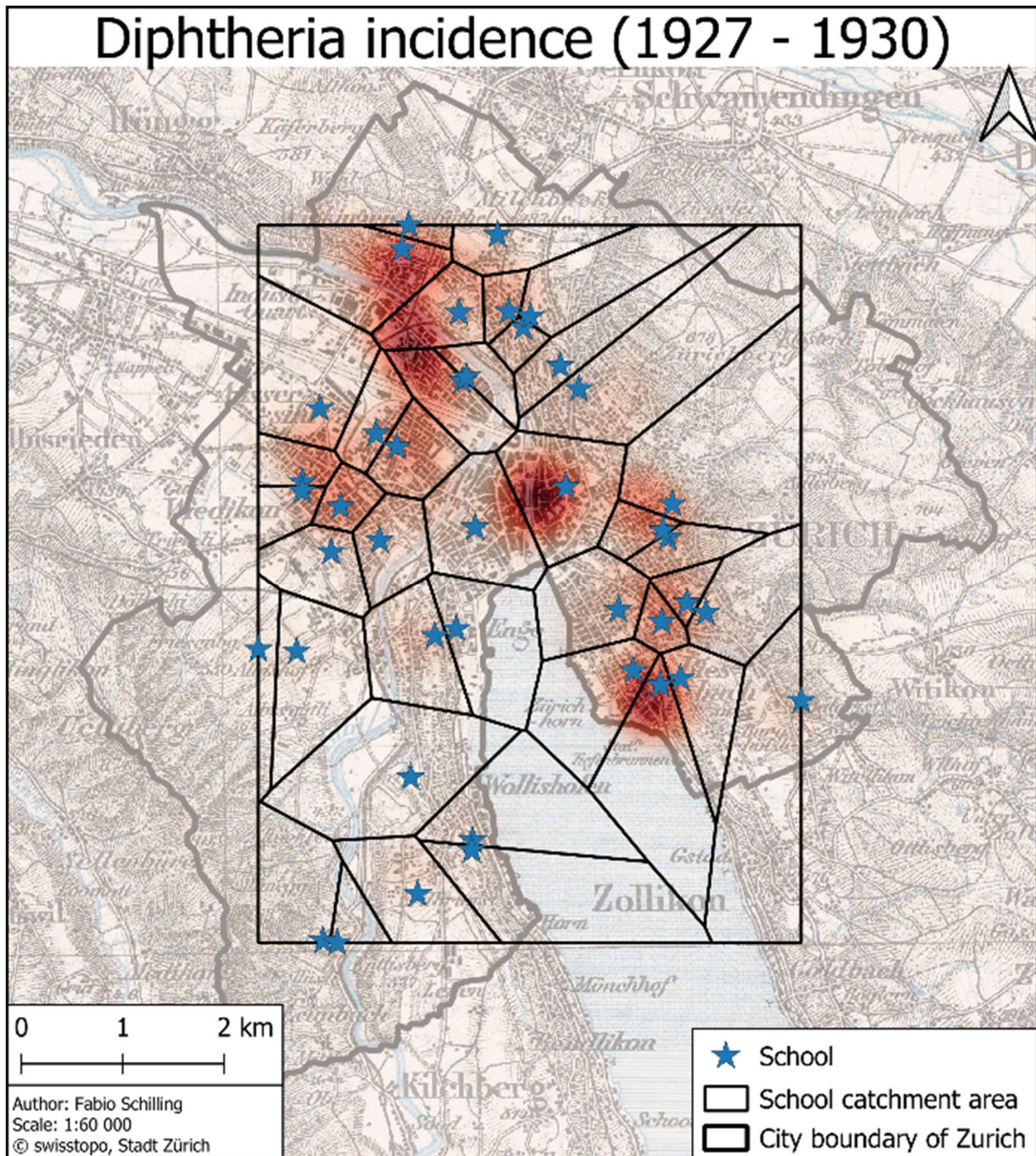


Figure 54: Diphtheria incidences overlayed with school catchment areas

The NNI of the diphtheria cases is 0.534. The z-score is -19.3, with a p-value that is almost equal to 0. The nearest neighbour distances range from 0 to 775m, with a median of 33m and a mean of 65m. The distribution, as illustrated below (figure 55), is predominantly concentrated at the shortest nearest neighbour distances, with the majority falling below 200m. Notably, a considerable proportion of approximately 200 nearest neighbour pairs are situated at the very shortest distance. The NNI of the diphtheria fatalities was not calculated due to the low number of fatalities and the resulting low significance of such a value.

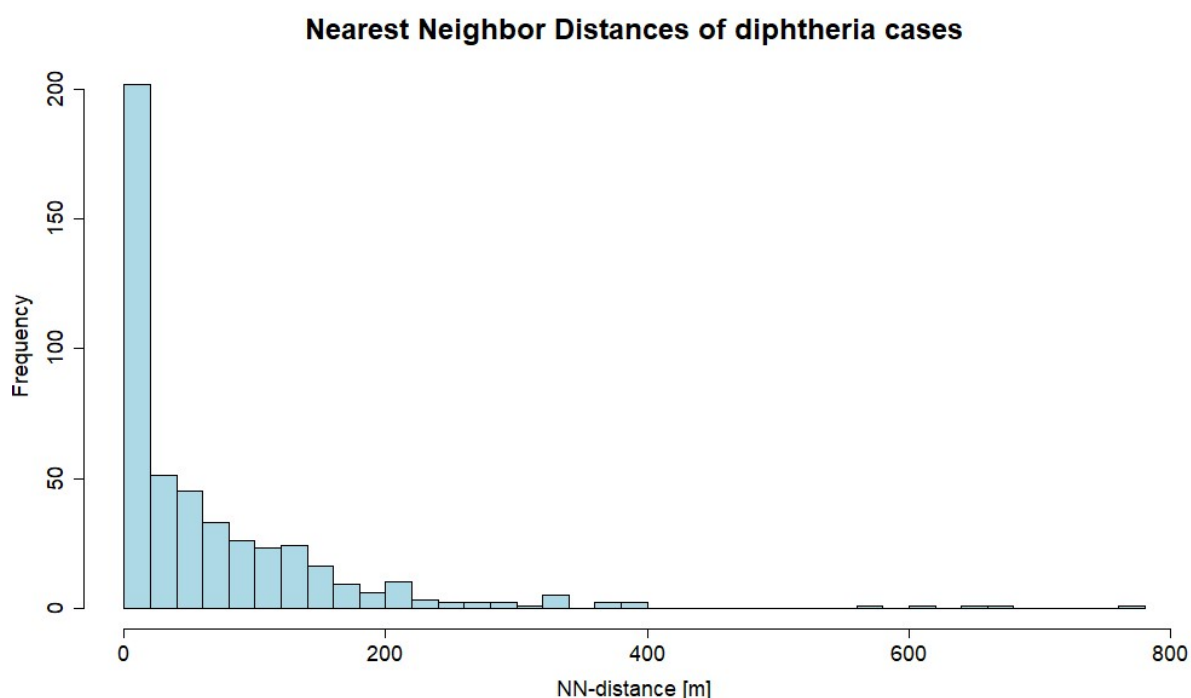


Figure 55: NN-distances of diphtheria cases

The G-function of the diphtheria cases (figure 56) exhibits a marked increase in the initial metres, where the cumulative distances rapidly approach 40% of the nearest neighbour pairs. Thereafter, on the graph line continuously rises to reach 80% at approximately 125m and crosses the 50% mark at approximately 40m. The black line representing the observed data consistently lies above the red reference line of a theoretical distribution throughout the entire graph. Again, I deemed it inappropriate to calculate the G-function for the diphtheria fatalities, due to the extremely limited data available, which would limit the significance of any result.

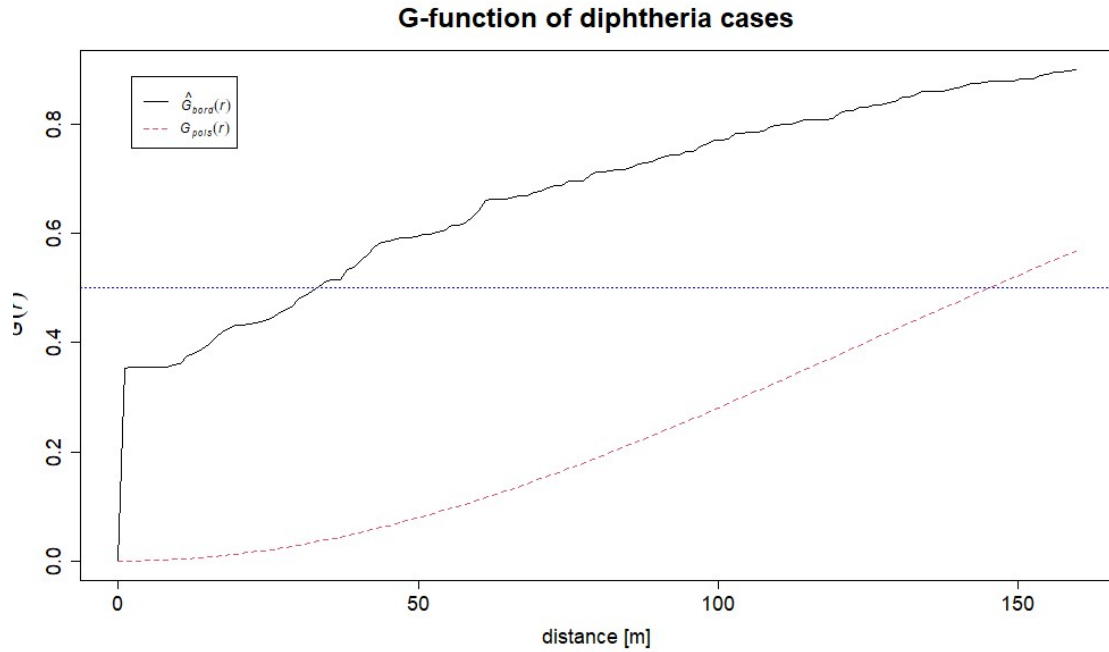


Figure 56: G-function of diphtheria cases

The result of Moran's I of diphtheria cases on district level is positive, as indicated by the positive slope of the graph (figures 57 and 58), with a value of 0.349. Conversely, the value of the index of diphtheria fatalities is negative at a value of -0.133.

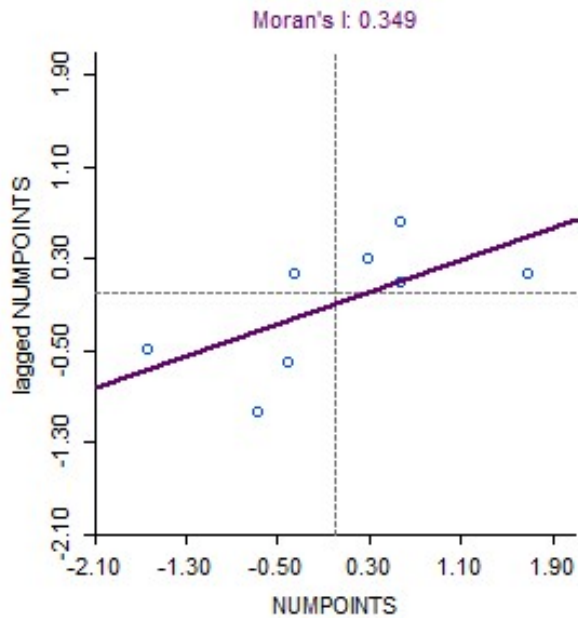


Figure 57: Moran's I of diphtheria cases

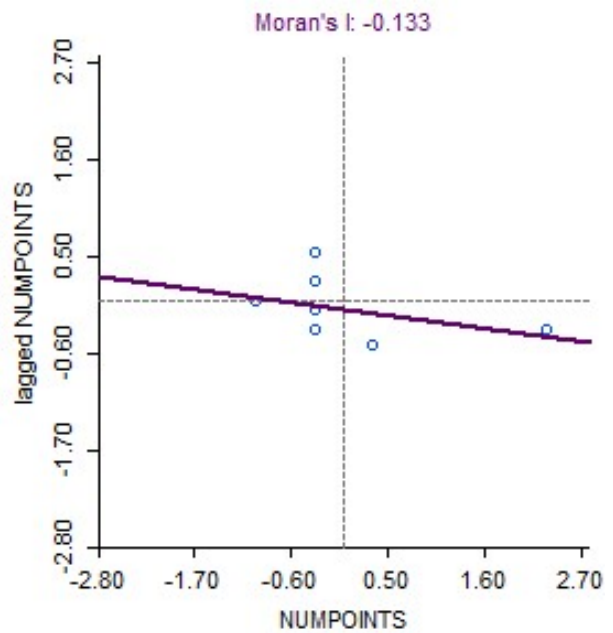


Figure 58: Moran's I of diphtheria fatalities

The result of local G^* for social class of diphtheria data (figure 59) demonstrates the presence of two distinct clusters: one in red, which corresponds to the upper-class demographic, and the other in blue, which corresponds to the lower-class demographic. The upper-class cluster is primarily located in district 8, with a small number of red points that are also located in districts 2 and district 6. Meanwhile, the lower-class cluster is predominantly situated in the working-class districts 4, 5 and 6, where only the neighbourhood of Wipkingen is affected. The remaining white points are of negligible significance and are not part of any cluster.

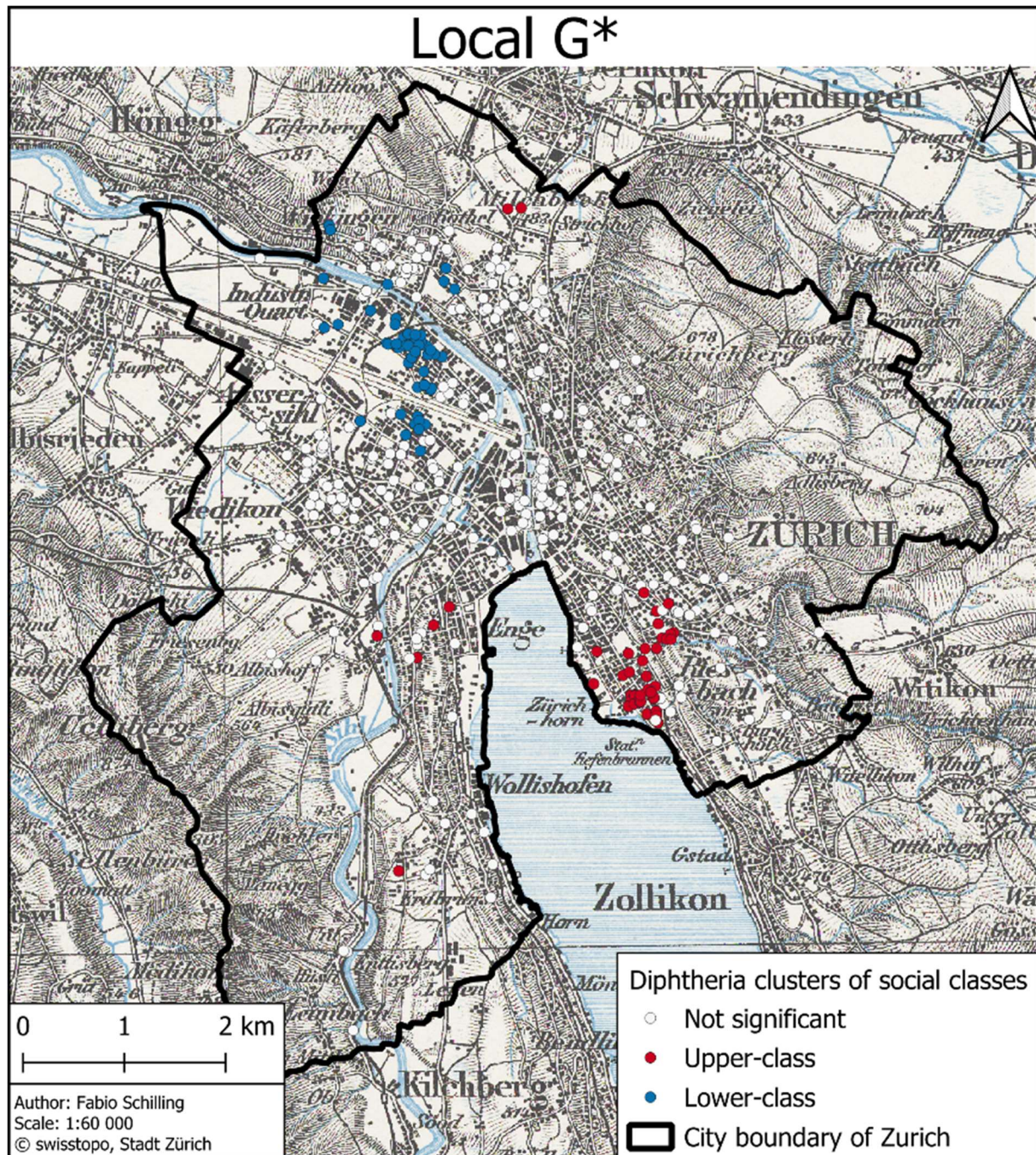


Figure 59: local G^* cluster map of social class in diphtheria data

The regression analysis of all diphtheria cases revealed that the attributes gender and age had a negative value, while the attributes hospital admission and belonging to higher social classes had positive values. The R-squared value is 0.009 and the adjusted R-squared is 0.001 with a p-value of 0.366.

The geographically weighted regression attests to spatially varying influences on all four investigated attributes (see table 7). The only positive median concerns the social class attribute, while all others have negative medians. In comparison to global regression, R-squared is increased from 0.006 to 0.325, while the adjusted R-squared increases from -0.006 to 0.024. The residual sum of squares (RSS) has been reduced from 11 to 7 and the Akaike information criterion (AIC) is -180 and smaller for GWR at -239. The results of the visualised coefficients are comparable, although reversed, when compared to the local G*.

