MODELLING HEALTH DATA ON A SMALL URBAN SCALE USING DETERMINISTIC ITERATIVE PROPORTIONAL FITTING

A Contribution to Setting up Citywide Health Monitoring Systems

PhD Thesis
by Evgenia Yosifova
September 2021

ACKNOWLEDGEMENTS

I would like to express my gratitude to my first supervisor Prof. Dr. Jörg Pohlan for his strong belief in the relevance of the chosen topic and for always showing me the bigger picture. My thanks go to my second supervisor Prof. Irene Peters, PhD as well, for her constant support and the unceasing flow of new ideas.

I would like to acknowledge my colleagues from the research project 'Gesunde Quartiere', who were part of my journey in the field of urban health from the very beginning. Thank you all for the many fruitful discussions and your curiosity about the urban planning perspective over the past four years. I especially want to thank Nele Mindermann, Johanna Buchcik, and Jana Borutta for providing me with the external data to validate my model.

I am also grateful to my interview partners Prof. Dr. Heiko Becher, Dr. Enno Swart, Prof. Dr. Susanne Busch, and Dr. Jobst Augustin for taking the time to contribute to my research with their perspective on the necessity of small-scale health data. Thank you for challenging my approach and pointing out the limitations, but still encouraging its further exploration.

My thanks go out to Dr. Ronny Kuhnert and Dr. Alice Nennecke for providing me with individual health data from the national representative survey GEDA 2012, and data about the prevalence of oncological diseases in Hamburg, respectively. Without these data sources, the adventure of modelling health data using spatial microsimulation would not have been possible.

In addition, I want to thank my fellow colleagues and friends – Tim, Venetsiya, Emiliya, Alessandro, Donald, and Maria for always asking how my dissertation is going and thus giving me the motivation to reach the finish line.

Last, but not least I would like to thank Ivan, who first told me about spatial microsimulation and never let me quit.

Table of Contents

1. INTRODUCTION

Cities, particularly those home of millions, are often viewed as unhealthy places. Poor air quality, noise pollution, and scarcity of public green space are among the first few things that come to mind when thinking about metropolises. Yet, does this picture apply to their entire territory? Are there no variations whatsoever in the characteristics of the urban environment in different neighbourhoods?

Obviously, these are rhetorical questions. Cities are not homogenous. In fact, they can be extremely diverse. There are various factors sculpturing the urban environment such as topography, proximity to water, street network, building structure, open public spaces, and so on. With this in view, the physical attributes of the urban environment usually vary significantly across neighbourhoods, thus resulting in unequal living conditions for the population. Someone living in an apartment located right next to an arterial road is going to have a different quality of life than someone living in a single-family house with private garden in a quiet residential area. This type of inequality is referred to as 'environmental injustice' and it is often the result of socio-economic inequality – while some have the luxury of choosing where in the city to live, others are left only with the affordable options. Usually, the latter are less attractive both in terms of housing and for reasons concerning the surrounding area.

Over time, the continuous exposure to noise and poor air quality may trigger chronic illnesses such as high blood pressure or asthma in those living in the apartment on the arterial road. Those enjoying the peacefulness of their private garden, on the other hand, will generally cope better with stress and will thus have higher odds of living a longer life. While these are merely a couple of examples to illustrate how contrasting the influence of zip code can be on health, they do manage to convey the diversity of metropolises.

Unfortunately, health-related dynamics like these usually remain hidden at the small urban scale. For reasons of data protection, the access to personal health records in Europe[1] is strictly regulated. Health data is published exclusively in the form of aggregated population counts or proportions for certain spatial units – regions, municipalities, cities, etc. These generally encompass tens of thousands, if not even more, inhabitants. While this approach ensures the anonymity of individuals and their health characteristics, it also impedes the identification of potential disease patterns at underlying spatial scales. Therefore, urban planning measures designed to facilitate disease prevention are not usually based on actual health data. Instead, the decision-making process relies on information sources such as geodata pointing to concentrations of multiple environmental risks, and regional health statistics. However, those cannot substitute disaggregated health data providing insight into the situation in the neighbourhoods across the city.

If such kind of data were available, a citywide small-scale health monitoring could be put in place. The latter would make it easier to identify population groups prone to developing specific medical conditions due to the combination of personal risk factors and unfavourable aspects of their living environment. Such a citywide monitoring can act as an early-warning system. At the same time, it could reveal hotspots of chronically ill inhabitants exposed to environmental

[1] The dissertation is limited to the European and more specifically the German spatial and political context.

threats such as noise or air pollution. Thus, necessary planning measures can be taken in more timely fashion. Additionally, the monitoring can provide reasoning for designing customised health-care packages for certain population groups. Although such course of action is in the realm of public health, urban planning can contribute by integrating the largely missing spatial ingredient into the picture.

While the idea of a citywide small-scale health monitoring sounds promising, one vital question remains unanswered – 'How to get small-scale health data?' Surveys are costly and time-consuming. Collecting representative sample data for every neighbourhood within a city on an annual basis is hardly realistic. At the same time, the General Data Protection Regulation (GDPR) of the European Union (EU) from 2018 makes it difficult for health insurance funds to provide data aggregated in a way that could potentially compromise the anonymity of insurants. Hence, obtaining data for areas encompassing only a couple of thousand inhabitants – as is generally the case with urban neighbourhoods – is practically impossible. With this in view, the acquisition of small-scale health data appears challenging, even more so for an entire city. Proposing a way to overcome this obstacle is where this dissertation aims.

2. GOAL, METHODS, AND OUTLINE

2.1. Goal and Research Questions

The goal of this dissertation is to explore a method for generating health data on a small urban scale – one corresponding to the perception of neighbourhoods. To that end, I used the city of Hamburg as case study. Taking into consideration its established levels of spatial division and available data sources, I investigated a potential strategy for developing a health-related model at the neighbourhood level.

Thus, I aspire to contribute to the academic discussion of modelling health data on a small scale. With the generated findings, I offer an objective perspective on the advantages and limitations related to using such data for setting up health monitoring systems in cities in general. To that end, I demonstrate how the modelled small-scale data can be used for the identification of spatial interactions between environmental and socio-economic factors on the one hand, and the prevalence of chronic disease on the other. Furthermore, I address the potential gains of small-scale health data availability in light of the COVID-19 pandemic.

Against this background, the research questions are as follows:

- How can health-related data be generated on a small urban scale?
- How can spatial interactions between environmental and/or socioeconomic factors and the prevalence of chronic disease be made evident using the modelled data?
- How can the generated small-scale health data facilitate the efforts of public health officials to combat the novel coronavirus?

2.2. Methods and Outline

Instead of analysing various available data modelling methods solely from theoretical point of view, I adopted a hands-on approach to demonstrate the steps comprising one particular method in greater detail. I applied spatial microsimulation, a well-established data modelling technique, to generate a so-called 'synthetic population' (see Chapter 2.3.2. 'Spatial microsimulation and population synthesis') and thus simulate the distribution of various non-communicable chronic illnesses at the neighbourhood level.

This dissertation is divided into several chapters. **Chapter 3. '**Cities and Health – A **Theoretical Overview'** provides scientific evidence of the relationships between social status, factors of the living environment, and individual health. The existing instruments for monitoring health and social status in Hamburg are also addressed. To conclude the theoretical part of the dissertation, **Chapter 4. 'A Spatial Microsimulation Approach'** clarifies the difference between modelling, simulation, microsimulation, and spatial microsimulation. Explaining the basic requirements for conducting a spatial microsimulation and describing the main existing approaches to population synthesis are the focus of this chapter. **Chapter 5. 'Modelling Health Data on a Small Urban Scale from the Perspective of Public Health Researchers'** offers a summary of the opinions of several local experts in the field of public health about the importance of small-scale health data and the challenges related to its generation. The latter were gathered through interviews, which I carried out via Zoom for the purpose of this research. **Chapter 6. 'Modelling Health-Related Data in Hamburg's Neighbourhoods'** demonstrates the entire process of generating synthetic population in detail – from choosing a suitable level

of spatial division, through selecting datasets and data pre-processing, to writing the actual population synthesis algorithm[2]. **Chapter 7. 'Model Validation'** is dedicated to the internal and external validation of the generated data. For the latter, two different external data sources are used – a sample survey conducted in six of Hamburg's neighbourhoods, and aggregated data from three health insurance funds in Hamburg. **Chapter 8. 'A Contribution to Setting up Citywide Health Monitoring Systems'** demonstrates a couple of application possibilities for the modelled data. First, the model is employed to identify hotspots of hypertensive individuals exposed to excessive levels of road and air noise. Second, the generated synthetic population is used to visualise where in the city live vulnerable individuals in terms of COVID-19. **Chapter 9. 'Summary and Discussion'** summarises the main findings and discusses the relevance of the model for future research purposes. **Chapter 10. 'Conclusion and Outlook'** draws the final conclusions and provides an outlook on the required prerequisites for applying the examined approach in other cities.

2.3. A Word on Terminology

Within the scope of this dissertation, there are some specific terms I am going to use often that need clarification upfront.

2.3.1. Small urban scale and levels of spatial division in Hamburg

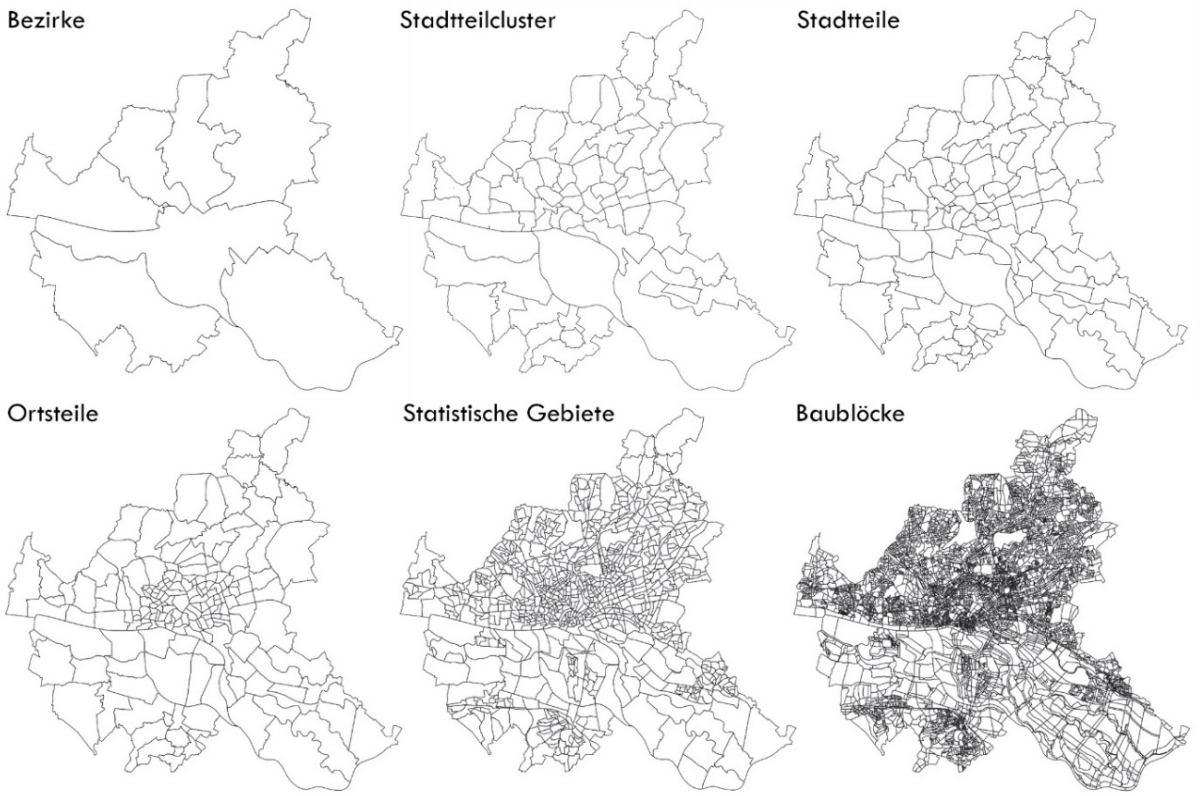
First of all, what is a *small urban scale*? Generally, every division in spatial units within the boundaries of a given city can be considered a small urban scale. To provide argumentation about the choice of spatial scale for the purposes of this dissertation, the next few paragraphs will shed more light onto the different levels of spatial division in Hamburg.

Generally, bigger cities are divided into administrative units for various purposes, such as voting, urban governance, and planning policies. In Hamburg, there are at least six different levels of spatial division, all illustrated in Figure 1. The highest level of spatial division is represented by the *Bezirke*, units with their own parliament and executive branch, which correspond to the English term *borough*. Hamburg has seven boroughs: Altona, Bergedorf, Eimsbüttel, Harburg, Hamburg-Mitte, Hamburg-Nord, and Wandsbek. The next level of spatial division is generally composed by the so-called *Stadtteile*, however there is an intermediate tier, not for administrative, but for research purposes – the *Stadtteilcluster*. This tier is based on the Stadtteile, but those of them with smaller populations are grouped into clusters. The Cancer Registry and the Morbidity Atlas of Hamburg, which I am going to address further below, provide health-related data aggregated at this spatial scale to ensure compliance with data protection regulations. The Stadtteile roughly correspond to *city quarters*. Hamburg has a total of 104 Stadtteile, ranging in population from 500 (Spadenland) to nearly 89.000 (Rahlstedt) inhabitants. Stadtteile, or city quarters, as I am going to refer to them henceforth, are administrative units, but without their own government. Densely populated city quarters are additionally divided in so-called *Ortsteile*. There are two further levels of spatial division – *Statistische Gebiete* and *Baublöcke*. Baublock is the German word for the term *urban block* – a spatial unit encompassed by streets on all sides. *Statistische Gebiete*, or *statistical areas* in English, were introduced following the Census in 1987 to obtain data for statistical purposes on a regular basis. They were shaped

[2] a set of subsequent commands applied to a string of data elements (e.g., vector, data frame, matrix, array) to transform them in a desired way.

with the aim to form homogeneous spatial units. It was therefore considered plausible for them to have an identical population size of around 2.000 inhabitants. The similarity of the building structure within the defined boundaries also played an important role (Loll 1991, p.92). Currently, the number of statistical areas in Hamburg is 941.

Figure 1. Levels of spatial division in Hamburg (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021; Statistisches Amt für Hamburg und Schleswig-Holstein 2017b)



Against this background, statistical areas come closest to the understanding of *neighbourhood*, bearing in mind that the individual perception of one's neighbourhood and its boundaries can vary significantly. How does one define a neighbourhood has long been discussed by geographers, epidemiologists, and other scholars whose research is based in the context of cities (Schnur 2008, p.22). To collect, analyse and compare urban data, one needs to spatially define the 'data holder'. Often, this decision must be practical. Especially when the aim is to develop a citywide small-scale monitoring, administrative spatial units are the most plausible choice. With their average population size of approximately 2.000 inhabitants, I considered the statistical areas the most suitable scale for conducting my analysis. Therefore, for the purposes of this dissertation, the terms *small urban scale*, *neighbourhood scale,* and *neighbourhoods* refer to Hamburg's statistical areas.

2.3.2. Spatial microsimulation and population synthesis

Moving on to the methodological part, I want to focus on the term *spatial microsimulation*. The method is based on the combination of population datasets (so-called micro datasets) without specific geographic dimension, and geographic datasets containing aggregated data related to specific geographic units. According to Tanton and Edwards (2013), '*the basic premise of microsimulation is that a more realistic picture of aggregate behaviour can be derived from*

looking at individual behaviour and modelling the interaction between the individual units in the system under consideration' (p.3). Generally, spatial microsimulation models can be applied for the following purposes: small area estimation, small area projection, and small area policy modelling (Tanton and Edwards 2013, p.5). Lovelace and Dumont (2016) describe spatial microsimulation as both a method to generate small-scale individual data, thus being '*roughly synonymous with 'population synthesis"* and as an '*approach to understanding multi-level phenomena based on spatial microdata – simulated or real*' (p.7).

Population synthesis represents the first step in setting up a spatial microsimulation model. It consists in generating a synthetic population by allocating individuals from sample surveys to spatial units. To that end, each individual is assigned a weight of representativity for each spatial unit and is then replicated there as many times as the integerised weight. The population synthesis procedure is discussed in detail in Chapter 6.

After this stage is completed, the generated population can be used to *simulate* changes in its composition based on modifying the parameters of certain variables. For instance, COVID-19 mortality rates available for specific population groups (e.g., in terms of age, sex, and comorbidities) can be used to project which neighbourhoods will have the highest number of deaths depending on the known combinations of individual characteristics of their populations. The simulation part of spatial microsimulation modelling is, however, beyond the scope of this dissertation.

2.3.3. Citywide health monitoring systems

Finally, I want to define the term *citywide health monitoring system.* A health monitoring system is designed to continuously observe health-related individual characteristics at a certain spatial scale (Robert Koch-Institute 2019). To that end, specific indicators are defined, and relevant data is collected for their estimation in predetermined time intervals – e.g., annually, biannually, etc. If a city wants to monitor variations in the prevalence of certain diseases across its territory, for instance, the corresponding indicators will be estimated at the city quarter, neighbourhood, or another level of spatial division suitable for analysis. A *citywide health monitoring system* would therefore encompass all the units within the chosen level of spatial division. This will allow comparing the situation in different parts of the city and potentially identifying existing spatial patterns of higher prevalence rates.

With the basic terminology now clarified, the next chapter will focus on the multidimensional effects the urban environment can have on health and how those can vary across neighbourhoods. To that end, I am going to introduce some basic theoretical concepts about urban health, social inequality, and environmental justice. I will then provide an overview of several spatial determinants of health, including air and noise pollution, heat load, and public green spaces. Finally, I am going to present the existing instruments for monitoring health and social inequality at the small urban scale in Hamburg.

3. CITIES AND HEALTH - A THEORETICAL OVERVIEW

3.1. Health-Related Inequalities within Cities

Human health depends on numerous factors including age, gender, genetic predispositions, health behaviour, social and community networks, living and working conditions as well as the characteristics of one's socioeconomic, cultural, and physical environment (Dahlgren and Whitehead 1991, p.11). Non-communicable diseases thus often result from the unfavourable combination of (some of) these factors.

In this regard, to '*ensure healthy lives and promote well-being for all at all ages*' was defined as one of the 17 Sustainable Development Goals within the 2030 Agenda for Sustainable Development (United Nations 2015, p.14). With more than half of the global population currently living in cities and demographic projections expecting this share to continue rising (United Nations 2018), many more individuals will be directly affected by the diverse characteristics of the urban living environment. The role of cities is therefore going to be increasingly important for reaching this goal.

Tiwari et al. (2015) define living environment as '*an assembly of the natural and built environment which is offered to the inhabitants of the place who perform various kinds of social, cultural, religious, economic, and political activities*' (p.153). Theoretical concepts about the influence of the living environment on the health situation in urban neighbourhoods thus assume an interplay between the socio-economic and ethnic composition of the population on the one hand, and the social and physical environment on the other (Westenhöfer et al. 2021, p.32). In this context, the social environment represents prevailing norms and values, social cohesion, social capital, levels of security and violence. The physical environment, on the other hand, covers aspects such as environmental pollution, housing quality, healthcare infrastructure, access to recreational and leisure facilities, food access, and other, aesthetic aspects (ibid.). Both the social and the physical environment affect the subjective perception of one's neighbourhood and thus influence individual health behaviour and well-being (Meijer 2013, p.31). Urban health thus '*reflects the outcomes of the physical and the social environment that impact residents' and communities' well-being and quality of life, within an urban setting*' (Wuerzer 2014, p.6835).

Research interest in urban health has been constantly rising over the past decades. Exploring the links between the physical characteristics of urban neighbourhoods, their social status, and the health of their inhabitants has become the focus of both urban planning and public health scholars (e.g. Yosifova and Pohlan 2021; Buchcik et al. 2021). During the COVID-19 pandemic, the topic of urban health has become even more prominent than usually. The current health crisis is increasingly recognised as an opportunity to re-evaluate the sustainability of existing spatial structures and uses and take measures to reduce exacerbating social inequalities (Akademie für Raumentwicklung in der Leibniz-Gemeinschaft 2021).

Social inequality usually has a clear spatial manifestation within cities as population groups of different socio-economic standing often live isolated from each other in different neighbourhoods (OECD 2018). This phenomenon is widely known as *social segregation,* and it puts low-income households at risk of becoming '*tied to neighbourhoods with characteristics that affect*

their present and future well-being' (OECD 2018, p.12). In essence, differences in living conditions make some urban neighbourhoods more attractive than others, which leads to higher real-estate and rental prices in certain parts of cities. As a result, low-income households have a limited choice of residence location. Hence, social status has a strong determining effect on the kind of neighbourhood characteristics people are exposed to and can thus influence their health and well-being.

Scientific evidence pointing to the variety of effects the urban living environment can have on individual health is mounting. Physical and social characteristics of urban neighbourhoods affect people's ability to practice health-promoting lifestyles. For example, the proximity of public green spaces can foster physical activity, whereas the perceived lack of security may impede this health-promoting effect. Poor access to healthy food in the vicinity of one's place of residence makes it more difficult to maintain a nutritious diet and can thus eventually lead to obesity, cardiovascular disease, and even death (Meijer 2013, pp.28–29). Continuous exposure to high levels of noise may cause cardiovascular problems such as coronary heart disease and myocardial infarction (Hahad et al. 2019, pp.246–247).

Socially deprived neighbourhoods are generally associated with health-damaging features of the urban living environment: high levels of traffic volume, poorly maintained public green spaces, limited number of playgrounds and possibilities for recreation, etc. (Gold et al. 2012, p.17). Therefore, socially disadvantaged people '*face a double burden: being socially marginalised and being subject to the inequities resulting from being located in poor social and physical environments*' (Frohlich 2013, p.49).

Against this background, the WHO offers definitions for both *inequality* and *equity* in terms of health:

'*"Inequality" reflects any differences and disparities in relation to environmental health inequality. It signals differences in exposure to environmental health risks and related health outcomes*'

(WHO Regional Office for Europe 2019, p.2).

'*"Equity" reflects the political goal of achieving equal conditions and equal opportunities, referring to equity in health outcomes as well as (environmental and other) health risks and determinants*'

(ibid.).

Both terms are closely related to the concept of environmental justice. The latter originated in the 1980s during protests in the United States of America (USA), which aimed to stop the allocation of polluting factories and waste sites exclusively in black neighbourhoods and indigenous people's reservations (Stephens and Church 2017). A decade later, the environmental justice movement came to Europe, where the focus shifted from racial minorities to socially disadvantaged people. More specifically, environmental justice activists were concerned with the disproportionate burden carried by those with lowest incomes (ibid.).

Noise and air pollution, heat load, public green spaces – being some of the most prominent health-related factors of the urban living environment – are usually unequally distributed across neighbourhoods. Often, there are concentrations of several unfavourable environmental risk

factors in deprived urban areas. Low-income households are thus generally more exposed to multiple pollutants in their immediate living environment. Compared to the general population, they are hence more susceptible to related unfavourable health effects (Sexton 2014).

In the context of the ongoing COVID-19 pandemic, socially deprived people are, once again, more vulnerable than other individuals. Those with lower income often have essential service jobs, which cannot be done from home. This, coupled with typically crowded housing conditions, increases their risk of virus exposure, and facilitates transmission (Upshaw et al. 2021). Work-related mobility, insufficient financial resources for taking protective measures, and difficult access to healthcare services are risk factors, which affect mainly socially deprived population groups (Dragano and Conte 2020). In this regard, a recent cohort study from Scotland found that patients from deprived areas '*had higher frequency of critical care admission and a higher adjusted 30-day mortality*' (Lone et al. 2021, p.1). Against this background, to achieve the Sustainable Development Goal of ensuring healthy lives for all, it is essential to combat social and environmental inequality across urban neighbourhoods.

With the main theoretical concepts regarding health-related inequalities within cities now introduced, the following paragraphs will focus on several spatial determinants of health. The multidimensional health effects that can be triggered by noise and air pollution, heat load, and public green spaces will be separately discussed. Those represent merely a subset of environmental aspects and thus do not attempt to exhaust all imaginable health-related effects of the urban living environment. Rather than that, the aim is to illustrate how health is influenced by factors that are unevenly distributed within the spatial realm of cities.

3.2. Spatial Determinants of Health

3.2.1. Noise pollution

According to the WHO, environmental noise is '*an important public health issue, featuring among the top environmental risks to health* [and being] *a growing concern among both the general public and policy-makers in Europe*' (World Health Organization Regional Office for Europe 2018, p.xiii). Continuous exposure to high levels of noise can have adverse effects on individual health and well-being. It can cause annoyance and thus trigger stress reactions in the body. As a result, cortisol levels could be constantly elevated, thus increasing the risk of developing cardiovascular illnesses, immunosuppression, and gastric ulcers (Kohlhuber and Bolte 2012, p.12). Additionally, noise pollution can cause cognitive impairment and sleep disturbance (Kohlhuber et al. 2012, p.88; Moshammer et al. 2002, p.246). Over time, this can have serious mental health implications.

Scientific evidence suggests that the '*health of those of lower socio-economic status can be disproportionately affected by noise*' (European Environmental Agency 2018, p.24). Several studies carried out in Germany, Switzerland, and the Netherlands have established a link between the exposure to road traffic noise and low socio-economic status (e.g. Kohlhuber et al. 2006; Hoffmann et al. 2003; Laussmann et al. 2013; Braun-Fahrländer 2004; Kruize and Bouwman 2004). Findings from the United Kingdom (UK) show that the access to tranquil open public space in deprived neighbourhoods is more limited compared to affluent urban areas (Battaner-Moro et al. 2010). Nevertheless, the results depend highly on the indicator used to describe social deprivation. The European Environmental Agency (EEA) therefore considers

socio-economic status alone insufficient for predicting noise exposure '*even if, in many places, people of lower socio-economic status live in areas with higher levels of noise*' (European Environmental Agency 2018, p.25). Still, this does not contradict the uneven distribution of noise within cities and the disproportionate burden carried by those living in the most severely affected neighbourhoods. If these neighbourhoods happen to be socially deprived as well, the adverse effects on health are amplified.

3.2.2. Air pollution

Within the EU, more than 1.000 premature deaths on average are attributed to air pollution each day. For comparison, this is more than 10 times the victims of road accidents (European Court of Auditors 2018, p.8). Fine particles ($PM_{2,5}$), followed by nitrogen dioxide (NO_2), and ozone (O_3) have by far the largest contribution: '*$PM_{2,5}$ concentrations in 2014 were responsible for about 428.000 premature deaths originating from long-term exposure in Europe* […]. *The estimated impacts on the population* […] *of exposure to NO_2 and O_3 concentrations in 2014 were around 78.000 and 14.400 premature deaths per year, respectively*' (European Environmental Agency 2017, p.9). These numbers encompass 41 European countries and are thus not limited to the EU.

Against this background, '*air pollution is the single largest environmental health risk in Europe* [increasing] *the incidence of a wide range of diseases, mainly respiratory and cardiovascular diseases*' (European Environmental Agency 2018, p.19). Furthermore, there is mounting scientific evidence for a link between air pollution and type 2 diabetes in adults, obesity, systemic inflammation, Alzheimer's disease, and dementia (ibid.). Adverse health effects are evident not only in the long term (e.g. over years) but also as a result of short term exposure to airborne particles (e.g. over hours or days) (World Health Organization Regional Office for Europe 2013, p.5).

People with lower socio-economic status are usually more severely affected by air pollution than other population groups. This is often attributed to their overall worse health status, resulting from poor diet, unhealthy lifestyle, lack of adequate healthcare, and stress (Khreis et al. 2017). Still, the European Environmental Agency (2018) report '*Unequal exposure and unequal impacts: social vulnerability to air pollution, noise and extreme temperatures in Europe*' introduces abundant evidence pointing to higher overall concentrations of air pollution in deprived areas (p.22). It is therefore unlikely that any adverse effects on health are merely the result of the overall poor health of the inhabitants. For instance, half of London's most deprived neighbourhoods are exposed to levels of NO_2 exceeding EU's maximum allowed values. In contrast, only 2% of the affluent neighbourhoods exhibit similar NO_2-concentrations (Aether 2017). Similar observations were registered for PM_{10} and NO_2 in Dortmund (Shrestha et al. 2016), Ostrava (Šlachtová et al. 2016), Wales (Brunt et al. 2017), Lille and Marseille (Padilla et al. 2016), Wallonia (Lejeune et al. 2016), and the Netherlands (Fecht et al. 2015).

Studies from Bristol and Rotterdam, however, point to similar levels of PM and NO_2 both in deprived and affluent areas. In Rome, the association between social status and exposure to air pollution is even reverse, due to the preference of people with higher social status to live in the city centre where traffic volume is higher (European Environmental Agency 2018, p.22).

All in all, scientific evidence suggests that there is a link between social status and exposure to air pollution, although there are some discrepancies in the observed trend in different European cities. In any case, concentrations of air pollutants differ across neighbourhoods, which implies varying effects on the health of their inhabitants.

3.2.3. Heat load

Excessive heat load can lead to heat stress and thus trigger symptoms such as fatigue, cognitive impairment, and blood circulation problems. In urban areas, the situation is especially problematic because cooling down during the night is more difficult due to higher building density. As a result, physical regeneration while sleeping is hindered, which additionally exacerbates the adverse health effects caused by heat load. Prolonged heatwaves thus often lead to circulatory collapses, heart attacks or even heat strokes ending in death (Katzschner and Bruse 2012, pp.102–103).

Certain population groups are at higher risk of heat-related mortality. The most prominent risk factors include old age (e.g. Paavola 2017; Urban et al. 2017), comorbidities such as electrolyte imbalances, cardiovascular, and respiratory diseases (Wolf et al. 2015), as well as socio-economic status (e.g. Arbuthnott and Hajat 2017; Fernandez Milan and Creutzig 2015). Living alone has also been identified as a factor increasing vulnerability during heatwaves (e.g. Seebaß 2017; McGeehin and Mirabelli 2001).

In Europe, there is a tendency for the more vulnerable population groups to live in dense, urban environments, thus being exposed to higher temperatures. Often, city centres are characterised by larger proportions of '*elderly, people in poor health, and those living alone*' (European Environmental Agency 2018, p.28). Evidence from the UK points to concentrations of poorer communities within urban heat islands (UHI) (Wolf and McGregor 2013). Nevertheless, people of higher socio-economic standing often choose to live in the more densely built city centres, thus being exposed to excessive heat load as well. Therefore, socio-economic standing alone cannot be used as predictor for exposure to heatwaves. Vulnerability towards heat load depends on several individual characteristics, which is why data aggregated at the district or city quarter level entails the risk of blurring the picture by assuming homogeneity.

3.2.4. Public green spaces

Contrary to noise, air pollution, and heat load, public green spaces can have health-promoting effects. White et al. (2013) found that '*controlling for individual and regional covariates,* […] *on average, individuals have both lower mental distress and higher well-being when living in urban areas with more green space*' (p.920).

Public green spaces have both passive and active effects independent of the actual use by the inhabitants. Passive effects manifest in three main ways: contributing to a better urban climate, improving air quality, and limiting the perception of traffic noise. Actively obtained gain from frequent use of the available green space includes improved physical and mental health as well as increased social interactions within the local community. The latter diminishes the risk of social isolation and is particularly important for those living alone (Federal Ministry for the Environment, Nature Conservation and Nuclear Safety 2015, pp.45–47).

However, public green spaces are not evenly distributed within cities. Densely built neighbourhoods offer less green space per person compared to other, more loosely built-up areas. Furthermore, in Germany, public green space in socially deprived neighbourhoods is around one quarter less than the city average (38 vs. 50 square metres per person) (Federal Ministry for the Environment, Nature Conservation and Nuclear Safety 2015, p.13). Wüstemann et al. (2017) investigated the access to urban green on household and individual level and identified strong disparities in green space provision across major German cities by applying the Gini coefficient: '*statistical analysis of the socio-economic background of households and individuals shows differences in urban green provision related to income, age, education and children in household*' (p.124). Against this background, the health-promoting effects of public green spaces appear to be more easily accessible for those living in affluent urban neighbourhoods.

The previous paragraphs introduced scientific evidence for the unequal distribution of both health-damaging and health-promoting factors of the urban living environment. Some of those factors, particularly noise and air pollution, often manifest in a highly localised manner. Their effects may hence be concealed if health-related trends are observed solely at higher spatial scales, such as city quarters. An approach like this would not allow identifying concentrations of vulnerable population groups in terms of specific environmental risks such as heatwaves either. Aggregated health data provides only the basic contours of the main picture, thereby keeping individual risk factor combinations hidden.

With this in view, the next section will shed light on current instruments for monitoring social inequality and health in the city of Hamburg.

3.3. Monitoring Social Inequality and Health in Hamburg

3.3.1. Hamburg's Social Monitoring

In 2010, in reaction to the growing inequality in living conditions across the city, Hamburg's State Ministry of Urban Development and Environment introduced a so-called *Sozialmonitoring* (engl.: Social Monitoring). Through a combined observation of seven central indicators, it offers an overview of the socio-demographic and socio-economic situation at the level of the statistical areas (Pohlan et al. 2010, p.10). The respective indicators are:

- Share of inhabitants with migration background of the population under 18 years,
- Share of the population under 18 years with single parents,
- Share of recipients of basic welfare benefits (SGB[3] II) of total population,
- Share of unemployed aged between 15 and 65 years,
- Share of the population under 15 years unfit for work (SGB II-recipients),
- Share of recipients of minimum welfare benefits in old age (SGB XII) of population aged 65 years and over,
- Share of school leavers without a school-leaving qualification or with a basic certificate or middle school certificate of all school leavers (three-year sum). (Pohlan and Strote 2017, p.2429)

[3] Sozialgesetzbuch (engl.: Social code)

After a z-standardisation[4], the indicators are summed up to compute a so-called *Statussumme* (engl.: status sum). The status sum is then used for the classification of the statistical areas into four different *Statusindexklassen* (engl.: status index classes) based on the standard deviation (Std Dev): high (< -1.00 Std Dev), average (-1.00 ≤ Std Dev ≤ 1.00), low (1.00 < Std Dev ≤ 1.50 SD), and very low (> 1.5 Std Dev) (Pohlan et al. 2010, p.46).

Simply put, those statistical areas where the share of inhabitants with migration background, the share of unemployed, the share of welfare benefits recipients, etc. is significantly below the city average are classified as areas with high social status. Those statistical areas, where the corresponding proportions of population are close to the city average, are classified as areas with average social status. The remaining statistical areas, where the share of socially deprived people, as described by the seven central indicators, is considerably larger than the average for Hamburg, are categorised as having either low, or very low status. It is important to note that income is not taken into consideration by the Social Monitoring.

Based on the defined status index, both the current social status and the development direction of each statistical area, compared to the city average, are tracked each year. Thus, Hamburg's Social Monitoring is intended to act as early-warning system for social deprivation. Additionally, it functions as an instrument for identifying areas with cumulated problems, which may thus be in urgent need for action (Pohlan et al. 2010, p.3).

3.3.2. Hamburg's Morbidity Atlas

Three years later, in 2013, Hamburg's Morbidity Atlas was published. Commissioned by the State Ministry of Health and Consumer Protection, its goal was to examine differences in the healthcare demand of the population with statutory health insurance across the 104 city quarters. More specifically, the following aspects were explored: number of patients treated for certain illnesses, scope of healthcare provided by registered doctors, frequency of hospital admissions, and differences in disease burden in relation to social deprivation (measured in terms of unemployment and average income).

The Morbidity Atlas is based on accounting data about all statutory health insurance (SHI) accredited physician services, provided to people with statutory health insurance, who visited a registered doctor at least once in 2011. Hence, there is no data available about those with private health insurance or those with statutory health insurance, who did not visit a doctor in 2011.

Table 1 provides an overview of the examined medical conditions. In general, the Morbidity Atlas delivers information about the proportion of individuals[5] suffering from each of the listed illnesses. The available data is aggregated at the level of the city quarters and city quarter clusters and divided into four age categories: 0-17 years, 18-64 years, 65-79 years, and 80+ years (Erhart et al. 2013, p.3).

[4] Defined as: (observed value – mean) / standard deviation

[5] Defined as: share of population with statutory health insurance who visited the doctor at least once in 2011

Table 1. Medical conditions examined in the Morbidity Atlas (with ICD-10 Codes) (Source: Erhart et al. 2013, p.4)

Common illnesses	ICD-10 Codes
Diabetes	E10% - E14%; H36.0
Hypertension	I10%; I12%; I15%; I67.4
Heart failure	I11%; I13%; I26%, I27%; I42%; I43%; I50%; I51.0 - I51.7; I09.2; I31.0 - I31.1; R57.0; T46.0
Depression	F32%; F33%; F34.1
Dementia	F00% - F02%; F03%; F04%; F05.1; F06.5 - F06.9; G30%; G31%
Specific Diagnoses	**ICD-10 Codes**
Pregnancy in females aged 15-44	O00-O08%; O09% - O16%; O20% - O26%; O28% - O31%; O34% - O36%; O43.0 - O43.1; O44%; O46% - O47%; O48%; Z32%; Z33%; Z34%; Z35%; Z36%
Early diagnosis of cervical carcinoma	Z12.4
Asthma in children aged < 15 years	J45%; J46%
Acute bronchitis in children aged < 15 years	J20%; J21%; J22%
Glaucoma	H40%-H42%
Prostate carcinoma	C61%
Conductive or dissociative hearing loss	H90%
Epilepsy	G40%
% stands for the inclusion of all subordinated end-numbers of the ICD-10 codes	

The city quarters differ significantly in terms of population size. To ensure that the number of people with statutory health insurance used for the computation of disease prevalence is sufficient, 53 city quarters were thus merged into 16 city quarter clusters for the purposes of the Morbidity Atlas. The necessary prerequisites for consolidation were for the city quarters to be adjacent and to exhibit similar social structure. Insel Neuwerk and HafenCity were excluded. Thus, altogether, the Morbidity Atlas encompasses 67 city quarters and city quarter clusters (Erhart et al. 2013, pp.5–6).

3.3.3. Hamburg's Health Reporting System

Unlike the Morbidity Atlas, which was conducted once and has not been updated since, the so-called *Gesundheitsberichterstattung* (engl.: Health Reporting System) has long tradition in Germany starting in the early 1990s. In Hamburg, there are Base Health Reports, Reports related to specific life phases, and Special Health Reports dedicated to single health-related topics. Some of the central indicators in the Base Health Reports include:

- Fertility rate,
- Mortality,
- Premature mortality,
- Preventable deaths,
- Hospital admissions,
- Infant mortality,
- Cot death,

- Cancer mortality,
- Suicide mortality,
- Cardiovascular disease mortality,
- Respiratory disease mortality,
- Mortality due to injuring or poisoning. (Sozialbehörde 2021)

The last two Base Health Reports are from the years 2009 and 2018 and thus have an approximately 10-year-long gap of updating. The presented results include temporal and, wherever possible, spatial comparisons (Saier 2020, p.7). As in the Morbidity Atlas, the spatial scale of analysis in Hamburg's Base Health Reports is limited to the city quarters and city quarter clusters.

During the COVID-19 pandemic, a necessity for small-scale data to successfully navigate demand, deficits, and improvement efforts has become evident (Akademie für Raumentwicklung in der Leibniz-Gemeinschaft 2021, p.10). The pandemic brought to light the problem of insufficient digital data availability, required for an evidence-based, integrated decision-making. Only few of Germany's cities are equipped with Health Reporting Systems providing sufficient level of spatial detail. Digitalisation of data collection and data processing in the public health sector lags behind. Integrating data from different departments at the same spatial scale is limited. With this in view, there is currently an urgent necessity for the spatial integration of different data sources to facilitate timely site planning for testing and medical service centres. Besides methodological competence, data and technology availability is required (Akademie für Raumentwicklung in der Leibniz-Gemeinschaft 2021, p.7). In this regard, data equipment standards must be agreed upon and measures for their compliance must be applied (ibid., p.10). Woock and Busch (2021) argue that only a timely, setting-based approach to health promotion can contribute to increasing the overall resilience of the public health sector.

Against this background, the next chapter will provide theoretical information about the proposed spatial microsimulation approach to generating individual health data at the small scale.

4. A SPATIAL MICROSIMULATION APPROACH

4.1. Introducing Spatial Microsimulation

This section describes the basic notion of spatial microsimulation and explains how the approach relates to more familiar concepts such as modelling, simulation, and microsimulation. Its advantages, fields of application, and basic implementation requirements are also covered.

4.1.1. Modelling, simulation, and microsimulation

The concepts of *modelling*, *simulation*, and *microsimulation* are central for understanding how *spatial microsimulation* works. Gilbert (2000) argues that *modelling* comes first because '*there is some 'real world' phenomenon in which the researcher is interested*', a so-called *target*, and '*the objective is to create a model of this target which is simpler to study than the target itself*' (p.3). The model therefore serves as a mean for drawing conclusions about the target because '*the two are sufficiently similar*' (ibid.).

Simulation, on the other hand, represents a necessary addition to the model, it builds upon it. Since '*the target is always a dynamic entity, changing over time and reacting to its environment* […], *the model must also be dynamic* […] *Simulation means 'running' the model forward through (simulated) time and watching what happens*' (Gilbert 2000, p.4).

Contrary to simulating change by assuming homogeneity of all units included in the model, *microsimulation* consists in '*simulating the passage of a large number of 'base' units (usually individuals, households or firms) through time, while applying transfer functions* […] *to each unit* […] *independently of the others and there* [being] *no direct interaction between the units*' (Gilbert 2000, p.7). The key notion of microsimulation is therefore simulating how certain events affect individual units rather than groups of units aggregated based on some specific rule (e.g., place of residence).

Microsimulation was first introduced as concept by Guy Henderson Orcutt – an American economist, academic, and researcher – in the 1950s. Around that time, he became convinced that '*data aggregated to the national accounts level simply could not provide sufficient information for discovering the elusive secrets of the economy with enough reliability to be useful for policy guidance*' (Watts 1991, p.173). For this reason, Orcutt aspired to develop models capturing the complex behaviour of multiple microeconomic units. He believed that the implications of policies depend on how their impact is '*distributed among non-homogeneous groups*' (ibid.). He therefore considered aggregated population data an unreliable foundation for estimating the effects of a certain policy and strived for a more sophisticated approach that would account for the heterogeneity of individuals and households.

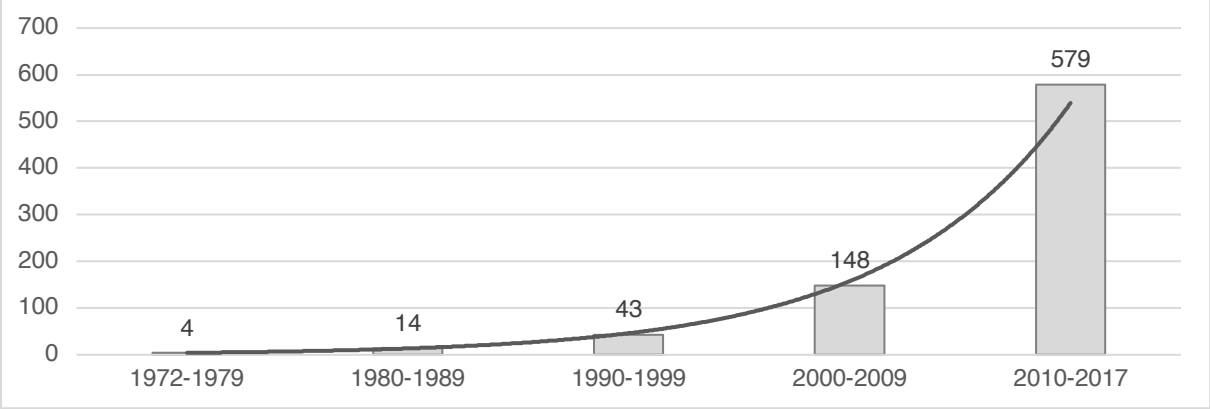
Orcutt's idea to estimate policy effects by modelling their impact on separate individuals rather than entire population groups came to fruition thanks to a couple of important developments happening at the time. First, university researchers had gotten an easier access to electronic digital computers, which had become powerful enough to process challenging computation amounts. As a result, the cost of manipulating and analysing large data files was reduced substantially. The other vital novelty was that the Survey Research Centre at the University of Michigan had begun to collect large data files containing household-related information that were made available to researchers. Orcutt thus recognised the opportunity '*of using large*

samples of microunits to estimate behavioral relations and the use of the same or similar samples to represent entire populations in simulations – both aspects making heavy use of the new computing machinery' (Watts 1991, p.174).

Orcutt first described his concept in the late 1950s and initialised its implementation with the help of several doctoral students – Martin Greenberger, John Korbel, and Alice Rivlin (Orcutt 1957). This first microsimulation model relied on a sample of 10.358 persons and simulated demographic processes (births, deaths, marriages, divorces, and ageing), labour supply, and education demand. The results of this work, which was referred to as *microanalytic modelling* at the time, are described in Orcutt et al. (1961) (Watts 1991, p.174).

Since the pioneering work of Guy Orcutt, microsimulation has been applied in various fields including social sciences, taxation, and health. The first health-related microsimulation model was developed just 15 years after Orcutt put his ideas into words. Nevertheless, the number of publications about using a microsimulation approach in the field of health research started to grow exponentially only after the millennium (Figure 2).

Figure 2. Number of publications related to the application of microsimulation in health research by decade (Source: Schofield et al. 2017, p.103)



In the early days of microsimulation, health researchers mostly adopted the approach to study topics related to family planning and the rate on conception (e.g. Mustafa 1973), fertility and breastfeeding (e.g. Roy 1984; Kono et al. 1983; Santow 1978). Other popular fields of health research that made use of this novel approach during the period 1975-1990 include cancer screening (e.g. Parkin 1985), health-insurance provision for employees (e.g. Chernick et al. 1987), health policy impact on individual behaviour (e.g. Yett et al. 1975), and transmission of vector-borne diseases (e.g. Plaisier et al. 1990). Most of those early microsimulation models were static rather than dynamic[6]. (Schofield et al. 2017, p.100).

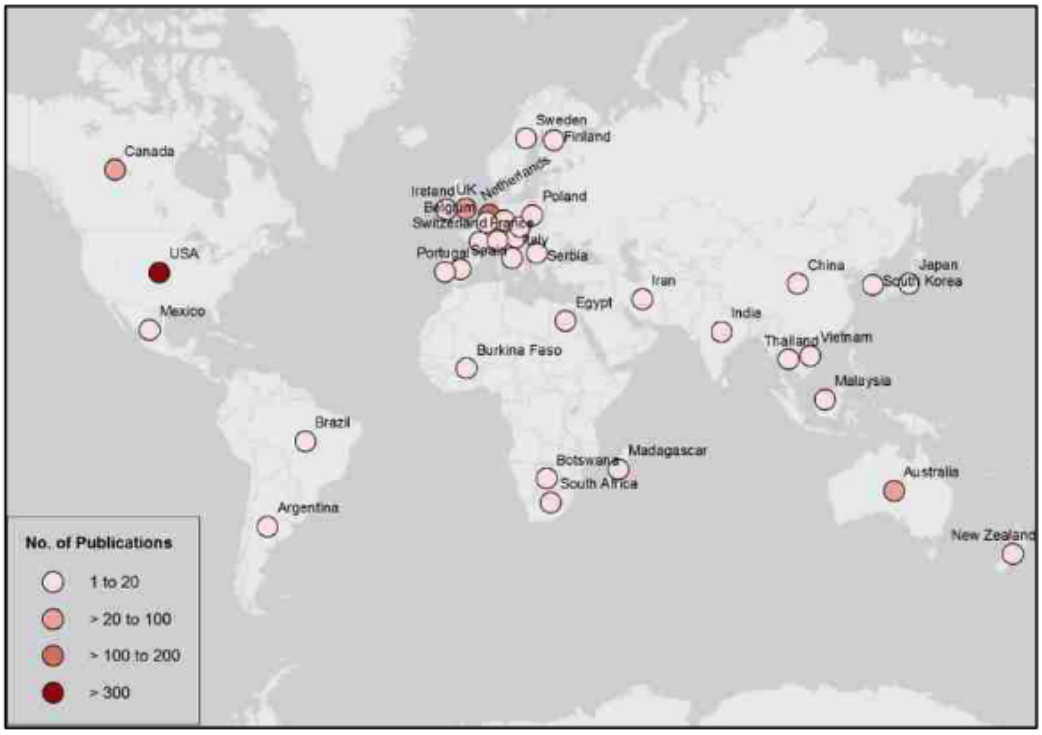
In the 1990s, the use of microsimulation in health research increased significantly, although most of the models were designed for other purposes. Criticism about those models '*not having the benefits of unit record health data and the capacity for distributional analysis that distinguished tax-benefit microsimulation models*' was very popular at the time (ibid., p.118). Using health survey data as base population proved to be the solution to overcoming this limitation.

[6] The difference between static and dynamic microsimulation models is explained in Chapter 4.2. 'Choosing a Spatial Microsimulation Method'.

Nowadays, the use of health surveys for the purpose of generating synthetic population has become the norm in microsimulation modelling '*with some models even being based on the collection of new primary data specific to the application, where appropriate survey data is not available*' (Schofield et al. 2017, p.118).

Applications of microsimulation methods for health research purposes have been constantly evolving over the years. Today, microsimulation models are being developed for '*the newest frontiers of medicine, such as genomic testing and personalised medicine*' (ibid.). Health researchers all over the world have been using microsimulation methods, whereby the most publications between 1972 and 2017 are from the USA, Australia, Canada, the UK, and the Netherlands (Figure 3).

Figure 3. Number of publications on the use of microsimulation for health research purposes 1972-2017, PubMed database (Source: Schofield et al. 2017, p.106)



The range of health-related topics, where microsimulation has found application is vast. Some examples include health expenditure and health policy (e.g. Lay-Yee and Cotterell 2015; Majstorovic et al. 2015), mortality (e.g. Carter et al. 2017), ageing and caregivers (e.g. Schneider and Kleindienst 2016; Singh et al. 2014), chronic illnesses such as cancer (e.g. Evans et al. 2013; Popadiuk et al. 2016) and diabetes (e.g. Willis et al. 2013; Clarke et al. 2004), transmission of disease (e.g. Cassels et al. 2008), cost-effectiveness of health interventions (e.g. Weycker et al. 2007), and spatial models (e.g. Rahman 2017; Campbell and Ballas 2016; Edwards and Clarke 2009). (Schofield et al. 2017, pp.108–112).

It is precisely the use of microsimulation methods for building spatial models that is going to be the focus of the next paragraphs. They will introduce the concept of spatial microsimulation – together with its advantages and various fields of application.

4.1.2. Spatial microsimulation: advantages and fields of application

Unlike traditional microsimulation, spatial microsimulation models contain geographical reference for each micro unit and thus enable estimating '*the characteristics of individuals within geographic zones about which only aggregate statistics are available*' (Lovelace and Ballas 2013, p.1). In this regard, Lovelace and Dumont (2016) define spatial microsimulation as the generation, the analysis, and the modelling of individual data allocated to geographic areas (p.7). The approach is based on the combination of population datasets (so-called *micro datasets*) without specific geographic dimension and geographic datasets containing aggregated data related to specific geographic areas. Therefore, micro datasets generated using a spatial microsimulation approach, typically refer to a certain geographic area. The latter may vary considerably in size and population count – it can be anything from an urban block to an entire region. Whatever the case may be, its physical boundaries are usually pre-defined for census purposes because aggregated place-specific data is required for setting up the model (Campbell and Ballas 2013, p.264).

A central advantage of adopting a spatial microsimulation approach over a traditional small area estimation approach is the generation of synthetic micro data for each spatial unit. While '*small area estimation methods produce a point estimate, spatial microsimulation can produce cross tabulations*' (Tanton 2014, p.5). This offers much more opportunities for in-depth analysis. If, for instance, the model aims to illustrate the spatial distribution of obesity, differences in terms of age, sex, income, education level, etc. – depending on the available variables – can be examined as well.

An additional major benefit of spatial microsimulation models is that the generated synthetic population can subsequently be updated using fertility and mortality rates (Ballas et al. 2007). Thus, demographic changes and their effect on the studied subject (e.g., heart failure) can also be explored at the spatial level chosen for analysis.

Another reason why spatial microsimulation models are so increasingly popular is that they allow transferring certain variables of interest (so-called *target variables*) from the micro dataset (e.g., national representative survey about health) to the synthetic population dataset. These target variables are not available in the geographic dataset. In other words, individuals from a survey sample can be allocated to the geographic areas at the desired spatial scale along with specific information from the survey (e.g., about suffering from a certain chronic disease). There are certain requirements for this procedure which I am going to address in more detail in Chapter 6.3. 'Selection of Constraint and Target Variables'.

The main reason why researchers are keen on implementing spatial microsimulation, however, is that the generated models can provide a solid base for decision-making and thus be successfully utilised by governments. Tanton and Edwards (2013) argue that the biggest strength of spatial microsimulation is its ability to model specific policies and test their potential effect at the small scale. Thus, decision makers can select geographic areas for intervention based on where the model predicts the strongest impact of the outlined measures.

With this in view, O'Donoghue et al. (2014) argue that despite speculations about microsimulation models being '*black box models, applicable only with caution where other methods are not available*', researchers are increasingly recognising spatial microsimulation as an important

instrument for analysis due to the rising awareness about the '*geographical impact of government policies, public and private investment and social networks*' (p.28). In this regard, Lovelace and Ballas (2013) point out that spatial microsimulation models '*cannot replace the 'gold standard' of real, small area microdata (Rees et al. 2002, p.4)*, [however] *the method's practical usefulness (see Tomintz et al. 2008) and testability (Edwards and Clarke 2009) are beyond doubt*' (p.2).

Against this background, spatial microsimulation has gained a lot of popularity over the past two decades and is currently being applied in various research fields including economic policy analysis (Campbell and Ballas 2013), welfare, poverty, and inequality (e.g. Tanton 2011; Chin et al. 2005; Harding et al. 2004; Ballas 2004), social policy (e.g. Ballas and Clarke 2001; Ballas et al. 2007), public policies related to education (e.g. Kavroudakis et al. 2012) and crime (e.g. Kongmuang et al. 2006), agriculture (e.g. Hynes et al. 2009; Ballas et al. 2006), regional development (e.g. van Wissen 2000), land use and spatial planning (e.g. Strauch et al. 2004), crisis planning and management (e.g. Chen et al. 2006), transport planning (e.g. Hollander and Liu 2008), influence assessment of demographics on heat consumption (e.g. Muñoz and Peters 2014) and last, but not least – health (e.g. An 2020; Campbell and Ballas 2016; Kosar and Tomintz 2014; Edwards and Clarke 2013).

Health was, in fact, one of the initial focal points of spatial microsimulation models. In the mid-1980s, Clarke et al. (1985) developed the so-called '*HIPS'* model (Health Information and Planning System) to facilitate the decision-making process of district health authorities in England and Wales. In its essence, the model represented a synthetic population with attributed demographic data instead of aggregated data for each district. Since then, the model has been updated annually (Tanton and Edwards 2013, p.4).

The decision to adopt a spatial microsimulation approach for health research purposes is usually motivated by (one of) several factors. First, disease data is often not available at the desired spatial scale. Using data aggregated at a larger scale can serve as alternative but is likely to provide a picture that lacks detail and assumes homogeneity of the observed population. Existing disease patterns at underlying spatial levels may thus remain hidden. Using sample surveys instead of spatially aggregated health data poses another risk. While they give great insight into the lives of the interviewed individuals, the public health perspective on entire population groups is limited. Against this background, spatial microsimulation offers a solution to providing more heterogeneity by generating synthetic populations for each geographic unit at a desired spatial scale. Furthermore, it is a cost-effective and time-saving alternative to conducting sample surveys (Edwards and Clarke 2013, pp.69–70).

4.1.3. Requirements for setting up a spatial microsimulation model

While this sounds promising, the quality of data needed to construct a spatial microsimulation model is decisive for its accuracy. As already pointed out, at least two data sources are necessary to set up this kind of model. One of the datasets must provide the micro data, i.e., individual data from a representative sample survey including wide range of information about the studied topic (e.g., health). The other dataset typically consists of data aggregated at a specific geographic scale (e.g., regions, districts, neighbourhoods, blocks, etc.) and can usually be obtained from local bureaus of statistics, or national census. The variables shared between the geographic and the micro dataset are used as so-called *constraints* or *benchmarks*.

They are essential because the modelled data has to be *constrained* to fit the known aggregates for each spatial unit (Cassells et al. 2013, pp.9–10).