Table 7: Comparison of regression results for diphtheria

	Global regression	Geographically weighted regression
Variable	Coefficient	Coefficient (median)
<i>Gender</i>	~0	-0.025
<i>Hospital</i>	~0	-0.023
<i>Age</i>	~0	-0.0003
<i>Social class</i>	~0	0.006
Observations	326	326
Multiple R ²	0.006	0.325
Adjusted R ²	-0.006	0.024
RSS	~11	~7
AIC	-180	-239
F-statistic	0.49 (4 & 321 DF)	
Note:	* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$	

More detailed results of the calculated regressions can be found in the appendix.

The influence of age in the case of diphtheria is either overwhelmingly negative or neutral, as indicated by the blue and white colours. However, a notable exception is observed in district 8, where a cluster of slightly red points is evident. Conversely, the influence of social class is predominantly positive, as indicated by the red points, or neutral, as indicated by the white points, although there are some blue points clustered in districts 2 and 8. The red points are also observed to be clustered in districts 4, 5 and 8 (see figure 60).

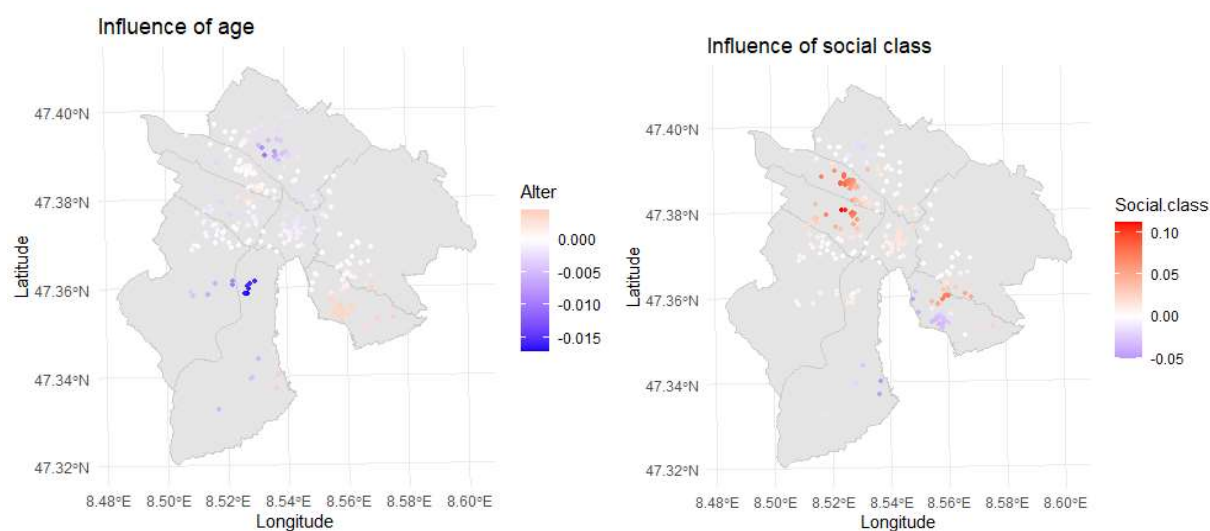


Figure 60: Spatial distribution of age and social class influences on diphtheria mortality

The temporal progression of diphtheria cases over time (figures 61 and 62) reveals two peaks rising from a baseline of approximately 10 to 20 monthly cases. The initial peak, which occurred around the turn of the year 1927 to 1928, reached approximately 25 monthly cases. From mid-1928 onwards, the baseline increases to approximately 20 cases per month, peaking in the winter of 1929 /1930 at over 40 cases per month.

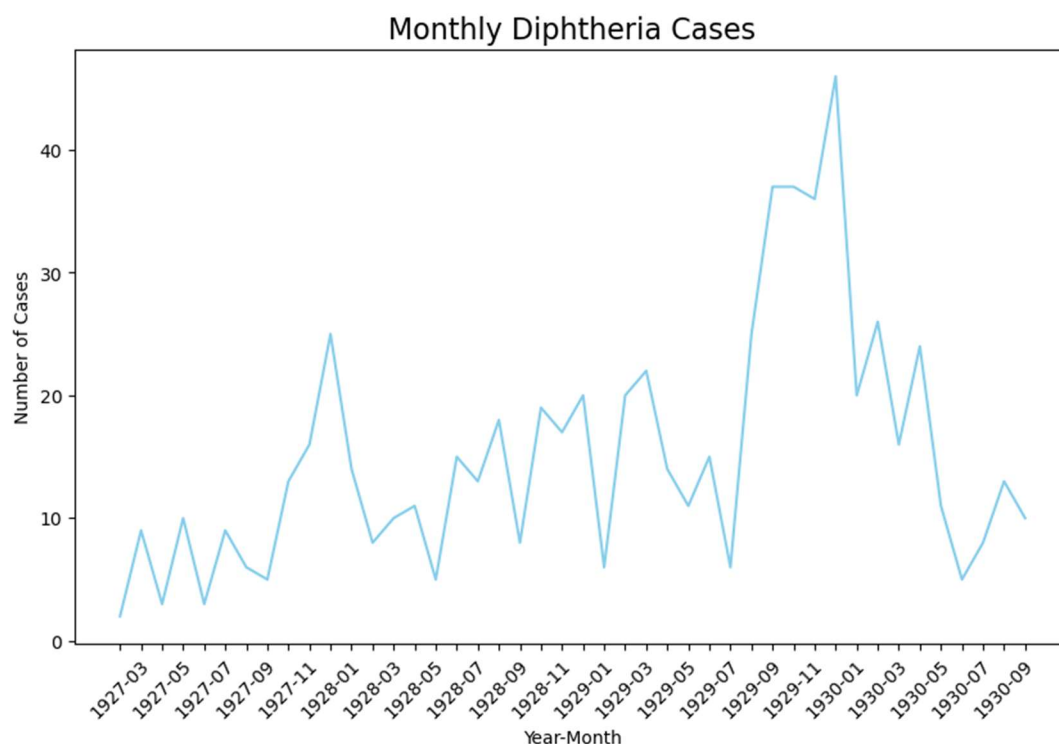


Figure 61: Diphtheria case graph

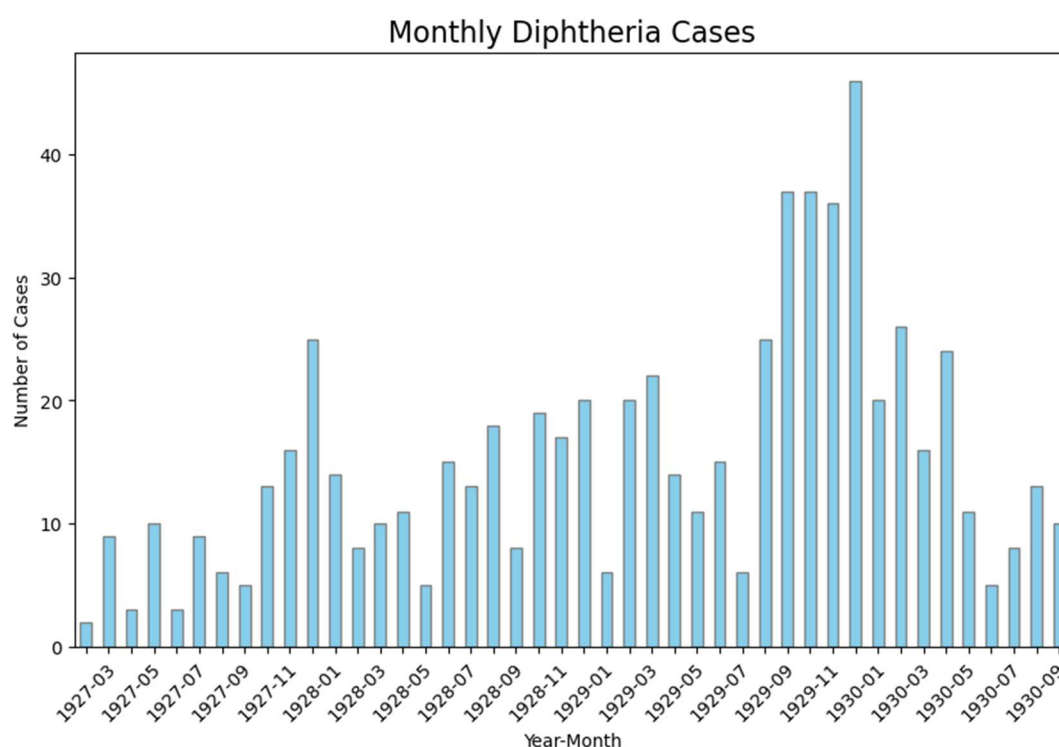


Figure 62: Diphtheria case histogram

The yearly diphtheria incidence (figure 63) varies considerably in both extent and density. Nevertheless, some general trends are still identifiable. These include high-density hotspots in the districts 1, 5, 6 and 8 throughout at least three of the four years in question, as for 1928 very little data was available. A medium-density hotspot has been observed in districts 3 and 4 in all years except 1928, where data is limited. The distribution of diphtheria displays greater variability than that observed for tuberculosis, manifesting as a greater degree of fluctuation in the individual hotspot locations across the city of Zurich. However, it is evident that districts 2 and 7, to a lesser extent, were predominantly unaffected by diphtheria cases. The animated map and GIF of monthly diphtheria heatmaps corresponds to the findings mentioned before. They are characterised by high variability in the hotspot locations and sizes. This underlines the temporal instability of the diphtheria case pattern.

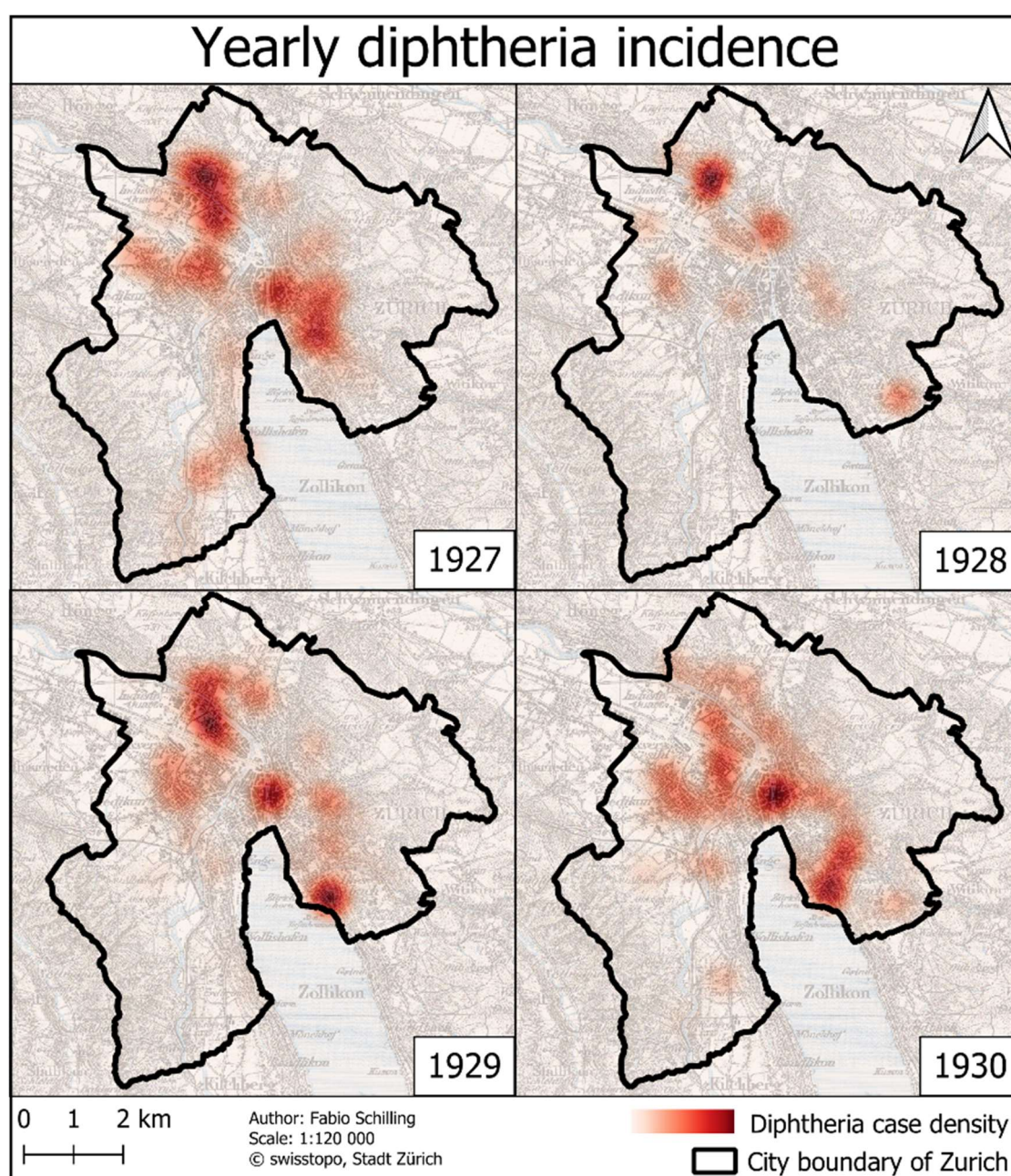


Figure 63: Temporal diphtheria incidence

6 Discussion

A key factor to consider in disease analysis is the population at risk. A comparison of the population at risk with the reported disease cases can facilitate the identification of where an unexpectedly high or low number of cases has occurred, indicating either an excess or a shortfall of infections. However, in order to formulate expectations, references are required to compile what disease numbers are to be expected. Numerous risk factors influence this number, such as age, sex, occupation, social class and socioeconomic status. As previously mentioned, higher infection numbers could be expected in neighbourhoods and districts with a low socioeconomic status, where the majority of the population is likely working-class (Coleman, 2018; Holloway et al., 2014, 2013; Kistemann et al., 2002; Müller et al., 2024). Furthermore, younger adults, predominantly those in their 20s and 30s, and males in general are more susceptible to tuberculosis infection due to a greater number of work-related social contacts (Horton et al., 2020; Humayun et al., 2022; Miller et al., 2021; Peer et al., 2023) and higher personal risks, often linked to unhealthy habits which are mainly perceived to be typical for males, e.g. smoking and drinking (Humayun et al., 2022; Nhamoyebonde and Leslie, 2014; Peer et al., 2023). School-age children represent the age group with the highest risk of diphtheria infection, due to their comparatively weaker immune systems and lack of natural immunisation (Byard, 2013; Truelove et al., 2020).

6.1 Tuberculosis

As indicated by the literature, the incidence of tuberculosis over the observed 4-year period is higher among males with 982 cases (53%) compared to 870 cases (47%) among females. This phenomenon is all the more noteworthy given that the city's population is majority female at 53.8%, or roughly 131'000 out of 244'000 inhabitants. It can be attributed to a number of factors including higher behavioural and physiological risks among men (Humayun et al., 2022; Nhamoyebonde and Leslie, 2014; Peer et al., 2023), or heightened exposure through social contacts and work (Horton et al., 2020; Humayun et al., 2022; Miller et al., 2021; Peer et al., 2023).

The relatively low mean and median ages of 31 and 30 years, respectively, indicate a prevalence of the disease among younger age groups. The age distribution of tuberculosis cases indicates excessive case numbers, especially among individuals aged 20-29, and less pronounced among those aged 30-39, mirroring the average and median age of infected persons discussed previously. The 20–29 age group accounts for a third (33.5%) of all tuberculosis cases, despite comprising less than a quarter (23.6%) of the overall population. In the 30–39 age group, the proportion of cases experienced by this age group was almost a quarter (23.2%), despite their population proportion being less than a fifth (18.6%). This finding suggests that these two age groups exhibit what can be termed excess morbidity, i.e. a higher proportion of tuberculosis cases than would be expected based on their relative share of the total population, if the disease distribution was random and did not exhibit clustering. The dominant explanation for this distribution is the

high number of social contacts this population group has due to work and social behaviour, making them more exposed and therefore more susceptible to infection (Chirenda et al., 2020; Corti, 2012; Gubéran, 1980; Shaweno et al., 2018). The age distribution of tuberculosis fatalities demonstrates that the proportion of the population younger than 20 years is negligible at just 3%. In older age groups the proportion of tuberculosis fatalities align with population shares but simultaneously increase with age. The mapped coefficients of the influence of age indicate that increasing age is associated with a heightened risk of fatal tuberculosis infection across most regions of the city. This phenomenon can be rationalised by the established correlation between age-related frailty and diminished resistance to infectious diseases, which in turn increases the probability of death. Notable exceptions to this pattern include Enge in district 2 and two spots in district 6, where the data suggest a decrease in mortality with increasing age

The case-fatality ratio displays a marked increase with age, as older age groups demonstrate excess mortality when contrasted with their population shares. The most affected age group is the over-60s, as people become more susceptible with age. The highest case-fatality ratios are observed among the oldest age group, with approximately 50% of cases. This observation could be indicative of heightened resistance due to peak fitness in younger age groups or, alternatively, it could be attributed to more generally improved living conditions. The suggested trend of mortality risk increasing with advancing age may be attributed to the well-documented increase in susceptibility to disease with increasing age.

The total of over 2000 registered cases is reasonably close to the expectations based on other data published in this context, as it amounts to approximately 500 cases per year in a population of 250'000 or roughly 200 per 100'000. For context, publications concerning nationwide, cantonal and city-level figures have estimated the tuberculosis mortality rate at 124,5 per 100'000 in 1930 and at 82,4 per 100'000 in 1938. The canton of Zurich has proudly been commended for its effective tuberculosis countermeasures (Ritzmann, 1998, pp. 36–37) and has reported a rate of 91,2 per 100'000 between 1930 and 1934, which corresponds to the timeframe under consideration, and 9,7 per 10'000 between 1926 and 1935. Consequently, the expected order of magnitude would be approximately 100 per 100'000 or 1 per 1000. In the 1930s, the canton of Zurich registered approximately 5000 tuberculosis fatalities per decade, or a yearly death rate of 8 per 10'000 living (80 per 100'000). On average, approximately 260 people died from tuberculosis in the city of Zurich alone each year (Kruker and Senti, 1932; Senti and Pfister, 1946). The mortality rate calculated from the dataset under consideration ranges from less than 30 to 50 per 100'000 inhabitants. This figure is lower than the reference suggests. A similar discrepancy was observed in the total number of fatalities, which, according to the available data, amounted to approximately 380 over a period of four years. This discrepancy can be partially attributed to missing entries of potentially undetected cases, which could not be located and thus were not integrated into the maps. However, this does not fully account for the observed inconsistency, which is likely attributable to undetected cases.