It is of utmost importance for the variables used as constraints from the micro dataset and the geographic dataset to be defined in the same way, that is, the characteristic attributes must be classified identically. If, for instance, age is used as constraint, the corresponding variable must either be metric, or otherwise the age intervals must be the same in both datasets. If one of the variables is metric, and the other one is ordinal, the problem can be easily solved by converting the metric variable into an ordinal one. Yet, if both variables are ordinal, but the age intervals are classified in a way that does not allow for a meaningful conversion, age cannot be used as constraint variable.

Furthermore, it must be ensured that the semantic content of the chosen constraint variables is the same. A variable concerning unemployment, for instance, may not provide the exact same information in two different datasets, as unemployment can differ in terms of duration, receipt of unemployment benefits, etc. Additionally, it may be important to clarify if the variable refers to self-perceived unemployment or registration in the local employment agency. This should serve as an example how the information provided by a variable referring to the same matter may vary in different datasets.

Last, but not least, to transfer selected target variables from the micro dataset into the synthetic population dataset, there should be a statistically significant correlation between the constraint variable(s) and each of the selected target variables. Moreover, the constraint variables should be able to explain (at least some part of) the variance of the target variable(s). Therefore, specific statistical tests must be carried out before deciding which variables from the micro dataset can the transferred to the synthetic population datasets along with the individuals.

Once these requirements are met, a synthetic population can be generated by applying spatial microsimulation. In this context, there are several available methods to choose from depending on the available data, the topic of interest, and its specific characteristics.

4.2. Choosing a Spatial Microsimulation Method

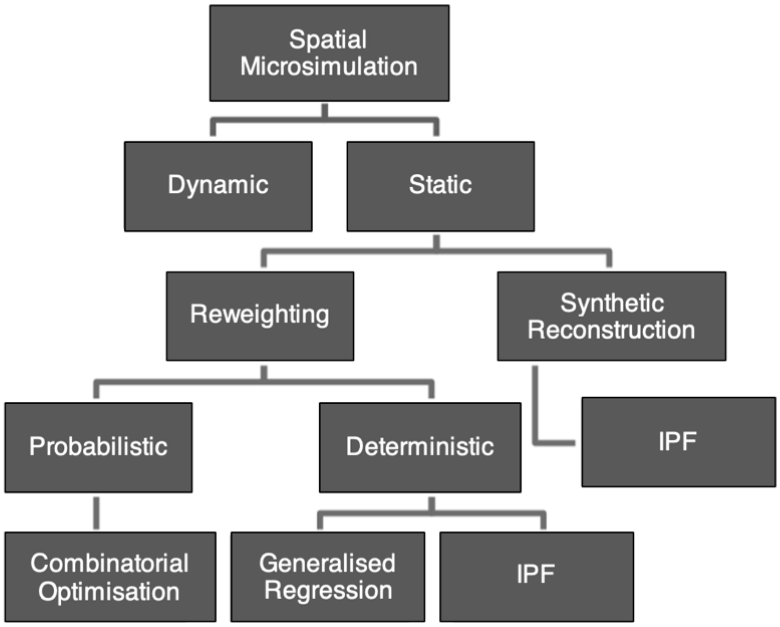
4.2.1. Static vs. dynamic methods

Spatial microsimulation methods are generally divided into two main categories – static and dynamic (Figure 4). Static methods do not account for changes in population over time. Instead, policy changes are modelled considering the *current* state of population. Therefore, static spatial microsimulation models are suitable for *next day analyses*, that is, if the change in policy is to be applied today, what would the outcome be tomorrow. In contrast, dynamic methods additionally consider demographic factors, such as births, deaths, and migration. The changes which are to result from a certain new policy are thus modelled for a *best guess* population for a specific time in the future (Tanton and Edwards 2013, pp.3–4). Dynamic models can therefore be regarded as an extension of static models. While they do take additional factors into account, the methodology for their generation is basically the same – it revolves around estimating the probability for each individual (or household) to be living in each area.

Both approaches have strengths and weaknesses. Static methods lack flexibility and provide limited time perspective. Dynamic methods offer insight into future population development patterns and are thus the more suitable choice when analysing long-term policy implications.

Nevertheless, dynamic methods rely on a demographic forecast, which may or may not prove to be fully correct. Therefore, depending on the available input data, static methods may prove to be more reliable.

Figure 4. Methods of spatial microsimulation (Source: Tanton 2014, p.7)



Static spatial microsimulation methods are further subcategorised in terms of data generation technique. In essence, there is the choice between synthetic reconstruction and reweighting.

4.2.2. Synthetic Reconstruction

Synthetic reconstruction operates by producing a list of individuals or households whose characteristics, when aggregated, match the already known aggregates at the spatial scale chosen for setting up the model. Usually, this process is carried out variable by variable so that individual '*characteristics are matched sequentially, rather than all at once*' (Tanton 2014, p.7). For example, the first variable to account for in the observed population can be age. Synthetic individuals are then going to be allocated to the areas considering solely the known age interval counts. The generated synthetic population will thus perfectly fit the observed population in terms of age. Next, the list with individuals will be adjusted to take another variable into account – e.g., gender, or income. This procedure continues until all variables are considered and an optimal composition of the synthetic population is achieved.

Synthetic Reconstruction using IPF

Tanton (2014) argues that while there are several methods used for synthetic reconstruction, the most widely accepted one is Iterative Proportional Fitting (IPF). It generates a synthetic population dataset by using census tables and iteratively estimating the probability distributions for each variable attribute in each area. The method should not be mistaken for the IPF used for deterministic reweighting because '*the reweighting IPF method starts with a record unit dataset, whereas the synthetic IPF method does not*' (Tanton 2014, p.8). A more detailed description of the IPF synthetic reconstruction method is offered by Birkin and Clarke (1988).

Not requiring a record unit dataset (also referred to as micro dataset) is exactly the advantage synthetic reconstruction offers compared to other spatial microsimulation methods. There can be instances when this kind of data is unavailable due to confidentiality restrictions or for other, unspecified, reasons. For example, studying indigenous disadvantage in Australia proved a challenging task because '*a record unit file was not available due to the scarcity of the Indigenous population in Australia, and concerns from the ABS* [Australian Bureau of Statistics] *about confidentialising the file*' (Vidyattama et al. 2013). The solution to this problem was to generate a synthetic unit record using a synthetic reconstruction method and then build a deterministic generalised regression spatial microsimulation model (Tanton 2014, p.8).

4.2.3. Reweighting

Whereas synthetic reconstruction methods do not rely on available micro data, reweighting methods generally make use of datasets containing individuals, or households[7]. These are often the product of representative surveys and can usually be obtained from local bureaus of statistics, research institutes, or some other data owners who have carried out a survey for a specific project-related purpose.

In essence, there are two types of approaches to reweighting a micro dataset: selecting individuals from the micro dataset to fill each area or adjusting the original weights on the micro dataset (ibid., p.9). The first type is referred to as *probabilistic*, and the second – as *deterministic* reweighting. Probabilistic algorithms produce different results each time they are run, because somewhere along the line they rely on random sampling. Basically, the algorithm runs until the optimal population composition is achieved so that it fits the observed aggregated counts defined by the constraint variables. This is also true for deterministic reweighting, but in the case of probabilistic algorithms, individuals are included in the simulated population, then taken out, then included again, and so on, until the algorithm stops when it achieves the perfect fit. Deterministic algorithms, on the other hand, are based on a predefined set of rules and thus provide the same set of individuals each time they are run (ibid., p.8).

The following paragraphs provide detailed information about the different types of reweighting algorithms and discuss their advantages and limitations.

Probabilistic Combinatorial Optimisation Method

Probabilistic combinatorial optimisation is the most common method among probabilistic reweighting algorithms. It takes at least two types of datasets as input data – a representative survey containing micro data, and small-area aggregates from a local bureau of statistics, or other, similar source.

The synthetic small-area population generated by probabilistic combinatorial optimisation is comprised by individuals, who are randomly selected from the sample survey. The selection process continues until the fit with the observed aggregates is optimal, that is, until it cannot be further improved. After each sampling step, the algorithm performs a quality check by calculating the total absolute difference between the estimated and the observed frequencies for each constraint category and for each area (Table 2). For the estimation of the goodness-of-

[7] Individuals and households are the most common 'units' in micro datasets. From here, for a more concise writing, I am going to omit households and refer to individuals only.

fit, other metrics can be used as well (e.g., relative error, root mean squared error, mean average percentage error, etc.).

Table 2. Example of assessing the fit of the synthetic micro data for zone XY (own representation)

Age intervals	Observed frequencies	Estimated frequencies	Absolute difference
18-44	18	20	2
45-64	12	10	2
65+	5	5	0
		Total Absolute Difference	4

To improve the fit, the algorithm starts swapping individuals. This does not necessarily mean that it substitutes one individual with another. Rather than that, the algorithm may choose to remove an individual completely or, on the contrary, to sample it more than once. Eventually, the total absolute difference must be as close to zero as possible. To that end, the algorithm may adopt a '*hill-climbing'* approach, when a swap is only accepted if it leads to improving the overall fit between the observed and the modelled frequencies. There are also other possible approaches to optimising the fit, which are more effective and thus more commonly used. Williamson (2013) argues that '*nearly all users of combinatorial optimisation prefer to adopt either a 'simulated annealing' or 'genetic' algorithm, in which swaps which adversely affect the fit might be accepted in order to avoid getting trapped with a suboptimal selection of households*' (p.25).

In any of these cases, however, the main limitation of probabilistic combinatorial optimisation consists in the computational intensity associated with the production of synthetic micro data. The '*computing overhead* […] *can run into CPU* [central processing unit] *days or weeks if whole country coverage is required*' (ibid., p.46). One of the reasons for this is the degree to which different combinations of households or individuals are tested in order to achieve a better fit (O'Donoghue et al. 2014, p.47).

Deterministic Iterative Proportional Fitting Method

Deterministic reweighting methods also rely on at least two types of datasets as input – a geographic dataset and a micro dataset. There are generally two main approaches to deterministic reweighting – IPF and generalised regression. The next few paragraphs will cover the application of IPF, '*the most widely used and mature deterministic method to allocate individuals to zones*' (Lovelace and Dumont 2016, p.70).

In essence, IPF operates in the following way: first, a weight matrix with the dimension of '`I x Z`' is created, where `I` refers to the number of individuals in the micro dataset and `Z` to the number of geographic zones. Each weight is initially assigned to 1. Then, the IPF algorithm starts running, proceeding zone by zone, updating the initial weight of *representativity* of each individual (row) for each given zone (column), based on the predefined constraint variables (Figure 5). The algorithm updates the matrix iteratively, for each constraint variable category. For instance, it first updates the matrix to account for age and thus adjusts the weight of each individual for zone 1, then for zone 2, etc. This continues until the weights of all individuals are updated for all zones to fit the population distribution in the given zones regarding age. The process is then repeated as many times as the number or remaining constraint categories.

Figure 5. Iterative updating process of the weight matrix (Source: Lovelace and Dumont 2016, p.74)

```
##          [,1] [,2] [,3]
## [1,] 1.333333    1    1
## [2,] 1.333333    1    1
## [3,] 4.000000    1    1
## [4,] 1.333333    1    1
## [5,] 4.000000    1    1
##          [,1]     [,2] [,3]
## [1,] 1.333333 2.666667    1
## [2,] 1.333333 2.666667    1
## [3,] 4.000000 1.000000    1
## [4,] 1.333333 2.666667    1
## [5,] 4.000000 1.000000    1
##          [,1]     [,2]     [,3]
## [1,] 1.333333 2.666667 1.333333
## [2,] 1.333333 2.666667 1.333333
## [3,] 4.000000 1.000000 3.500000
## [4,] 1.333333 2.666667 1.333333
## [5,] 4.000000 1.000000 3.500000
```

These can relate to different types of gender, income groups, levels of education, etc. The adjustment of the weights is carried out by multiplying each weight by a specific coefficient. In the case of age, for example, this coefficient equals the division of the total number of observed individuals belonging to a certain age category (e.g., 18-44 years) by the equivalent cell aggregated version of the micro data, that is, the sum of all people in the micro dataset belonging to the same age category. Usually, the coefficient will be different for each age category, or in other words, for each characteristic attribute of the given constraint. This iterative updating of the weights is why the method is referred to as *iterative proportional fitting*.

As opposed to probabilistic combinatorial optimisation, deterministic reweighting with IPF delivers the same results every time. This, combined with its robustness, reliability, speed, and simplicity are the major advantages of the method. The main flaw of this approach is that the initial form of the generated micro data is that of fractional weights. To carry out further analyses, researchers must therefore apply a method of integerisation so that they can deal with *whole* individuals rather than abstract fractions (Lovelace et al. 2014, p.287). The combined use of the processes *integerisation* and *expansion*, which will be addressed in further detail in Chapter 6, allows for converting the generated weight matrix into a population dataset – the same output format produced directly by probabilistic combinatorial optimisation (Lovelace and Dumont 2016, p.67).

Deterministic Generalised Regression Method

Deterministic reweighting using generalised regression is similar to the previously introduced IPF method as it also adjusts the weights of individuals from a micro dataset based on available small area constraints. Nevertheless, the procedure used for computing the weights is different. The first step necessary for applying the generalised regression method is to take the weights available in the micro dataset and, for each area, to divide them by the respective population counts. Thus, a '*reasonable starting weight required for the generalised regression procedure*' is provided (Tanton 2014, p.13).

The new set of weights is calculated using a regression model based on the constraints available for the geographic areas. Since the '*weights are limited to being positive weights only,* […] *the procedure may iterate a number of times if positive weights aren't achieved for every record in the first run*' (ibid.). Each of the initial weights from the survey is continually adjusted and the procedure stops only if either reasonable results are achieved, or a predefined maximum number of iterations is reached.

A key advantage of generalised regression methods is that '*projections are very easy to create, either by inflating the weights; or inflating the benchmarks* [another popular term for the constraint variables] *and reweighting to new benchmarks*' (ibid.). One potential limitation of the

method can be a slight decreasing of the model accuracy when adding more constraint variables. This can go hand in hand with an increasing number of areas failing a certain accuracy criterion (Tanton and Vidyattama 2010).

Each of the introduced spatial microsimulation methods has certain advantages over the others. The choice of one specific approach depends mostly on the available data sources, the computing power at hand, the desired level of complexity, and the preferred format of the generated data. In my case, micro data was available, which is why I did not have to use synthetic reconstruction – a more complex, and time-consuming approach than the reweighting methods. Choosing between probabilistic and deterministic reweighting methods, I opted for the latter because I preferred not having a random component in the reweighting algorithm. Thus, the generated dataset remains consistent no matter how many times the algorithm is run. Finally, I favoured the IPF approach over the generalised regression model for adjusting the weights. I considered the slight loss of accuracy caused by the integerisation of the generated fractional weights to be the more acceptable alternative than being limited in adding more constraint variables, which is a possible downside of using a generalised regression model.

All steps necessary for implementing the IPF approach to generating a synthetic population for the purposes of this dissertation are described in detail in Chapter 6. ‘Modelling Health-Related Data in Hamburg’s Neighbourhoods’. Before diving deep into the methodology, the next chapter is going to introduce the perspective of several public health researchers, based in Germany, on modelling health data at the urban neighbourhood level. The importance of small-scale health data, the reliability of health models, and their potential for identifying hotspots of vulnerable population groups, also in terms of the ongoing COVID-19 pandemic, will be the main topics brought to attention in the following pages.

5. MODELLING HEALTH DATA ON A SMALL URBAN SCALE FROM THE PERSPECTIVE OF PUBLIC HEALTH RESEARCHERS

Hamburg has already gathered considerable experience in monitoring social deprivation and health as it became clear in Chapter 3.3. 'Monitoring Social Inequality and Health in Hamburg'. Nevertheless, while social status is tracked annually at the scale of the statistical areas, exploring health-related patterns at the urban neighbourhood level lags behind because of unavailable data. Against this background, I carried out several interviews with researchers in the field of public health in Germany, to find out whether they consider modelling health data an approach worth adopting.

Prof Dr Heiko Becher, Director of the Institute for Medical Biometry and Epidemiology at the University Medical Centre Hamburg-Eppendorf (UKE Hamburg) stated that the existence of considerable disparities across Hamburg's city quarters in terms of social structure is well known. According to him, it is therefore not new information that there is a disproportionate health burden carried by certain population groups resulting from social inequality. Nevertheless, Becher considered it a meaningful endeavour to use the already available data and adopt a methodological approach allowing to reveal existing disparities in more detail. Thus, the specific factors contributing to the aggravation of social inequality can be outlined. Since the latter is central to the Health Reporting System, Becher emphasized the importance to capture data in the most detailed way possible. Still, he expressed uncertainty whether a small-scale health monitoring would, in fact, contribute to reducing existing inequalities.

Dr rer. biol. hum. habil. Enno Swart from the Institute of Social Medicine and Health Systems Research (ISMHSR) at Otto-von-Guericke University Magdeburg, generally expressed his support about the further exploration of the suggested spatial microsimulation approach. He pointed out that one can surely assign certain socio-economic characteristics to Hamburg's city quarters and then look for correlations with the prevalence of certain diseases available at this spatial scale. Nevertheless, he emphasised the risk of ecological fallacy because it is not clear and it cannot be explained how such a correlation will manifest in an isolated case, e.g., at the individual level. Therefore, he deems the spatial scale of the city quarters suitable for hypothesis development, and maybe for detecting some abnormalities but not for analysing any cause-effect relationships:

'In the realm of spatial statistics, or in the social epidemiology realm, one is aware that it is especially the individual factors that influence health: social status, health behaviour, individual risk factors, and at the same time factors of the environment, such as availability and accessibility of health services, public green spaces, noise, etc. By observing the scale of the city quarters, all these factors are measured by the same yardstick. This can be avoided only if one is able to obtain and use individual data and integrate it into multivariate and hierarchical models to establish how individual factors affecting health interact with small-scale environmental determinants.' (own translation from German, approved by the interviewee)[8]

Swart stated that it would be ideal if there was individual data at the level of the statistical areas, but for the time being this is not the case: *'This gap can possibly be filled by using health*

[8] Hereinafter applicable to all interview statements in *italics* (see Appendix for the approval confirmations)

insurance data but at the scale of the statistical areas one is quickly faced with data protection problems because the number of cases is too small'.

This, in his opinion, is a problem, which the suggested modelling approach may be able to address. He thus advocated testing it out using the already available data for Hamburg. Swart suggested carrying out external validation of the model with health insurance data obtained in the course of the research project 'Healthy Neighbourhoods'[9]:

'Provided that the modelled data proves to be reliable we can maybe even spare ourselves future efforts to obtain micro data. Currently, obtaining data from health insurance funds is a complex and challenging process, which usually takes a lot of time.'

Swart believes it is possible to establish a common health research data centre in the not so far future and thus commit health insurance funds to deliver data for research purposes on a regular basis. Step by step, the data pool can be expanded. However, it is still unclear how fine-grained the data from health insurance funds is across regions:

'*At the moment, the thinking is rather in the direction of delivering data at the federal state or at the Landkreise (engl.: rural district, county) level. Therefore, the progress would not be as big as it would be if there is health data aggregated at the level of the statistical areas, but it is generally possible to ask health insurance funds to deliver their data at the zip code level, for instance. Provided that the data is temporally rather than just spatially aggregated, that is, morbidity numbers are averaged over a period of three years, for instance, data protection issues could be overcome*.'

According to Swart, the suggested spatial microsimulation model can prove to be a suitable alternative to obtaining this kind of *real* individual health data, because it would still take time until the health insurance funds are convinced to deliver data at a smaller spatial scale:

'Provided that the public health impact of such models or small-scale visualisations becomes evident, health insurance funds as well as other data providers may be motivated to deliver data to health research data centres because they would see that this is not only additional effort for them, but they may, indeed, win something out of it too'.

Swart sees the potential use of small-scale health data mainly for the purposes of structural prevention:

'Generally, we know that a large proportion of cancer or diabetes cases are behaviourally induced, that is, there are individual behaviourally induced causes – poor diet, or unbalanced diet, too little physical activity, smoking, excessive alcohol consumption. It is therefore clear what can be done to counteract these factors at the individual level – raising people's awareness, providing them with guidance, offering individual consultations about weight reduction or quitting smoking. Nonetheless, there are certain limits as some population groups cannot be as easily reached, although they may be the ones who especially need such services. Regardless of the reason – be it education, or a language barrier, when dealing with migrants, for

[9] Research project cooperation between the University of Applied Sciences (HAW Hamburg), University Medical Center Hamburg-Eppendorf (UKE Hamburg), HafenCity University (HCU Hamburg) and Otto-von-Guericke-University Magdeburg (2017-2021). As a research associate at the HCU Hamburg, I participated in the project from its start in July 2017. More information can be found on the project's website: http://www.gesundequartiere.de/

instance, if we turn our attention from the individual constellation to the living environment, and ensure the living environment offers sufficient supply of public green space so that people can engage in sporting activity, where they feel safe to walk and bike, then we may be able to reach population groups, which are generally more difficult to reach with classic behavioural prevention measures. Thus, it may be possible to change something. If we think about noise – there are people with small income, who cannot afford renting the nice apartment in the more peaceful residential neighbourhoods, but are, instead, forced to rent a more affordable flat on a busy arterial road. In such cases, noise abatement measures may lead to improving people's health without them having to change their individual behaviour. In such instances, the small-scale perspective can be valuable. One should look at what is really happening – if, based on a socio-spatial monitoring at the small scale, there are, indeed, disease-related hotspots. Let us assume that we know there is a cardiovascular disease hotspot, and we know that cardiovascular disease is associated with noise, therefore we see a highly exposed population in area XY, and we are doing something against it. If this really brings something, that is unknown, but it is at least worth trying because there are plenty of opportunities to address existing issues, and structural prevention is overall believed to be quite promising.'

In terms of utilising a small-scale health model for a more efficient distribution of structural prevention measures and thus possibly saving available resources, Swart is rather sceptical:

'There are different actors depending on who is responsible for behavioural and structural prevention. Behavioural prevention generally shall be carried out by health insurance funds or by general practitioners, in other words, actors in the realm of public health. Structural prevention, on the other hand, can be addressed by actors promoting various policies – educational systems – e.g., schools offering healthy meals, or enough facilities for physical training; transport planning measures aimed at providing secure and bike-friendly ways to school; urban planning in terms of noise pollution, etc. Still, the problem with structural prevention is that results are not usually visible in the short-term. This, of course, is a problem, especially in terms of politics, because politicians, who are thinking from one election to the next, cannot score high enough with such policies as the positive results rarely manifest during their term. Therefore, saving in the realm of prevention is always difficult – it is necessary to invest first in order to spare resources in other areas in the mid- to long-term.'

From his position of *'an observer rather than an expert'* in terms of the COVID-19 pandemic, Swart considers the spatial scale of decision-making being bigger than that of the statistical areas:

'There may be a difference in the risk of infections, if looking south or north of the Elbe River, or if looking at the level of the city quarters, but I do not think one has to go so much further down for COVID-19 as for other things. I would say that for the usual chronic illnesses, which, for the most part, are behaviourally and environmentally induced, it can be interesting to look at the level of the statistical areas, that is, at the immediate residential environment, but for COVID-19, I would intuitively say that the scale of the city quarters must be sufficient.'

Prof Dr Susanne Busch from the Department of Nursing and Management at the University of Applied Sciences (HAW Hamburg) considers modelled data a suitable alternative to *real* health data at the small scale, but only as an intermediate step rather than a permanent solution:

'I think it is brilliant to model this data, but one has to look at the results with humility. I find it absolutely important to try it out and then look at the whole thing with expert knowledge, ideally within expert panels with experts of various backgrounds, even if it leads to tearing it all apart, even if it means discussing it very critically. And if something of the initial big thinking is left at the end, that would be terrific progress, absolutely helpful. Each advancement of knowledge that we can have, albeit small, is relevant and precisely the notion that "we should take social inequality into account in the realm of expanded social policies" obliges us to use every small additional information we may be able to get. In any case, I would try to present this approach as what it really is – an attempt to achieve a better small-scale management of health-promoting and healthcare services, to provide argumentation depending on certain factors, to confess that there is a multitude of limitations, and yet the approach still offers a certain knowledge gain. I would, indeed, highlight this knowledge gain, albeit in a humble manner, but underline that the knowledge gain is there and that we do need such gains because otherwise we would not attain socially equal, or better living conditions. And we need them even more so now, because COVID-19 is going to create an even wider social gap than before.'

Against this background, Busch considers modelling individual health-related data more meaningful at the level of the statistical areas than the city quarters because the latter are too heterogeneous. In this context, she suggests that there is a necessity to review the statistical areas and the current social indicators from the Social Monitoring. Like Swart, she recommends using data from health insurance funds for the external validation of the model:

'For me, that would be the most elegant solution – to validate the model from the researcher perspective, in an exemplary manner, complying with the existing data protection guidelines.'

In terms of using the data obtained from three health insurance funds in Hamburg for the purposes of the research project 'Healthy Neighbourhoods', Busch pointed out that there is a certain data bias one should be aware of:

'We tried to validate the representativity of the data we obtained from the health insurance funds and we most probably have an imbalance, which, however, cannot be identified directly from studying the data. The thing is that the different health insurance funds have a certain population composition, which is not always diverse, and in that sense, it is not necessarily representative of Hamburg's entire population. This is a certain limitation of this kind of data that we constantly have to deal with, and I would simply address this issue as a limitation, which has to be taken as a given.'

PD Dr rer. nat. Jobst Augustin from the Centre for Psychosocial Medicine, and Institute of Health Care Research in Dermatology and Nursing (IVDP[10]) at UKE Hamburg, generally encouraged testing out the proposed modelling approach:

'I have some experience with data disaggregation, and I know this is a complex matter. It goes hand in hand with a certain loss of accuracy, which is the main difficulty. Still, I deem it interesting and necessary to try it out.'

[10] German: Institut für Versorgungsforschung in der Dermatologie und bei Pflegeberufen

Having worked with different types of health data, both spatially aggregated, and individual data from sample surveys, Augustin has experienced their advantages and limitations:

'Each dataset has strengths and weaknesses, or characteristics, regardless of whether it contains individual or aggregated data, for instance. The use depends strongly on the research question. If you are using individual data, you have a significantly smaller sample, which is not the case when working with aggregated data. However, you can address a whole other range of research questions using individual data instead of aggregated data. Therefore, one cannot generally say which type of data is better, that is simply very much dependent on the specific context.'

While Augustin generally advocated the use of aggregated health data at the city quarters level for answering various research questions, he named some instances, where this scale of aggregation would not be sufficient:

'The city quarters scale is most probably sufficient for visualising the spatial distribution of the prevalence of certain chronic diseases in Hamburg. If, however, the aim was to check for a correlation between asthma, or chronic obstructive pulmonary disease (COPD), and air pollution, for instance, the city quarters level would not suffice. Air pollutants are highly dynamic, both in spatial and in temporal terms, and the information available at the city quarters level is simply not detailed enough to address all possible factors affecting their fluctuations. It is, of course, possible to take the air pollution constellation at this scale, or at a 1x1-kilometre scale, but this would lead to unreliable results in this context. There are approaches making this possible, but they would raise a fair deal of justified criticism. To test for correlations between disease prevalence and social status, however, you would have to go down one spatial scale below as the city quarters are quite heterogeneous.'

In the context of the proposed spatial microsimulation model as a mean to generate such small-scale disease prevalence data, Augustin emphasized the validation of the model being the main necessity:

'Modelled data can certainly be used as starting point for further in-depth analyses. However, statements, such as the model being able to represent the population's health status, should be avoided. Instead, the generated data must be critically observed and used only for generating hypotheses. In my opinion, it is not suitable for more than that. While the necessity to test modelling approaches is there, it is imperative to validate the generated data before making statements about its quality. Testing for correlations between the modelled disease prevalence and different spatial factors, such as noise pollution, is of course, interesting, that, for instance, is something everyone wants to know. Nevertheless, one should be careful how good the data is for carrying out such tests. The noise data available for Hamburg is modelled, another model is generating the disease data, and if the two models are not good enough, the validity of the results will be lost.'

Augustin also pointed out the existing bias in the data from health insurance funds. He thus considered it greatly beneficial that there is available data for external validation of the proposed model from several health insurance funds:

'There is a health insurance bias as certain health insurance funds tend to insure certain population groups. The AOK has another population composition than the TK, for instance. Therefore, it is a good thing you are going to use data from three different health insurance funds to validate your model.'