Multiple methods, including NNI, G-function, Moran's I and HDBSCAN indicate the presence of clustering in the data by their outcomes. The dot maps reveal a higher density of cases and fatalities in districts 1, 3, 4, 5 and 6, which represent the traditionally working-class districts. Clustering in the data is identified by the NNI and Moran's I, which are within the thresholds that indicate clustering and are significant. The G-function also indicates clustering, as the data graph consistently exceeds the red reference line of a random distribution. Furthermore, the degree of clustering is more pronounced in the cases, as their indicators point towards stronger clustering by means of a lower NNI and shorter nearest neighbour distances. However, Moran's I contradicts this, as the smaller value for cases suggests higher clustering in fatalities on district-level. HDBSCAN shows a strong resemblance between the data clusters and the districts themselves, suggesting a potential correlation between low socioeconomic status and tuberculosis incidence in working-class districts (Ritzmann, 2010a, 1998, pp. 36–37). Furthermore, the upper class exhibited the lowest number of cases, while tuberculosis appeared to be more prevalent among the middle and lower classes. However, the clusters also align with the areas of highest population density. A comparison of the number of cases per district to the district's population share of the city demonstrates that districts 1, 3 and 4 still exhibit above-average values. Conversely, in the affluent districts 2, 7 and 8 the case numbers were generally lower. These findings suggest that in densely populated working-class districts, the spread of disease may be facilitated due to the lower socioeconomic status of the inhabitants, which is supported by literature and includes bad working and living conditions (Chirenda et al., 2020; Gubéran, 1980; Hermans et al., 2015; Kanturk, 2007; Liu et al., 2012; Setiawan et al., 2021; Shaweno et al., 2018; Sun et al., 2015; Tiwari et al., 2006; Vaughan, 2018).

The distribution of tuberculosis cases and fatalities by social class is not immediately apparent, as it is characterised by a high degree of heterogeneity in the dot maps. This challenges the notion of socially homogeneous districts, which are inhabited by one single, dominant social class. However, when including social class into clustering methods such as HDBSCAN, local G^* , and regression coefficients, socioeconomically motivated clusters become discernible. HDBSCAN unveiled a greater number of lower-class clusters for cases and fatalities, compared to only one cluster for the middle- and upper-classes, suggesting higher spatial case density and more local variability in the lower-class. The location of multiple clusters per working-class district supports this interpretation. This hypothesis appears feasible due to the existence of distinct working-class districts and neighbourhoods and given that the lower-class population is significantly affected by tuberculosis. However, the middle class is affected even more strongly. The findings could also indicate greater spatial dispersion in higher social classes. It is important to note that with minor adjustments of the input parameters, the middle-class clusters could also have been subdivided into two or more clusters, suggesting a similar, albeit weaker effect than in the working-class. Local G^* also clearly uncovered a lower-class cluster in much of districts 3 and 4, while simultaneously exhibiting upper-class clusters in parts of the districts 1, 2, 6, 7 and 8. The regression coefficient of social class is very

similarly distributed to the results of local G^* , supporting this finding. However, the observed variability of the influence of social class may suggest a reduced significance, in comparison to age.

The morbidity and mortality rates by districts represented in cartograms exhibit higher morbidity and mortality rates in most districts with a significant working-class population (i.e. 1, 3, 4, 5 and 6), with some exceptions. For example, district 5 has a comparatively low morbidity rate, but has one of the highest mortality rates, while district 8 has one of the highest mortality rates, while having a comparatively low morbidity rate. This indicates that working-class demographics are more susceptible to tuberculosis infection. Moran's I indicated the presence of weak negative spatial autocorrelation, which indicates the presence of high-incidence and high-mortality districts next to such with low numbers, as can also be seen from the cartograms. This phenomenon is partially attributable to the geographic location of the upper-class districts, which are adjacent to one or more working-class districts.

The findings of the regression analyses suggest that age is the sole highly significant factor influencing tuberculosis mortality. The estimated likelihood of death rises by 0.006 by year of age at high confidence (p-value approximately 0) and by 0.02 to 0.05 by social class at relatively lower confidence (p-value approximately 0.027). The social class is also a significant factor, with membership of higher social classes paradoxically increasing the likelihood of death. The findings reveal that hospital admission and gender do not exert a significant influence on the outcome. However, being male and being admitted to a hospital increases the likelihood of death. Consequently, the findings suggest that advanced age and elevated social class are significant contributors to the fatality rate in cases of tuberculosis. While the influence is very limited in normal regression at just 5 to 6% variance explained (due to an R-squared value of 0.05 and a p-value of approximately 0 in the linear model and an R-squared value of 0.063 and a p-value of approximately 0 in the global regression, respectively), it rises to 35% with space included in the regression (and an R-squared value of 0.358). Additionally, the reduced RSS and AIC further supports the better match of the GWR. The models are all significant, as evidenced by the significance thresholds of 0.05 and 0.01 being underbid by the very small p-values. Nevertheless, this still leaves 65% of the variance unexplained, indicating, that not all of the relevant factors have been included in the analysis. Conversely, the majority of influences appear to have remained undetected, resulting in a significant number of research gaps that require further exploration in future studies. The paradoxical effect of social class may be attributed to members of higher social classes living longer and therefore being more exposed to diseases at older age. This assumption is, however, not supported by evidence. Another possibility is that there is simply a higher number of cases in the middle- and upper-classes combined, which could potentially mask the influence of social class on mortality.

The temporal analysis is limited in its ability to draw meaningful conclusion due to the short time span of only four years. However, the monthly and even yearly patterns of citywide distribution of cases do appear to indicate a persistent hotspot of tuberculosis cases in the districts 3 and 4, as well as a less clear one in district 1. These districts also

happen to be the most densely populated working-class districts, thereby underscoring a correlation between socioeconomic status and tuberculosis infections. The heatmap of yearly tuberculosis incidence demonstrates remarkable persistence in the individual clusters, particularly concerning the hotspots in the working-class districts 3 through 6 and the Niederdörfli neighbourhood of district 1. Conversely, the coldspots, situated in the affluent districts 2 and 7, also demonstrate a remarkable persistence of low case densities.

These density heatmaps and animated maps (see storymap at <https://arcg.is/0CzKDP>) provide clear evidence that there is a significantly higher tuberculosis incidence density in the working-class districts, which largely coincide with local hotspots. Meanwhile, the affluent districts 2 and 7 exhibit a contrasting pattern of persisting low tuberculosis incidence density. The pattern is remarkably stable over time, indicating a continued correlation between socioeconomic status and tuberculosis incidence. While population density, which is typically higher in lower-class neighbourhoods, probably exerts some influence on the distribution, the distribution arguably extends beyond this influencing factor, as demonstrated through comparison with district populations. This phenomenon was previously highlighted through a comparison of population shares with incidence shares, which revealed that incidence rates amplify existing disparities in population.

6.2 Diphtheria

The approximately 20 addresses referring to hospitals could indicate that a greater number of infected patients are either admitted to or registered by the hospitals. A comparison of the hospital admission rates for tuberculosis and diphtheria reveals that while 77% of tuberculosis patients were admitted to hospitals, only 63% of diphtheria patients were. Consequently, the assertion that diphtheria patients are more likely to be hospitalised cannot be sustained.

The results of the analysis of diphtheria cases have shown that, in agreement with the literature (Byard, 2013; Truelove et al., 2020), the main victims of diphtheria were children, with over 80% of cases occurring in individuals of to 19 years of age. The mean age of diphtheria cases is just 12 years, with the median even lower at 8 years, which confirms the young ages of the victims. The proportion of the very youngest age group is even higher in fatalities than in cases. This observation, when considered in conjunction with the elevated numbers of both diphtheria cases and fatalities within the youngest age group, serves to reinforce the prevailing concept of diphtheria as a childhood disease.

The mortality rates for diphtheria are comparatively low, indicating a lower severity of diphtheria, as has been previously discussed in the literature (Brunner and Senti, 1937; Byard, 2013). This phenomenon may be attributable to the early development of medical interventions (Gubéran, 1980; Kaba, 2010; Müller et al., 2024; Ritzmann, 2015). Furthermore, the mortality rate for diphtheria remains consistently below 5%, in comparison to over 10% for tuberculosis, with only one exception. The mean mortality rate across all age groups is 2.7% of all recorded cases, compared to approximately 20% for tuberculosis.

These findings underscore two noticeable points. Firstly, diphtheria is characterised by a lower severity than tuberculosis. Secondly, diphtheria exhibits a lower mortality rate than tuberculosis. Unfortunately, precise numbers for diphtheria, akin to those available for tuberculosis, could not be sourced.

The presence of clustering is indicated by the results of dot maps, NNI, Moran's I and HDBSCAN. The dot maps show clustering in districts 1, 5, 7 and 8, but these are only in part working-class districts. The working-class districts 3 and 4 are not strongly affected by diphtheria, contrary to expectations. The low incidence in districts 3 and 4 suggests a potentially less robust association between diphtheria and low socioeconomic status. The NNI and G-function lend further support to the presence of clustering in the data, by their respective threshold values. The degree of clustering is, according to the NNI similar to the degree of clustering observed in tuberculosis. The HDBSCAN clusters do reflect the shape and location of the administrative districts. This alignment suggests a potential correlation between social classes and the districts they characterise with the diphtheria clusters (Ritzmann, 1998, pp. 36–37). The distribution pattern which is not restricted to working-class districts may be attributable to the spread of diphtheria, which occurs mainly between school-age children. Children may exhibit reduced mobility and social contact in comparison to adults, which potentially reduces the influence of social class. Consequently, the probability of infection is likely to occur within the school environment or in close proximity to their place of residence, which are often situated in close proximity to children's homes.

The distribution of cases in relation to social class exhibits a high degree of heterogeneity. However, if social class is included in HDBSCAN multiple lower-class clusters can be observed, located in districts 1, 4, 5 and 6. The three clusters for the lower-class are likely due to the same reason as in the tuberculosis data: higher spatial case density and more local variability. Local G^* findings suggest the presence of a lower-class cluster in districts 4, 5 and 6, while upper-class clusters are situated mainly in districts 7 and 8. Regression coefficients cluster similarly to local G^* , although they are not statistically significant and should thus not be overestimated. The clustering in correlation with socioeconomic status appears to be weaker when compared with tuberculosis cases. Furthermore, the high proportion of upper-class individuals among the infected does support the finding of weaker correlation with socioeconomic status. However, the three clusters found within the lower-class could either indicate that a greater proportion of these cases were concentrated in distinct neighbourhoods or that greater spatial dispersion of cases exists in higher social classes, hindering the creating of multiple clusters.

The cartograms have shown that high morbidity and mortality rates are not confined to the traditional working-class districts, as districts 3 and 4 exhibit low rates, while districts 1, 5, 7 and 8 exhibit high rates. However, the mortality cartogram may be of limited significance due to the low number of recorded fatalities. Given the expected correlation between social class and high diphtheria incidence, these findings are intriguing. They suggest a weak correlation of diphtheria incidence with socioeconomic status, as was

mentioned previously. Moran's I of diphtheria cases suggests weak positive spatial autocorrelation, as districts with high morbidity rates are located together, while those with low morbidity are also situated close to each other, as the city is split into two halves of high and low morbidity. Moran's I for the diphtheria fatalities indicates very weak negative spatial autocorrelation, possibly due to the low number of fatalities, which makes analysis more difficult.

The regression analyses have demonstrated that there are no significant influences on death rates. The most influential variable, albeit not significant, is gender. The remaining independent variables have negligible influence on mortality, as their values are well below any threshold of significance. The model, too, was found to be non-significant, as evidenced by its p-value of 0.366, irrespective of its very low R-squared value of 0.001. However, due to the limited number of fatalities, it is challenging to draw definitive conclusions from these forms of analysis. The GWR did enhance the model's fit, explaining approximately 32% of the variability (R-squared of 0.325), while the RSS decreased together with the reduced AIC, potentially attributable to the limited number of fatalities, which renders estimation challenging.

Concerning the temporal evolution of the diphtheria incidence during the four years on record, the trend of increasing case numbers towards 1930 are indicative of a diphtheria wave hitting Zurich around the beginning of the 1930s, bringing with it higher infection rates. The density pattern on a yearly basis is more in flux than for tuberculosis, indicating more fluctuations here. Nevertheless, the presence of persistent hotspots in districts 1, 5, 6 and 8 is evident, with these areas displaying notable and clearly distinguishable concentrations of cases throughout the observed period. A less clearly defined hotspot of lower intensity is visible in the districts 3 and 4. Conversely, the upper-class districts 2 and 7 have largely been spared, as evidenced by the absence of reported cases, excluding cases reported at the children's hospital and minimal influence from district 8. Consequently, these districts can be designated as coldspots.

In an analogous manner to the tuberculosis heatmaps and animated maps (see storymap at <https://arcg.is/0CzKDP>), it can be speculated that a discernible influence of social class on the geographical distribution of diphtheria case hotspots exists, though this influence is comparatively less pronounced than that observed for tuberculosis. The temporal stability of the heatmaps over time is low, as the hotspots regularly change location and few of them are stable over time. This phenomenon may be attributed to the fact that diphtheria primarily spreads among children who attend different schools within their respective districts, and therefore, are possibly less segregated in terms of social class compared to adults. The previously discussed schools as possible locations of infection, i.e. transmission hotspots, have been shown to remain as such over the years and are affiliated with either the lower- or middle-class. Meanwhile, affluent districts 2 and 7 have exhibited minimal incidence of diphtheria cases. The observed variations in the number of individual hotspots may be indicative of a less chronic nature of diphtheria, as suggested by literature, in comparison to tuberculosis. The role of population differences in

this context is also a salient consideration; however, in contrast to tuberculosis, where disparities in case shares are more pronounced than differences in population, the disparities observed in diphtheria are comparatively less significant.

6.3 Synthesis

Returning to the original research question and its subquestions, I am now in a position to provide answers. In the interest of summarizing the key points, a brief repetition of the research question and subquestions will be offered, prior to providing a response based on my findings. The original research question was as follows:

How did socioeconomic factors influence the spatial patterns of tuberculosis and diphtheria cases in Zurich in the late 1920s and early 1930s, and where were the hotspots and coldspots of these two diseases, respectively?

- *How did tuberculosis and diphtheria cases spread spatially and temporally in Zurich during the late 1920s and early 1930s?*

The city of Zurich experienced widespread tuberculosis, a chronic disease, and its prevalence was attributed to its inherent characteristics. However, the districts that were typically considered to be working-class or lower-class exhibited higher morbidity and mortality rates than the middle- and upper-class districts. These hotspots remained consistent over time, displaying remarkable stability. Similarly, the coldspots also demonstrated remarkable stability over time. These were situated in affluent districts, such as districts 2 and 7, and exhibited low morbidity and mortality rates.

The temporal stability of diphtheria is weak, as evidenced by the shifting of clusters over time, with infrequent persistence in a given location over multiple years. A notable increase in cases was observed during the late 1920s, which may be attributed to the introduction of a new wave of the disease. However, even within this context, a discernible pattern emerges, albeit one that encompasses districts beyond those classified as working-class, extending to more affluent districts as well. Districts 5 and 6, as well as 1 and 8, are examples of this phenomenon. Conversely, districts 3 and 4 have been observed to demonstrate a notable absence of high morbidity and mortality. The coldspots demonstrate some stability district 2. However, a notable concentration of cases is observed in districts 7 and 8, a phenomenon that is potentially influenced by the site of the children's hospital in district 7 and the Balgrist clinic in district 8. In comparison with tuberculosis, diphtheria demonstrates a less pronounced spatial pattern and exerts a lesser influence on mortality. However, these findings are difficult to ascertain due to the limited sample size of only 12, respectively 19, fatalities.

- *Where are hotspots and coldspots of tuberculosis and diphtheria cases in Zurich, and how did disease prevalence vary across the city?*

The distribution of hotspots reveals notable disparities, with concentrations observed in the traditionally working-class districts 3 and 4, 5 and the Niederdörfli neighbourhood of district 1. Conversely, the coldspots are located in districts 2 and 7, which are traditionally affluent.

The distribution of diphtheria cases is concentrated in districts 1, 5, 6, 7 and 8, similar to an axis, running north to south. The potential correlation with specific schools, which have potentially been identified as sources of infection, warrants further investigation. Districts 2, 3 and 4 have been identified as coldspots.

- *What was the relationship between socioeconomic status and social class and the spatial distribution of tuberculosis and diphtheria cases?*

The distribution of tuberculosis cases is significantly influenced by socioeconomic status, with cases clustered in densely populated working-class districts that are characterised by a poorer hygienic situation and where the inhabitants are of lower social classes and socioeconomic status. Contrarily, affluent, middle- and upper-class districts have been observed to exhibit a correlation with coldspots.

The correlation between socioeconomic status and diphtheria cases appears to be relatively weak, as no definitive influence could be observed at the district level. The findings indicate that socioeconomic status does not appear to exert a substantial influence on diphtheria cases, as no significant disparities in case numbers were observed between typically working-class districts and other areas. While a definitive correlation between hotspots and social class, including socioeconomic status, remains elusive, the potential influence of schools on disease patterns is noted.