Augustin was rather reserved about using modelled small-scale disease data to identify clusters of vulnerable population groups within the city in the context of the COVID-19 pandemic:

'It is more or less evident where vulnerable population groups live based on various, already available statistics. Local resident registration offices have information about how old everyone is, by address. As for vulnerability in terms of comorbidities, I believe the public health authorities are already aware of spatial distributions, based on data from the Morbidity Atlas, etc. Therefore, if you can validate your modelled data, this may be a gain, but it is already evident where the problematic hotspots are, and where special attention must be paid. This is why I do not think there is necessarily a demand for this, then again, I am not directly involved in the public health system, I may be wrong.'

In summary, all the interviewed public health researchers agreed that testing out the proposed spatial microsimulation approach to generate small-scale health data is a meaningful endeavour. One of the main reasons, emphasised by everyone, was the heterogeneity of Hamburg's city quarters. The latter hinders the identification of factors of the living environment, which may be influencing the health of inhabitants at the smaller, neighbourhood scale. This, in turn, impedes developing the full potential of structural prevention.

All interviewees expressed the need to observe critically the generated results. External validation of the model, e.g., with health insurance data, was highly recommended. At the same time, the bias of the latter was brought to attention as an existing limitation to be considered. Still, further exploration of the suggested modelling approach was encouraged from all.

The next chapter is going to introduce the selected spatial microsimulation approach and present the necessary steps for setting up a small-scale health model for the city of Hamburg. Data selection, choice of constraint and target variables, writing a population synthesis algorithm, and eventually compiling a synthetic population dataset are the main topics that are going to be covered in detail. The choice of software is also going to be addressed.

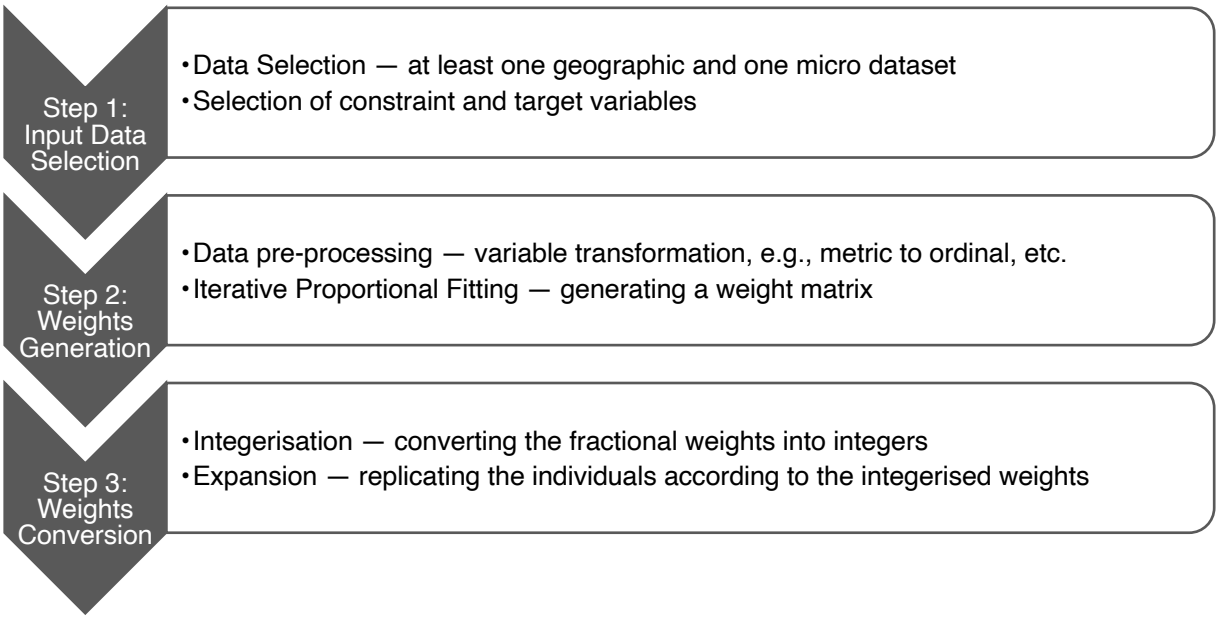
6. MODELLING HEALTH-RELATED DATA IN HAMBURG'S NEIGHBOURHOODS

6.1. Two-tier Modelling Strategy

To create a synthetic population with health-related attributes for each of Hamburg's neighbourhoods, I adopted a two-tier modelling strategy, which integrates data sources available at the level of the city quarters and the statistical areas. Aggregated health data is available only at the scale of the city quarters and the city quarter clusters. Going down to the level of the statistical areas, health data no longer exists due to data protection regulations. With this in view, I decided to divide the task of setting up a small-scale health model into two stages. I began with constraining the synthetic population at the first tier of the city quarters using the available aggregated health data. Thus, the counts of the individuals modelled at the city quarter level are guaranteed to match the observed counts from the available datasets regarding specific medical conditions. Knowing the exact count of people suffering from diabetes, for instance, will ensure that there are exactly as many diabetics – no more, and no less – in the synthetic population generated for each city quarter (cluster). The same goes for all other illnesses available in aggregated form, including hypertension, heart failure, cancer, and depression. The available socio-demographic data referring to age and gender is integrated in the constraining process as well. Thus, the first modelling tier enables generating reliable individual counts regarding different kinds of combinations at the scale of the city quarters and city quarter clusters, such as number of males older than 65 years and suffering from diabetes and hypertension, or number of females aged between 18 and 45 years without any comorbidities, and so on. Starting from there, the synthetic population which is already constrained according to health data available at the scale of the city quarters, is constrained again – this time based on aggregated socio-demographic data available for the underlying statistical areas.

With the general reasoning behind the choice of a two-tier modelling strategy explained, the next paragraphs are going to address all necessary steps for generating a synthetic population with health-related attributes. These steps are illustrated in their respective order in Figure 6.

Figure 6. Steps for generating synthetic population (own representation)



6.2. Data Selection

Four different datasets were used for the two-tier population synthesis. To constrain the synthetic population, I used a geographic dataset containing aggregated socio-demographic data available at the level of the statistical areas[11], two further geographic datasets containing aggregated health-related data available at the level of the city quarters and city quarter clusters, and one micro dataset with individual health data from a national representative health survey.

6.2.1. The socio-demographic geographic dataset

The Bureau of Statistics for Hamburg and Schleswig-Holstein provides data about various socio-demographic characteristics, aggregated at different levels of spatial division including the city quarters and the statistical areas. The list of available variables is long and includes, among other things, number of males and females, number of people belonging to different age groups, number of employed people, number of recipients of unemployment benefits, number of single households, etc. I used socio-demographic data aggregated at the level of the statistical areas to constrain the synthetic population to the observed population counts regarding age, gender, employment, and living situation. Since I started working at this dissertation in 2018, I used the most current dataset at the time, which was referenced to the date 31.12.2017 (Statistisches Amt für Hamburg und Schleswig-Holstein 2018).

6.2.2. The health-related geographic datasets

Next to the available socio-demographic data, I used two sources of aggregated health data at the level of the city quarter (clusters) to constrain the modelled population at the first tier. One of them was the Morbidity Atlas published in 2013, which, among other things, contains data about the prevalence of hypertension, diabetes, heart failure, and depression divided into age and gender categories (Erhart et al. 2013).

The other source was Hamburg's Cancer Registry, which provided me with aggregate counts of the individuals who were diagnosed with any type of cancer between 2008 and 2018 at the city quarter clusters level (Hamburg Cancer Registry 2020).

6.2.3. The micro dataset

The micro dataset, which contains individual health data, originated from the national representative health survey '*Gesundheit in Deutschland aktuell*' (engl.: '*Current health situation in Germany*') carried out by Robert Koch-Institute (RKI) between March 2012 and March 2013. For brevity, it is often referred to as '*GEDA 2012*'. It contains data about over 200 health-related variables for approximately 26.000 people living in Germany. All respondents, who participated in the survey, were older than 18 years and were interviewed over the phone. The questionnaire encompasses the following topics:

- Subjectively perceived health,
- Chronic (and other) illnesses,
- Accidents and injuries,
- Psychological health,

[11] The population counts of the city quarters equal the cumulated population counts of the statistical areas within their boundaries, which is why I only needed data aggregated at the smaller scale of the statistical areas.

- Illness consequences and disability,
- Health behaviour and prevention measures such as vaccination, diet, and physical activity,
- Health-related risk factors, such as alcohol consumption and smoking,
- Use of healthcare services,
- Health-related support and burdens,
- Socio-demographic characteristics such as education, occupational status, and migration background. (Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

6.3. Selection of Constraint and Target Variables

Having selected the input datasets, the next step towards generating the synthetic population was to define the so-called *constraint* and *target variables*.

In essence, the purpose of the *constraint variables* is to *constrain* the model. With this in view, these are variables shared between the geographic datasets and the micro dataset. Some typical examples of constraint variables when modelling individuals include age, gender, income, and education. Ultimately, they serve to determine how representative each individual is for each spatial unit. Assuming that in neighbourhood X*,* there are more males than females, more adults in working age than elderly or infants and more high-school graduates than university graduates or people without degree, an adult man with a high-school degree would be highly representative for this neighbourhood. The population synthesis algorithm is therefore going to replicate such individuals in this particular neighbourhood more times than any other individual who has a different set of characteristics.

Target variables, on the other hand, are the variables of interest, which are not available in the geographic dataset. Simply put, the target variables are the reason for generating a synthetic population in the first place. For the purposes of this dissertation, the targets[12] will be variables regarding different types of chronic disease, health behaviour, and the like.

Coming back to the constraint variables, they generally fit into two categories: *optimising* and *elective constraints*. To include a certain variable as optimising constraint, it must exhibit a statistically significant relationship with one or more of the targets. Furthermore, it should *explain* the variance of the targets at least to some extent because the generated synthetic population will only be as good as the underlying associations. Elective constraints, on the other hand, are useful for examining the studied targets for specific population groups. Gender and age, for instance, can serve for defining such groups. Edwards and Clarke (2013) argue that while such variables may not necessarily explain the targets, it still makes sense to include them in the model as elective constraints.

After examining the micro dataset and identifying possible targets, I singled out the variables shared between the micro dataset and the geographic datasets. At the same time, I ensured that if a variable transformation is necessary for the characteristic attributes to match, the way the variables are coded is going to allow it.

[12] For the sake of brevity, target and constraint variables will sometimes be referred to as *targets* and *constraints*.

While good constraint data is necessary to generate a synthetic population well aligned to the observed population counts, there are no specific recommendations how many constraint variables to use. In fact, Lovelace and Dumont (2016) argue that the number of characteristic attributes is more important. Fewer constraints with more than just two categories can constrain a model better than a larger number of dichotomous constraints. However, this decision is often not at all there to make because of scarce data. In my case, the only available non-dichotomous constraint variable was age[13].

Before deciding which variables to model as targets, I checked to what extent the initially identified constraints shared between the micro dataset and the geographic datasets manage to explain their variance. If the constraints have little prediction power, the synthetic population, together with its health-related attributes, is not going to provide a reliable representation of the actual health situation.

Prior to carrying out the required statistical tests, I had to decide whether to filter the micro data based on the population size of the respondents' place of residence to get a more representative sample for a big city like Hamburg. Since the number of survey participants suffering from different types of chronic illnesses was relatively small, I opted for using the entire sample. I thus ensured that there are enough cases to generate valid results.

Next, I applied a predefined weight variable available in the micro dataset, as recommended in the GEDA 2012 documentation:

'Regression analyses should be carried out with weighted cases in order to take into account the sample design and response behaviour. […] Unweighted analyses can lead to clearer statements, i.e., higher significance than weighted regression analyses. […] Nevertheless, the results of an unweighted analysis cannot be considered representative for Germany' (own translation from German. Robert Koch-Institute 2015, p.12).

The main factors for adjusting the weights are sample design and the related selection probability of respondents. Respondents with lower selection probability represent more people in the total population than respondents with higher selection probability. People's willingness to participate in the survey is another factor influencing the weights. Participation willingness is generally measured as the proportion of a certain population group in the survey sample divided by the respective proportion in the total observed population.

A note on the statistical tests

For the final selection of constraints and targets, I carried out several logistic regression tests. Thus, I was able to establish to what extent the constraint variables can predict the targets. The decisive factor for the selection of a suitable statistical test was the type of the target and constraint variables – continuous or categorical. In the micro dataset, most of the variables regarding different medical conditions are dichotomous, as they provide information whether the respondent has (had) any of them (or not). The constraints in the micro dataset are also either dichotomous (gender, being currently employed, living alone) or categorical (age, divided into several age intervals). I therefore chose to rely on binomial logistic regression.

---

[13] The entire compilation of constraint variables is presented in Table 22.

To interpret the results provided by a binomial logistic regression, several metrics must be considered. *Pearson's Chi-Square* coefficient estimates the goodness-of-fit of the regression model. With Sig. = 0,000 it can be concluded that the model is good enough for testing the relationship between the dependent and the independent variables. *Pseudo R^2* shows what share of the variance of the dependent (i.e., target) variable is explained by the independent variable(s) (the potential constraints). The statistical software I used – IBM's SPSS, provides three alternative Pseudo R^2 values: *Cox & Snell*, *McFadden*, and *Nagelkerke*.

In general, Cox & Snell and McFadden compare the ways in which the modelled values of the dependent variable are distributed with and without taking the explanatory variable(s) into account. The values generated for Pseudo R^2 using these approaches are between 0 and 1 like classic R^2 values. However, neither Cox & Snell nor McFadden values could ever reach 1, regardless of how good the regression model is. Nagelkerke, on the other hand, corrects the Cox & Snell values so that the estimated Pseudo R^2 value could theoretically reach 1. Therefore, Nagelkerke values are typically higher than Cox & Snell and McFadden (Brosius 2011, pp.616–617). With this in view, I decided to consider only the corrected version of Pseudo R^2 – Nagelkerke – for the purposes of this dissertation.

The following sections are going to provide brief theoretical information about each of the variables identified as potential targets within the micro dataset. Additionally, the results from the conducted statistical tests are going to be presented. These will serve as basis for deciding whether to include the examined variables as targets in the spatial microsimulation model.

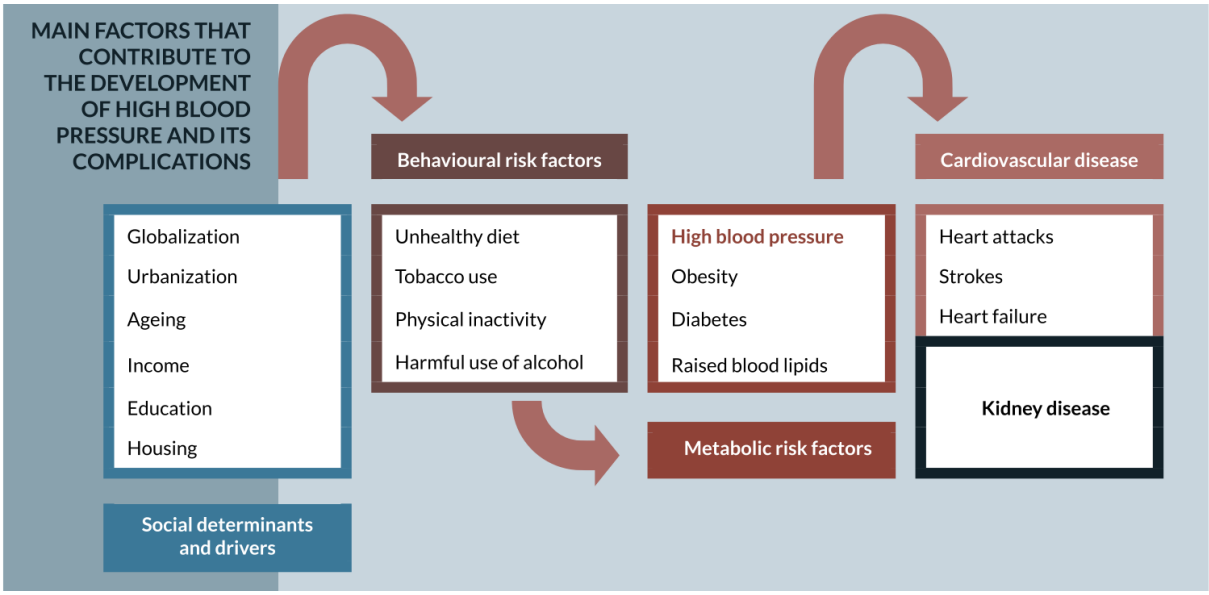
6.3.1. Hypertension

Every heartbeat pumps blood into the blood vessels for the blood to reach all parts of the body. The so-called blood pressure arises from blood pushing against the walls of the arteries when being pumped from the heart. High blood pressure, referred to as hypertension, is a condition characterised by constantly raised pressure in the blood vessels. As a result, the heart must continuously work harder to pump the blood. Since hypertension rarely causes symptoms, it can remain unnoticed for a long period of time (World Health Organization 2013, p.20). Nonetheless, it causes significant repercussions for human health and is thus referred to as a 'silent killer'. Left uncontrolled, that is, without changes in the lifestyle and/or medication intake, '*hypertension can lead to a heart attack, an enlargement of the heart and eventually heart failure*' (ibid., p.17). Another possibly fatal consequence of hypertension can be stroke caused by blood leakage in the brain. This can happen because of the development of weak spots in the blood vessels due to the continuous high pressure. Kidney failure, blindness, rupture of blood vessels, and cognitive impairment can also manifest as results of untreated hypertension (ibid.). Against this background, hypertension is considered a key determinant for the most common causes of death in adults (Robert Koch-Institute 2014, p.123). Globally, a total of 9 million deaths are attributed to it each year (World Health Organization 2013, p.5).

According to recent statistics, 1.13 billion people suffer from high blood pressure: approximately one in four men and one in five women (World Health Organization 2019). In a survey conducted by RKI, nearly one in three adults living in Germany reported physician-diagnosed hypertension (Neuhauser et al. 2017, p.53).

Hypertension is most often triggered by a combination of factors, such as genetic predisposition, age, gender, nutrition, elevated salt consumption, alcohol consumption, lack of physical activity, stress, etc. (Robert Koch-Institute 2014, p.123). Factors of the living environment such as '*higher levels of noise and air pollution have been related to higher blood pressure levels*' as well (Schulz et al. 2018, p.8). The World Health Organization (2013) distinguishes social determinants and drivers such as urbanisation, ageing, and income to have an impact on several behavioural risk factors, which foster the development of hypertension, obesity, diabetes, and other, metabolic risk factors (Figure 7). Depending on the individual case, either kidney disease or cardiovascular disease (such as heart attacks, strokes, and heart failure) may develop over the course of time.

Figure 7. Main factors contributing to hypertension and its complications (Source: World Health Organization 2013, p.18)



Nevertheless, early diagnosis of hypertension can limit the damage on heart and blood vessels and thus lower the incidence of more serious chronic illnesses if lifestyle changes occur: '*Doing so is far less costly and far safer for patients, than interventions like cardiac bypass surgery and dialysis that may be needed when hypertension is missed and goes untreated*' (World Health Organization 2013, p.5).

Empirical evidence points to a strong connection between hypertension and age – while the 12-month prevalence of physician-diagnosed high blood pressure among 18- to 44-year-olds is only 5% for females and 10% for males, is rises to over 30% after the age of 45. In comparison, more than half of both females and males older than 65 years, report to have been diagnosed with hypertension (Robert Koch-Institute 2014, p.123).

Within the scope of the study GEDA 2012, currently having hypertension is defined by the affirmative answer to the question '*Have you ever been diagnosed with hypertension by a physician?*' and any other of the further two questions '*Did you have hypertension in the past 12 months as well?*' and '*Are you currently on medication for treating hypertension?*'. No actual measurements of the respondents' blood pressure levels were carried out during the survey (ibid.).

Against this background, there are some limitations of the available micro data to be considered. First, only clinically diagnosed hypertension is taken into account. Since early stages of this condition usually do not cause symptoms, many cases may have remained unreported. Therefore, it is possible that the actual prevalence of hypertension is underrepresented in this dataset. Furthermore, there is a common trend in Germany, and many other countries as well, that more cases of hypertension in men remain unreported than in women (Robert Koch-Institute 2014, p.123).

Keeping these considerations in mind, I proceeded with carrying out statistical tests to ascertain the effects of the available constraint variables on the likelihood that the individuals in the micro dataset have hypertension. If the constraints manage to explain (at least some part of) the variance of hypertension, it can be included as target variable and its prevalence can thus be modelled at the level of the statistical areas.

All statistical tests were carried out in SPSS. First, I created a new variable to account only for current hypertension cases, that is, people who replied affirmatively to having had high blood pressure over the past 12 months and/or being on medication for its treatment. My goal was to test to what extent are the constraint variables at hand – age, gender, employment, and living situation – able to explain the variance of hypertension. All of these constraints, except for age, are dichotomous: gender is coded in 'males' and 'females'; employment in 'yes' and 'no'; and living situation in 'single household' and 'non-single household'. Age is coded into thirteen 5-year-intervals, ranging from 18-24 years to 80+ years. I thus proceeded with a binomial logistic regression.

Without considering the independent variables, the logistic regression model was able to correctly classify 73,6% of the cases for the newly defined variable 'currently having hypertension' by assigning all cases to the category 'no'. With nine standardised residuals (> ±2,5 Std Dev), the model was statistically significant, $\chi^2(4) = 3946{,}372$, $p < 0.0005$ and explained 27,1% (Nagelkerke R^2) of the variance in hypertension. After considering the four independent variables, it managed to classify 75,6% of the cases correctly. Sensitivity[14] was 39,0%, specificity[15] was 88,7%, positive predictive value[16] was 55,4% and negative predictive value[17] was 80,2%. All predictor variables except for living situation were statistically significant (Table 3).

Table 3. Logistic regression predicting the likelihood of hypertension (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,312	,007	1755,082	1	,000	1,366
	Gender	-,210	,038	30,988	1	,000	,811
	Employment status	-,153	,046	11,043	1	,001	,858
	Living situation	,036	,045	,638	1	,425	1,037
	Constant	-2,964	,097	930,293	1	,000	,052

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

[14] Share of correctly assigned 'yes'-cases out of all cases.

[15] Share correctly assigned 'no'-cases out of all cases.

[16] Share of correctly assigned 'yes'-cases out of all predicted 'yes'-cases.

[17] Share of correctly assigned 'no'-cases out of all predicted 'no'-cases.

Since the results suggested that age has the highest predictive power, I carried out the test again, this time only using age as independent variable. It turned out that it accounts for explaining 26,8% (Nagelkerke R^2) of the variance and classifying correctly all the 75,6% of the cases. Against this background, the other independent variables barely play a role for predicting the likelihood of hypertension. Nevertheless, they can be included as elective constraints.

The results showed that people aged over 80 years are 39 times more likely to suffer from hypertension than people under 25 years. For those older than 65 years the odds of having high blood pressure are 5,5 times higher than for younger adults. This is a clear statement. I controlled for obesity and overweight defined by a Body Mass Index (BMI) $\geq$ 30 and $\geq$ 25, respectively. BMI was measured using self-reported weight in kilograms and height in centimetres. I found that overweight under 25-year-olds have 26 times smaller odds of having hypertension compared to overweight people older than 80 years. For people, who are not overweight, the odds are 41 times smaller. Obese people younger than 25 years are 13 times less likely to suffer from hypertension compared to obese over 80-year-olds. In contrast, for non-obese individuals, the same likelihood is 46 times smaller. If we take 65 years as age delimiter instead, the odds of having hypertension for adults younger than 65 are 4 times smaller when being overweight and 10 times smaller if this is not the case. If obese, these odds are 3 times smaller, compared to 7 times smaller for non-obese individuals.

With this in view, both overweight and obesity suggest an increase in the chances of having high blood pressure for younger adults. Still, the results do not suggest that the effect age has on the likelihood of hypertension is dependent on weight – on the contrary, getting older still increases those odds regardless of BMI.

All things considered, the independent variable age manages to explain more than a quarter of the variance of having hypertension and it helps increase the proportion of correctly classified cases. I therefore decided to include hypertension as a target variable in the spatial microsimulation model.

6.3.2. Heart failure

Approximately one in every five adults develops heart failure in the course of their life. Heart failure, also known as cardiac insufficiency, is a chronic medical condition caused by dysfunction of the heart muscle. It manifests either as weakening of the muscle so that the heart cannot keep up with pumping blood at the pace necessary for maintaining regular body functions, or as increased stiffness of the muscle preventing the heart from relaxing while filled with blood. In any case, typical symptoms of heart failure include shortness of breath, fatigue, swelling of the legs, and inability to engage in physical exercise (Horwich and Fonarow 2017, p.116).

Like in the case of hypertension, heart failure is generally caused by a combination of risk factors, such as older age, family history of heart disease, overweight or obesity, unhealthy diet, lack of physical exercise, smoking, and binge drinking. Hypertension, type 2 diabetes, and cholesterol build-up in the arteries are other major risk factors (ibid.).

Not only does heart failure significantly decrease life quality, but it is also a growing economic problem, with high prevalence rates worldwide. Europe and the USA, for instance, spend up to 2% of their annual healthcare budget for its treatment. Globally, the economic burden of

heart failure is estimated at 108 billion USD per year. Europe accounts for nearly 7% of the total global costs (Lesyuk et al. 2018, pp.1–2).

Some of the factors contributing to the development of heart failure, such as maintaining balanced diet and regular physical activity, can be supported by the living environment. Therefore, a small-scale distribution of its prevalence may point to areas where interventions promoting healthy eating habits and daily exercise would have biggest impact.

To include heart failure as target variable in the spatial microsimulation model, I tested how much of its variance can be explained by the four constraints: age, gender, employment, and living situation. Since heart failure is generally an irreversible chronic condition, I chose the variable 'heart failure (12-month prevalence)' instead of creating a new one to account only for current cases like I did for hypertension. A cross tabulation with the dichotomous variable 'heart failure in the past 12 months' confirmed that there were no discrepancies – all respondents, who had answered the question regarding 12-month prevalence affirmatively, confirmed they had the disease in the past 12 months as well. The illness was therefore not back in the past, but it was a medical condition they currently had to deal with.

I carried out a binomial logistic regression and inserted the dichotomous variable 'heart failure (12-month prevalence)' as dependent variable and age, gender, employment status, and living situation as independent variables. Since most of the respondents answered negatively to having heart failure, the model classified 96,6% of all cases correctly by assigning them to the category 'no', without considering any of the independent variables.

Of n=19.162 individuals included in the analysis, the model generated n=155 standardised residuals with a value of > ±2,5 Std Dev. Nevertheless, it was statistically significant: $\chi^2(4) = 902{,}487$, $p < 0.0005$ and explained 17,9% (Nagelkerke R^2) of the variance in the dependent variable. The percentage of correctly classified cases, however, remained unchanged after the input of the independent variables.

Even though all independent variables, except for living situation, were statistically significant (Table 4), they did not contribute much to explaining the variance of heart failure. Therefore, using them for constraining the model is not going to contribute to a non-random allocation of the positive cases.

Table 4. Logistic regression predicting the likelihood of heart failure (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,282	,019	216,174	1	,000	1,326
	Gender	-,269	,086	9,867	1	,002	,764
	Employment status	-1,001	,134	55,633	1	,000	,367
	Living situation	,128	,093	1,921	1	,166	1,137
	Constant	-5,019	,246	417,484	1	,000	,007

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Still, I preferred to include heart failure as a target variable and then use external data to validate the modelled distribution at the level of the statistical areas, instead of leaving it out altogether.

6.3.3. Diabetes mellitus

Diabetes mellitus is a relatively common metabolic disorder marked by elevated blood sugar levels. There are generally two types of diabetes – type 1 is characterised by an autoimmune destruction of insulin-producing cells and mainly affects children and adolescence, whereas type 2 usually develops in adults and occurs either because of insufficient insulin secretion or due to impairment of insulin action. There is also a so-called gestational diabetes, which initially occurs during pregnancy and generally reverses back (Robert Koch-Institute 2011, p.1). It does, however, '*carry a lifetime risk of progression to type 2 diabetes of up to 60%*' (Noctor and Dunne 2015, p.234).

Lack of physical exercise, unbalanced diet and obesity are the major modifiable risk factors for developing type 2 diabetes. The growing prevalence of sedentary lifestyle combined with unhealthy eating habits and rising life expectancy due to medical progress are the reasons for the increasing incidence of diabetes worldwide (Robert Koch-Institute 2014, p.65). Although early diagnosis and available disease treatment contribute to the longer lifespan of diabetics, there are still long-term complications, which considerably lower their life quality. These include partial disability, coronary heart disease, heart and kidney failure, blindness, and amputation of the lower limbs (Robert Koch-Institute 2011, p.1).

Against this background, '*diabetes is one of Germany's most expensive chronic diseases*' (ibid., p.4). According to estimations of the Federal Statistical Office, direct costs of caring for diabetes patients in Germany amount to 2,5% of health expenditure for all other diseases (ibid.). Jacobs et al. (2017) found that *'one in 10 Euros of healthcare expenses is spent on people with type 2 diabetes in Germany*' (p.855). With this in view, diabetes prevention should clearly be prioritised to reduce healthcare expenditure.

Findings from the empirical evidence gathered in the study GEDA 2012 do not suggest a significant connection between diabetes and gender – 7,5% of all females and 7,9% of all males reported to have been diagnosed with the disease. The 12-month-prevalence of diabetes under the age of 45 is below 2% for both men and women. It rises to 6,5% for females and 9,4% for males in the age group 45-65 years – this being the only age interval where gender significantly correlates with diabetes. After the age of 65, diabetes prevalence increases again for both genders: 17,4% of the interviewed females and 18,6% of the males answered affirmatively to having diabetes (Robert Koch-Institute 2014, p.65).

To gather the data these findings are based on, respondents were asked three diabetes-related questions: '*Have you ever been diagnosed with diabetes by a physician?*', '*Was it during pregnancy?*', and '*Did you have diabetes in the past 12 months?*'. In the micro dataset, there are four variables – one for each of these questions and one regarding 12-month-prevalence of diabetes, which is created based on having been diagnosed with diabetes and having had symptoms over the last year. All of those, who answered affirmatively to the first question reported having the disease in the past 12 months as well. Therefore, all respondents, who had been diagnosed with diabetes, are counted as positive cases for the estimation of the 12-month-prevalence. The GEDA 2012 study relied on self-reported cases – that is, no additional laboratory tests were carried out to validate them. Furthermore, the gathered data does not allow for a differentiation between type 1 and type 2 diabetes. These are the main limitations of the available micro data to be considered.

I carried out a binomial logistic regression taking '12-month-prevalence of diabetes' as dependent variable and age, gender, employment, and living situation as explanatory variables. Out of n=19.280 cases, included in the test, the model generated 166 standardised residuals with a value of > ±2,5 Std Dev. Without considering the independent variables, it managed to correctly classify 92,3% of the cases by assigning all of them to the category 'no'. While the model – $\chi^2(4) = 1374{,}344$, $p < 0.0005$ – as well as the included independent variables were statistically significant (Table 5), they were only able to explain 16,4% (Nagelkerke R^2) of the variance of diabetes. Moreover, the four variables did not contribute to a better prediction of the cases.

Table 5. Logistic regression predicting the likelihood of diabetes (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,235	,012	397,697	1	,000	1,265
	Gender	-,367	,059	39,017	1	,000	,693
	Employment status	-,634	,079	64,712	1	,000	,530
	Living situation	,240	,065	13,661	1	,000	1,271
	Constant	-3,612	,153	553,791	1	,000	,027

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Against this background, the same conclusion as for heart failure applies here – the selected constraints do not contribute sufficiently to explaining the variance of diabetes. Hence, if diabetes is included as target variable, the cases with positive outcomes are going to be distributed to the statistical areas randomly. Still, I decided to include diabetes as target variable in the spatial microsimulation model and rely on external validation to establish to what extent it manages to depict reality.

6.3.4. Cancer

Schlander et al. (2018) argue that cancer '*is the second leading cause of mortality in Germany* […] *and accounts for almost one fifth of the total burden of disease* […], *as measured by means of disability-adjusted life years*' (p.332). According to official statistics, cancer-related healthcare expenditures amounted to nearly 20 billion EUR in 2015 (ibid.).

For the purposes of the study GEDA 2012, respondents were asked if they have ever been diagnosed with cancer by a physician. To test the predictive power of the four predefined constraints for having an oncological disease, I used the corresponding dichotomous variable to carry out a binomial logistic regression. Without considering the independent variables, 92,7% of the cases were classified correctly by assigning all of them to the category 'no'. The model generated 175 standardised residuals with a value of > ±2,5 Std Dev. It was statistically significant – $\chi^2(4) = 1042{,}025$, $p < 0.0005$ and explained 13% (Nagelkerke R^2) of the variance in the dependent variable. However, the explanatory variables did not contribute to increasing the proportion of correctly predicted cases.

Since there are different morbidity risk factors depending on the type of cancer, it is hardly surprising that age, gender, employment, and living situation cannot sufficiently predict whether an individual is likely to have (had) cancer. Moreover, the results pointed that gender and living situation are not even statistically significant (Table 6).

Table 6. Logistic regression predicting the likelihood of cancer (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,229	,012	373,996	1	,000	1,257
	Gender	,081	,060	1,866	1	,172	1,085
	Employment status	-,420	,078	28,756	1	,000	,657
	Living situation	-,071	,068	1,087	1	,297	,931
	Constant	-4,282	,159	722,600	1	,000	,014

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Against this background, I chose, as with the other chronic illnesses discussed above, to still include this variable as a target in the spatial microsimulation model. Although the allocation of positive outcomes to the statistical areas is going to be random, this is the only way to have at least some distribution of the generated synthetic population to different groups in terms of cancer. Data provided by health insurance funds was then used to verify the modelled disease prevalence.

6.3.5. Depression

In Germany, and in western countries in general, depression is the most common mental disorder. It is associated with high levels of suffering and imposes significant economic burden estimated at 15.6 billion EUR per year (Krauth et al. 2014, p.1).

Participants in the GEDA 2012 survey were asked whether they have ever been diagnosed with depression by a physician and if this has been in the past 12 months. I filtered out only the cases of 'active' depression, that is, having been diagnosed in the last year. Then, I carried out a binomial logistic regression to estimate the predictive power of age, gender, employment, and living situation.

Without considering these variables, all cases were classified to the category 'no'. Thus, 92% of them were correctly predicted as this was the proportion of people who replied negatively to having been recently diagnosed with depression. The regression model was statistically significant: $\chi^2(4) = 259{,}600$, $p < 0.0005$. Nonetheless, it generated 78 standardised residuals with a value of > ± 2,5 Std Dev. Moreover, it only managed to explain 3,1% of the variance in the dependent variable. While all independent variables were statistically significant (Table 7), they did not contribute to increasing the proportion of correctly predicted cases.

Table 7. Logistic regression predicting the likelihood of depression (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,049	,009	33,426	1	,000	,952
	Gender	,403	,056	51,796	1	,000	1,497
	Employment status	-,523	,061	73,197	1	,000	,593
	Living situation	,626	,064	95,807	1	,000	1,869
	Constant	-2,617	,120	473,415	1	,000	,073

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

The model suggested that of all four explanatory variables age has the weakest impact on the likelihood of depression. Gender plays slightly bigger role as females have 1,5 times higher odds of being diagnosed with depression than males. Being unemployed and living alone also increase those odds by approximately 1,7 and 1,9 times, respectively. While these certainly are some indications of potentially existing trends, they do not suffice to reliably predict the likelihood of depression.

Considering the generated residuals, the small proportion of explained variance by the independent variables, and their failure to increase the number of correctly predicted cases, it would be logical to leave depression out of the spatial microsimulation model. There is a wide variety of factors which can trigger depression, which is why attempting to model its distribution based solely on age, gender, employment, and living situation, may seem unjustified. Nevertheless, I had external data about depression at my disposal. Therefore, instead of narrowing down the scope of the model from the very beginning, I preferred to include depression as target variable to test out how the model will perform despite the unconvincing results of the statistical tests.

6.3.6. Subjectively perceived health, chronic medical condition(s), and impairment due to illness

Next to the specific medical conditions introduced above, the micro dataset contains variables regarding the overall health of the respondents: subjectively perceived health, suffering from chronic medical condition(s), and impairment in everyday activities due to illness.

Subjectively perceived health

Literature reviews and meta-analyses of scientific evidence show that subjective well-being can have a beneficial effect on health and longevity (Diener et al. 2017, p.133). Particularly immune, cardiovascular, and endocrine measures correlate with some types of subjective well-being (ibid., p.139).

Participants in the study GEDA 2012 were asked to evaluate their overall health by answering the question '*How would you evaluate your overall health – very good, good, average, poor or very poor?*'. The micro dataset contains a dichotomous variable corresponding to this question, coded into the categories 'very good/good' and 'average/bad/very bad'.

Without considering the four independent variables – age, gender, employment, and living situation, 70,4% of the cases were classified correctly by all of them being assigned to the category 'very good/good'. The binomial logistic regression model did not generate any residuals and was statistically significant: $\chi^2(4) = 2347{,}377$, $p < 0.0005$. It explained 16,3% (Nagelkerke R^2) of the variance in subjectively perceived health and classified 71,3% of the cases correctly. Sensitivity was 86,6%, specificity 34,8%, positive predictive value was 75,9% and negative predictive value was 47,8%. All predictor variables, except for gender, were statistically significant (Table 8).

Table 8. Logistic regression predicting the likelihood of subjectively perceived health as average/bad (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,152	,006	696,722	1	,000	1,164
	Gender	-,027	,034	,625	1	,429	,973
	Employment status	-,638	,039	262,552	1	,000	,528
	Living situation	,284	,042	45,091	1	,000	1,329
	Constant	-1,611	,079	416,720	1	,000	,200

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Out of all explanatory variables, employment status turned out to have the biggest impact on how respondents evaluate their own health. The results showed that employed people are approximately half as likely to view their health as average, bad, or very bad compared to people, who are currently not employed (regardless of whether they are unemployed or retired). Age and living alone, on the other hand, slightly increased the odds of not feeling in good health, by 1,2 and 1,3 times, respectively.

At the city quarter level, there are further constraint variables which can serve to better explain the variance in subjectively perceived health. Hypertension, diabetes, heart failure, depression, and cancer, can turn out to be good predictors. I thus carried out the binomial logistic regression once again, this time considering all constraint variables available at the first tier.

The model was statistically significant: $\chi^2(9) = 4078{,}360$, $p < 0.0005$, and generated only three standardised residuals > ± 2,5 Std Dev. The consideration of the additional health-related constraints increased the proportion of explained variance from 16,3% to 27,2% (Nagelkerke R^2). Overall, the model managed to classify 76,5% of all cases correctly. Sensitivity was 92%, specificity 39,6%, positive predictive value was 78,4% and negative predictive value was 67,4%. In this case, again, gender was the only explanatory variable, which was not statistically significant (Table 9).

Table 9. Logistic regression predicting the likelihood of subjectively perceived health as average/bad when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,101	,007	238,065	1	,000	1,107
	Gender	-,014	,037	,152	1	,697	,986
	Employment status	-,456	,042	117,968	1	,000	,634
	Living situation	,192	,045	18,102	1	,000	1,212
	Hypertension	,689	,041	282,556	1	,000	1,991
	Diabetes mellitus	1,046	,064	265,671	1	,000	2,845
	Heart failure	1,270	,102	156,168	1	,000	3,560
	Cancer	,636	,064	99,959	1	,000	1,888
	Depression	1,547	,061	650,746	1	,000	4,695
	Constant	-1,904	,084	509,200	1	,000	,149

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

The results clearly showed that chronic illnesses have a much more pronounced effect than age, gender, employment, and living situation on how survey respondents view their health. Depression turned out to have the biggest impact as it increased the odds of having a negative perception of one's own health almost five times. Heart failure and diabetes increased this likelihood approximately three times, followed by hypertension and cancer, with 2,0- and 1,9-times higher odds, respectively. Against this background, modelling subjectively perceived health at the city quarter level promises to provide more reliable results compared to considering only the four basic constraints available at the statistical areas level.

Chronic medical condition(s)

For the analysis of the overall chronic disease prevalence, regardless of the type of medical condition, respondents of the survey GEDA 2012 answered if they have one or many long-lasting illnesses requiring constant treatment and monitoring. Without considering the four basic independent variables, the model classified 59,2% of the cases correctly by assigning all of them to the category 'no'. The logistic regression model did not generate any residuals with a value > ±2,5 Std Dev and it was statistically significant: $\chi^2(4) = 1692{,}293$, $p < 0.0005$. Overall, it managed to explain 11,3% (Nagelkerke R^2) of the variance in the dependent variable and thus increased the proportion of correctly predicted cases to 63,7%. Sensitivity was 40,6%, specificity was 79,6%, positive predictive value was 57,8% and negative predictive value was 66,1%. All predictor variables were statistically significant (Table 10).

Table 10. Logistic regression predicting the likelihood of chronic illness(es) (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,143	,005	751,525	1	,000	1,153
	Gender	,074	,031	5,605	1	,018	1,077
	Employment status	-,308	,036	73,838	1	,000	,735
	Living situation	,149	,040	13,618	1	,000	1,161
	Constant	-1,300	,071	332,159	1	,000	,273

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

The model showed that age and living alone were factors, which very slightly increase the odds of suffering from a chronic illness, while gender barely plays a role. Being employed decreases the likelihood of being chronically ill by approximately 1,4 times and is thus the variable with the biggest prediction power.

As in the case of subjectively perceived health, I conducted the logistic regression a second time to account for the available health data aggregated at the city quarter level. The model was statistically significant: $\chi^2(9) = 4080{,}295$, $p < 0.0005$, and generated 13 standardised residuals > ±2,5 Std Dev. Compared to the previous case of only considering age, gender, employment, and living situation as predictors of chronic disease, the model managed to increase the proportion of explained variance in the dependent variable more than twice – from 11,3% to 25,8% (Nagelkerke R^2). The correctly predicted cases were 71,6% compared to 59,2% when not taking any explanatory variables into account, and 63,7% when considering the four basic ones. Sensitivity was 54,5%, specificity was 83,4%, positive predictive value was 69,3% and negative predictive value was 72,7%. All predictor variables, except for living alone, were statistically significant (Table 11).

Table 11. Logistic regression predicting the likelihood of chronic illness(s) when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,071	,006	142,923	1	,000	1,074
	Gender	,129	,033	14,924	1	,000	1,138
	Employment status	-,090	,039	5,362	1	,021	,914
	Living situation	,054	,044	1,528	1	,216	1,056
	Hypertension	1,055	,040	686,755	1	,000	2,872
	Diabetes mellitus	1,873	,083	510,079	1	,000	6,510
	Heart failure	1,108	,110	101,126	1	,000	3,027
	Cancer	,453	,065	48,748	1	,000	1,572
	Depression	1,389	,063	484,050	1	,000	4,012
	Constant	-1,588	,077	429,085	1	,000	,204

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

The results indicated that people with diabetes are six times more likely to answer affirmatively to the question whether they suffer from a chronic disease, which is, of course, logical. Depression, followed by heart failure, hypertension, and cancer also increased those odds. With this in view, modelling chronic disease at the city quartier level should deliver reliable results. Nevertheless, it is rather obsolete because there is data available about the specific chronic illnesses.

Impairment in everyday activities due to illness

Being impaired in everyday activities can sometimes be the consequence of suffering from chronic illness. The corresponding variable in the micro dataset is dichotomous and accounts for the state of feeling constantly impaired for at least six months.

The logistic regression model generated valid results, without any residuals. It was statistically significant: $\chi^2(4) = 2564{,}907$, $p < 0.0005$ and explained 17,4% (Nagelkerke R^2) of the variance in the dependent variable. Overall, it classified 69,8% of the cases correctly (compared to 67,1% without considering the independent variables). Sensitivity was 41,3%, specificity was 83,8%, positive predictive value was 55,6% and negative predictive value was 74,4%. All predictor variables were statistically significant (Table 12).

Table 12. Logistic regression predicting the likelihood of impairment due to illness (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,150	,006	726,379	1	,000	1,161
	Gender	-,100	,034	8,909	1	,003	,904
	Employment status	-,694	,038	332,103	1	,000	,500
	Living situation	,332	,042	62,918	1	,000	1,393
	Constant	-1,290	,076	287,946	1	,000	,275

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

The model suggested that males and older people have slightly higher odds of being impaired in their everyday activities. Employed people, on the other hand, appeared half as likely to experience everyday impairment due to illness. Living alone suggested approximately 1,4 times higher odds.

Considering the available disease data aggregated at the city quarter level, the logistic regression model provided more promising results. It was statistically significant: $\chi^2(9) = 4035{,}437$, $p < 0.0005$ and generated only six standardised residuals > ± 2,5 Std Dev. Overall, it managed to explain 26,4% of the variance in the dependent variable as opposed to 17,4% when only considering the four basic constraints. Furthermore, the explanatory variables contributed to increasing the proportion of correctly predicted cases to 74,5%. Sensitivity was 45,4%, specificity was 88,7%, positive predictive value was 66,4% and negative predictive value was 76,8%. All predictor variables were statistically significant (Table 13).

Table 13. Logistic regression predicting the likelihood of impairment due to illness when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,115	,006	331,663	1	,000	1,122
	Gender	-,117	,035	10,898	1	,001	,890
	Employment status	-,537	,040	178,040	1	,000	,584
	Living situation	,247	,044	31,362	1	,000	1,280
	Hypertension	,447	,041	121,070	1	,000	1,564
	Diabetes mellitus	,677	,063	113,871	1	,000	1,968
	Heart failure	1,476	,110	179,585	1	,000	4,376
	Cancer	,692	,064	118,386	1	,000	1,997
	Depression	1,625	,062	695,733	1	,000	5,078
	Constant	-1,528	,080	360,784	1	,000	,217

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

According to the results, depression and heart failure contributed most to increasing the odds of being impaired in daily activities – by 5,1 and 4,4 times, respectively. Cancer and diabetes mellitus made it twice as likely, whereas hypertension had the weakest effect.

In all those cases – regarding overall health, chronic medical conditions, and impairment due to illness – the four basic constraint variables available at the statistical areas level did not explain much of the variance in the dependent variables (11-17%). Considering chronic disease data available at the city quarter level, however, nearly doubled the explained variance to 26-27%. I thus decided to model all of them as targets at both spatial scales. The results at the level of the city quarters are more reliable, which must be considered when interpreting the modelled spatial distributions. Nonetheless, modelling these variables at the first tier increases the reliability of the modelled population at the underlying spatial scale as well because the population is constrained to the realities in the corresponding city quarter.

6.3.7. Overweight and obesity

While there is no aggregated data about overweight and obesity at the level of the city quarters, I wanted to include them as target variables and thus model their distribution at the statistical areas level as they are major risk factors for triggering many chronic diseases such as hypertension and diabetes mellitus.

In the individual dataset, overweight and obesity are operationalised by BMI. The latter is calculated by dividing the weight in kilograms by the squared height in metres. A person who is 1,70 metres tall and weighs 55 kilograms, for instance, would have a BMI = 19 and would thus be categorised as having *normal* weight. In this context, the WHO has established the following categories: 'underweight' (BMI < 18,5), 'normal weight' (18,5 – 24,9), 'overweight' (25 – 29,9), 'first-grade obesity' (30 – 34,9), 'second-grade obesity' (35 – 39,9), and 'third-grade obesity' (≥ 40) (Robert Koch-Institute 2014, p.93).

In the study GEDA 2012, BMI is based on self-reported weight and height. Self-reported weight tends to be underestimated, as opposed to height, which is often overestimated. The BMI in the micro dataset may therefore be lower than if it were based on actual measurements. This is one possible flaw of the available micro data, which must be considered.

To estimate the extent to which age, gender, employment, and living situation can explain the variance in being overweight, or obese (regardless of obesity grade), I ran a binomial logistic regression. Both dependent variables are dichotomous: 'overweight according to WHO classification' and 'obesity according to WHO classification'.

Of n=18.901 individuals tested for being overweight, no standardised residuals > ±2,5 Std Dev were generated. The model was statistically significant: $\chi^2(4) = 1581.408$, $p < 0.0005$ and explained 10,7% (Nagelkerke R^2) of the variance in the dependent variable. Without considering the explanatory variables, it managed to correctly classify 52,6% of the cases by assigning all of them to the category 'yes', thereby assuming that all respondents are overweight. Considering the four independent variables, the percentage of correctly classified cases increased to 62,9%. Sensitivity was 69,9%, specificity was 55,2%, positive predictive value was 63,4% and negative predictive value was 62,3%. All predictor variables were statistically significant (Table 14).

Table 14. Logistic regression predicting the likelihood of overweight (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,167	,005	1018,654	1	,000	1,182
	Gender	-,629	,031	409,974	1	,000	,533
	Employment status	,232	,036	40,732	1	,000	1,261
	Living situation	-,208	,041	25,833	1	,000	,812
	Constant	-,136	,069	3,887	1	,049	,873

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

In the case of obesity, the results were less convincing. Without considering the explanatory variables, 83,5% of all cases were correctly classified by assigning all of them to the category 'no'. The logistic regression model was statistically significant – $\chi^2(4) = 338.207$, $p < 0.0005$,

and did not generate any standardised residuals > ±2,5 Std Dev. Nevertheless, it only managed to explain 3% (Nagelkerke R^2) of the overall variance in the dependent variable. Furthermore, only age and gender were statistically significant (Table 15), but they did not contribute enough to explaining the variance. The model thus failed to increase the percentage of correctly classified cases.

Table 15. Logistic regression predicting the likelihood of obesity (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,095	,007	186,485	1	,000	1,099
	Gender	-,104	,041	6,648	1	,010	,901
	Employment status	-,060	,048	1,549	1	,213	,942
	Living situation	,078	,051	2,379	1	,123	1,081
	Constant	-2,113	,095	493,500	1	,000	,121

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

At the level of the city quarters, overweight and obesity can be modelled using the additional health-related constraint variables: hypertension, heart failure, diabetes, depression, and cancer. I carried out another set of logistic regression tests to account for them and found that they further increase the explained variance for both dependent variables: from 10,7% to 16,1% for overweight, and from 3% to 11,4% for obesity. In both cases, the regression model was statistically significant and generated no standardised residuals, which suggested valid results.

In the case of overweight, the proportion of correctly predicted cases increased from 62,9%, when accounting only for the four basic constraints, to 66,2% when considering the health-related variables available at the first tier. Sensitivity was 66,7%, specificity was 65,6%, positive predictive value was 68,3%, and negative predictive value was 63,9%. Apart from cancer and heart failure, all explanatory variables were statistically significant (Table 16). Hypertension and diabetes mellitus increased the odds of being overweight more than twice. This is logical as overweight is a significant risk factor for both chronic illnesses.

Table 16. Logistic regression predicting the likelihood of overweight when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,118	,006	433,821	1	,000	1,126
	Gender	-,613	,032	372,180	1	,000	,542
	Employment status	,364	,038	92,436	1	,000	1,439
	Living situation	-,274	,042	41,671	1	,000	,760
	Hypertension	,917	,041	492,956	1	,000	2,502
	Diabetes mellitus	,903	,073	154,681	1	,000	2,468
	Heart failure	,080	,095	,701	1	,402	1,083
	Cancer	-,062	,063	,964	1	,326	,940
	Depression	,306	,059	26,855	1	,000	1,358
	Constant	-,214	,071	9,167	1	,002	,807

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

For obesity, the consideration of all constraint variables available at the city quarter level contributed to increasing the proportion of correctly predicted positive cases from 0% to 7%. Still, the overall share of correctly classified cases remained 83,5%. Sensitivity was 7%, specificity was 98,5%, the positive predictive value was 48,3%, and the negative predictive value was 84,3%. All explanatory variables except for gender and living alone were statistically significant (Table 17). Like in the case of being overweight, hypertension and diabetes mellitus turned out to have the most pronounced effect on increasing the likelihood of obesity.

Table 17. Logistic regression predicting the likelihood of obesity when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	,018	,008	4,997	1	,025	1,018
	Gender	-,036	,042	,749	1	,387	,964
	Employment status	,112	,051	4,871	1	,027	1,119
	Living situation	,015	,053	,081	1	,776	1,015
	Hypertension	1,080	,048	516,757	1	,000	2,946
	Diabetes mellitus	1,039	,062	276,536	1	,000	2,827
	Heart failure	,326	,095	11,740	1	,001	1,385
	Cancer	-,162	,077	4,389	1	,036	,851
	Depression	,285	,069	16,807	1	,000	1,330
	Constant	-2,279	,099	530,749	1	,000	,102

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

Against this background, I decided to model overweight and obesity both at the city quarter and at the statistical areas level. In the case of overweight, the distribution at both levels would be relatively equally reliable. In the case of obesity, the matter is more complex. Age is not as reliable predictor. Instead, there seem to be other powerful stimuli for triggering obesity. Still, there was available external data from a sample survey carried for the purposes of the research project 'Healthy Neighbourhoods', which was used to validate the modelled spatial distribution of both overweight and obesity.

6.3.8. Health behaviour

Next to the illness-related variables, the micro dataset contains variables regarding individual health behaviour, such as frequency of physical and sporting activity, diet, smoking, and alcohol consumption. I carried out several statistical tests and found that the variables available at the statistical areas level – age, gender, employment, and living situation – can predict only sporting activity and smoking to a certain degree.

Sporting activity

Regular physical exercise has many positive effects, such as maintaining normal weight, reducing stress, lowering blood pressure, etc. (Monteiro and Sobral Filho 2004). Regarding their sporting activity, participants in the study GEDA 2012 responded whether they had done any type of physical exercise in the past three months. I used the corresponding variable in the micro dataset as dependent variable and carried out a binomial logistic regression to estimate the explained variance.

Without considering the four explanatory variables, 65,8% of all cases were classified correctly by assigning all of them to the category 'yes', thereby assuming all survey participants had engaged in some type of sporting activity over the past three months. The logistic regression model was statistically significant - $\chi^2(4) = 1079.241$, $p < 0.0005$, and did not generate any standardised residuals. It managed to explain 7,5% (Nagelkerke R^2) of the variance in the dependent variable and slightly increased the proportion of correctly predicted cases to 67,3%. Sensitivity was 92,2%, specificity was 19,4%, positive predictive value was 68,7% and negative predictive value was 56,4%. All predictor variables were statistically significant (Table 18).

Table 18. Logistic regression predicting the likelihood of sporting activity in the past 3 months (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,124	,005	530,395	1	,000	,884
	Gender	,108	,032	11,599	1	,001	1,115
	Employment status	,141	,037	14,290	1	,000	1,151
	Living situation	-,215	,040	28,381	1	,000	,806
	Constant	1,296	,073	312,419	1	,000	3,656

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Accounting for the health-related variables available at the city quarter level slightly increased the proportion of explained variance from 7,5% to 8,3%. The logistic regression model was statistically significant – $\chi^2(9) = 1188.536$, $p < 0.0005$, and did not generate any standardised residuals. Overall, it managed to increase the proportion of correctly predicted cases to 67,8%. Sensitivity was 93%, specificity was 19,2%, positive predictive value was 68,9% and negative predictive value was 58,8%. Except for cancer, all explanatory variables were statistically significant (Table 19). The results indicated that chronic illnesses, albeit slightly, decrease the likelihood of engaging in sporting activity.

Table 19. Logistic regression predicting the likelihood of sporting activity in the past 3 months when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,108	,006	345,296	1	,000	,898
	Gender	,092	,032	8,179	1	,004	1,096
	Employment status	,085	,038	5,061	1	,024	1,089
	Living situation	-,198	,041	23,614	1	,000	,820
	Hypertension	-,125	,039	10,534	1	,001	,882
	Diabetes mellitus	-,376	,058	41,925	1	,000	,687
	Heart failure	-,439	,085	26,980	1	,000	,644
	Cancer	-,106	,059	3,230	1	,072	,899
	Depression	-,152	,057	7,096	1	,008	,859
	Constant	1,349	,074	332,183	1	,000	3,853

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

Smoking

Unlike physical exercise, smoking is a type of health behaviour associated with negative repercussions such as increasing the odds of lung cancer, infertility, etc. (e.g. Remen et al. 2018; American Society for Reproductive Medicine 2018).

Participants in the study GEDA 2012 were asked whether they smoke, even if only occasionally. I transformed the corresponding variable, which had four categories – 'yes', 'occasionally', 'not anymore', and 'never smoked', into a dichotomous one to account only for smokers and non-smokers. To that end, smokers and occasional smokers were assigned to the category 'yes', as opposed to 'non-smokers' and 'ex-smokers', who were put in the category 'no'.