- *To what extent did socioeconomic disparities correlate with the spatial distribution of tuberculosis and diphtheria cases in Zurich, and how do disease patterns reflect socioeconomic inequalities in the city?*

The correlation between hotspots and social class, including socioeconomic status, is relatively clear. A greater number of cases, higher morbidity and mortality rates were registered in the typically working-class districts of the city of Zurich. The morbidity and mortality rates are generally higher in working-class districts than in middle- and upper-class districts. This finding is consistent with other studies that have identified tuberculosis hotspots in typically working-class districts of Zurich characterised by low socioeconomic status (Ritzmann, 1998, pp. 36–37).

The correlation between socioeconomic status and social class with the spatial distribution of diphtheria cases and fatalities is found to be relatively weak. The analysis revealed that not all typical working-class districts exhibited high morbidity and mortality rates. Consequently, it can be posited that diphtheria exhibits a comparatively weak correlation

between socioeconomic status and incidence, especially in comparison with tuberculosis.

This study suggests that spatial clustering of disease is not a uniform phenomenon but is rather conditioned by a complex interdependence of socioeconomic factors, demographic vulnerability, and disease type. This perspective has the potential to inform future GIS-based epidemiological modelling. Table 8 offers a systematic comparison of the most important points found in the analysis. It highlights that tuberculosis is significantly influenced stronger by socioeconomic differences, a phenomenon that may be attributed to the specific age group under consideration, as evidenced by the less pronounced disparities observed in the population group including children, which were mainly affected by diphtheria.

Table 8: Comparative synthesis of tuberculosis and diphtheria data analysis

Dimension	Tuberculosis	Diphtheria	Interpretation
Age group	(Young) working-age adults	Children	Socioeconomic vs. school-based spread
Spatial clustering	Yes	Yes	Linked to socioeconomic status / social class
Socioeconomic link	Strong	Medium-weak	Class-based vs. universal exposure
Mortality distribution	Socioeconomically shaped	More uniformly shaped	Class-based vs. universal exposure

Answering the original research question, I can state that the spatial distribution of tuberculosis cases in Zurich during the late 1920s and early 1930s was closely intertwined with socioeconomic conditions. The geographical distribution of tuberculosis exhibited a marked prevalence in densely populated, working-class districts, including districts 3, 4, 5, and the Niederdörfli neighbourhood of district 1. These areas were distinguished by low socioeconomic status and low social class negatively impacting living and hygienic conditions. Conversely, districts 2 and 7, which were affluent and exhibited low morbidity and mortality rates throughout the study period, were identified as coldspots. This was demonstrated by the findings of various analytical methods employed. The findings highlight a clear and persistent correlation between social inequality and disease burden, with the spatial patterns of tuberculosis reflecting broader socioeconomic disparities within the urban fabric of Zurich, as described in health geography literature (Moon, 2020; Shaweno et al., 2018).

The influence of socioeconomic factors on the spatial pattern of diphtheria is much less clearly visible and less significant, than for tuberculosis. Nevertheless, the influence of social class is discernible to a certain extent, possibly in connection with specific schools that may have functioned as transmission hotspots. While districts 5 and 6, in conjunction with the Niederdörfli neighbourhood of district 1 and district 8, were identified as

hotspots, districts 2, and notably 3 and 4 exhibited low incidence and therefore only a weak correlation with socioeconomic status and social class. While tuberculosis displayed comparatively stronger and more stable spatial patterns aligned with socioeconomic disparities, diphtheria was more spatially variable and less clearly associated with socioeconomic status. Instead, they may be associated with the presence of other factors, such as the location of schools or hospitals, including the children's hospital. Conversely, the coldspots, particularly in district 2, demonstrated relative stability.

Appropriate measures and health interventions should have been implemented on a grand scale to provide greater protection from tuberculosis to the lower-class demographic, given the knowledge at the time that they exhibited higher risk and vulnerability to tuberculosis. An example for such a measure is the Zürichberg forest school, but it only offered limited capacity. For diphtheria, protection measures to inhibit the spread at schools could have been taken. The implementation of targeted health interventions could have resulted in the preservation of numerous lives, in addition to the mitigation of the economic impact of tuberculosis on the working-age population and the increased childhood mortality from diphtheria. This applies even today, when resources are stretched thin by rapid population growth, especially in developing countries, where tuberculosis and diphtheria are a persistent health issue. It is important to acknowledge the significance of these demographics in the future development of any nation and to invest in its health and wellbeing, as their contributions are in turn likely to be substantial and long-lasting.

6.4 Limitations

Limitations concerning the datasets utilised, epidemiological constraints, methodological constraints and specific challenges to the scale and aggregation of spatial data, namely the modifiable areal unit problem (MAUP), must be considered.

Initially, it is conceivable that some original data was misread, given that the records were created manually, and the handwriting is at times not clearly legible. Furthermore, it is entirely possible that any attribute entry may be missing or be unreadable. Due to the nature of the data, which was collected almost 100 years ago, it is impossible to determine its completeness. While the literature mentions both the national obligations of the cause-of-death statistics and the cantonal obligation to report tuberculosis cases, the ability and willingness of the reporting physicians is the greatest unknown influencing factor (Ritzmann, 2010b, 1998, pp. 24–27). Despite the potential incompleteness of the data, it nevertheless offers valuable insights into the past and is the sole source available for this study.

While both datasets exhibit instances of missing addresses, the impact of this variation in size is discernible. As in the diphtheria data, approximately 200 addresses are missing, this has a greater impact, as the dataset only has 670 entries. The missing addresses appear to be all from 1928, which reduces the accuracy of data concerning this year. Meanwhile, the tuberculosis data also includes approximately 200 missing addresses, yet this

is mitigated by the fact that, of the roughly 2000 addresses, 1800 are still available for analysis. Similarly, the limited number of diphtheria fatalities poses a significant challenge to the accurate representation of mortality rates and mortality-based findings, which should be interpreted cautiously.

A number of addresses refer to three different hospitals in the city of Zurich. The old site of the children's hospital (Kinderspital) in the neighbourhood of Hottingen in district 7 was registered 15 times, the cantonal hospital (Kantonsspital), which today is the university hospital (Universitätsspital) in district 1, was registered twice and the Balgrist clinic in district 8 was registered 5 times. The conspicuous concentration of diphtheria cases in just these few hospitals could potentially skew the spatial distribution and introduce bias into the analysis. This is particularly likely to be the case for the children's hospital, where 15 cases were reported, which may have influenced the distribution pattern of the 468 located addresses, given that there are relatively few other cases in the surrounding area.

As indicated by literature, an increased share of people went to get treated but also passed away in sanatoria by the 1930s (Corti, 2012; Holloway et al., 2013; Kruker and Senti, 1932; Rucker and Kearny, 1913; Senti and Pfister, 1946; Silberschmidt, 1930). It is reasonable to assume that these individuals would tend to be of a higher socioeconomic status and higher social class, as the treatment would prevent them from working and could also be quite expensive. The subsequent loss of income and additional costs could have deterred working-class people from seeking treatment in sanatoria. Consequently, a bias may have emerged, influenced by individual financial resources, which are closely associated with socioeconomic status and social class. However, the legal framework stipulating universal access to treatment (Holloway et al., 2013, p. 81) has the potential to mitigate the impact of this bias by ensuring equitable access to treatment irrespective of financial circumstances. Nevertheless, the quality and effectiveness of treatment at the cantonal sanatorium in Wald might have fallen short of that achieved in the high-end, privately run sanatoria in the Swiss Alps, as evidenced by the persistent disparities in the quality of treatment between public and private health institutions that persist to this day.

Another point that could be questioned, is the accuracy and relevance of the HISCO classification. However, given the scope of the project, it encompasses a wide variety of occupations across multiple languages, giving it a solid foundation (van Leeuwen et al., 2002), thereby ensuring its implementation is firmly grounded. While HISCO was originally developed for slightly earlier time periods, it nevertheless retains a reasonable degree of relevance to the data under consideration.

The complex structure of social class, according to the HISCO classification, challenges the assumption of clearly defined correlations between social classes and districts. The high heterogeneity displayed in maps of social classes may indicate that social classes were intermingling by the late 1920s and early 1930s. This lacking crispness of the spatial distribution of social classes may impede analyses based on that presumption.

The two datasets utilised in this thesis are, although temporally close, not equal, as the diphtheria data ranges from 1927 to 1930 while the tuberculosis data ranges from 1932 to 1935. While the disparity of two years may appear negligible, it is imperative to acknowledge the potential for subtle alterations that may have occurred during this period.

The registration of asymptomatic patients, by definition, was not possible; however, this is a problem that still exists today and cannot really be dealt with. The question that arises is whether a bias has been created through the underreporting of certain population groups.

A further, albeit largely inconsequential, challenge is presented by the second city unification of 1934, wherein the City of Zurich was expanded to include several neighbouring communities, including what would subsequently become district 9 (Albisrieden and Altstetten), Höngg, the future districts 11 and 12 (Seebach, Oerlikon, Schwamendingen) and Witikon. However, from 1934 onwards the reports incorporated the new neighbourhoods and districts, which thereafter had to be filtered out, to ensure they would not influence the analysis.

Colours designed to be compatible with colour-blindness were not available for all intended use cases. This was the case for the HDBSCAN maps, which relied on distinctly different colours, drawn from ColorBrewer2 (Brewer et al., 2013).

Due to temporal constraints and the necessity to maintain the scope and length of the thesis, choices regarding the methods employed had to be made, and some potentially interesting methods were excluded (e.g. SaTScan). It is important to note that each method possesses its own advantages and disadvantages, thus enabling only a specific set of conclusions to be drawn, as no method is universally applicable. The focus was thus directed towards methods that I considered both interesting and up-to-date according to the latest research. It is also crucial to emphasize that the selection of methods had to align with the analytical scope, specifically to investigate clustering and the influence of social class.

The capacity of HDBSCAN to manage varying densities, as in the data at hand, renders it well-suited for this analysis. However, it focuses on high-density areas, potentially overlooking clusters in less densely populated regions, as the underlying population at risk is not taken into account. While nearest neighbour analysis is a relatively simple method, it does help to determine the presence and degree of clustering in the data, which also makes it a valuable addition to this study. Moran's I, despite its limited applicability to data aggregated to specified areas and on a global scale, such as administrative districts, can assist in determining clustering and autocorrelation. LISA would have been a valuable addition, providing a deeper insight. The combination of methods employed in this study was well-suited to exploratory spatial analysis of disease data, as contemporary research suggests. However, all methods are subject to inherent limitations, which must be considered when interpreting the results.

Several limitations of the methodologies utilised during the course of this thesis apply. Of particular note is the MAUP (modifiable areal unit problem), which invariably arises when dealing with data that has been aggregated to artificial boundaries. Had the level of analysis been changed to the neighbourhoods that constitute the districts, it is reasonable to hypothesise that the analysis would have returned an entirely different result, as different district borders would have. However, given that socioeconomic data was only available at the district level, conducting a neighbourhood-level analysis would have been unfeasible and thus was not pursued. Fortunately, the districts of Zurich largely coincide with the inner homogeneity of populations (Ritzmann, 1998, pp. 36–37), thereby eliminating the necessity for more detailed neighbourhood-level analysis.

A critique on district level analysis, despite the availability of point-level data, is warranted. It may not have been necessary at all to aggregate the data at the district level, as the data was available as individual points, which could have rendered district-level analysis superfluous. However, I decided to do so anyway, because some of the methods I intended to apply were not designed for point data, and because relevant socioeconomic data was only available at this level.

As previously mentioned, it is not possible to ascertain the precise location of an infection; only the place of residence is known (Shaweno et al., 2018). This limitation must be borne in mind. Despite these limitations, I believe that the situation is not as dire, given the constrained mobility that characterised the era.

7 Conclusions and future work

Considering the research objectives, the geographical locations of official tuberculosis and diphtheria case data from the late 1920s and early 1930s have been geolocated from address data. Utilising the coordinates extracted from the geolocation process, the disease data were visualised in distinct methods, thereby highlighting the spatial distribution and hotspots. The influence of socioeconomic status and social class on the occurrence of both tuberculosis and diphtheria was analysed. While a significant correlation was established between socioeconomic patterns of social class and tuberculosis, a similar correlation was found for diphtheria cases, albeit comparatively weaker and not statistically significant.

The findings of the analyses conducted indicate that both diseases, tuberculosis and diphtheria, were distributed throughout the entirety of the city of Zurich in the respective time periods. However, this distribution was not uniform but rather varied across different geographical areas and clearly exhibited clustering behaviour. This variation was expected and can be attributed to Tobler's first law (Miller, 2004; Tobler, 1970; Waters, 2017). The distribution of tuberculosis cases exhibited a stronger correlation with socioeconomic status and social class, while the correlation for diphtheria cases was weaker and its total case numbers were lower. Tuberculosis, being a chronic disease, exhibited a consistent prevalence throughout the study period (1932-1935), with no significant fluctuations. In contrast, diphtheria demonstrated an increase in cases between 1927 and 1930, suggesting a possible disease outbreak.

Concerning the distribution of tuberculosis in space, and with reference to the observed hotspots and their relation to social class, I can state that there is a clearly observable hotspot of tuberculosis cases located in the characteristically working-class districts 3 and 4, but also in the Niederdörfli neighbourhood of district 1. Conversely, the working-class district 5 exhibited the highest mortality rate of the entire city, closely followed by districts 3 and 1, as expected from the literature (Ritzmann, 1998, pp. 36–37). With regard to the social class and closely related socioeconomic status, a clearly observable correlation was found between the shares of high case numbers and high fatalities and the socioeconomic characteristics of districts influenced by the working-class. The results of HDBSCAN and local G^* indicated the presence of clusters of different social classes in districts associated with said social classes. Specifically, working-class clusters were identified in districts 3, 4 and even 5, while upper-class clusters were located in districts 7 and 8, as would have been anticipated. Intriguingly, a notable presence of upper-class clusters was also observed in district 6, despite this district's historical association with the working-class. Contrarily, districts characterised by the upper-class have exhibited a reduced incidence of tuberculosis, both in terms of infections and fatalities. These findings are consistent with research suggesting that males are at higher risk of contracting a tuberculosis infection and that young adults, who are the economically most active age group, are at higher risk of infection too (Gwitira et al., 2021; Horton et al., 2020, 2016;

Humayun et al., 2022; Nhamoyebonde and Leslie, 2014; Peer et al., 2023). Consequently, the present study offers a compelling response to the research questions posed, affirming the negative impact of low socioeconomic status on health outcomes concerning tuberculosis.

The examination of the distribution of diphtheria reveals divergent patterns when compared to those observed for tuberculosis. The districts 1, 5, 6, 7 and 8 were identified as hotspots of diphtheria, and these areas were less evidently associated with low socioeconomic status and social class. While districts 5 and 6, and the Niederdörfli neighbourhood of district 1 exhibit characteristics indicative of working-class areas, the affluent districts 7 and 8 also demonstrated high incidence. However, the analysis may be influenced by the presence of a children's hospital in district 7, which could introduce bias. The highest morbidity and mortality rates, although generally lower than for tuberculosis, were observed in districts 5 and 8, while districts 1 and 4 exhibited high mortality rates only. However, the findings are only of limited significance due to the limited number of fatalities, rendering the analysis vulnerable to outliers. The findings of HDBSCAN and local G* demonstrate clusters of social classes in different districts, which, as previously mentioned, are not always clearly correlating with low socioeconomic status. While affluent districts 7 and 8 are affected by diphtheria, district 2 is not, thus marking it as a coldspot, together with district 3, which, as a working-class district, would not have been expected to be a coldspot. The findings align with the established knowledge that diphtheria primarily affects children (Byard, 2013; Truelove et al., 2020). Consequently, the present study provides no evidence to support the hypothesis that socioeconomic status exerts significant influence on health outcomes concerning diphtheria. While the observed trends indicate a slight negative tendency, the findings must be interpreted with caution due to the lack of conclusive evidence.

In conclusion, this master's thesis explored the distribution of tuberculosis and diphtheria cases in the City of Zurich in the late 1920s and early 1930s in depth and offered a geographical perspective of the events. It contributes to the field of health geography by applying spatial clustering methods to historical disease data, offering insights into how socioeconomic patterns influenced disease distribution. Unlike previous studies, which focused on other cities or diseases, this thesis offers a novel spatial and temporal perspective on Zurich's fight against tuberculosis and diphtheria in the late 1920s and early 1930s in relation to socioeconomic indicators. It aims to offer new insights into its course, which could be central to understanding further developments in health policy and health geography of Zurich, with implications for modern health geographic and policy. Recognizing how historical disease patterns disproportionately affected socioeconomically disadvantaged demographics highlights the continued need for targeted public health strategies and interventions that address spatial inequality and socioeconomic status in contemporary cities.

It is evident that further research is required, particularly in the context of diphtheria, which has received comparatively less research attention due to its lower prevalence.

Future research could concentrate on this aspect. Additionally, it would be worthwhile to direct future research towards the identification of the remaining factors that influence the distribution of both diseases. Ultimately, these findings provide not only historical insights but also a foundation for understanding how spatial inequalities continue to shape urban health outcomes today.