Without accounting for the explanatory variables age, gender, employment, and living situation, 72,4% of all cases were predicted correctly by classifying all of them as non-smokers. The binomial regression model was statistically significant – $\chi^2(4) = 1084.623$, $p < 0.0005$, and did not generate any standardised residuals. The independent variables managed to explain 7,9% of the variance and while they did not contribute to increasing the total proportion of correctly predicted cases, they did predict at least 2,3% of the positive cases correctly. Sensitivity was 2,3%, specificity was 99%, positive predictive value was 47,8%, and negative predictive value was 72,7%. All independent variables were statistically significant (Table 20).

Table 20. Logistic regression predicting the likelihood of smoking (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,124	,006	502,489	1	,000	,883
	Gender	-,338	,034	101,187	1	,000	,713
	Employment status	,350	,039	80,607	1	,000	1,420
	Living situation	,521	,045	134,701	1	,000	1,684
	Constant	-,011	,073	,022	1	,881	,989

a. Variable(s) entered on step 1: age, gender, employment status, and living situation.

Including the health-related constraints available at the first tier as explanatory variables in the logistic regression model barely altered the results. While the proportion of explained variance slightly increased to 8,4%, the overall share of correctly predicted cases remained the same. The model was statistically significant – $\chi^2(9) = 1155.844$, $p < 0.0005$, and did not generate standardised residuals. Sensitivity was 2,8%, specificity was 98,9%, positive predictive value was 48,3% and negative predictive value was 72,8%. Apart from cancer and diabetes, all independent variables were statistically significant (Table 21). Out of these, being depressed and living alone had the most pronounced effect on increasing the odds of smoking. Currently being employed also suggested higher chances of being a smoker. People with heart failure, on the other hand, were approximately half as likely to smoke.

Table 21. Logistic regression predicting the likelihood of smoking when accounting for comorbidities (own representation, Source: Robert Koch-Institute, Department of Epidemiology and Health Monitoring 2014)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,111	,006	337,572	1	,000	,895
	Gender	-,359	,034	112,280	1	,000	,699
	Employment status	,336	,040	72,459	1	,000	1,400
	Living situation	,506	,045	124,896	1	,000	1,658
	Hypertension	-,140	,045	9,546	1	,002	,869
	Diabetes mellitus	-,136	,075	3,299	1	,069	,873
	Heart failure	-,594	,129	21,091	1	,000	,552
	Cancer	-,015	,073	,043	1	,836	,985
	Depression	,415	,059	48,931	1	,000	1,515
	Constant	-,027	,073	,137	1	,711	,973

a. Variable(s) entered on step 1: age, gender, employment status, living situation, hypertension, diabetes mellitus, heart failure, cancer, and depression.

6.3.9. Overview of the selected constraint and target variables

The choice of constraint and target variables slightly differs at the two tiers. At the level of the city quarter (clusters), there is a broader choice of constraints due to data availability. Thus, not only socio-demographic, but also health-related variables were used to constrain the synthetic population (Table 22).

Table 22. Constraint variables overview (own representation)

Spatial scale		Constraint variables	Characteristic attributes
Tier 1: city quarter (clusters)	Tier 2: statistical areas	age	Tier 1: 18-64; 65+ Tier 2: 18-24; 25-29 …; 80+
		gender	male; female
		currently employed	yes; no
		living in a single household	yes; no
		hypertension (12-month prevalence)	yes; no
		heart failure (12-month prevalence)	yes; no
		diabetes (12-month prevalence)	yes; no
		depression (12-month prevalence)	yes; no
		cancer (in the past 10 years)	yes; no

The health-related variables used as constraints at the first tier do not count as targets at this level, as they are already known. At the level of the statistical areas, however, they do not play the role of constraints anymore because the corresponding population counts are no longer available. Thus, their distribution must be modelled based on the remaining constraints. As a result, they change their role from constraints to targets at the second tier. Additionally, there are other health-related variables, such as subjectively perceived health, impairment in daily activities, sport, and smoking, which are modelled as targets both at the city quarter and the statistical areas level. Table 23 provides a summary of the target variables at both modelling tiers.

Table 23. Target variables overview (own representation)

Spatial scale		Target variables	Characteristic attributes
	Tier 2: statistical areas	hypertension (12-month prevalence)	yes; no
		heart failure (12-month prevalence)	yes; no
		diabetes (12-month prevalence)	yes; no
		cancer (in the past 10 years)	yes; no
Tier 1: city quarter (clusters)		subjectively perceived health	Very good/good; average/bad/very bad
		chronic illness	yes; no
		impairment of everyday activities	yes; no
		overweight	yes; no
		obesity	yes; no
		smoking	yes; no
		sporting activity (in the past 3 months)	yes; no

With that, the stage of selecting constraint and target variables was completed. The following paragraphs are dedicated to describing the further path towards generating a synthetic population.

6.4. Population Synthesis

6.4.1. Data pre-processing

Before the actual synthetic population generation, the input data originating from different datasets needed to be transformed so that the micro data and the aggregated geographic data have the same format. This initial step is referred to as *data pre-processing*.

Against this background, I carried out several variable transformations. At the level of the city quarters no adjustments were necessary, as the health-related constraints were defined as dichotomous variables (e.g., hypertension – 'yes'/'no'). I constrained age to just two intervals at this tier, '18-64 years' and '65+ years' because the health-related variables were subcategorised into those two age groups. The last variable used as constraint at this scale was gender, which was also coded dichotomously. Like age, it served for the further subdivision of the health constraints.

At the scale of the statistical areas several modifications were necessary. While age was coded into 5-year-intervals both in the micro dataset and in the geographic dataset, in the latter, the final category was '80+' as opposed to '85+' in the micro dataset. Thus, I assigned all individuals aged between 80-84 years as well as those older than 85 years to the category '80+'.

Another necessary input data adjustment regarded the constraint 'living in a single household'. While the geographic dataset contained single household counts for each statistical area, in the micro dataset there was a metric variable regarding household size. For the data format to become compatible, I converted the metric variable about household size in the micro dataset into a dichotomous one. Thus, it only provided information whether one lives in a single household or not. I then created a new variable in the geographic dataset to account for the number of non-single households. To that end, I subtracted the number of single households from the

total count of private households. The latter was available as a separate variable in the geographic dataset.

The last constraint variable, which required adjustments of the initial data format, was 'employment'. In the geographic dataset, there were aggregated counts of employed people subject to compulsory insurance for each statistical area, whereas the micro dataset contained a dichotomous variable about current employment. To account for all individuals, who are currently not employed, I subtracted the count of employed people subject to compulsory insurance from the total count of adults. While this means putting retired people in the same category as unemployed, my aim was to consider the effects of employment on health – being active in a working environment every day. Therefore, the opposite category did not have to be limited to unemployed people only. Rather than that, I intended to encompass all individuals, who, for whatever reasons there may be, were not employed at the time of participating in the survey. I did not account for full-time and part-time employment because such detailed information was not available in the geographic dataset.

The next step was to *flatten* the micro dataset. Flattening refers to taking one column in the dataset, which represents a single variable with multiple attributes and converting it into multiple columns, each of them referring to the attributes of the given variable. For instance, flattening an age-related variable coded into three categories, e.g., '18-44', '45-64', and '65+', is going to result in three new columns, each corresponding to one of these age intervals (Table 24). Flattening therefore does not alter the available information, but merely modifies the way it is stored.

Table 24. Example of flattening individual data (own representation)

Person ID	Age
1	18-44
2	45-64
3	65+

→

Person ID	18-44	45-64	65+
1	1	0	0
2	0	1	0
3	0	0	1

By the means of flattening, variables of type *character* are converted into multiple *numeric* columns, which can store only two values – either 0, or 1. In a single row, only one of the newly generated columns can contain the value 1, because an individual can only be assigned a single attribute relating to one variable. For instance, if a person is between 18 and 44 years old, this means that he or she cannot also be older than 65 years.

With this in view, I flattened all constraint variables in the micro dataset and then combined the newly generated columns, each referring to a single variable attribute, into a single matrix. Thus, each row represented one individual, and each column contained attribute information about that person – their age, gender, whether they live in a single household, whether they are currently employed, and whether they suffer from hypertension, heart failure, diabetes, cancer, or depression.

Last, I ensured that the column order is the same in both matrices – the one from the geographic dataset, and the one containing the flattened individual data. For the population synthesis algorithm to run smoothly, the number of columns and their order must match perfectly. After this final data pre-processing step, I proceeded with the generation of the synthetic population.

6.4.2. The Iterative Proportional Fitting approach

In the previous section, I introduced all constraint and target variables – both at the scale of the city quarters, and the statistical areas. As the constraints regarding health were available solely at the higher spatial scale, I wrote an algorithm that allows constraining the synthetic population at both spatial scales simultaneously. In essence, the algorithm assigns a weight to the individuals in the micro dataset depending on how representative they are for each city quarter, and for each statistical area within. The representativity for each zone is determined based on the aggregate count of individuals with their attribute characteristics. The weight assigned to each individual is updated iteratively after accounting for each category of the constraint variables available at both spatial scales. Hence, the term *iterative proportional fitting*.

The first step was to create an initial weight matrix `'weights'` with dimensions equal to the number of individuals (rows), and the number of geographic zones (columns). For the latter, I used the count of the statistical areas rather than the city quarters because my goal was to generate synthetic population at the small urban scale. Generally, when creating an initial weight matrix, all cells are assigned the value 1. The weights are thus updated iteratively starting from a neutral initial position. However, I opted for the weights computed by RKI for correcting the GEDA 2012 survey sample as initial weight values in the created matrix. The RKI weights refer to each individual, regardless of the zone, as they were computed in relation to the total population of Germany and its composition. Therefore, all cells belonging to the same row in the initial weight matrix store the same value, that is, the weight related to the individual represented by this row. Thus, before adjusting the weight of each individual to ensure perfect fit for each zone, the representativity of the given individual for Germany is taken into account.

Next, I generated the vector `'tier1_id'` to store the ID of the corresponding city quarter for each statistical area. The vector length hence equals the count of the statistical areas and has the following format:

```
1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3…
```

It becomes clear from this excerpt that the first nine statistical areas belong to the first city quarter, the next nine to the second city quarter, and so on. This vector was created as separate object for the purpose of an aggregation carried out within the IPF algorithm.

To grasp the logic of the population synthesis algorithm, here is an explanation of each line of code:

`for(iter in 1:50){` *# I began by opening a* `for loop`[18] *for the number of iterations. I set 50 as a starting point, but the algorithm stops running before the 50th iteration if it reaches the breaking point defined at the end of the code. The latter is included to avoid unnecessary computations and thus speed up the process of population synthesis.*

`for(i in 1:ncol(cons_tier1)){` *# Within the first* `for loop`*, I opened another* `for loop` *so that the algorithm can run through each constraint category –* `i`*, available at the first tier and thus stored as column in the matrix* `cons_tier1`*.*

[18] 'For loop is a programming language conditional iterative statement which is used to check for certain conditions and then repeatedly execute a block of code as long as those conditions are met.' (Techopedia 2014)

```
index <- ind_cat[, i] == 1
```
 # Next, I defined an index used for selecting individuals belonging to each constraint category `i`*, that the algorithm is running through in iterative manner.* `ind_cat` *is the matrix containing the flattened individual data, where individuals (represented by the rows) are assigned either 1 or 0 in the different columns depending on whether the corresponding constraint category applies to them. For instance, when the algorithm is running through the constraint category males, only those individuals who were assigned 1 in the corresponding column* `i` *within* `ind_cat` *will be selected, as opposed to those who were assigned 0 (because they are females and thus should not be selected).*

```
weights_sum <- t(aggregate(t(weights * ind_cat[, i]), by =
list(tier1_id), FUN = sum))[2:nrow(ind),]
```
 # Next, I created a new matrix `weights_sum` *with dimensions equal to the number of individuals (rows) x the number of city quarters (columns). To that end, I aggregated the values stored in the initial weight matrix* `weights` *by city quarter ID. The vector* `tier1_id` *created earlier was used as aggregation reference as it contains the ID of the respective city quarter for each statistical area. In the generated matrix* `weights_sum`*, each cell represents the weight of a given individual for a certain city quarter, whereby this weight is updated iteratively with the* `for loop` *for each constraint category* `i`*.*

```
weights_sum <- colSums(weights_sum)
```
 # Next, I aggregated the weights by column so that `weights_sum` *becomes a vector with a length that equals the number of city quarters. Accordingly, each value represents the total number of individuals in the given city quarter based on the aggregated weights.*

```
weights_corr <- cons_tier1[,i] / weights_sum
```
 # Here, the actual reweighting begins. I created a new vector, `weights_corr`*, where* `corr` *stands for 'corrected' as the weights aggregated in the previous step must be corrected by the observed counts of the constraints at the first tier. To that end, I divided the observed count of individuals in each city quarter by the corresponding aggregated sum of individual weights. Again, this is carried out iteratively for each constraint category* `i`*.*

```
weights_corr[is.nan(weights_corr)] <- 0
weights_corr[is.infinite(weights_corr)] <- 0
```
 # In case the aggregated sum of individual weights for a certain city quarter turns out to be zero, this and the previous line of code intend to substitute with zero the values `NaN` *or* `Inf` *assigned by the programme in cases of dividing by zero. This step ensures the algorithm will keep on going even if there are no individuals for a certain constraint category in the micro dataset.*

```
weights_corr_exp <- weights_corr[tier1_id]
```
 # Next, I created a new, expanded vector, `weights_corr_exp`*, containing the corrected weights. Its length equals the number of statistical areas. The sum of corrected weights for each city quarter is thus assigned to each statistical area that belongs to it.*

```
weights[index, ] <- t(t(weights[index, ]) * weights_corr_exp)
```
 # Finally, I updated the initial weight matrix `weights` *by multiplying the initial weights (= RKI weight) of the indexed individuals (the individuals belonging to the constraint category* `i`*) with the weights corrected for each city quarter based on the constraints observed at Tier 1. Thus, after iterating through all constraint categories available at the city quarter level,*

each cell in the matrix will represent the weight of each individual for a given statistical area based on their representativity for the city quarter the statistical area is part of.

`}` *# This* `for loop` *continues until all constraint categories available at Tier 1 are considered for correcting the weights.*

`for(i in 1:ncol(cons_tier2)){` *# Next, I opened another* `for loop` *for the IPF algorithm to run through each column in the matrix* `cons_tier2`*. This matrix contains the constraint categories available at Tier 2, that is, the statistical areas.*

`index <- ind_cat[, ncol(cons_tier1) + i] == 1` *# Like in the case with the city quarters, I defined an index to select individuals belonging to each constraint category* `i` *that the algorithm runs through in iterative manner. The matrix* `ind_cat` *contains the flattened individual data for all constraint categories, both at Tier 1 and Tier 2. The indexing at Tier 2 must therefore begin from the first column succeeding the columns referring to Tier 1. The column indexing is hence defined as* `ncol(cons_tier1) + i`*.*

`weights_sum <- colSums(weights[index,])` *# Next, I defined a vector* `weights_sum` *with a length equal to the number of statistical areas. Each value in this vector represents the sum of the weights generated with the previous* `for loop` *for Tier 1, whereby only the weights of the individuals indexed in the previous step are summed up.*

`weights_corr <- cons_tier2[,i] / weights_sum` *# Next, the sum of the weights computed on the basis of the constraint categories at the city quarter level is corrected in order to optimise the fit for the level of the statistical areas. To that end, the observed count of individuals for each constraint category* `i` *in each statistical area is divided by the corresponding sum of individual weights in the vector generated in the previous step.*

```
weights_corr[is.nan(weights_corr)] <- 0
weights_corr[is.infinite(weights_corr)] <- 0
```
 # In case the aggregated sum of individual weights for a certain statistical area turns out to be zero, this and the previous line intend to substitute with zero any `NaN` *or* `Inf` *values, which are assigned by default when trying to divide by zero. This step ensures the algorithm will keep running even if there are no individuals for a certain constraint category in the micro dataset.*

`weights[index, ] <- t(t(weights[index, ]) * weights_corr)` *# The last line of code takes the* `weights` *matrix, which was updated iteratively in the previous* `for loop` *for Tier 1 and is thus adjusted to fit perfectly the constraint categories available at the city quarter level, and updates the weights to optimise the fit at the statistical areas level. To that end, each weight is adjusted iteratively, for each constraint category* `i` *available at Tier 2. The algorithm runs category by category, indexing the individuals belonging to each one of them at the time. Each weight of the indexed individuals, which was last adjusted for the spatial context at Tier 1 is multiplied by the corresponding corrected weight from* `weights_corr`*.*

`}` *# This for loop continues until all constraint categories available at Tier 2 are considered for correcting the weights.*

 # Within the iteration-related `for loop 'for(iter in 1:50)'` *I defined a new object to store the Total Absolute Error (TAE) representing the total sum of differences between the observed and simulated constraint categories for all statistical areas. TAE is updated*

after each iteration and the algorithm stops running only when TAE becomes smaller than 0.001 or when the 50th iteration is completed. The last iteration may thus not be the 50th, but the one which produces a TAE < 0.001:

```
    total_abs_error_t2 = 0
    for(j in 1:ncol(cons_tier2)){
      index <- ind_cat[, ncol(cons_tier1) + j] == 1
      abserror_t2 = sum(abs(colSums(weights[index,]) - cons_tier2[,j]))
      total_abs_error_t2 <- total_abs_error_t2 + abserror_t2
      }

  # Break the iteration for loop if TAE at Tier 2 < 0.001
    if (total_abs_error_t2 < 0.001)  break
}
```

To make the process of reweighting more transparent, I am going to bring in an example with a reduced number of individuals, zones, and constraint categories:

Table 25. Individuals, zones, and constraint categories for a reweighting example (own representation)

Individuals with distinct characteristics	N = 4
Zones at Tier 1	N = 2
Zones at Tier 2	N = 4
Constraint categories at Tier 1	N = 2
Constraint categories at Tier 2	N = 4

Table 26. Example individual data (own representation)

Person ID	hyp.[19]	no hyp.	male	female	18-64	65+
1	0	1	1	0	1	0
2	0	1	0	1	1	0
3	1	0	0	1	0	1
4	1	0	1	0	0	1

Table 27. Example aggregated geographic data (own representation)

		Cons_tier1		Cons_tier2			
Zone ID Tier 1	Zone ID Tier 2	hyp.	no hyp.	males	females	18-64	65+
1	1	13	39	10	12	15	7
1	2			15	15	17	13
2	3	15	33	12	8	16	4
2	4			18	10	14	14

[19] hypertension

Now that the basic framework is clear, I am going to proceed with the steps:

Step 1: Creating initial weight matrix (own representation)

```
weights <- matrix(data = 1, nrow = nrow(ind_cat), ncol = nrow(cons_tier2))
```

```
      [,1]  [,2]  [,3]  [,4]
 [1,]    1     1     1     1
 [2,]    1     1     1     1
 [3,]    1     1     1     1
 [4,]    1     1     1     1
```

When creating an initial weight matrix, the weight of each individual for each zone is generally set to 1. While I assigned the RKI weight as initial individual weight for the purposes of this dissertation, I am going to show the simplified approach in this example.

Step 2: Generating a vector holding the IDs of the spatial units at the higher scale (own representation)

```
tier1_id <- cons_tier2$tier1_ID
```

```
 [1]  1 1 2 2
```

Next, I generate a vector, the length of which equals the number of spatial units at the lower scale. In this example, they are four. The first two belong to the first spatial unit at Tier 1, and the latter two belong to the second unit at Tier 1. Thus, the ID of the first spatial unit is repeated twice and then followed by the ID of the second, again, repeated twice.

Step 3: Updating the weight matrix according to the constraints available at Tier 1 (own representation)

```
weights_sum <- t(aggregate(t(weights * ind_cat[, i]), by = list(tier1_id), FUN =
sum))[2:nrow(ind),]
```

```
     [,1]  [,2]  [,3]  [,4]        ind_cat[,hyp]        [,1]  [,2]  [,3]  [,4]
[1,]    1     1     1     1              0              0     0     0     0
[2,]    1     1     1     1     *        0        →     0     0     0     0
[3,]    1     1     1     1              1              1     1     1     1
[4,]    1     1     1     1              1              1     1     1     1

     [,1]  [,2]  [,3]  [,4]         aggregate by        [,1]  [,2]
[1,]    0     0     0     0            tier1_id          0     0
[2,]    0     0     0     0               →              0     0
[3,]    1     1     1     1                              2     2
[4,]    1     1     1     1                              2     2
```

```
weights_sum <- colSums(weights_sum)
```

```
  [1]    4     4
```

```
weights_corr <- cons_tier1[,i] / weights_sum[20]

 [1]   13    15     /   [1]   4     4      →   [1]   3.25  3.75

weights_corr_exp <- weights_corr[tier1_id]

 [1]   3.25  3.25  3.75  3.75

weights[index, ] <- t(t(weights[index, ]) * weights_corr_exp)

      [,1]  [,2]  [,3]  [,4]       [,1]  [,2]  [,3]  [,4]
 [1,]   1     1     1     1          1     1     1     1
 [2,]   1     1     1     1     →    1     1     1     1
 [3,]   1     1     1     1         3.25  3.25  3.75  3.75
 [4,]   1     1     1     1         3.25  3.25  3.75  3.75
```

The weights must then be updated one more time at Tier 1 to account for the other constraint category ‘no hypertension’. Eventually, the weight matrix will look like this:

```
      [,1]  [,2]  [,3]  [,4]
 [1,] 9.75  9.75  8.25  8.25
 [2,] 9.75  9.75  8.25  8.25
 [3,] 3.25  3.25  3.75  3.75
 [4,] 3.25  3.25  3.75  3.75
```

Step 4: Updating the weight matrix according to the constraints at Tier 2 (own representation)

```
weights_sum <- colSums(weights[index, ])

      [,1]  [,2]  [,3]  [,4]
 [1,] 9.75  9.75  8.25  8.25
 [2,] 9.75  9.75  8.25  8.25   →   [1] 26.00 26.00 24.00 24.00
 [3,] 3.25  3.25  3.75  3.75
 [4,] 3.25  3.25  3.75  3.75

weights_corr <- cons_tier2[,i] / weights_sum[21]

 [1]  10   15   12   18   /   [1]  26   26   24   24   →   [1]  0.38  0.58  0.50  0.75

weights[index, ] <- t(t(weights[index, ]) * weights_corr)

      [,1]  [,2]  [,3]  [,4]       [,1]  [,2]  [,3]  [,4]
 [1,] 9.75  9.75  8.25  8.25       3.71  5.66  4.13  6.19
 [2,] 9.75  9.75  8.25  8.25   →   9.75  9.75  8.25  8.25
 [3,] 3.25  3.25  3.75  3.75       3.25  3.25  3.75  3.75
 [4,] 3.25  3.25  3.75  3.75       1.24  1.89  1.88  2.81
```

[20] In this example, I am updating the weights for the constraint category ‘hypertension’

[21] In this example, I am updating the weights according to the constraint category ‘males’

This is what the weight matrix is going to look like after updating it to account for the constraint category 'males' at Tier 2. Only the first and the last row are updated because the male individuals are stored in those two rows. Next, the weights must be adjusted based on the observed counts of females, people aged 18-64 years, and people older than 65 years. After that, the algorithm keeps on running until it produces the perfect fit, that is, until the aggregated version of the weights matches the observed counts for each category in each zone at both spatial tiers. Eventually, the weight matrix will look like this:

```
      [,1]  [,2]  [,3]  [,4]
[1,]  6.8   8.5   9.6    9
[2,]  8.2   8.5   6.4    5
[3,]  3.8   6.5   1.6    5
[4,]  3.2   6.5   2.4    9
```

To generate a synthetic population for the last zone, we must therefore replicate the first and the last individual nine times, and the second and the third individual five times. If we look at the individual data used for the purpose of this example, we quickly establish that this would result in assigning 18 males, 10 females, 14 people aged 18-64, and 14 people aged over 65 to zone 4. This distribution perfectly matches the observed counts presented in Table 27. The hypertensive individuals assigned by the algorithm to this zone are 14. The third zone, which is part of the same spatial unit at Tier 1, has a weight sum of 2.4 + 1.6 = 4 individuals with hypertension. Thus, the total count of simulated individuals with hypertension for the second zone in Tier 1 equals 18. The corresponding observed count is 15. There are two reasons for this outcome. First, the algorithm finishes with the reweighting at the second tier, which means that the fit will always be better at the lower spatial scale, as it should be because the aim is to generate synthetic population at the smaller scale. Since hypertension is available as constraint only at Tier 1, it has a smaller influence on the final distribution compared to the constraint variables available at Tier 2. Second, the number of individuals included in this example was too small. The lacking diversity of individual characteristics therefore does not allow the algorithm to achieve a perfect fit at both spatial scales. In contrast, for the purposes of the actual model, 5.853 unique[22] individuals both in terms of constraint and target variables characteristics were fed as input data.

All in all, this is how the two-tier IPF algorithm operates. To put this code together, I used two main references: the extensive work of Lovelace and Dumont (2016) on the subject of spatial microsimulation with R, and more specifically the chapter about Population synthesis and the IPF in theory; and the paper of Konduri et al. (2016), which provides insight into adopting a reweighting approach at multiple spatial scales for the purpose of population synthesis.

[22] With unique combinations of individual characteristics

6.4.3. Integerisation and Expansion

While the IPF represents the core of population synthesis, the generated fractional weights are practically useless on their own. Allocating individual 'X' 1,2 times and individual 'Y' 0,8 times to zone 'Z', for instance, would not result in allocating two *whole* individuals, as each of them has differing characteristics. To generate a synthetic population, the weights must therefore be converted from decimal numbers into integers. This procedure is called *integerisation*.

One possible method to integerise the generated fractional weights is to simply round them up. Nevertheless, this would imply that a weight of 0,51 and a weight of 0,99 are treated the same way and would thus result in losing information detail. There is another problem related to simply rounding decimal weights as well. Assuming that the integer is rounded up when the decimal remainder $\geq 0{,}5$, or else rounded down, individuals with a weight $< 0{,}5$ will end up not being included in the sample at all. This causes a significant problem in cases when the IPF algorithm generates weights smaller than 0,50 only, which is not necessarily an exception. Such cases depend on the ratio of the observed counts for each constraint category to the corresponding population count in the micro dataset. If, for instance, there are only 10 males in zone 'Z', but 100 males in the micro dataset, the generated weight for this constraint category will be 0,1. The latter is naturally going to be updated to account for all constraint categories. Still, the basic notion is that when computing weights for a zone with a total population much smaller than the individuals available in the micro dataset, these are generally going to be smaller than 1. In such cases, population synthesis may fail to produce adequate results when implementing a crude integerisation technique such as simple rounding.

Against this background, scholars in the field of spatial microsimulation have developed more sophisticated solutions to this problem. One of them relies on introducing a so-called '*inclusion threshold*'. The latter is initially '*set to 1 and then iteratively reduced (by 0.001 each time), adding extra individuals with incrementally lower weights*' (Lovelace and Ballas 2013, p.4). Below the exit value of this threshold in each zone, no more individuals can be included.

Another integerisation method, referred to as '*counter-weight approach'* sorts the individuals in ascending order according to their weight and then assigns them a counter. Then, an algorithm iterates over all individuals in the order of their counter and computes a new integer weight, which '*is set as the rounded weight plus the rounded sum of its decimal weight plus the decimal weight of the next individual, until the desired total population is reached'* (ibid.). The advantage of this solution, compared to the former one relying on an inclusion threshold, is that individuals with smaller weights of down to 0,25 may be selected, or, in other words, integerised to 1. Still, both approaches assume that an individual with a weight of 0,2 and another individual with a weight of 0,0001 have the same chance to be selected – that being no chance at all. Failing to include individuals with lower weights compromises the diversity of the sample. Those with a less common combination of personal characteristics are completely excluded, whereas those, who are more representative for a certain zone, are over-replicated. Furthermore, an individual with a weight of 0,2 is 2.000 times more representative than an individual with a weight of 0,0001. Therefore, putting them both into the group of individuals, who will never be selected, means oversimplifying the integerisation procedure.

The '*proportional probabilities approach'* to integerisation addresses the problem of the total exclusion of weights below a certain value. It relies on probabilistic selection sampling, where

the likelihood of being selected depends on the ratio of the weight to the total sum of weights. Thus, individuals with high weights should be replicated more times, and individuals with low weights should be replicated fewer times or not appear at all. While this does sound logical, the problem is that '*because all weights are treated as probabilities, there is non-zero chance that an individual with a low weight (e.g., 0.3) is replicated more times than an individual with a higher weight (e.g., 3.3)*' (Lovelace and Ballas 2013, p.5). Moreover, if all weights are lower than zero, the choice of who gets selected becomes relatively random.

Against this background, Lovelace and Ballas (2013) introduced a new integerisation method, referred to as '*Truncate, replicate, sample'* (TRS). They argue that the ideal method '*would build upon the simplicity of the rounding method, select the correct simulated population size (as attempted by the threshold approach and achieved by using 'proportional probabilities'), make use of all the information stored in IPF weights and reduce the error introduced by integerisation to a minimum*' (ibid., p.5). The developed TRS integerisation method therefore aims to integrate the strengths of the different approaches introduced above. It is based on the notion that the generated weights do not simply represent the odds of an individual to be selected. Rather than that, they indicate how many times an individual should be replicated in a certain zone (provided that the weight is higher than 1). If, on the other hand, the weight is smaller than 1, it then represents the probability of the individual being selected in a representative sampling strategy, as in the case of the proportional probabilities approach. Hence, instead of interpreting the IPF weights as inaccurate count data, the TRS approach considers them to be a complex information source about both replication and selection probability. As its name suggests, the method is comprised of three separate steps. The first one – *truncate* – consists in cutting the decimal remainder to the right of the decimal point and thus keeping the integer part, which determines how many times the individual should be replicated in the given zone. The second step – *replicate* – refers to the replication of the individuals according to the integer weights defined in the previous step. Nevertheless, these two steps alone are not sufficient for generating a synthetic population with the same size as the observed one. Since the replication is based on integers, which were *truncated* rather than *rounded up*, the simulated population is always going to be smaller than the real one at this point. Therefore, the final step – *sample* – ensures that the desired population size is achieved by adding more individuals. The sampling is carried out based on selection probabilities equal to the decimal remainders cut in the first part of the TRS algorithm (ibid., pp.5-6).

To integerise the fractional weights generated by the two-tier IPF algorithm for the purposes of this dissertation, I used the TRS integerisation algorithm available in the book 'Spatial Microsimulation with R' by Lovelace and Dumont (2016, p.91):

```
int_trs <- function(x){ # the TRS integerisation algorithm int_trs is defined as function
    xv <- as.vector(x) # The IPF-generated weight matrix is fed as x and converted into a vector xv
    xint <- floor(xv) # The integer part of the weights is extracted and stored into a new vector xint
    r <- xv - xint # Then, the decimal remainder of the weights is stored into a new vector r
    def <- round(sum(r)) # The deficit population is defined as the total sum of the decimal remainders. Deficit population refers to the population that must be filled in after replicating the individuals based on the integer part of the weights to reach the total observed population count.
    # To that end, the weights must be 'topped up' (+ 1 applied: e.g., 1,6 becomes 2,0)
    topup <- sample(length(x), size = def, prob = r) # to select weights for topping up, a sample with the length of the vector x containing the weights, and the size of the deficit population is generated. The probability for a weight to be selected from this sample depends on its decimal remainder.
    xint[topup] <- xint[topup] + 1 # the initially extracted integer parts of the weights selected in the previous step are topped up.
    dim(xint) <- dim(x) # the vector xint, which contains the integerised weights, is converted into a matrix with the same dimensions as the IPF-generated weight matrix.
    dimnames(xint) <- dimnames(x) # the newly created integer matrix xint is assigned the same row and column names as those of the IPF-generated weight matrix x.
}
```

Here is a TRS integerisation example with five random weights:

0.90	1.20	2.40	3.60	4.90

First, we store the integer part of the weights – *Truncate* →

0	1	2	3	4

Next, we store the decimal remainders →

9	2	4	6	9

The total observed population is *filled up* when the fractional weights are summed up:

$$0.90 + 1.20 + 2.40 + 3.60 + 4.90 = 13$$

If we use the integer parts of the fractional weights to replicate the individuals, we get a total population of:

$$0 + 1 + 2 + 3 + 4 = 10$$

This means that there is a deficit population of three people that must be filled in by topping up three weights. Obviously, the decimal remainders 0.9 and 0.6 have higher odds to be selected

than 0.2 and 0.4. The TRS algorithm is therefore most likely going to choose the weights 0.90, 3.60, and 4.90 for topping up. The integerised version of the IPF-generated fractional weights will thus most probably look like this:

```
1  1  2  4  5
```

This is how the TRS approach, which I adopted for integerising the fractional weights generated by the two-tier IPF algorithm, works.

With that, there was just one final step left to generate the synthetic population and it is widely known as *expansion*. In essence, each integerised weight corresponds to the number of times the individual should be replicated in the given zone. With this in view, the process of expansion consists in replicating the individuals as many times as the integerised weights. To that end, the following function was defined by Lovelace and Dumont (2016, p.95):

```
int_expand_vector <- function(as.vector(x)){ # the matrix containing the integerised
    weights is fed into the function and converted into a vector
    index <- 1:length(x) # an index with the length of the vector is created so that each
    integer is assigned a unique index
    rep(index, round(x)) # each index is replicated as many times as the integer weight it
    refers to
}
```

To simplify this with an example, once a vector containing the following integers

```
[1] 1 2 1 5
```

is expanded, it will look like this:

```
[1] 1 2 2 3 4 4 4 4 4
```

Each index in this vector refers to a certain row in the individual dataset – a row, which represents an individual with unique characteristics. Hence, this newly generated *expanded* vector is used to index the rows in the micro dataset as many times as necessary for populating a given zone. Basically, entire rows from the micro dataset including both the constraint and target variables are extracted and allocated to a new synthetic population dataset and replicated if needed. The big advantage of the synthetic population dataset is that it contains both individual data and spatial reference about the city quarter and the statistical area each individual lives in. With that final step, the population synthesis was completed (see Appendix, Table 44, for a sample of the generated synthetic population).

6.4.4. Choice of software environment

While software choice is barely addressed in papers on the subject of spatial microsimulation, I consider it an important aspect for researchers aspiring to get to know this method first-hand. The next few paragraphs are therefore dedicated to my journey regarding the selection of programming language and software environment.

Naturally, I was looking for a programming language, which is appropriate and *mature* for the intended purpose of setting up a spatial microsimulation model. Before I embarked upon the path of studying spatial microsimulation methods, I barely had experience with programming languages. Choosing an intuitive software, which would allow me to learn fast, was therefore essential. Another important aspect in my decision-making process was cost. For flexibility reasons, I preferred to use an open-source programme. Furthermore, open-source software products generally offer the advantage of a large user community sharing their knowledge and providing solutions to various programme-specific problems on the web.

With this in view, I chose the software environment for statistical computing *R* (R Core Team 2020). It ended up meeting all my needs as a researcher, who was already experienced in the field of statistics, but less so in the field of programming, and spatial microsimulation. I used R throughout the entire process – input data pre-processing, writing the two-tier IPF algorithm, integerising the generated fractional weights, applying expansion to convert the integers into actual synthetic population, and finally carrying out internal and external model validation (to be addressed in the next chapter).

At the same time, I profited enormously from an already existing body of work on the topic of spatial microsimulation with R. Especially the book of Lovelace and Dumont (2016) served as my navigation into this fascinating field. Fortunately, R offers great flexibility and enabled me to take parts of the code, available in the book for teaching purposes, to adjust and expand them in order to finally set up the algorithm I needed for reaching my research goal.

Since R was initially designed to operate for statistical purposes, many functions necessary for statistical analysis are already available in the default installation package. Hence, users do not need to define them as functions. In the context of population synthesis, carrying out a statistical analysis of the results is just as important as generating the individuals. Therefore, the integrated statistical analysis features of R were another factor, which influenced my decision-making.

Being introduced back in 1993, R's functionalities have grown enormously due to the continuously expanding community of programmers contributing to the software environment. As a result, there are already predefined functions enabling to run the IPF algorithm without having to 'hard-code' it by yourself. Examples of such functions are *ipfp* and *mipfp*, which are available for installation in additional packages with the same names. With this in view, R offers even more comfort for novices in the field of deterministic reweighting, and with that in the field of spatial microsimulation models. Personally, I found it more rewarding to hard-code each part of the algorithm instead of using these 'black-box' functions, as this allowed me to develop an understanding of how IPF works. Nonetheless, having multiple options to select from was yet another advantage that made R the right choice for me.

With the generation of the synthetic population, the main task towards setting up a small-scale health model for Hamburg was completed. Still, there was one crucial step left – the model validation. The next chapter is going to explain its purpose and address the difference between internal and external validation. Moreover, it is going to present the results from validating the generated model so that the extent to which it manages to depict reality can be evaluated.

7. MODEL VALIDATION

Model validation is an integral part of applying spatial microsimulation because it serves the purpose of evaluating the reliability of the generated synthetic data. In general, there are two types of validation – internal and external.

Internal validation is also referred to as '*model checking*' and it consists in '*comparing model results against a priori knowledge of how they should be*' (Lovelace and Dumont 2016, p.143). This generally means that the modelled data is aggregated at a chosen level of spatial division (e.g., city quarters and statistical areas) and compared with the observed population counts from the geographic dataset used for setting up the model. Therefore, only the constraint variables can be *checked*, but not the targets.

External validation, on the other hand, also known as *model evaluation*, uses data external to the model for its verification. Depending on the external data available, this process can be executed at the aggregate and/or at the individual level. Nevertheless, the main reason for setting up spatial microsimulation models is there being no data available at the desired spatial level. Therefore, external validation often cannot be carried out at all. According to Lovelace and Dumont (2016), '*in such cases internal validation, combined with proxy variables for which external datasets are available, may be the best approach to model evaluation*' (p.144).

Internal and external validation have different purposes. While internal validation tests the coherency of the model and reveals errors in the algorithm or discrepancies in the input data, external validation determines the extent to which the modelled data corresponds to reality.

If the internal validation fails, there may be a problem with the input micro data. Common problems include the excess of empty cells or the insufficient representativity of the survey sample. The latter is often manifested as lacking combinations of individual characteristics impeding an accurate estimation of the population composition. Another possible reason for a failed internal validation may be an error in the algorithm or the use of contradictory constraints (e.g., two dichotomous variables related to both employment and unemployment with differences in the definitions). There may also be a discrepancy in the total number of individuals for the various constraint variables in the geographic dataset (ibid., pp.144-145). In any case, internal validation results show whether the input data, or the algorithm need to be re-examined. If the differences between the modelled and the observed population aggregates for the constraint categories (e.g., total number of males) are minimal, the model is generally considered coherent. Against this background, internal validation is considered '*the bare minimum in terms of model evaluation*' (ibid., p.145). It should be regarded as standard procedure and thus carried out for each spatial microsimulation model because it requires only the used input data.

External validation, on the other hand, relies on data *external* to the model, which was not used as input to set it up. Hence, there may be plenty of reasons for a failed external validation. This makes fixing the model much harder compared to simply correcting the algorithm or checking the input data for inconsistencies. Nevertheless, if external data is available, regardless of its form and geographic coverage, it should be used for model evaluation. While internal validation may help to rule out faulty methods, external validation provides actual insight into how good of a proxy the modelled data is (ibid., p.144).

With this in view, I carried out both internal and external validation. The next sections are going to present the results and describe in detail the applied methods and the specifics of the used external data.

7.1. Internal Validation Results

For the purposes of internal validation, it is generally recommended to use several different metrics. I therefore estimated Pearson's correlation, the total absolute error (TAE), the relative error (RE), the root mean squared error (RMSE), and the mean absolute percentage error (MAPE) for each zone at the city quarter and at the statistical areas level – both before and after integerisation. My intention was to combine several differing estimation approaches, which are transparent, easy to interpret, and simple enough to communicate to a wider audience. Further metrics, which can generally be used for the purpose of internal validation, include the mean relative error and Chi-squared.

Pearson's Correlation (*r*) quantifies the linear correlation between the modelled and the observed population counts for each variable category in each zone. It is estimated using Formula 1:

Formula 1. Pearson's Correlation (Source: Lovelace and Dumont 2016, p.146)

$$r = \frac{s_{XY}}{S_X S_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2}}$$

Here, `X` and `Y` correspond to the observed and the modelled matrices (accounting both for zones and variable categories), which are converted into vectors for the purpose of this estimation. The formula divides the covariance by the product of the standard deviation of each vector. Thus, the covariance is standardised.

The better the fit between the observed and the modelled values, the closer r would be to 1. If the values match perfectly, the covariance will be equal to the product of the standard deviations. Generally, '*r values greater than 0.9 should be sought in spatial microsimulation and in many cases r values exceeding 0.99 are possible, even after integerisation*' (Lovelace and Dumont 2016, p.146). The main flaw of Pearson's Correlation is its high sensitivity towards outliers.

TAE is the sum of the absolute differences between the modelled and the observed population counts for each constraint variable category in each zone. It is estimated using Formula 2

Formula 2. Total Absolute Error (Source: Lovelace and Dumont 2016, p.147)

$$TAE = \sum_{ij} |e_{ij}|$$

where $e_{ij} = obs_{ij} - sim_{ij}$.

Here, e refers to error, whereas obs and sim represent the observed and simulated (or modelled) values for each zone (i) and each constraint variable category (j), respectively.

It is important to emphasize on the *absolute* part of this measure, meaning that it takes the absolute values of the error so that differences cannot be compensated. For instance, having

10 males fewer, but 10 females more in zone Z, would result in an error of 20, rather than 0, as TAE counts each difference between observed and modelled cases. Therefore, albeit not particularly refined, the simplicity of this measure facilitates its comprehension. Still, the interpretation of the results can be difficult because the calculated error is a number that cannot be related to anything in order to estimate whether TAE it is large, small, expected, or unexpected.

RE is closely related to TAE as it is calculated by dividing TAE with the population of the respective zone. If the results are compared for all constraint variables altogether, the total population must be additionally multiplied by the number of variables (Formula 3). RE may therefore be viewed as the percentage of error. Its main advantage over TAE is that it is not sensitive to the number of individuals and variable categories included in the model.

Formula 3. Relative Error (Source: Lovelace and Dumont 2016, p.147)

$$RE = \frac{TAE}{total\_pop * n\_var}$$

RMSE is also partly based on TAE. However, it relies on squaring the errors rather than simply adding them up. Basically, RMSE estimates the square root of the mean of all squared errors (Formula 4). Thus, larger differences in the fit between the observed and the modelled population become evident more easily. According to Chai and Draxler (2014), when errors have an approximately normal distribution RMSE is the more suitable choice than TAE and RE.

Formula 4. Root Mean Squared Error (Source: Lovelace and Dumont 2016, p.147)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n} e_i^2}$$

MAPE is the '*mean or average of the absolute percentage error of forecasts*' (Swamidass 2000). It is calculated by estimating the average of the sum of the absolute error for all zones and constraint variable categories divided by the corresponding observed population counts (Formula 5).

Formula 5. Mean Absolute Percentage Error (Source: Stellwagen 2019)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{obs_i - sim_i}{obs_i}\right| * 100$$

The biggest advantage of MAPE is that it is easy to comprehend because the error is presented as percentage, as opposed to an abstract number that cannot be related to anything for the sake of interpretation (as in the case of TAE and RMSE). Therefore, MAPE is commonly used in forecasting: the smaller MAPE is, the better the forecast fit. For an even simpler understanding as to how good the modelled cases depict reality, one can subtract MAPE from 1 and thus calculate something along the line of a *Fitting Score*, that is, the extent to which the modelled cases fit the observed cases. If MAPE = 5%, for instance, that would result in a Fitting Score of 95%.

One downside of MAPE is that it is not suitable in cases when the observed value equals zero. Therefore, if the geographic dataset contains areas, where the observed population count for a certain constraint variable category is zero, MAPE cannot be applied for estimating the model

fit. Still, there are ways to work around such instances. The simplest option is to transform the given constraint variable by combining multiple variable categories into one and thus increase the observed population count. In an age-related constraint variable, for example, 5-year age intervals may be converted into 10- or even 20-year intervals. Such an approach would lead to losing some detail, but it would enable the estimation of MAPE. Nevertheless, some variables do not allow for such transformations. Dichotomous categorical variables, accounting for either/or conditions, such as having or not having certain medical condition, entirely exclude the possibility of variable transformation. In such cases, the only other option for using MAPE to estimate the model fit, would be spatial unit aggregation, that is, combining several smaller areas into one. Such an approach, however, could raise questions regarding the underlying logic of the spatial aggregation – is it based on administrative boundaries, such as combining urban blocks based on neighbourhood delineations, is it based on reaching a predefined population count, etc. In any case, such transformations will always lead to losing detail.

Before exploring the fit between the observed and the modelled cases, I examined the extent to which the modelled cases for each constraint variable category coincide in their sum at both modelling tiers. For example, I checked whether the number of males allocated to the statistical areas within city quarter *A* (cases modelled at Tier 2) equals the number of males allocated to city quarter *A* (cases modelled at Tier 1), and so on. Only those constraint variables, available at both modelling tiers (age and gender), were considered because only they determine the fit. The results of this test are presented in Table 28.

Table 28. Testing the convergence between Tier 1 and Tier 2 (own representation)

	Min	1st Quarter	Median	Mean	3rd Quarter	Max
r	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
TAE	0	0	0	0	0	0
RE	0	0	0	0	0	0
RMSE	0	0	0	0	0	0
MAPE	0	0	0	0	0	0
Fitting Score	100%	100%	100%	100%	100%	100%

All metrics indicate a perfect fit between the two tiers, that is, the individual counts regarding age and gender fully coincide at the city quarter and the statistical areas level. Simply put, the sum of all males allocated to statistical areas, belonging to a certain city quarter, equals the number of males allocated to the same city quarter. The same goes for females as well as for the two age intervals: 18-64 years and 65+ years. This outcome suggests that the IPF algorithm, which constrained the model at two different spatial scales, did not cause any discrepancies in the population counts belonging to corresponding constraint variable categories.

The internal validation results for both modelling tiers, considering the observed and modelled cases for all constraint variable categories, are summarised in Table 29. The five metrics are compared before and after the integerisation of the fractional weights generated by the IPF algorithm. Thus, the extent to which integerisation negatively affects the fit, becomes evident.

Table 29. Internal validation results (own representation)

	Tier 1: city quarters					
	Before integerisation			After integerisation		
	Min	Mean	Max	Min	Mean	Max
r	1,0000	1,0000	1,0000	0,9996	0,9998	1,0000
TAE	0,1641	3,1963	14,8555	363	2236,6	7124
RE	~ 0,0000	~ 0,0000	~ 0,0000	~ 0,0000	~ 0,0000	~ 0,0000
RMSE	0,0091	0,1776	0,8120	12,89	72,95	228
MAPE	~ 0,0000	~ 0,0000	~ 0,0000	2,6%	4,5%	7%

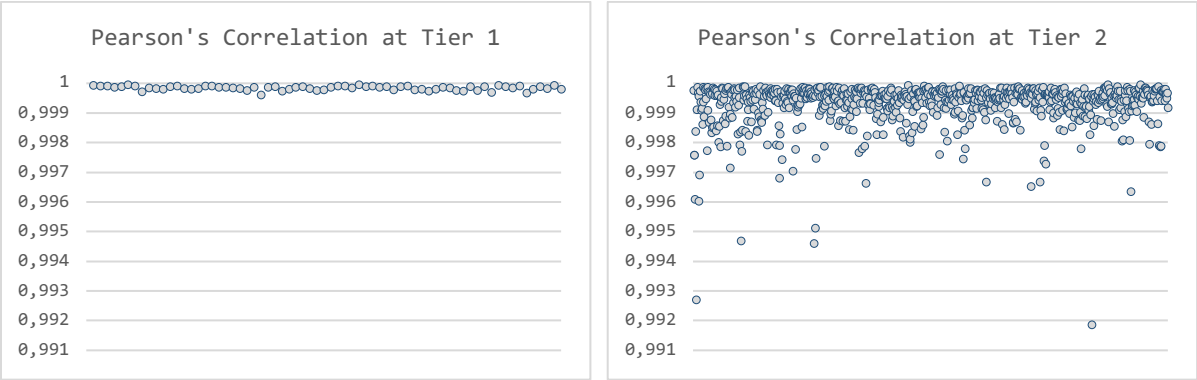
	Tier 2: statistical areas					
	Before integerisation			After integerisation		
	Min	Mean	Max	Min	Mean	Max
r	1,0000	1,0000	1,0000	0,9919	0,9993	1,0000
TAE	~ 0,0000	~ 0,0000	~ 0,0000	52	198,7	356
RE	~ 0,0000	~ 0,0000	~ 0,0000	1,3%	4,1%	14,2%
RMSE	~ 0,0000	~ 0,0000	~ 0,0000	2,45	9,36	20,35
MAPE	0,0000	0,0000	0,0000	2,6%	8,6%	29,3%

Looking at the results, it becomes clear that non-integerised weights provide a better fit between the observed and the modelled cases at both tiers. However, whereas the model fit at the level of the statistical areas is perfect before integerisation, this is not the case at the level of the city quarters. The reason for this is that the algorithm starts constraining the synthetic population at the city quarter level but finishes at the statistical areas level. Since the constraint variables are different at both tiers, optimising the fit at the smaller scale leads to a slightly worse fit at the larger scale. Still, non-integerised weights are useless on their own, as they do not allow to carry out the expansion necessary for generating synthetic population. I am therefore going to focus on interpreting the results about the model fit after integerisation.

Generally, Pearson's Correlation, RE, and MAPE point to a better fit at the first tier, whereas TAE and RMSE suggest more satisfying convergence at the second tier. To explore this outcome in more detail, the distributions of the metrics after integerisation are illustrated in figures 8-12.

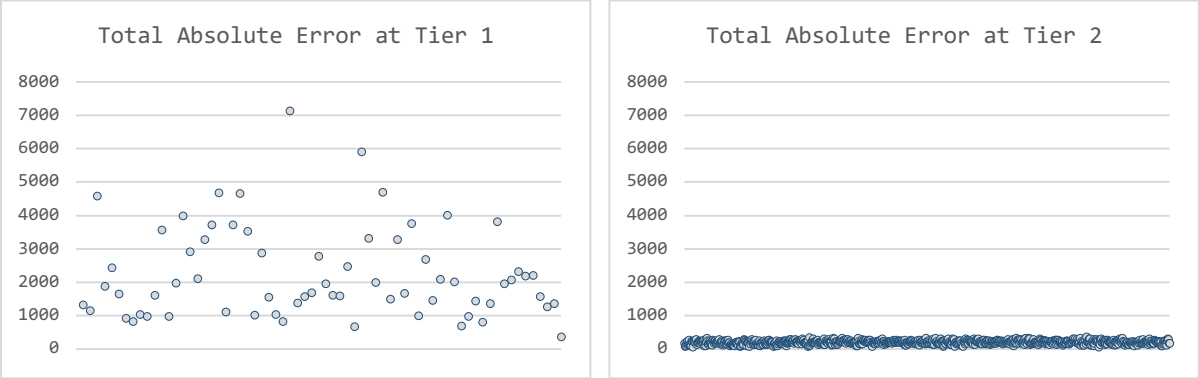
Overall, the metrics point to a satisfying fit at both modelling tiers. Pearson's Correlation is at least 0,991 after integerisation, which speaks for a remarkably close convergence between observed and modelled values. The results are slightly better at the level of the city quarters with all zones being in the range 0,999 – 1,000. In contrast, approximately 10% of the statistical areas fall below 0,999 (Figure 8). Still, those are minimal differences, and altogether this metric suggests more than an adequate fit between observed and modelled cases at both spatial scales.

Figure 8. Pearson's Correlation regarding the convergence between observed and modelled population at both spatial tiers (own representation)



TAE is significantly larger at the first modelling tier – more than 80% of all city quarters have a TAE of at least 1.000, whereas the statistical areas have considerably lower TAE within the range of 50 to 350 (Figure 9). This difference results mostly from the metric's sensitivity to the number of individuals per zone. The city quarter with largest TAE (= 7.124) has a total population of over 50.000 people – significantly larger than the average population of ca. 23.000 people at this spatial scale. In fact, there is a positive correlation of 0,81 between population size and TAE at the city quarter level. At the level of the statistical areas, on the other hand, the correlation is weaker: 0,62. In this context, the larger variation in population size at the city quarter level (min = 2.871, max = 75.647) as opposed to the statistical areas level (min = 223, max = 5.761), supports the assumption that TAE tends to be larger for more populous areas. Since city quarters generally encompass several statistical areas, their populations are naturally larger, which is why TAE is larger as well.

Figure 9. Total absolute error regarding the convergence between observed and modelled population at both spatial tiers (own representation)



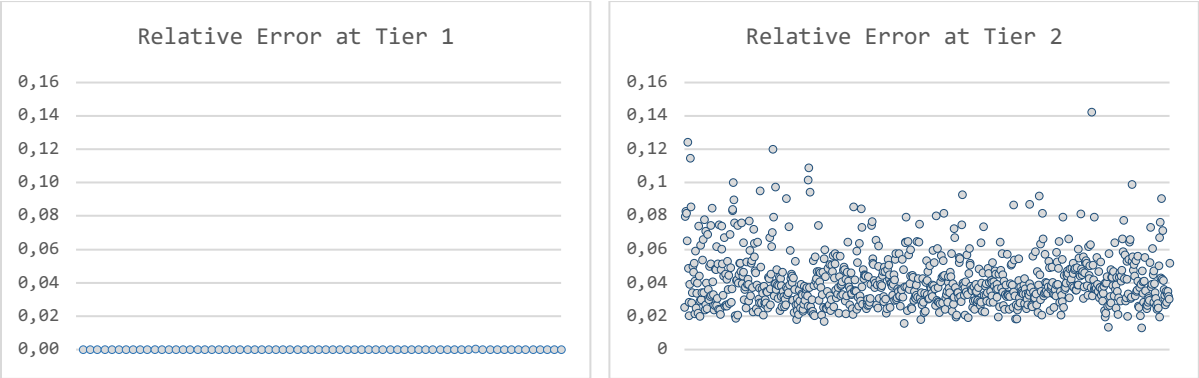
TAE also depends on the number of variable categories. At the first modelling tier, there are 40 categories, whereas at the second, there are 30. The larger TAE at the city quarter level is therefore not difficult to grasp.

Another contributing factor may be the number of missing individual combinations in the micro dataset, impeding the algorithm from reaching the perfect fit. The total number of possible combinations at the first modelling tier is 128. Out of those, 102 are available in the micro dataset. This is not necessarily problematic, as cases of people having all five chronic diseases serving as constraints at this spatial scale, i.e., hypertension, heart failure, diabetes, cancer, and depression, shall be close to zero. The latter being simply an example for combinations,

which are not that necessary as they do not manifest that often in reality. Still, this 'deficiency' of the micro dataset may be another reason for higher TAE at the city quarter level.

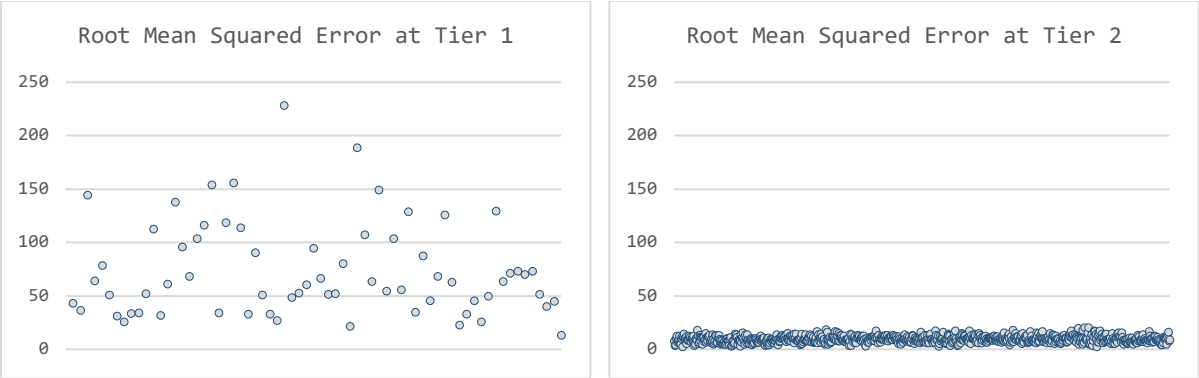
Looking at RE, the picture is different – all city quarters have a relative error of approximately zero. At the level of the statistical areas, on the other hand, RE falls into the range of 0,01 – 0,14 (Figure 10). Hence, the percentage of error is clearly larger at the second modelling tier. In other words, the *error per person* is larger at the scale of the statistical areas. This results from a larger number of smaller differences, which is revealed after *controlling* for population size and number of constraint variable categories.