8 References

- Agbor, A.S., 2014. Using GIS to map the spatial and temporal occurrence of cholera epidemic in Camaroon.
- Amt für Städtebau, 2008. Schulhäuser der Stadt Zürich. Zürich.
- Anselin, L., Ibnu, S., Youngihn, K., 2006. GeoDa: An Introduction to Spatial Data Analysis. *Geogr. Anal.* 38, 5–22.
- Behrens, N., Motschi, A., Schultheiss, M., 2015. Zürich (Gemeinde). *Hist. Lex. Schweiz*.
- Bernhard, M., Leuch, C., Kordi, M., Gruebner, O., Matthes, K.L., Floris, J., Staub, K., 2023. From pandemic to endemic: Spatial-temporal patterns of influenza-like illness incidence in a Swiss canton, 1918–1924. *Econ. Hum. Biol.* 50, 101271. <https://doi.org/10.1016/j.ehb.2023.101271>
- Bertin, J., 2011. *Semiology of graphics : diagrams, networks, maps*, *Semiology of graphics diagrams, networks, maps*. Esri Press, Redlands, Calif.
- Bertin, J., Berg, W.J., 1983. *Semiology of graphics : diagrams, networks, maps*, *Semiology of graphics diagrams, networks, maps*. The University of Wisconsin Press, Madison.
- Bhatia, N., 2010. *Survey of Nearest Neighbor Techniques* 8.
- Bingham, P., Verlander, N.Q., Cheal, M.J., 2004. John Snow, William Farr and the 1849 outbreak of cholera that affected London: a reworking of the data highlights the importance of the water supply. *Public Health* 118, 387–394. <https://doi.org/10.1016/j.puhe.2004.05.007>
- Birkhölzer, I., 2023. *Reconstructing the 1855 Cholera Epidemic in Basel Using Geographic Information Visualization*.
- Brewer, C., Harrower, M., Sheesley, B., Woodruff, A., Heyman, D., 2013. *ColorBrewer: Color Advice for Maps* [WWW Document]. URL <https://colorbrewer2.org/#> (accessed 3.20.25).
- Brodbeck, A., Hermann, A., 2012. *Stadtentwicklung von Zürich 1850-2000*.
- Brody, H., Rip, M.R., Vinten-Johansen, P., Paneth, N., Rachman, S., 2000. Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *The Lancet* 356, 64–68. [https://doi.org/10.1016/S0140-6736\(00\)02442-9](https://doi.org/10.1016/S0140-6736(00)02442-9)
- Brunner, H., Senti, A., 1937. Ansteckende Krankheiten in Zürich. *Zür. Stat. Nachrichten* 14, 177–233.
- Bundesamt für Gesundheit, 2024a. *Infektionskrankheiten melden* [WWW Document]. URL <https://www.bag.admin.ch/bag/de/home/krankheiten/infektionskrankheiten-bekaempfen/meldesysteme-infektionskrankheiten/meldepflichtige-ik/meldeformulare.html> (accessed 11.15.24).
- Bundesamt für Gesundheit, 2024b. *Tuberkulose* [WWW Document]. URL <https://www.bag.admin.ch/bag/de/home/krankheiten/krankheiten-im-ueberblick/tuberkulose.html> (accessed 1.30.25).
- Bundesamt für Gesundheit, 2024c. *Diphtherie* [WWW Document]. URL <https://www.bag.admin.ch/bag/de/home/krankheiten/krankheiten-im-ueberblick/diphtherie.html> (accessed 1.30.25).
- Bundesamt für Statistik, 2024. *Lebenserwartung bei Geburt* [WWW Document]. URL <https://datawrapper.dwcdn.net/a103f3c5005b39101d78fbe8dd83a3fb/6/> (accessed 1.15.25).
- Burkhard, C., 2023. *Geovisualization of Swiss Birth Weights at the Beginning of the 20th Century*.

- Bwire, G., Ali, M., Sack, D.A., Nakinsige, A., Naigaga, M., Debes, A.K., Ngwa, M.C., Brooks, W.A., Orach, C.G., 2017. Identifying cholera “hotspots” in Uganda: An analysis of cholera surveillance data from 2011 to 2016. *PLoS Negl. Trop. Dis.* 11, e0006118. <https://doi.org/10.1371/journal.pntd.0006118>
- Byard, R.W., 2013. Diphtheria – ‘The strangling angel’ of children. *J. Forensic Leg. Med.* 20, 65–68. <https://doi.org/10.1016/j.jflm.2012.04.006>
- Caplan, J.M., Kennedy, L.W., Neudecker, C.H., 2020. Cholera deaths in Soho, London, 1854: Risk Terrain Modeling for epidemiological investigations. *PLOS ONE* 15, e0230725. <https://doi.org/10.1371/journal.pone.0230725>
- CDC, 2024a. About Tuberculosis [WWW Document]. Tuberc. TB. URL <https://www.cdc.gov/tb/about/index.html> (accessed 10.15.24).
- CDC, 2024b. About Diphtheria [WWW Document]. Diphtheria. URL <https://www.cdc.gov/diphtheria/about/index.html> (accessed 10.15.24).
- Cheeseman, E.A., Martin, W.J., Russell, W.T., 1939. Diphtheria: A suggested explanation of the relative change in age incidence. *Epidemiol. Infect.* 39, 181–202. <https://doi.org/10.1017/S0022172400011803>
- Chirenda, J., Gwitira, I., Warren, R.M., Sampson, S.L., Murwira, A., Masimirembwa, C., Mateveke, K.M., Duri, C., Chonzi, P., Rusakaniko, S., Streicher, E.M., 2020. Spatial distribution of Mycobacterium Tuberculosis in metropolitan Harare, Zimbabwe. *PLOS ONE* 15, e0231637. <https://doi.org/10.1371/journal.pone.0231637>
- Coleman, S., 2018. The association between tuberculosis and diphtheria. *Epidemiol. Infect.* 146, 940. <https://doi.org/10.1017/S0950268818000936>
- Corti, F., 2012. Tuberkulose. *Hist. Lex. Schweiz.*
- Dauer, C.C., 1950. Trends in Age Distribution of Diphtheria in the United States. *Public Health Rep.* 1896-1970 65, 1209–1218. <https://doi.org/10.2307/4587476>
- ETH-Bibliothek, 2025. E-Pics Baugeschichtliches Archiv [WWW Document]. URL <https://baz.e-pics.ethz.ch/#> (accessed 4.10.25).
- Fabrikant, S.I., Hespanha, S.R., Hegarty, M., 2010. Cognitively Inspired and Perceptually Salient Graphic Displays for Efficient Spatial Inference Making. *Ann. Assoc. Am. Geogr.* 100, 13–29. <https://doi.org/10.1080/00045600903362378>
- Gatrell, A.C., Löytönen, M., 1998. GIS and health, GISDATA 6. Taylor & Francis, London.
- Gaudart, J., Rebaudet, S., Barraï, R., Boncy, J., Faucher, B., Piarroux, M., Magloire, R., Thimothe, G., Piarroux, R., 2013. Spatio-Temporal Dynamics of Cholera during the First Year of the Epidemic in Haiti. *PLoS Negl. Trop. Dis.* 7, e2145. <https://doi.org/10.1371/journal.pntd.0002145>
- Getis, A., Ord, J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geogr. Anal.* 24, 189–206. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>
- Griffin, A.L., MacEachren, A.M., Hardisty, F., Steiner, E., Li, B., 2006. A Comparison of Animated Maps with Static Small-Multiple Maps for Visually Identifying Space-Time Clusters. *Ann. Assoc. Am. Geogr.* 96, 740–753. <https://doi.org/10.1111/j.1467-8306.2006.00514.x>
- Gubéran, E., 1980. Tendances de la mortalité en Suisse. 2. Maladies infectieuses 1876-1977. *Schweiz. Med. Wochenschr.* 110, 574–583.
- Gwitira, I., Karumazondo, N., Shekede, M.D., Sandy, C., Siziba, N., Chirenda, J., 2021. Spatial patterns of pulmonary tuberculosis (TB) cases in Zimbabwe from 2015 to 2018. *PLOS ONE* 16, e0249523. <https://doi.org/10.1371/journal.pone.0249523>

- Harrower, M., Fabrikant, S., 2008. The Role of Map Animation for Geographic Visualization, in: Dodge, M., McDerby, M., Turner, M. (Eds.), *Geographic Visualization*. Wiley, pp. 49–65. <https://doi.org/10.1002/9780470987643.ch4>
- Hermans, S., Jr, C.R.H., Wood, R., 2015. A Century of Tuberculosis Epidemiology in the Northern and Southern Hemisphere: The Differential Impact of Control Interventions. *PLOS ONE* 10, e0135179. <https://doi.org/10.1371/journal.pone.0135179>
- Hernandez, J.B.R., Kim, P.Y., 2024. Epidemiology Morbidity And Mortality, in: StatPearls. StatPearls Publishing, Treasure Island (FL).
- Holloway, K.L., Henneberg, R., Lopes, M.D.B., Staub, K., Link, K., Rühli, F., Henneberg, M., 2013. Secular Trends in Tuberculosis during the Second Epidemiological Transition: A Swiss Perspective. *Adv. Anthropol.* 03, 78–90. <https://doi.org/10.4236/aa.2013.32011>
- Holloway, K.L., Staub, K., Rühli, F., Henneberg, M., 2014. Lessons From History Of Socio-economic Improvements: A New Approach To Treating Multi-Drug-Resistant Tuberculosis. *J. Biosoc. Sci.* 46, 600–620. <https://doi.org/10.1017/S0021932013000527>
- Horton, K.C., Hoey, A.L., Béraud, G., Corbett, E.L., White, R.G., 2020. Systematic Review and Meta-Analysis of Sex Differences in Social Contact Patterns and Implications for Tuberculosis Transmission and Control. *Emerg. Infect. Dis.* 26, 910–919. <https://doi.org/10.3201/eid2605.190574>
- Horton, K.C., MacPherson, P., Houben, R.M.G.J., White, R.G., Corbett, E.L., 2016. Sex Differences in Tuberculosis Burden and Notifications in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. *PLOS Med.* 13, e1002119. <https://doi.org/10.1371/journal.pmed.1002119>
- Humayun, M., Chirenda, J., Ye, W., Mukeredzi, I., Mujuru, H.A., Yang, Z., 2022. Effect of Gender on Clinical Presentation of Tuberculosis (TB) and Age-Specific Risk of TB, and TB-Human Immunodeficiency Virus Coinfection. *Open Forum Infect. Dis.* 9, ofac512. <https://doi.org/10.1093/ofid/ofac512>
- IEM, I. of E.M., 2024. Welcome! | Institute of Evolutionary Medicine (IEM) | UZH [WWW Document]. URL <https://www.iem.uzh.ch/en.html> (accessed 11.15.24).
- Kaba, M., 2010. ‘The Geneva serum is excellent!’ Autonomy and Isolation in the Swiss Cantons During the Early Years of Diphtheria Serum: The Case of Geneva, in: Gradmann, C., Simon, J. (Eds.), *Evaluating and Standardizing Therapeutic Agents, 1890–1950*. Palgrave Macmillan UK, London, pp. 105–117. https://doi.org/10.1057/9780230285590_7
- Kanturk, G., 2007. Using Gis Technology To Analyse Tuberculosis Incidence In Izmir.
- Khan, K., Rehman, S.U., Aziz, K., Fong, S., Sarasvady, S., 2014. DBSCAN: Past, present and future, in: *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*. Presented at the The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), pp. 232–238. <https://doi.org/10.1109/ICADIWT.2014.6814687>
- Kirby, R.S., Delmelle, E., Eberth, J.M., 2017. Advances in spatial epidemiology and geographic information systems. *Ann. Epidemiol., GIS and Spatial Methods in Epidemiology Symposium* 27, 1–9. <https://doi.org/10.1016/j.annepidem.2016.12.001>

- Kistemann, T., Munzinger, A., Dangendorf, F., 2002. Spatial patterns of tuberculosis incidence in Cologne (Germany). *Soc. Sci. Med.*, Selected papers from the 9th International Symposium on Medical Geography 55, 7–19. [https://doi.org/10.1016/S0277-9536\(01\)00216-7](https://doi.org/10.1016/S0277-9536(01)00216-7)
- Koch, T., 2017. Cartographies of disease : maps, mapping, and medicine, New expanded edition. ed, Cartographies of disease maps, mapping, and medicine. Esri Press, Redlands, California.
- Koch, T., Denike, K., 2009. Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854. *Soc. Sci. Med.* 69, 1246–1251. <https://doi.org/10.1016/j.socscimed.2009.07.046>
- Kruker, M., Senti, A., 1932. Tuberkulosesterblichkeit in Zürich. *Zür. Stat. Nachrichten* 9, 289–325.
- Lawson, A.B., 2006. Statistical methods in spatial epidemiology, 2nd ed. ed, Wiley series in probability and statistics. Wiley, Hoboken, NJ.
- Leuch, C., 2021. In the Footsteps of the “Mother of all Pandemics”: Spatio-Temporal Analysis of the 1918 Flu Pandemic in the Canton of Berne, Switzerland.
- Liu, Y., Li, X., Wang, W., Li, Z., Hou, M., He, Y., Wu, W., Wang, H., Liang, H., Guo, X., 2012. Investigation of space-time clusters and geospatial hot spots for the occurrence of tuberculosis in Beijing. *Int. J. Tuberc. Lung Dis.* 16, 486–491. <https://doi.org/10.5588/ijtld.11.0255>
- Lowe, R.K., 2003. Animation and learning: selective processing of information in dynamic graphics. *Learn. Instr.* 13, 157–176. [https://doi.org/10.1016/S0959-4752\(02\)00018-X](https://doi.org/10.1016/S0959-4752(02)00018-X)
- Matthijs, H., 2024. Die Waldschule beim Zürichberg. *Med. Zür.* URL <https://dlf.uzh.ch/sites/medizingeschichte/die-waldschule-beim-zuerichberg/> (accessed 2.15.25).
- Meliker, J.R., Sloan, C.D., 2011. Spatio-temporal epidemiology: Principles and opportunities. *Spat. Spatio-Temporal Epidemiol.* 2, 1–9. <https://doi.org/10.1016/j.sste.2010.10.001>
- Miller, H.J., 2004. Tobler's First Law and Spatial Analysis. *Ann. Assoc. Am. Geogr.* 94, 284–289.
- Miller, P.B., Zalwango, S., Galiwango, R., Kakaire, R., Sekandi, J., Steinbaum, L., Drake, J.M., Whalen, C.C., Kiwanuka, N., 2021. Association between tuberculosis in men and social network structure in Kampala, Uganda. *BMC Infect. Dis.* 21, 1023. <https://doi.org/10.1186/s12879-021-06475-z>
- Moon, G., 2020. Health Geography, in: Kobayashi, A. (Ed.), *International Encyclopedia of Human Geography* (Second Edition). Elsevier, Oxford, pp. 315–321. <https://doi.org/10.1016/B978-0-08-102295-5.10388-9>
- Müller, S.-M., Matthes, K.L., Lang, P., Fehr, J., Chirolera, A., Krauer, F., van Wijhe, M., Rühli, F., Staub, K., 2024. From non-persistent natural to sustainable vaccine immunity: Morbidity and mortality from diphtheria in Switzerland from 1877 until today.
- Ngwa, M.C., Ihekweazu, C., Okwor, T., Yennan, S., Williams, N., Elimian, K., Karaye, N.Y., Bello, I.W., Sack, D.A., 2021. The cholera risk assessment in Kano State, Nigeria: A historical review, mapping of hotspots and evaluation of contextual factors. *PLoS Negl. Trop. Dis.* 15, e0009046. <https://doi.org/10.1371/journal.pntd.0009046>
- Nhamoyebonde, S., Leslie, A., 2014. Biological Differences Between the Sexes and Susceptibility to Tuberculosis. *J. Infect. Dis.* 209, S100–S106. <https://doi.org/10.1093/infdis/jiu147>

- Nusrat, S., Kobourov, S., 2016. The State of the Art in Cartograms. *Comput. Graph. Forum* 35, 619–642. <https://doi.org/10.1111/cgf.12932>
- Osei, F.B., 2014. Current Statistical Methods for Spatial Epidemiology: A Review.
- Osei, F.B., Duker, A.A., 2008. Spatial and demographic patterns of Cholera in Ashanti region - Ghana. *Int. J. Health Geogr.* 7, 44. <https://doi.org/10.1186/1476-072X-7-44>
- Pape, S., Karki, S.J., Heinsohn, T., Brandes, I., Dierks, M.-L., Lange, B., 2024. Tuberculosis case fatality is higher in male than female patients in Europe: a systematic review and meta-analysis. *Infection* 52, 1775–1786. <https://doi.org/10.1007/s15010-024-02206-z>
- Pavia, D., Pesaresi, C., Vito, C.D., 2019. The re-elaboration of John Snow's map in a GIS environment. Input for transferring methodological and applied skills being inspired by a virtuous practical example of social utility. *J-Read. J. Res. Didact. Geogr.* 2.
- Peer, V., Schwartz, N., Green, M.S., 2023. Gender differences in tuberculosis incidence rates—A pooled analysis of data from seven high-income countries by age group and time period. *Front. Public Health* 10, 997025. <https://doi.org/10.3389/fpubh.2022.997025>
- Peter, E., 2019. Altitude therapy at the sanatorium for the treatment of tuberculosis. *Main Libr. Blog*. URL <https://www.uzh.ch/blog/hbz/2019/10/28/altitude-therapy-at-the-sanatorium-for-the-treatment-of-tuberculosis/?lang=en> (accessed 2.14.25).
- Peterson, L.E., 2009. K-nearest neighbor. *Scholarpedia* 4, 1883. <https://doi.org/10.4249/scholarpedia.1883>
- Pezeshki, Z., Tafazzoli-Shadpour, M., Mansourian, A., Eshrati, B., Omid, E., Nejadqoli, I., 2012. Model of cholera dissemination using geographic information systems and fuzzy clustering means: Case study, Chabahar, Iran. *Public Health* 126, 881–887. <https://doi.org/10.1016/j.puhe.2012.07.002>
- Picken, R.M.F., 1937. CHANGE IN THE AGE OF MORTALITY FROM DIPHTHERIA. *The Lancet*, Originally published as Volume 1, Issue 5938 229, 1145–1451. [https://doi.org/10.1016/S0140-6736\(00\)98096-6](https://doi.org/10.1016/S0140-6736(00)98096-6)
- Rebsamen, H., Bauer, C., Capol, J., 1992. Zürich. INSA Inventar Neuer. Schweiz. Archit. 1850-1920 Städte Inventaire Suisse D'architecture 1850-1920 Villes Inventar. Sviz-zero Archit. 1850-1920 Città 10, 197–456. <https://doi.org/10.5169/seals-10931>
- Ritzmann, I., 2017. Sanatorien. *Hist. Lex. Schweiz*.
- Ritzmann, I., 2015. Impfung. *Hist. Lex. Schweiz*.
- Ritzmann, I., 2010a. Heilsame Höhenluft? : die Höhenkliniken als Wallfahrtsorte, in: *Zauber Berge. hier + jetzt*, Zürich, pp. 61–65.
- Ritzmann, I., 2010b. Schauderberge : die Schweiz als Krankheitslandschaft, in: *Zauber Berge. hier + jetzt*, Zürich, pp. 163–167.
- Ritzmann, I., 1998. Hausordnung und Liegekur: vom Volkssanatorium zur Spezialklinik : 100 Jahre Zürcher Höhenklinik Wald. *Chronos*, Zürich.
- Rucker, W.C., Kearny, R.A., 1913. Tuberculosis in Switzerland: Results of the Campaign against the Disease. *Public Health Rep.* 1896-1970 28, 2815–2829. <https://doi.org/10.2307/4570294>
- Ruiz-Moreno, D., Pascual, M., Emch, M., Yunus, M., 2010. Spatial clustering in the spatio-temporal dynamics of endemic cholera. *BMC Infect. Dis.* 10, 51. <https://doi.org/10.1186/1471-2334-10-51>

- Sasaki, S., Suzuki, H., Igarashi, K., Tambatamba, B., Mulenga, P., 2008. Spatial Analysis of Risk Factor of Cholera Outbreak for 2003–2004 in a Peri-urban Area of Lusaka, Zambia. *Am. J. Trop. Med. Hyg.* 79, 414–421. <https://doi.org/10.4269/ajtmh.2008.79.414>
- Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* 42, 1–21. <https://doi.org/10.1145/3068335>
- Schulamamt Stadt Zürich, 2016. Die Stadtvereinigung.
- Senti, A., 1928. Zürichs Bevölkerung im Jahre 1928. *Zür. Stat. Nachrichten* 5, 235–253.
- Senti, A., Pfister, H.O., 1946. Tuberkulosesterblichkeit in Zürich. *Zür. Stat. Nachrichten* 23, 133–180.
- Setiawan, A., Hendrati, L.Y., Mirasa, Y.A., 2021. The Mapping And Analysis Of Diphtheria Cases In Surabaya (2017-2018). *J. Biom. Dan Kependud.* 10, 45. <https://doi.org/10.20473/jbk.v10i1.2021.45-52>
- Shaweno, D., Karmakar, M., Alene, K.A., Ragonnet, R., Clements, A.C., Trauer, J.M., Denholm, J.T., McBryde, E.S., 2018. Methods used in the spatial analysis of tuberculosis epidemiology: a systematic review. *BMC Med.* 16, 193. <https://doi.org/10.1186/s12916-018-1178-4>
- Shiode, N., Shiode, S., Rod-Thatcher, E., Rana, S., Vinten-Johansen, P., 2015. The mortality rates and the space-time patterns of John Snow's cholera epidemic map. *Int. J. Health Geogr.* 14, 21. <https://doi.org/10.1186/s12942-015-0011-y>
- Shiode, S., 2012. Revisiting John Snow's map: Network-based spatial demarcation of cholera area. *Int. J. Geogr. Inf. Sci. - GIS* 26, 133–150. <https://doi.org/10.1080/13658816.2011.577433>
- Silberschmidt, W., 1930. The Acute Infectious Diseases In Switzerland. *J. State Med.* 1912-1937 38, 109–113.
- Slocum, T.A., McMaster, R.B., Kessler, F.C., Howard, H.H., 2022. Thematic Cartography and Geovisualization, 4th ed. CRC Press, Boca Raton. <https://doi.org/10.1201/9781003150527>
- Stadt Zürich, 2025a. Stadtkreise und Quartiere Zürich und Winterthur (OGD) - Stadt Zürich [WWW Document]. URL <https://www.stadt-zuerich.ch/geodaten/download/278> (accessed 1.16.25).
- Stadt Zürich, 2025b. Statistische Quartiere - Stadt Zürich [WWW Document]. URL https://www.stadt-zuerich.ch/geodaten/download/Statistische_Quartiere (accessed 1.16.25).
- Stadt Zürich, 2025c. Öffentliche Oberflächengewässer (OGD) - Stadt Zürich [WWW Document]. URL <https://www.stadt-zuerich.ch/geodaten/download/45> (accessed 1.16.25).
- Stadt Zürich, 2025d. Schulanlagen [WWW Document]. URL <https://www.stadt-zuerich.ch/geodaten/download/Schulanlagen> (accessed 3.21.25).
- Stadt Zürich, 2024. Statistisches Jahrbuch der Stadt Zürich [WWW Document]. URL <https://www.stadt-zuerich.ch/de/politik-und-verwaltung/statistik-und-daten/publikationen-und-dienstleistungen/publikationen/jahrbuch.html> (accessed 11.1.24).
- Stadt Zürich, 2018a. Zürich wird zur Grossstadt [WWW Document]. URL <https://www.stadt-zuerich.ch/de/stadtleben/stadtportraet/zuerichs-geschichte/digitale-zeitreise/zuerich-wird-zur-grossstadt.html> (accessed 2.6.25).

- Stadt Zürich, 2018b. Bevölkerung | Bevölkerungsentwicklung über Zeit [WWW Document]. URL <https://www.stadt-zuerich.ch/de/stadtleben/stadtportraet/zuerichs-geschichte/digitale-zeitreise/zuercher-bevoelkerung.html> (accessed 2.6.25).
- Statistik Stadt Zürich, 1936. Statistisches Jahrbuch der Stadt Zürich 1935.
- Statistik Stadt Zürich, 1931. Statistisches Jahrbuch der Stadt Zürich 1930.
- Staub, K., Jüni, P., Urner, M., Matthes, K.L., Leuch, C., Gemperle, G., Bender, N., Fabrikant, S.I., Puhani, M., Rühli, F., Gruebner, O., Floris, J., 2021. Public Health Interventions, Epidemic Growth, and Regional Variation of the 1918 Influenza Pandemic Outbreak in a Swiss Canton and Its Greater Regions. *Ann. Intern. Med.* 174, 533–539. <https://doi.org/10.7326/M20-6231>
- Sun, W., Gong, J., Zhou, J., Zhao, Y., Tan, J., Ibrahim, A.N., Zhou, Y., 2015. A Spatial, Social and Environmental Study of Tuberculosis in China Using Statistical and GIS Technology. *Int. J. Environ. Res. Public Health* 12, 1425–1448. <https://doi.org/10.3390/ijerph120201425>
- Sweller, J., 1994. Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* 4, 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- swisstopo, 2024a. Swiss Map Vector 1000 [WWW Document]. URL <https://www.swisstopo.admin.ch/de/landeskarte-swiss-map-vector-1000> (accessed 3.22.25).
- swisstopo, 2024b. Digitale Siegfriedkarte 1:25'000 [WWW Document]. URL <https://www.swisstopo.admin.ch/de/digitale-siegfriedkarte-1-25000> (accessed 1.16.25).
- swisstopo, 2024c. Digitale Dufourkarte [WWW Document]. URL <https://www.swisstopo.admin.ch/de/digitale-dufourkarte> (accessed 1.28.25).
- swisstopo, 2023. Einfach Koordinaten von Adressen in der Schweiz generieren und auf map.geo.admin.ch anzeigen [WWW Document]. URL <https://www.geo.admin.ch/de/koordinaten-von-adressen-generieren> (accessed 10.28.24).
- Tiwari, N., Adhikari, C., Tewari, A., Kandpal, V., 2006. Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic. *Int. J. Health Geogr.* 5, 33. <https://doi.org/10.1186/1476-072X-5-33>
- Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* 46, 234–240. <https://doi.org/10.2307/143141>
- Truelove, S.A., Keegan, L.T., Moss, W.J., Chaisson, L.H., Macher, E., Azman, A.S., Lessler, J., 2020. Clinical and Epidemiological Aspects of Diphtheria: A Systematic Review and Pooled Analysis. *Clin. Infect. Dis.* 71, 89–97. <https://doi.org/10.1093/cid/ciz808>
- Tsatsaris, A., Kalogeropoulos, K., Stathopoulos, N., 2023. Chapter 1 - Geoinformatics, spatial epidemiology, and public health, in: Stathopoulos, N., Tsatsaris, A., Kalogeropoulos, K. (Eds.), *Geoinformatics for Geosciences, Earth Observation*. Elsevier, pp. 3–29. <https://doi.org/10.1016/B978-0-323-98983-1.00002-8>
- van Leeuwen, M.H.D., Maas, I., Miles, A., 2002. History of Work - HISCO | IISG [WWW Document]. URL <https://iisg.amsterdam/en/data/data-websites/history-of-work> (accessed 1.20.25).
- Vaughan, L., 2018. *Mapping Society: The Spatial Dimensions of Social Cartography*. UCL Press. <https://doi.org/10.2307/j.ctv550dcj>
- Walford, N.S., 2020. Demographic and social context of deaths during the 1854 cholera outbreak in Soho, London: a reappraisal of Dr John Snow's investigation. *Health Place* 65, 102402. <https://doi.org/10.1016/j.healthplace.2020.102402>

- Waters, N., 2017. Tobler's First Law of Geography. <https://doi.org/10.1002/9781118786352.wbieg1011>
- Wood, S.M., Alston, L., Beks, H., Mc Namara, K., Coffee, N.T., Clark, R.A., Wong Shee, A., Versace, V.L., 2023. Quality appraisal of spatial epidemiology and health geography research: A scoping review of systematic reviews. *Health Place* 83, 103108. <https://doi.org/10.1016/j.healthplace.2023.103108>
- Ziegler, E., Matthes, K.L., Middelkamp, P.W., Schünemann, V., Althaus, C.L., Rühli, F., Staub, K., 2024. (Re-)modelling of the disease and mortality burden of the 1918-1920 influenza pandemic in Zurich, Switzerland. *medRxiv* 2024.03.14.24304276. <https://doi.org/10.1101/2024.03.14.24304276>

Appendix

Figure sources references

Figure 14

Dreamstime.com, illustration of a street sign by Sidip Farhutochman, image ID349584828, 2025. URL: “<https://www.dreamstime.com/street-sign-street-name-sign-icon-illustration-logo-vector-image349584828>”, accessed on April 10.

The Noun Project, Geocoding icon by ProSymbols, ID 3467401, licensed under Creative Commons Attribution 4.0, 2025. URL: “<https://thenounproject.com/icon/geocoding-3467401/>”, accessed on April 10.

Vecteezy: Vecteezy, Data Visualization icon by Freepik, licensed under Freepik License, URL: “<https://www.vecteezy.com/vector-art/23752910-data-visualization-icon-in-vector-illustration>”, accessed on April 10, 2025.

Flaticon: Flaticon, Analyse icon by Freepik, licensed under Flaticon License, URL: “https://www.flaticon.com/de/kostenloses-icon/analyse_1006585”, accessed on April 10, 2025.

Geocoding methodology

The geocoding of the residential addresses commenced with the linkage of street names, the suffix “strasse” (German for street), house numbers and the addition of the city name “Zürich” in a Microsoft Excel spreadsheet. This was achieved by employing the &-operator which facilitates the concatenation of individual cells and text entered in quotation marks. The formula is expressed as follows:

$$= H4 \& "strasse " \& I4 \& ", Zürich"$$

where the H and I columns correspond to the street and house number columns, respectively. The result for the first entry is therefore “Lubsstrasse 39, Zürich” for the entries “Lubs” and “39”. However, it should be noted that there are numerous exceptions to this rather elementary implementation, as in some cases the street name was entered already complete without having to add the “strasse”. Consequently, it is necessary to define exceptions to the aforementioned formula. The IF-function fulfils this requirement, in conjunction with the ISNUMBER- and FIND-functions. By utilising these functions in succession, exceptions can be delineated where the “strasse” element can be omitted, as demonstrated below.

$$= IF(ISNUMBER(FIND("strasse"; H4)); H4 \& " " \& I4 \& ", Zürich"; H4 \& "strasse " \& I4 \& ", Zürich")$$

It is necessary to add all the exceptions where the street name does not contain the element “strasse”, for which the OR-function is required. This includes a multitude of streets

and squares whose names contain “hof” (yard, e.g. Münsterhof), “markt” (market, e.g. Rindermarkt), “gasse” (alley), “quai” (quay), “graben” (ditch, e.g. Hirschengraben), “platz” (square) and some more peculiarities of German street names which lack a direct translation (“steig”, “rain”) and are at times case sensitive. Other intentionally included street names contain “im” (in, e.g. Im Sydefädeli), “lände” (lands, e.g. Schiff lände), “Schipfe” and [Am] “Wasser”. Combined with a check for entries where the street name could not be deciphered and was entered as “???”, “?” “-” or void (“”), the first and second IF-statements could be formulated and create an empty result if the street name could not be read (first line) or does not add the “strasse” element if any of the exceptions applies respectively. The final IF-statement is employed to verify exceptions to the aforementioned exceptions as the “Waffenplatzstrasse” would be reduced to “Waffenplatz” due to the second IF-statement. The same principle applies to the “Bahnhofstrasse” and “Farbhofstrasse”.

$$= IF(OR(H4 = ???; H4 = "?"; H4 = "-"; H4 = ""); "", H4 \&$$

$$IF(OR($$

$$ISNUMBER(FIND("hof"; H4));$$

$$ISNUMBER(FIND("markt"; H4));$$

$$ISNUMBER(FIND("gasse"; H4));$$

$$ISNUMBER(FIND("Gasse"; H4));$$

$$ISNUMBER(FIND("weg"; H4));$$

$$ISNUMBER(FIND("quai"; H4));$$

$$ISNUMBER(FIND("Im"; H4));$$

$$ISNUMBER(FIND("im"; H4));$$

$$ISNUMBER(FIND("graben"; H4));$$

$$ISNUMBER(FIND("steig"; H4));$$

$$ISNUMBER(FIND("strasse"; H4));$$

$$ISNUMBER(FIND("str."; H4));$$

$$ISNUMBER(FIND("g."; H4));$$

$$ISNUMBER(FIND("lände"; H4));$$

$$ISNUMBER(FIND("Strasse"; H4));$$

$$ISNUMBER(FIND("Schipfe"; H4));$$

$$ISNUMBER(FIND("Wasser"; H4));$$

$$ISNUMBER(FIND(platz; H4));$$

```
ISNUMBER(FIND("rain"; H4));

);

IF(OR(H4 = "Waffenplatz"; H4 = "Bahnhof"; H4 = "Farbhof"); "strasse"; "");

"strasse") & " " & I4 & ", Zürich")
```

GeoDa options

For the HDBSCAN clustering of tuberculosis cases a MINCLUSTERSIZE of 50 and MINPOINTS of 5 were utilised. For clustering including social class the MINCLUSTERSIZE was adapted to 20 and the MINPOINTS to 3

For the HDBSCAN clustering of tuberculosis fatalities a MINCLUSTERSIZE of 15 and MINPOINTS of 5 were utilised. For clustering including social class a MINCLUSTERSIZE of 12 and a MINPOINTS of 5 were utilised.

For the HDBSCAN clustering of diphtheria cases a MINCLUSTERSIZE of 20 and MINPOINTS of 5 were utilised. For clustering including social class the MINCLUSTERSIZE was adapted to 10 and the MINPOINTS to 5.

R code

TB analysis

Fabio Schilling

2024-04-08

Setup libraries

Data loading & preparation

```
# Read in TB case location data & keep only ID & coordinate columns
TB_case <- read.csv("../Data/TB_locations_old_ZH.csv", header = TRUE, sep = ",") %>% select
(ID, N, E, lon, lat)

# Read in TB mortality location data
TB_death <- read.csv("../Data/TB_mortality_old_ZH.csv", header = TRUE, sep = ",") %>% selec
t(ID, N, E, lon, lat)

# Read in general TB data & drop unnecessary columns
TB <- read.csv("../Data/Tuberculosis_V2_old-ZH.csv", header = TRUE, sep = ";") %>% select(
-c(16:20))

# Remove the rows with NA values for ID from the data set
TB <- TB[!is.na(TB$ID), ]
```

Geometries

```
# Read geometries of Zurich
ZH_city <- read_sf("../Data/Stadtkreise Zürich Rekonstruktion/ganze Stadt bis 1934.shp") %>
% select(geometry)
```

```
# Read geometries of Zurich's districts
ZH_districts <- read_sf("../Data/Stadtkreise Zürich Rekonstruktion/Stadtkreise bis 1934_ohne See_Korrektur.shp")

# Read in counts shapefile
TB_counts <- st_read("../Data/Results/TB_morbidity_old-ZH_counts_in_polygon.shp")
```

Prepare data for spatial analysis

```
# Join the two dataframes of locations and further detail
TB_case_all <- left_join(TB_case, TB, by = "ID")

# Convert coordinates to numeric
TB_case_all$N <- as.numeric(TB_case_all$N)
TB_case_all$E <- as.numeric(TB_case_all$E)
TB_case_all$lon <- as.numeric(TB_case_all$lon)
TB_case_all$lat <- as.numeric(TB_case_all$lat)

# Join the two dataframes of locations and further detail
TB_death_all <- left_join(TB_death, TB, by = "ID")

# Convert coordinates to numeric
TB_death_all$N <- as.numeric((TB_death_all$N))
TB_death_all$E <- as.numeric((TB_death_all$E))
TB_death_all$lon <- as.numeric((TB_death_all$lon))
TB_death_all$lat <- as.numeric((TB_death_all$lat))

# Define the spatial window (study area) using min and max coordinates
min_x <- min(TB_case_all$E)
max_x <- max(TB_case_all$E)
min_y <- min(TB_case_all$N)
max_y <- max(TB_case_all$N)

study_window <- owin(xrange = c(min_x, max_x), yrange = c(min_y, max_y))

# Convert to point pattern object (ppp)
TB_case_ppp <- ppp(x = TB_case_all$E, y = TB_case_all$N, window = study_window)

TB_death_ppp <- ppp(x = TB_death_all$E, y = TB_death_all$N, window = study_window)
```

NN function (G-function) (TB)

```
# Calculate nearest neighbor distances
nn_distances <- nndist(TB_case_ppp)

# Summary statistics of the distances
summary(nn_distances)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00   10.92   28.69   38.68   49.66   615.13

# Histogram of nearest neighbor distances
hist(nn_distances, breaks=40, main="Nearest Neighbor Distances of tuberculosis cases", cex.
main = 2, cex.lab = 1.5, cex.axis = 1.5, col = "lightblue", xlab = "NN-distance [m]")
```

```
# Estimate the G-function
G <- Gest(TB_case_ppp, correction = "border")

# Plot the G-function
plot(G, main="G-function of tuberculosis cases", cex.main = 2, cex.lab = 1.5, cex.axis = 1.
5, xlab = "distance [m]")
abline(h = 0.5, lty = 3, col = "blue")
```



```

# Calculate nearest neighbor distances
nn_distances_rip <- nndist(TB_death_ppp)

# Summary statistics of the distances
summary(nn_distances_rip)

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      0.00   43.01   78.36  110.14  143.20  789.35

# Histogram of nearest neighbor distances
hist(nn_distances_rip, breaks=32, main="Nearest Neighbor Distances of tuberculosis fatalities", cex.main = 2, cex.lab = 1.5, cex.axis = 1.5, col = "lightblue", xlab = "NN-distance [m]")

```

```

# Estimate the G-function
G_rip <- Gest(TB_death_ppp, correction = "border")

## Plot the G-function
plot(G_rip, main="G-function of tuberculosis fatalities", cex.main = 2, cex.lab = 1.5, cex.axis = 1.5, xlab = "distance [m]")
abline(h = 0.5, lty = 3, col = "blue")

```

Nearest Neighbour Analysis (TB)

```

# Convert to sf object
TB_case_sf <- st_as_sf(TB_case_all, coords = c("E", "N"), crs = 2056)

# Add distances back to the original sf object
TB_case_sf$nearest_neighbor_distance <- nn_distances

# Convert to sf object
TB_death_sf <- st_as_sf(TB_death_all, coords = c("E", "N"), crs = 2056)

# Add distances back to the original sf object
TB_death_sf$nearest_neighbor_distance <- nn_distances_rip

# Summary statistics for nearest neighbor distances
summary_stats <- TB_case_sf %>%
  summarize(
    mean_distance = mean(nearest_neighbor_distance),
    sd_distance = sd(nearest_neighbor_distance),
    min_distance = min(nearest_neighbor_distance),
    max_distance = max(nearest_neighbor_distance)
  )

print(summary_stats)

## Simple feature collection with 1 feature and 4 fields
## Geometry type: MULTIPOINT
## Dimension:      XY
## Bounding box:  xmin: 2679922 ymin: 1242638 xmax: 2686560 ymax: 1251158
## Projected CRS: CH1903+ / LV95
##   mean_distance sd_distance min_distance max_distance
## 1      38.68136   48.80465           0       615.1324
##
##               geometry
## 1 MULTIPOINT ((2679922 124978...

```

Nearest Neighbour Index (NNI) (TB)

```
# Calculate NNI (easier & robuster)
NNI_case <- nni(TB_case_sf)

# Interpret result of NNI
print(paste("The NNI is:", round(NNI_case$NNI, 3)))

## [1] "The NNI is: 0.518"

print(paste("The z score is:", round(NNI_case$z.score, 2)))

## [1] "The z score is: -38.72"

print(paste("The p value is:", round(NNI_case$p, 3)))

## [1] "The p value is: 0"

# Calculate NNI (easier & robuster)
NNI_death <- nni(TB_death_sf)

## Warning: data contain duplicated points

# Interpret result of NNI
print(paste("The NNI is:", round(NNI_death$NNI, 3)))

## [1] "The NNI is: 0.735"

print(paste("The z score is:", round(NNI_death$z.score, 2)))

## [1] "The z score is: -9.3"

print(paste("The p value is:", round(NNI_death$p)))

## [1] "The p value is: 0"

# NNI per social class
lower_class <- TB_case_sf %>% filter(Social.class == "1")
middle_class <- TB_case_sf %>% filter(Social.class == "2")
upper_class <- TB_case_sf %>% filter(Social.class == "3")

NNI_lower <- nni(lower_class)

NNI_middle <- nni(middle_class)

NNI_upper <- nni(upper_class)

# Interpret result of NNI
print(paste("The NNI of the lower class is:", round(NNI_lower$NNI, 3)))

## [1] "The NNI of the lower class is: 0.61"

print(paste("The z score is:", round(NNI_lower$z.score, 2)))

## [1] "The z score is: -17"

print(paste("The p value is:", round(NNI_lower$p, 3)))

## [1] "The p value is: 0"

print(paste("The NNI of the middle class is:", round(NNI_middle$NNI, 3)))

## [1] "The NNI of the middle class is: 0.615"

print(paste("The z score is:", round(NNI_middle$z.score, 2)))

## [1] "The z score is: -17.69"
```

```

print(paste("The p value is:", round(NNI_middle$p, 3)))
## [1] "The p value is: 0"
print(paste("The NNI of the upper class is:", round(NNI_upper$NNI, 3)))
## [1] "The NNI of the upper class is: 0.716"
print(paste("The z score is:", round(NNI_upper$z.score, 2)))
## [1] "The z score is: -11.51"
print(paste("The p value is:", round(NNI_upper$p, 3)))
## [1] "The p value is: 0"

```

Logistic regression (TB)

```

# Convert 'Gestorben' column to binary
TB$Gestorben_binary <- ifelse(TB$Gestorben == "ja", 1, 0)

# Drop the outdated "Alter" column
TB <- TB %>% select(-Alter)

# Rename the recalculated "Alter" column
TB <- TB %>% rename_with(~ "Alter", .cols = 4)

# Format age as continuous variable
TB$Alter <- as.numeric(TB$Alter)

## Warning: NAs durch Umwandlung erzeugt

# Format gender as factor variable
TB$Geschlecht <- factor(TB$Geschlecht, levels = c("m", "w"))

# Format hospital admission as variable
TB$Spital <- ifelse(TB$Spital == 1, 1, 0)

# Format social class as variable
TB$Social.class <- factor(TB$Social.class, levels = c("1", "2", "3"))

# Simple regression
model <- lm(Gestorben_binary ~ Geschlecht + Alter + Spital + Social.class, data = TB)
summary(model)

##
## Call:
## lm(formula = Gestorben_binary ~ Geschlecht + Alter + Spital +
##     Social.class, data = TB)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48549 -0.22392 -0.15770 -0.08053  1.03587
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0539917  0.0376792  -1.433   0.1521
## Geschlechtw  -0.0046086  0.0200232  -0.230   0.8180
## Alter         0.0060490  0.0007038   8.595 <2e-16 ***
## Spital        0.0166807  0.0231933   0.719   0.4721
## Social.class2  0.0541677  0.0235751   2.298   0.0217 *
## Social.class3  0.0555638  0.0249996   2.223   0.0264 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3938 on 1597 degrees of freedom
## (249 Beobachtungen als fehlend gelöscht)

```

```
## Multiple R-squared:  0.05066,    Adjusted R-squared:  0.04769
## F-statistic: 17.04 on 5 and 1597 DF,  p-value: < 2.2e-16

# Global regression
global_model <- glm(Gestorben_binary ~ Geschlecht + Alter + Spital + Social.class, data = T
B, family = binomial)
summary(global_model)

##
## Call:
## glm(formula = Gestorben_binary ~ Geschlecht + Alter + Spital +
##       Social.class, family = binomial, data = TB)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.985350   0.255962 -11.663  < 2e-16 ***
## Geschlechtw   -0.048864   0.130331  -0.375   0.7077
## Alter          0.035430   0.004353   8.140 3.95e-16 ***
## Spital         0.112941   0.151535   0.745   0.4561
## Social.class2  0.357028   0.155678   2.293   0.0218 *
## Social.class3  0.375074   0.166080   2.258   0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 1624.6  on 1602  degrees of freedom
## Residual deviance: 1546.0  on 1597  degrees of freedom
## (249 Beobachtungen als fehlend gelöscht)
## AIC: 1558
##
## Number of Fisher Scoring iterations: 4

exp(coef(global_model))

##      (Intercept)      Geschlechtw          Alter          Spital Social.class2
##      0.05052182      0.95231087      1.03606527      1.11956642      1.42907605
## Social.class3
##      1.45509901
```

GWR

```
# Drop the outdated "Alter" column
TB_case_sf <- TB_case_sf %>% select(-Alter)

# Rename the recalculated "Alter" column
TB_case_sf <- TB_case_sf %>% rename_with(~ "Alter", .cols = 6)

# Format age as continuous variable
TB_case_sf$Alter <- as.numeric(TB_case_sf$Alter)

## Warning: NAs durch Umwandlung erzeugt

TB_case_sf$died_binary <- ifelse(TB_case_sf$Gestorben == "ja", 1, 0)
TB_case_sf$hospital <- as.numeric(TB_case_sf$Spital)

## Warning: NAs durch Umwandlung erzeugt

TB_case_sf$Social.class <- as.numeric(TB_case_sf$Social.class)

TB_case_sf <- TB_case_sf[!is.na(TB_case_sf$hospital), ]
TB_case_sf <- TB_case_sf[!is.na(TB_case_sf$Alter), ]
TB_case_sf <- TB_case_sf[!is.na(TB_case_sf$Social.class), ]

# Run GWR
TB_gwr <- gwr.basic(died_binary ~ Geschlecht + hospital + Alter + Social.class,
                    data = TB_case_sf,
```



```

        bw = 50, # bandwidth
        adaptive = TRUE)

# Visualize local coefficients
print(TB_gwr)

## *****
## *                               Package    GWmodel                               *
## *****
## Program starts at: 2025-04-08 20:58:03.233143
## Call:
## gwr.basic(formula = died_binary ~ Geschlecht + hospital + Alter +
##   Social.class, data = TB_case_sf, bw = 50, adaptive = TRUE)
##
## Dependent (y) variable: died_binary
## Independent variables: Geschlecht hospital Alter Social.class
## Number of data points: 1747
## *****
## *                               Results of Global Regression                               *
## *****
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4962 -0.2145 -0.1505 -0.0355  1.0344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0183123  0.0372099  -0.492   0.6227
## Geschlechtw  -0.0150181  0.0185001  -0.812   0.4170
## hospital     -0.0140598  0.0217716  -0.646   0.5185
## Alter         0.0059946  0.0006052   9.906 <2e-16 ***
## Social.class  0.0210293  0.0095104   2.211  0.0272 *
##
## ---Significance stars
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.3829 on 1742 degrees of freedom
## Multiple R-squared:  0.06314
## Adjusted R-squared:  0.06099
## F-statistic: 29.35 on 4 and 1742 DF, p-value: < 2.2e-16
## ***Extra Diagnostic information
## Residual sum of squares: 255.3937
## Sigma(hat): 0.3825669
## AIC: 1610.554
## AICc: 1610.602
## BIC: -58.8582
## *****
## *                               Results of Geographically Weighted Regression                               *
## *****
## *****Model calibration information*****
## Kernel function: bisquare
## Adaptive bandwidth: 50 (number of nearest neighbours)
## Regression points: the same locations as observations are used.
## Distance metric: Euclidean distance metric is used.
##
## *****Summary of GWR coefficient estimates:*****
##              Min.      1st Qu.      Median      3rd Qu.      Max.
## Intercept  -0.7530304 -0.1917156 -0.0262222  0.1324224  1.1872
## Geschlechtw -0.4715119 -0.1009327  0.0022413  0.0978906  0.4192
## hospital    -0.6420763 -0.1377262 -0.0320757  0.0744382  0.6727
## Alter       -0.0150123  0.0025209  0.0067505  0.0098991  0.0261
## Social.class -0.3323870 -0.0344568  0.0149660  0.0617136  0.2388
## *****Diagnostic information*****
## Number of data points: 1747

```

```
## Effective number of parameters (2*trace(S) - trace(S'S)): 551.6668
## Effective degrees of freedom (n-2*trace(S) + trace(S'S)): 1195.333
## AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): 2083.54
## AIC (GWR book, Fotheringham, et al. 2002, GWR p. 96, eq. 4.22): 1368.754
## BIC (GWR book, Fotheringham, et al. 2002, GWR p. 61, eq. 2.34): 2400.974
## Residual sum of squares: 175.076
## R-square value: 0.3577683
## Adjusted R-square value: 0.06111924
##
## *****
## Program stops at: 2025-04-08 20:58:04.070608

# GWR-Modell bereits berechnet (Ergebnis = gwr_model)
# Die Koeffizienten extrahieren
gwr_results_TB <- as.data.frame(TB_gwr$SDF) # Spatial Data Frame aus GWR
gwr_results_TB$geometry <- TB_case_sf$geometry # Geometrie hinzufügen
gwr_TB_sf <- st_as_sf(gwr_results_TB) # In sf-Objekt umwandeln

# Funktion zum Erstellen einer Karte für einen bestimmten Prädiktor
plot_gwr_coef <- function(var_name, title) {
  ggplot() +
    geom_sf(data = ZH_districts$geometry, color = "gray80") + # Hintergrund für Kontext
    geom_sf(data = gwr_TB_sf, aes(color = .data[[var_name]]), size = 1) +
    #scale_color_distiller(palette = "RdBu", direction = -1) +
    scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
    labs(title = title, fill = "Koeffizient") +
    xlab("Longitude") +
    ylab("Latitude") +
    theme_minimal()
}

# Karten für die Variablen erstellen
plot_gwr_coef("Geschlechtw", "Influence of gender")

plot_gwr_coef("Alter", "Influence of age")

plot_gwr_coef("Social.class", "Influence of social class")
```

DT analysis

Fabio Schilling

2025-04-08

Setup libraries

Data loading & preparation

```
# Read in TB case Location data & keep only ID & coordinate columns
DT_case <- read.csv("../Data/DT_locations_old_ZH.csv", header = TRUE, sep = ",") %>% select
(ID, N, E, lon, lat)

# Read in TB mortality Location data
DT_death <- read.csv("../Data/DT_mortality_old_ZH.csv", header = TRUE, sep = ",") %>% selec
t(ID, N, E, lon, lat)

# Read in general TB data & drop unnecessary columns
DT <- read.csv("../Data/Diphtheria_old-ZH.csv", header = TRUE, sep = ";") %>% select(-c(20)
)

# Remove the rows with NA values for ID from the data set
DT <- DT[!is.na(DT$ID), ]
```

Geometries

```
# Read geometries of Zurich
ZH_city <- read_sf("../Data/Stadtkreise Zürich Rekonstruktion/ganze Stadt bis 1934.shp") %>
% select(geometry)

# Read geometries of Zurich's districts
ZH_districts <- read_sf("../Data/Stadtkreise Zürich Rekonstruktion/Stadtkreise bis 1934_ohne See_Korrektur.shp")

# Read in counts shapefile
DT_counts <- st_read("../Data/Results/DT_morbidity_old-ZH_counts_in_polygon.shp")
```

Prepare data for spatial analysis

```
# Join the two dataframes of locations and further detail
DT_case_all <- left_join(DT_case, DT, by = "ID")

# Convert coordinates to numeric
DT_case_all$N <- as.numeric(DT_case_all$N)
DT_case_all$E <- as.numeric(DT_case_all$E)
DT_case_all$lon <- as.numeric(DT_case_all$lon)
DT_case_all$lat <- as.numeric(DT_case_all$lat)

# Join the two dataframes of locations and further detail
DT_death_all <- left_join(DT_death, DT, by = "ID")

# Convert coordinates to numeric
DT_death_all$N <- as.numeric(DT_death_all$N)
DT_death_all$E <- as.numeric(DT_death_all$E)
DT_death_all$lon <- as.numeric(DT_death_all$lon)
DT_death_all$lat <- as.numeric(DT_death_all$lat)

# Define the spatial window (study area) using min and max coordinates
min_x <- min(DT_case_all$E)
max_x <- max(DT_case_all$E)
min_y <- min(DT_case_all$N)
max_y <- max(DT_case_all$N)

study_window <- owin(xrange = c(min_x, max_x), yrange = c(min_y, max_y))

# Convert to point pattern object (ppp)
DT_case_ppp <- ppp(x = DT_case_all$E, y = DT_case_all$N, window = study_window)

DT_death_ppp <- ppp(x = DT_death_all$E, y = DT_death_all$N, window = study_window)
```

NN function (G-function) (TB)

```
# Calculate nearest neighbor distances
nn_distances_DT <- nndist(DT_case_ppp)

# Summary statistics of the distances
summary(nn_distances_DT)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.00    0.00   33.29   65.98   97.11  775.74

# Histogram of nearest neighbor distances
hist(nn_distances_DT, breaks=40, main="Nearest Neighbor Distances of diphtheria cases", cex
.main = 2, cex.lab = 1.5, cex.axis = 1.5, col = "lightblue", xlab = "NN-distance [m]")

# Estimate the G-function
G <- Gest(DT_case_ppp, correction = "border")
```

```
# Plot the G-function
plot(G, main="G-function of diphtheria cases", cex.main = 2, cex.lab = 1.5, cex.axis = 1.5,
xlab = "distance [m]")
abline(h = 0.5, lty = 3, col = "blue")
```

Nearest Neighbour Analysis (DT)

```
# Convert to sf object
DT_case_sf <- st_as_sf(DT_case_all, coords = c("E", "N"), crs = 2056)

# Add distances back to the original sf object
DT_case_sf$nearest_neighbor_distance <- nn_distances_DT

# Summary statistics for nearest neighbor distances
summary_stats <- DT_case_sf %>%
  summarize(
    mean_distance = mean(nearest_neighbor_distance),
    sd_distance = sd(nearest_neighbor_distance),
    min_distance = min(nearest_neighbor_distance),
    max_distance = max(nearest_neighbor_distance)
  )

print(summary_stats)

## Simple feature collection with 1 feature and 4 fields
## Geometry type: MULTIPOINT
## Dimension: XY
## Bounding box: xmin: 2680686 ymin: 1242199 xmax: 2686207 ymax: 1250321
## Projected CRS: CH1903+ / LV95
##   mean_distance sd_distance min_distance max_distance
## 1      65.97979   98.01249           0      775.7378
##               geometry
## 1 MULTIPOINT ((2680686 124982...
```

Nearest Neighbour Index (NNI) (DT)

```
# Calculate NNI (easier & robuster)
NNI_case_DT <- nni(DT_case_sf)

## Warning: data contain duplicated points

# Interpret result of NNI
print(paste("The NNI is:", round(NNI_case_DT$NNI, 3)))

## [1] "The NNI is: 0.534"

print(paste("The z score is:", round(NNI_case_DT$z.score, 2)))

## [1] "The z score is: -19.3"

print(paste("The p value is:", round(NNI_case_DT$p, 3)))

## [1] "The p value is: 0"

# NNI per social class
lower_class <- DT_case_sf %>% filter(Social.class == "1")
middle_class <- DT_case_sf %>% filter(Social.class == "2")
upper_class <- DT_case_sf %>% filter(Social.class == "3")

NNI_lower <- nni(lower_class)

## Warning: data contain duplicated points
```



```

NNI_middle <- nni(middle_class)

## Warning: data contain duplicated points

NNI_upper <- nni(upper_class)

## Warning: data contain duplicated points

# Interpret result of NNI
print(paste("The NNI of the lower class is:", round(NNI_lower$NNI, 3)))

## [1] "The NNI of the lower class is: 0.776"

print(paste("The z score is:", round(NNI_lower$z.score, 2)))

## [1] "The z score is: -4.94"

print(paste("The p value is:", round(NNI_lower$p, 3)))

## [1] "The p value is: 0"

print(paste("The NNI of the middle class is:", round(NNI_middle$NNI, 3)))

## [1] "The NNI of the middle class is: 0.575"

print(paste("The z score is:", round(NNI_middle$z.score, 2)))

## [1] "The z score is: -9.65"

print(paste("The p value is:", round(NNI_middle$p, 3)))

## [1] "The p value is: 0"

print(paste("The NNI of the upper class is:", round(NNI_upper$NNI, 3)))

## [1] "The NNI of the upper class is: 0.652"

print(paste("The z score is:", round(NNI_upper$z.score, 2)))

## [1] "The z score is: -8.52"

print(paste("The p value is:", round(NNI_upper$p, 3)))

## [1] "The p value is: 0"

```

Logistic regression (DT)

```

# Convert 'Gestorben' column to binary
DT$Gestorben_binary <- ifelse(DT$Tod == "ja", 1, 0)

# Format age as continuous variable
DT$Alter <- as.numeric(DT$Alter)

## Warning: NAs durch Umwandlung erzeugt

# Format gender as factor variable
DT$Geschlecht <- factor(DT$Geschlecht, levels = c("m", "w"))

# Format gender as factor variable
DT$Spital <- ifelse(DT$Spital == 1, 1, 0)

# Format social class as variable
DT$Social.class <- factor(DT$Social.class, levels = c("1", "2", "3"))

model_DT <- lm(Gestorben_binary ~ Geschlecht + Alter + Spital + Social.class, data = DT)
summary(model_DT)

```

```
##
## Call:
## lm(formula = Gestorben_binary ~ Geschlecht + Alter + Spital +
##      Social.class, data = DT)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05660 -0.04115 -0.02397 -0.01659  0.98622
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0242063   0.0164229   1.474   0.1410
## Geschlechtw  -0.0235305   0.0141192  -1.667   0.0961 .
## Alter        -0.0001164   0.0006783  -0.172   0.8638
## Spital        0.0136890   0.0142947   0.958   0.3386
## Social.class2  0.0176459   0.0176263   1.001   0.3172
## Social.class3  0.0188187   0.0165878   1.134   0.2571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.167 on 587 degrees of freedom
## (75 Beobachtungen als fehlend gelöscht)
## Multiple R-squared:  0.009174, Adjusted R-squared:  0.0007346
## F-statistic: 1.087 on 5 and 587 DF, p-value: 0.3663

# Global regression
global_model_DT <- glm(Gestorben_binary ~ Geschlecht + Alter + Spital + Social.class, data
= DT, family = binomial)
summary(global_model_DT)

##
## Call:
## glm(formula = Gestorben_binary ~ Geschlecht + Alter + Spital +
##      Social.class, family = binomial, data = DT)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.88716    0.69484  -5.594 2.21e-08 ***
## Geschlechtw  -0.88875    0.55490  -1.602   0.109
## Alter        -0.00484    0.02714  -0.178   0.858
## Spital        0.46418    0.50055   0.927   0.354
## Social.class2  0.72975    0.71781   1.017   0.309
## Social.class3  0.78913    0.68988   1.144   0.253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 154.28  on 592  degrees of freedom
## Residual deviance: 148.66  on 587  degrees of freedom
## (75 Beobachtungen als fehlend gelöscht)
## AIC: 160.66
##
## Number of Fisher Scoring iterations: 7
```

GWR

```
# Format age as continuous variable
DT_case_sf$Alter <- as.numeric(DT_case_sf$Alter)

## Warning: NAs durch Umwandlung erzeugt

DT_case_sf$died_binary <- ifelse(DT_case_sf$Tod == "ja", 1, 0)
DT_case_sf$hospital <- as.numeric(DT_case_sf$Spital)

## Warning: NAs durch Umwandlung erzeugt
```

```

DT_case_sf$Social.class <- as.numeric(DT_case_sf$Social.class)

DT_case_sf <- DT_case_sf[!is.na(DT_case_sf$hospital), ]
DT_case_sf <- DT_case_sf[!is.na(DT_case_sf$Alter), ]
DT_case_sf <- DT_case_sf[!is.na(DT_case_sf$Social.class), ]

# Run GWR
DT_gwr <- gwr.basic(died_binary ~ Geschlecht + hospital + Alter + Social.class,
                    data = DT_case_sf,
                    bw = 50, # bandwidth
                    adaptive = TRUE)

# Visualize local coefficients
print(DT_gwr)

## *****
## *                               Package    Gwmodel                               *
## *****
## Program starts at: 2025-04-08 20:53:42.780645
## Call:
## gwr.basic(formula = died_binary ~ Geschlecht + hospital + Alter +
##   Social.class, data = DT_case_sf, bw = 50, adaptive = TRUE)
##
## Dependent (y) variable: died_binary
## Independent variables: Geschlecht hospital Alter Social.class
## Number of data points: 326
## *****
## *                               Results of Global Regression                               *
## *****
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05874 -0.04309 -0.03251 -0.02399  0.98082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.437e-02  3.718e-02   0.925   0.356
## Geschlechtw  -6.359e-03  2.074e-02  -0.307   0.759
## hospital      -1.725e-02  2.096e-02  -0.823   0.411
## Alter         9.706e-05  9.673e-04   0.100   0.920
## Social.class  1.274e-02  1.086e-02   1.174   0.241
##
## ---Significance stars
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 0.1814 on 321 degrees of freedom
## Multiple R-squared:  0.00612
## Adjusted R-squared: -0.006265
## F-statistic: 0.4941 on 4 and 321 DF,  p-value: 0.7401
## ***Extra Diagnostic information
## Residual sum of squares: 10.56379
## Sigma(hat): 0.1805666
## AIC: -180.8579
## AICc: -180.5945
## BIC: -449.4151
## *****
## *                               Results of Geographically Weighted Regression                               *
## *****
## *****Model calibration information*****
## Kernel function: bisquare
## Adaptive bandwidth: 50 (number of nearest neighbours)
## Regression points: the same locations as observations are used.
## Distance metric: Euclidean distance metric is used.
##

```

```
## *****Summary of GWR coefficient estimates:*****
##           Min.      1st Qu.      Median      3rd Qu.      Max.
## Intercept -0.41696548 -0.06687345  0.00000000  0.10863437  0.5689
## Geschlechtw -0.23961963 -0.07893113 -0.02498960  0.04076336  0.5033
## hospital -0.21839262 -0.06267230 -0.02332865  0.05765624  0.2766
## Alter -0.01711387 -0.00143124 -0.00025983  0.00063420  0.0044
## Social.class -0.05047057 -0.00014591  0.00632893  0.02345894  0.1120
## *****Diagnostic information*****
## Number of data points: 326
## Effective number of parameters (2trace(S) - trace(S'S)): 100.38
## Effective degrees of freedom (n-2trace(S) + trace(S'S)): 225.62
## AICc (GWR book, Fotheringham, et al. 2002, p. 61, eq 2.33): -105.5815
## AIC (GWR book, Fotheringham, et al. 2002, GWR p. 96, eq. 4.22): -239.9132
## BIC (GWR book, Fotheringham, et al. 2002, GWR p. 61, eq. 2.34): -186.9285
## Residual sum of squares: 7.172351
## R-square value: 0.3251988
## Adjusted R-square value: 0.02363813
##
## *****
## Program stops at: 2025-04-08 20:53:42.810344

# GWR-Modell bereits berechnet (Ergebnis = gwr_model)
# Die Koeffizienten extrahieren
gwr_results_DT <- as.data.frame(DT_gwr$SDF) # Spatial Data Frame aus GWR
gwr_results_DT$geometry <- DT_case_sf$geometry # Geometrie hinzufügen
gwr_DT_sf <- st_as_sf(gwr_results_DT) # In sf-Objekt umwandeln

# Funktion zum Erstellen einer Karte für einen bestimmten Prädiktor
plot_gwr_coeff <- function(var_name, title) {
  ggplot() +
    geom_sf(data = ZH_districts$geometry, color = "gray80") + # Hintergrund für Kontext
    geom_sf(data = gwr_DT_sf, aes(color = .data[[var_name]]), size = 1) +
    #scale_color_distiller(palette = "RdBu", direction = -1) +
    scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
    labs(title = title, fill = "Koeffizient") +
    xlab("Longitude") +
    ylab("Latitude") +
    theme_minimal()
}

# Karten für die Variablen erstellen
plot_gwr_coeff("Alter", "Influence of age")

plot_gwr_coeff("Social.class", "Influence of social class")
```

Python code

Analysis for MSc thesis

Setup

```
import pandas as pd
import numpy as np
import geopandas as gpd
import libpysal as ps
import matplotlib.pyplot as plt
from matplotlib.ticker import MaxNLocator
```

Read in data from excel files

```
# Load data
TB = pd.read_excel('Tuberculosis_V2_old-ZH.xlsx', header = 1)
```



```

# Drop first row
TB = TB.drop(0)

# Drop all unnamed and unneeded columns
TB = TB.drop(columns = ['Unnamed: 0', 'Alter', 'Unnamed: 16', 'Unnamed: 17', 'Foto',
'Adresse',
    'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22', 'Statistics',
    'Unnamed: 24', 'Unnamed: 25', 'Unnamed: 26',
    'Unnamed: 27', 'Unnamed: 28', 'Unnamed: 29', 'Unnamed: 30'])
# Basic statistics
TB["Anzeigedatum"] = pd.to_datetime(TB["Anzeigedatum"], errors='coerce')
TB["Anzeigedatum"] = TB["Anzeigedatum"].dt.date
TB["Alter"] = TB["Alter neu"].astype("Int64")
C:\Users\Fabio\AppData\Local\Temp\ipykernel_5776\121820076.py:2: UserWarning: Pars-
ing dates in DD/MM/YYYY format when dayfirst=False (the default) was specified. This
may lead to inconsistently parsed dates! Specify a format to ensure consistent
parsing.
    TB["Anzeigedatum"] = pd.to_datetime(TB["Anzeigedatum"], errors='coerce')
# Load data
DT = pd.read_excel('Diphtheria_old-ZH.xlsx', header = 1)

```

```

# Drop first row
DT = DT.drop(0)
# Drop all unnamed and unneeded columns
DT = DT.drop(columns = ['Krankheitsbeginn', 'Foto', 'Adresse', 'Unnamed: 20', 'Un-
named: 21', 'Unnamed: 22',
    'Unnamed: 23', 'Unnamed: 24', 'Unnamed: 25', 'Unnamed: 26',
    'Unnamed: 27', 'Unnamed: 28'])
# Basic statistics
DT["Anzeigedatum"] = pd.to_datetime(DT["Anzeigedatum"], errors='coerce')
DT["Anzeigedatum"] = DT["Anzeigedatum"].dt.date
#DT["Alter"] = DT["Alter"].astype("Int64")

```

Histograms

```

# Drop rows with NaT values
TB_hist = TB.dropna(subset=['Anzeigedatum'])
TB_hist['Anzeigedatum'] = pd.to_datetime(TB_hist['Anzeigedatum'], errors='coerce')

# Check for any NaT values after conversion
if TB_hist['Anzeigedatum'].isna().any():
    print("Warning: Some dates could not be converted to datetime and are set as
NaT.")
C:\Users\Fabio\AppData\Local\Temp\ipykernel_5776\966915518.py:3: SettingWithCopy-
Warning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy
    TB_hist['Anzeigedatum'] = pd.to_datetime(TB_hist['Anzeigedatum'], errors='coer-
ce')
# Set the 'date' column as the DataFrame index (useful for resampling)
TB_hist.set_index('Anzeigedatum', inplace=True)

# Resample by month and count occurrences
monthly_counts = TB_hist.resample('M').size()

# Remove entries for the year 1931 (first 2 rows)
monthly_counts = monthly_counts.iloc[321:]

```

```
# Format the index to display only Year-Month
monthly_counts.index = monthly_counts.index.strftime('%Y-%m')
# Plotting the monthly histogram
monthly_counts.plot(kind='bar', figsize=(10, 6), color='skyblue', edgecolor='grey')
plt.title('Monthly Tuberculosis Cases', fontsize=16)
plt.xlabel('Year-Month')
plt.xticks(ticks=range(len(monthly_counts)), labels=[label if i % 2 == 0 else ''
for i, label in enumerate(monthly_counts.index)], rotation=45) # Skip every second
tick
plt.ylabel('Number of Cases')
#plt.xlim(319, 367)
plt.show()
```

```
# Plotting the monthly histogram
monthly_counts.plot(kind='line', figsize=(10, 6), color='skyblue')
plt.title('Monthly Tuberculosis Cases', fontsize=16)
plt.xlabel('Year-Month')
plt.xticks(ticks=range(len(monthly_counts)), labels=[label if i % 2 == 0 else ''
for i, label in enumerate(monthly_counts.index)], rotation=45) # Skip every second
tick
plt.ylabel('Number of Cases')
#plt.xlim(319, 367)
plt.show()
```

```
# Drop rows with NaT values
DT_hist = DT.dropna(subset=['Anzeigedatum'])
DT_hist['Anzeigedatum'] = pd.to_datetime(DT_hist['Anzeigedatum'], errors='coerce')
```

```
# Check for any NaT values after conversion
if DT_hist['Anzeigedatum'].isna().any():
    print("Warning: Some dates could not be converted to datetime and are set as
NaT.")
C:\Users\Fabio\AppData\Local\Temp\ipykernel_5776\92631.py:3: SettingWithCopyWarn-
ing:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
DT_hist['Anzeigedatum'] = pd.to_datetime(DT_hist['Anzeigedatum'], errors='coer-
ce')
# Set the 'date' column as the DataFrame index (useful for resampling)
DT_hist.set_index('Anzeigedatum', inplace=True)
```

```
# Resample by month and count occurrences
monthly_counts_DT = DT_hist.resample('M').size()
```

```
# Remove entries for the year 1931 (first 2 rows)
#monthly_counts_DT = monthly_counts_DT.iloc[321:]
```

```
# Format the index to display only Year-Month
monthly_counts_DT.index = monthly_counts_DT.index.strftime('%Y-%m')
# Plotting the monthly histogram
monthly_counts_DT.plot(kind='bar',      figsize=(10,      6),      color='skyblue',
edgecolor='grey')
plt.title('Monthly Diphtheria Cases', fontsize=16)
plt.xlabel('Year-Month')
```

```

plt.xticks(ticks=range(len(monthly_counts_DT)), labels=[label if i % 2 == 0 else ''
for i, label in enumerate(monthly_counts_DT.index)], rotation=45) # Skip every
second tick
plt.ylabel('Number of Cases')
#plt.xlim(319, 367)
plt.show()

```

```

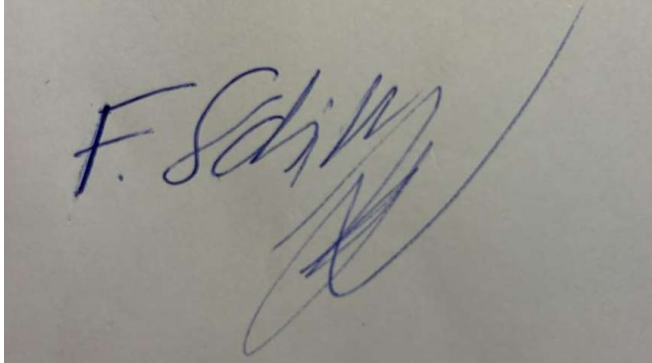
# Plotting the monthly histogram
monthly_counts_DT.plot(kind='line', figsize=(10, 6), color='skyblue')
plt.title('Monthly Diphtheria Cases', fontsize=16)
plt.xlabel('Year-Month')
plt.xticks(ticks=range(len(monthly_counts_DT)), labels=[label if i % 2 == 0 else ''
for i, label in enumerate(monthly_counts_DT.index)], rotation=45) # Skip every
second tick
plt.ylabel('Number of Cases')
#plt.xlim(319, 367)
plt.show()

```

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.

A handwritten signature in blue ink on a light-colored background. The signature appears to be 'F. Schim' followed by a stylized, cursive flourish that extends upwards and to the right.

Embrach, 17 April 2025