Figure 10. Relative error regarding the convergence between observed and modelled population at both spatial tiers (own representation)



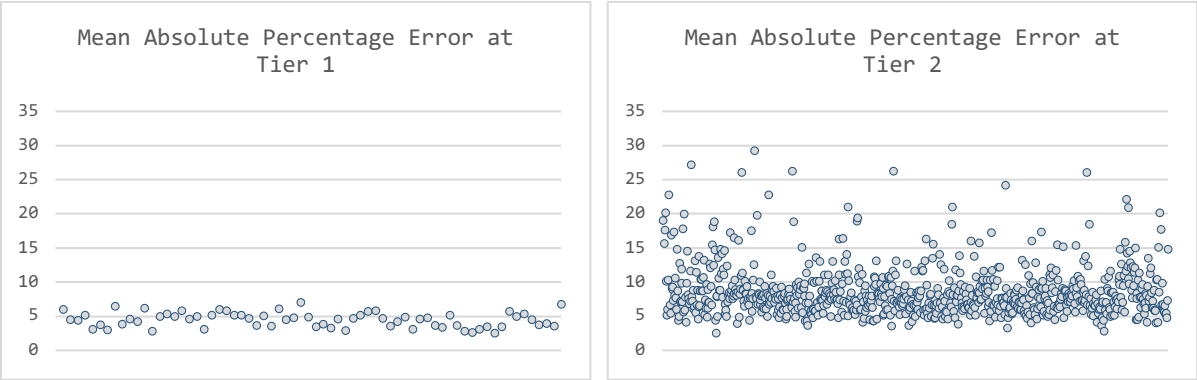
RMSE is larger at the city quarter level than at the statistical areas level (Figure 11). At the first tier, all but one city quarters have a RMSE > 20, whereby one quarter have a RMSE > 100. At the second tier, on the other hand, all areas have RMSE ≤ 20. With this in view, RMSE at the first tier has a wider range and a broader distribution than at the second tier. This suggests that differences between observed and modelled cases may be fewer, but larger at the city quarter level than at the statistical area level. As already discussed, TAE follows a similar pattern and thus supports this assumption. The main reasons for this are the larger number of constraints and the missing combinations in the micro dataset, impeding a perfect fit.

Figure 11. Root mean squared error regarding the convergence between observed and modelled population at both spatial tiers (own representation)



MAPE, similarly to RE, is higher at the level of the statistical areas (Figure 12). While the minimum MAPE is the same at both modelling tiers (=2,6%), the maximum is more than four times larger at the scale of the statistical areas. The MAPE therefore reinforces the assumption that factoring in population size, the error is larger at the second modelling tier.

Figure 12. Mean absolute percentage error regarding the convergence between observed and modelled population at both spatial tiers (own representation)



Overall, the internal validation results suggest a satisfying fit between the observed and the modelled populations at both spatial scales – the city quarters and the statistical areas. Pearson's Correlation is higher than 0,99 even after integerisation for all spatial units, which suggests an almost ideal convergence. The remaining four metrics – TAE, RE, RMSE, and MAPE point to an existing pattern in the differences between the two modelling tiers. There seem to be fewer, but larger differences between the observed and modelled cases at the city quarter level, leading to higher TAE and RMSE. The lack of certain individual combinations in the micro dataset appears to play the pivotal role here. In contrast, RE and MAPE are higher at the statistical areas level, suggesting a greater *error per person* due to a larger number of smaller differences.

7.2. External Validation Results

While good internal validation results verify the coherency of the model, they do not confirm that the generated data accurately describes reality. To that end, external validation is necessary. For this purpose, representative samples of the population in the small areas are recommended (Lovelace and Dumont 2016, p.159).

To conduct external validation of my spatial microsimulation model, I used two data sources – a small representative survey for six statistical areas with differing social status; and total population counts regarding the prevalence of several chronic illnesses categorised by age, gender, and social status, provided by three of Hamburg's health insurance funds: AOK Rheinland/Hamburg, BKK Mobil Oil, and DAK-Gesundheit. Both the sample survey data and the health insurance data were obtained in the course of the research project 'Healthy Neighbourhoods'.

Before diving into the external validation, it should be noted that neither of the available external data sources was perfect: the survey data only covered a limited number of statistical areas, whereas the insurance data completely lacked geographical reference. Nevertheless, I consider the use of this data a better alternative than skipping external validation altogether. Moreover, if there already was small-scale data about the prevalence of chronic disease available for the entire city of Hamburg, there would have been no demand for setting up this model in the first place.

7.2.1. External validation with survey data

An integral part of the research project ‘Healthy Neighbourhoods' was primary data collection through standardised sample surveys. The aim was to compare the health situation across Hamburg’s neighbourhoods in terms of socio-economic standing. To that end, individual health-related data was obtained and analysed for six statistical areas with differing social status according to the classification of the Social Monitoring.

The six statistical areas were selected randomly, after applying several filters upfront: population of at least 2.000 inhabitants, stable social dynamic (as in Social Monitoring 2016), and no health-related projects with a funding exceeding 10.000 EUR per annum conducted in the three years prior to the research project ‘Healthy Neighbourhoods’. The selection was carried out in four separate clusters, each representing a different social status: high, average, low, and very low. Of the six selected areas, one had high social status, one had average status, two had low status, and the remaining two had very low status. Part of the research project was dedicated to the development and implementation of health-promoting and prevention measures in two study areas belonging to each of the low social status classes. One of the areas selected within the classes *low* and *very low* was therefore intended as *implementation area* and the other one as *control area* (Yosifova 2021, pp.50–51).

At the start of the project in July 2017, the population size in the study areas was within the range of 2.200 – 6.200 inhabitants (Statistisches Amt für Hamburg und Schleswig-Holstein 2017) and was thus considered sufficient for reaching the intended sample of 150 survey respondents per area. It took a bit longer than one year to conduct the interviews – from May 2018 until July 2019. However, it proved quite difficult to reach the desired sample in all six areas, and data collection was thus terminated after the completion of 815 interviews. Following the inspection of the input data for mistakenly checked boxes, misstatements, typing errors, etc., 799 of those were included in the final dataset (Buchcik et al. 2021, p.54).

The survey covered three main topics of interest: individual health, health behaviour and health competence, and living environment – the focus being on the *neighbourhood* setting. Some, albeit not so many of the variables available in the dataset, directly corresponded to the variables I modelled at the level of the statistical areas using a spatial microsimulation. This allowed the external validation of several of the modelled individual health characteristics, including hypertension, heart failure, diabetes, cancer, depression, obesity, overweight, and smoking. Furthermore, the survey contained variables, which I was able to use as *proxy* for validating sporting activity, subjectively perceived health, and impairment in daily activities due to illness.

Having access to external data is essential for model evaluation. Nevertheless, any possible limitations, or flaws of this data must be taken into consideration before issuing the final statement how good the model is. The available external data may not be as good of a representation of reality either. This is especially the case with survey data, as there is almost always some kind of bias. Whenever possible, this must be accounted for by applying additional weight variables. With this in view, Table 30 illustrates the marginal distribution of the sampled population for age and gender (in absolute and relative terms) compared to the corresponding marginal distribution of the observed population in the six statistical areas just before the start of the primary data collection (Statistisches Amt für Hamburg und Schleswig-Holstein 2018).

Table 30. Observed vs. sample cross tabulation: gender by age (own representation)

			18-64		65+	
			Count	% of Total Count	Count	% of Total Count
Statistical Area #66004 in Sasel	Females	Sample	47	32,2%	28	19,2%
		Observed	663	33,1%	449	22,4%
	Males	Sample	44	30,1%	27	18,5%
		Observed	609	30,4%	284	14,2%
Statistical Area #43010 in Stellingen	Females	Sample	79	56,4%	8	5,7%
		Observed	884	35,2%	459	18,3%
	Males	Sample	47	33,6%	6	4,3%
		Observed	866	34,5%	303	12,1%
Statistical Area #9005 in Hamm	Females	Sample	56	42,1%	12	9%
		Observed	813	38,9%	160	7,7%
	Males	Sample	57	42,9%	8	6%
		Observed	979	46,9%	137	6,6%
Statistical Area #75019 in Lohbrügge	Females	Sample	55	49,1%	13	11,6%
		Observed	1.358	38,7%	399	11,4%
	Males	Sample	39	34,8%	5	4,5%
		Observed	1.473	41,9%	283	8,1%
Statistical Area #16023 in Wilhelmsburg	Females	Sample	34	58,6%	3	5,2%
		Observed	1.844	41,4%	452	10,2%
	Males	Sample	21	36,2%	0	0%
		Observed	1.780	40%	376	8,4%
Statistical Area #74024 in Rahlstedt	Females	Sample	66	48,9%	13	9,6%
		Observed	847	40,5%	228	10,9%
	Males	Sample	44	32,6%	12	8,9%
		Observed	835	39,9%	182	8,7%

Clearly, several adjustments were necessary. The differences between the sample and the observed distribution for some population groups were considerable. This was especially true for the relative distribution of age, as the survey only included adults. The sample distribution was therefore naturally distorted when compared to the observed one, which considered the entire population.

To compute the weights for adjusting the marginal distribution in the sample, the observed percentage of the total count for a given category had to be divided by the corresponding percentage in the sample. For instance, the weight for a male, older than 65, and living in Sasel would be 14,2/18,5 = 0,77. Since the proportion of such individuals in the sample population was larger than in reality, it had to be scaled down to improve the representativity of the sample. Naturally, weights smaller than 1 serve the purpose of scaling down, whereas weights larger than 1 do the opposite.

While this is how the general procedure works, Table 30 shows that there are no males, older than 65, and living in Wilhelmsburg in the sample. It was therefore not possible to validate the modelled characteristics for this population group using the survey in its original format. The solution to overcoming this problem was to merge the data for all six statistical areas. In so doing, the missing male respondents older than 65 years in Wilhelmsburg were compensated

for by the same type of individuals living in the other five areas. Thus, despite the loss of spatial detail, I was still able to use the survey data for model validation. Merging the data offered another advantage as well – it resulted in one bigger sample instead of six small ones, which increased the reliability of the data as it eliminated empty categories and reduced the number of categories with few individuals. Table 31 provides a comparison of the cross-tabulation *gender* by *age* between the merged sample population and the observed population. To compensate for the loss of spatial detail, I divided age into three instead of two categories.

Table 31. Observed vs. merged sample cross tabulation: gender by age (own representation)

		18-44		45-64		65+	
		Count	% of Total Count	Count	% of Total Count	Count	% of Total Count
Females	Sample	204	28,2%	133	18,4%	77	10,6%
	Observed	3.749	22,5%	2.660	16,0%	2.147	12,9%
Males	Sample	145	20,0%	107	14,8%	58	8,0%
	Observed	3.924	23,5%	2.618	15,7%	1.565	9,4%

Besides age and gender, the other two constraint variables at the statistical areas level – employment and living situation played an equal part in calculating the weights (Table 32).

Table 32. Observed vs. sample frequencies of living situation and employment status (own representation)

		Yes		No	
		Count	% of Total Count	Count	% of Total Count
Living in a single household	Sample	122	17,5%	575	82,5%
	Observed	5.336	32%	11.327	68%
Currently employed	Sample	430	64,6%	236	35,4%
	Observed	7.522	45,1%	9.141	54,9%

The total observed population in the six statistical areas at the time chosen for reference (31.12.2017), was 16.663 people. To compute the weights, I adopted the Iterative Proportional Fitting approach, which was introduced earlier. As the sample data for the six statistical areas is merged into one, each individual must be assigned one single weight accounting for the total observed population in these six areas. Initially, this weight was set to 1 and it was then updated iteratively using a *for loop* to account for age, gender, employment, and living situation. In essence, the weight was updated by dividing the observed count referring to a certain variable category (e.g., males) by the corresponding count in the survey. This action was repeated for each variable category. First, the algorithm updated the weights to account for gender (male and female), then, it took the adjusted weight and updated it again to account for age (18-44 years, 45-64 years, 65+ years), and so on. After the weights of all individuals were updated considering each of the variable categories, the first iteration was completed. If the algorithm were to stop here, the weights would have been ideally adjusted to the last variable category accounted for, but they would not have been as accurate for the previous ones. Therefore, the algorithm continued running from the beginning, that is, from the first variable category. This process was repeated until the aggregated version of the weights for each variable category ideally fitted the corresponding count in the observed population. Generally, the number of iterations depends on how fast the algorithm manages to reach the perfect fit, whereby the more variable categories there are, the more iterations are needed. The computation of the

weights necessary for optimising the representativity of the survey required four iterations. I then integerised the generated weights and expanded the individual indices to generate the final weighted sample survey dataset. The procedure was thus identical to the population synthesis described in the previous chapter.

The last step before proceeding with the validation was to recode several variables in the sample survey dataset to ensure they fit the corresponding variable categories in the synthetic population dataset (Table 33).

Table 33. Recoding of survey dataset variables to fit the synthetic population dataset variables (own representation)

Survey Dataset	Synthetic Population Dataset	Recoded Categories
'How do you evaluate your overall health?' 'Excellent' 'Very good' 'Good' 'Less good' 'Bad'	**Subjectively perceived health** 'Very good/Good' 'Average/Bad/Very bad'	'Excellent' → 'Very good/ Good' 'Very good' 'Good' 'Less good' → 'Average/ Bad/ Very bad' 'Bad'
'**Do you have none of those chronic illnesses?**' 'Yes' 'No'	**Chronic medical condition(s)** 'Yes' 'No'	'No' → 'Yes' 'Yes' → 'No'
'To what extent did pain impair you from carrying out your usual daily activities?' 'Not at all' 'A little' 'Moderately' 'Quite' 'A lot'	**Impairment in daily activities due to illness** 'Yes' 'No'	'Moderately' → 'Yes' 'Quite' 'A lot' 'Not at all' → 'No' 'A little'
'How often do you engage in any type of sporting activity?' 'Not at all' 'Less than 1hr a week' '1-2 hrs a week' '2-4 hrs a week' '> 4 hrs a week'	**Sporting activity over the past three months** 'Yes' 'No'	'Less than 1hr a week' → 'Yes' '1-2 hrs a week' '2-4 hrs a week' '> 4 hrs a week' 'Not at all' → 'No'
'Do you smoke, even if only occasionally?' 'Yes' 'No' 'Not anymore'	**Smoking** 'Yes' 'No'	'Yes' → 'Yes' 'No' → 'No' 'Not anymore'

Table 33-continued. Recoding of survey dataset variables to fit the synthetic population dataset variables (own representation)

Survey Dataset	Synthetic Population Dataset	Recoded Categories
'What is your current living situation?' 'Alone' 'With partner' 'With other people'	**Living in a single household** 'Yes' 'No'	'Alone' → 'Yes' 'With partner' 'With other people' → 'No'
'What is your current employment?' 'Not employed' 'Mini job' 'Part-time job' 'Full-time job'	**Employment** 'Yes' 'No'	'Mini job' 'Part-time job' 'Full-time job' → 'Yes' 'Not employed' → 'No'
Body Mass Index (BMI) (Metric variable)	**Overweight: 25 ≤ BMI < 30** 'Yes' 'No'	25 < BMI < 30 → 'Yes' ELSE → 'No'
Body Mass Index (BMI) (Metric variable)	**Obesity: BMI ≥ 30** 'Yes' 'No'	BMI > 30 → 'Yes' ELSE → 'No'

Ideally, the external validation with the available survey data should account for variations in the model fit according to the four constraints applied at the statistical areas level – age, gender, employment, and living situation. However, the sample is relatively small, and thus more than half of the disease categories remain empty when grouping the individuals according to those constraints. Using the sample data for model validation in such a format would therefore lead to poor results because of insufficient sample diversity. To support this statement, Table 34 illustrates the unweighted sample disease data classified by age and gender.

Table 34. Unweighted sample disease data classified by age and gender (own representation)

	18-44		45-64		65+	
	female	male	female	male	female	male
hypertension	10	10	30	26	31	27
no hypertension	151	113	78	65	36	27
heart disease (incl. heart failure)	3	0	7	13	13	11
no heart disease	158	123	101	78	54	43
diabetes	7	2	9	6	9	12
no diabetes	154	121	99	85	58	42
cancer	1	0	5	4	4	1
no cancer	160	123	103	87	63	53
depression	27	15	13	8	6	1
no depression	134	108	95	83	61	53

Especially in the case of cancer, diabetes, and heart disease, the observation counts are extremely small, which diminishes the reliability of the data. With this in view, I considered the survey data unsuitable for external validation when being additionally classified by any of the

constraint variables – be it age, gender, employment, or living situation. Rather than that, it was more plausible to compare total population counts, such as total number of people with hypertension in the survey as opposed to in the synthetic population.

There were relatively many missing cases for the target variables 'chronic illness' (n=43; 6,6%), 'impairment due to illness' (n=51; 7,9%), 'sporting activity in the past three months' (n=55; 8,5%), and 'BMI' (n=69; 10,7%) in the survey. I therefore excluded them from the evaluation of the overall model fit and, instead, estimated their fit separately. Thus, I still managed to assess how well the simulated population illustrates the patterns of these variables, while at the same time I avoided a potential negative influence of the missing cases on the final evaluation result.

Against this background, I estimated the goodness-of-fit of the model by comparing the aggregate sums of the individuals having hypertension, heart failure, diabetes, cancer, and/or depression in the weighted survey and in the synthetic population. I stored the respective counts as integers in two separate vectors, each with a length = 10. In these vectors, each integer represents the sum of individuals belonging to one of the ten variable categories[23]. For the estimation, I used four different metrics introduced earlier: TAE, RE, RMSE, and MAPE (Table 35).

Table 35. External validation with survey data: Overall fit results (own representation)

TAE	RE	RMSE	MAPE
7.074	9,2%	861	17%

The TAE of 7.074 suggests that approximately 700 individuals are mistakenly allocated by the model to each one of the ten variable categories. To bring a little perspective into the matter, the total observed population for each couple of categories is 16.663 people. For instance, the number of people not having hypertension equals 16.663 – the count of individuals with hypertension. The same goes for the four remaining sets of variable categories. With this in view, TAE appears relatively small. RE goes one step further into clarifying the results. Basically, it can be interpreted as the model assigning approximately 9% of the population in each constraint category incorrectly. While RMSE appears slightly abstract, it generally provides an idea about how big the absolute differences between the categories are. Here, RMSE suggests rather smaller than larger differences between the individual sums in both vectors. By far the most important metric in terms of the overall evaluation of the model fit is MAPE. With 17%, it points to a less than perfect, and yet way above-average result.

To find out how well the model performs when simulating the patterns of individual target variables, I estimated TAE and RE for each of them separately. This is intriguing as the targets are not constrained to observed data but are simply *inferred* from the available constraints. The latter is still a legitimate approach as the constraints manage to explain part of the targets' variance (see 6.3. 'Selection of Constraint and Target Variables'). The results are summarised in Table 36.

[23] Positive and negative outcomes of the variables hypertension, heart failure, diabetes, cancer, and depression.

Table 36. External validation with sample survey data: Characteristics-specific fit results (own representation)

	Total Count (Model)	Total Count (Sample)	% of Total Count (Model)	% of Total Count (Sample)	TAE	RE	Missing cases in sample in %
Hypertension	4.802	3.672	28,8%	22%	1.130	30,8%	8%
Heart failure	998	1.491	6%	8,9%	493	33,1%	8%
Diabetes mellitus	1.839	1.277	11%	7,7%	562	44%	8%
Cancer	618	637	3,7%	3,8%	19	3%	8%
Depression	2.907	2.242	17,4 %	13,5%	665	29,7%	8%
Sporting activity (past 3 months)	10.015	10.328	60,1%	62%	313	3%	8,5%
Subjectively perceived health[24]	9.891	12.253	59,4%	73,5%	2.362	19,3%	1,7%
Impairment due to illness	7.364	4.113	44,2%	24,7%	3.251	79%	7,9%
Chronic medical condition(s)	8.766	6.208	52,6%	37,3%	2.558	41,2%	6,6%
Smoking	5.627	5.676	33,8%	34,1%	49	0,9%	2,3%
Overweight	5.948	4.714	35,7%	28,3%	1.234	26,2%	12,2%
Obesity	3.496	2.983	21%	17,9%	513	17,2%	12,2%

Overall, the external validation with survey data points to a higher-than-average goodness of fit for the different target variables as well. The estimations suggest that the model provides excellent fit for the variables ‘cancer’, ‘sporting activity’, and ‘smoking’, and a much less adequate match for other variables, such as ‘impairment in daily activities due to illness’. The results for ‘hypertension’, ‘depression’, ‘subjectively perceived health’, ‘obesity’, and ‘overweight’ indicate a rather satisfying fit. ‘Chronic medical condition(s)’, ‘diabetes mellitus’, and ‘heart failure’, on the other hand, are less well represented in the synthetic population, at least as far as the sample survey data is considered a reliable illustration of reality. Against this background, there are certain details about the survey, which may have an ameliorating effect on the discouraging results for some of the variables.

First, I would like to address ‘impairment due to illness’. Here, the proportion of affected population in the model is considerably larger compared to the sample survey. There are several questions in the survey that relate to this matter, such as ‘*Do you have difficulty climbing stairs?*’, ‘*Do you have difficulty moving heavier objects?*’, and the one I took as a reference ‘*To what extent did pain impair you from carrying out your usual daily activities?*’. Taking all these variables into account, may deliver a more encouraging result. Nonetheless, such an approach may also result in over-representing impairment in daily activities.

Next, I am going to set the focus on ‘diabetes mellitus’. The number of individuals suffering from diabetes in the simulated population is larger than in the weighted survey dataset. One possible explanation for this may be that the survey relies on self-reported diabetes, whereas the data from the Morbidity Atlas used for constraining the model includes cases of diabetes

[24] as *good*

types 1 and 2, diabetes resulting from malnutrition, '*other, specific types of diabetes*' as well as '*other, unspecified types of diabetes*', all represented by the ICD-10 Codes: E10-E14 (Erhart et al. 2013, p.4). The number of individuals in the sample is relatively small. It is therefore not unlikely for it to fall short of covering the entire spectrum of diabetes diagnoses and thus deliver a picture which underestimates actual diabetes prevalence.

Finally, I would like to consider the specifics of the variable 'heart failure'. In the survey, respondents were asked whether they (have) had any type of heart disease, including heart attack, chronic condition because of a heart attack, coronary heart disease, angina pectoris, stroke, chronic condition as a result of a stroke, *or* heart failure. It is therefore not surprising that the number of individuals with heart disease in the sample survey dataset is larger than the corresponding synthetic population group because the latter accounts solely for heart failure.

The considerations regarding 'diabetes mellitus' and 'heart failure' are reason to believe that the overall goodness-of-fit of the model may be better than indicated by the metrics in Table 35. With this in view, I regard the external validation results as encouraging rather than the opposite. Still, the survey data taken as reference covers only six statistical areas. Furthermore, the sample is relatively small, which makes the data suitable for comparison but not so much for validation. As the survey itself may not be a completely reliable representation of reality, the results do not necessarily say how well the model manages to illustrate existing disease patterns. Therefore, while the estimated outcome suggests that the model provides an adequate picture of reality, a humble interpretation of the results is advisable. To delve deeper into the model evaluation, the next section will present the results from the external validation with data from three of Hamburg's health insurance funds.

7.2.2. External validation with health insurance data

In the course of the project 'Healthy Neighbourhoods' (2017-2021), health-related data was obtained for research purposes from three health insurance funds in Hamburg: AOK Rheinland/Hamburg, BKK Mobil Oil, and DAK-Gesundheit, covering approximately 30% of the total population (Mindermann et al. 2021, p.116). For reasons of data protection, the dataset I was provided with by the project partners responsible for the health insurance data acquisition (HAW Hamburg), did not allow data classification by insurance fund. The data was aggregated at the level of the social status index classes defined by the Social Monitoring.

To evaluate my spatial microsimulation model, I used the total counts of people with hypertension, heart failure, diabetes mellitus, cancer, and depression divided in terms of gender and 5-year age intervals. Since the generated synthetic population encompassed only adults, I did not use insurance data about children and adolescence.

Before proceeding with the model validation, I had to scale up the available population counts because, as already noted, they covered less than one third of the population in Hamburg. I therefore calculated a weight for each age-gender category: males aged 18-24 years; females aged 18-24 years; males aged 25-29 years, and so on. The last age category was 80+. The weight was estimated by dividing the observed population count for each of those age-gender categories (at the time of 31.12.2017) by the corresponding population count in the health insurance data, which was also related to 2017. To account for possible differences in the

demographic structure of neighbourhoods with differing social status, the weights for the age-gender categories were computed for each status index class separately.

The obtained health insurance data was aggregated in seven, rather than in four status index classes as it is customary for the Social Monitoring. The reason for this was the uneven population distribution in the four established status index classes 'high', 'average', 'low', and 'very low' (Table 37). With over 63% of Hamburg's total population living in statistical areas with average social status, aggregating the health insurance data based on this classification would lead to a great loss of detail.

Table 37. Distribution of the statistical areas and their population in four status index classes (own representation, Source: Statistisches Amt für Hamburg und Schleswig-Holstein 2018)

Status index class	Standard deviation	Number of statistical areas	Total population
high	< -1,00	156	301.078
average	-1,00 to 1,00	542	1.141.480
low	1,01 to 1,50	67	157.410
very low	> 1,50	82	209.265
Total		847	1.809.233

The methodology used for the allocation of the statistical areas to the status index classes was briefly introduced in Chapter 3.3.1. 'Hamburg's Social Monitoring'. It is explained in more detail in the pilot report of the Social Monitoring (Pohlan et al. 2010, pp.44–46). In essence, the social status index is comprised of seven main indicators which are standardised using a z-transformation (Formula 6).

Formula 6. z-Transformation (Pohlan et al. 2010, p.6)

$$z_i = \frac{x_i - \bar{x}}{s}$$

where:

- x_i = the value to be standardised
- $\bar{x}$ = the mean
- s = the standard deviation

Through the z-transformation, the mean value of each indicator is set to 0, and the standard deviation is set to 1. This allows comparing the variance of indicators with differing dimensions (e.g., measured in absolute vs. relative terms). The status index equals the sum of the z-values of all seven indicators. Finally, the statistical areas are allocated to one of the four status index classes based on standard deviation. If, for instance, the status index falls within -1,00 to 1,00 standard deviations of the average, the statistical area is classified as having 'average' social status (Table 37).

To solve the problem with the unequal population distribution within the four established status index classes, the class 'average' was further divided into four categories. The range '-1,00 to 1,00 SD' was split into the categories '-1,00 to -0,50 SD', '-0,49 to 0,00 SD', '0,01 to 0,50 SD', and '0,51 to '1,00 SD'. This resulted in a more balanced distribution of the population based

on the social status of their statistical area of residence (Table 38). The obtained health insurance data was aggregated according to the illustrated, slightly refined status index classification (Mindermann et al. 2021, p.112).

Table 38. Refined distribution of the statistical areas and their population in seven status index classes (Source: Mindermann et al. 2021, p.112)

Status index class	Standard deviation	Number of statistical areas	Total population
high	< -1,00	156	301.078
average 1	-1,00 to -0,50	163	321.520
average 2	-0,49 to 0,00	152	334.192
average 3	0,01 to 0,50	125	262.837
average 4	0,51 to 1,00	102	232.931
low	1,01 to 1,50	67	125.410
very low	> 1,50	82	209.265
Total		847	1.809.233

Before going into further detail about the interpretation of the external validation results, I want to address some limitations of the obtained health insurance data. First, it covers solely inhabitants with a statutory health insurance. Those who are privately insured are hence not represented. Second, the data encompasses just about one third of the population with statutory health insurance. Third, the socio-demographic composition in terms of age and gender in the obtained insurance data slightly differs from that of Hamburg's total population (Tables 39-40).

Table 39. Comparison of gender distribution in 2017 (absolute/relative) between the obtained insurance data and the total observed population in Hamburg (Source: Mindermann et al. 2021, p.117)

	Obtained insurance data		Hamburg Total*		Deviation**
	Absolute count	Relative count (%)	Absolute count	Relative count (%)	in (%)
Females	255.902	52,8	955.103	50,8	2,0
Males	229.086	47,2	925.894	49,2	-2,0
Total	484.988	100,0	1.880.997	100,0	
* Hamburg's total population including people with private health insurance					
** from the reference data 'Hamburg Total'					

Table 40. Comparison of age distribution in 2017 (absolute/relative) between the obtained insurance data and the total observed population in Hamburg (Source: Mindermann et al. 2021, p.117)

	Obtained insurance data		Hamburg Total*		Deviation**
	Absolute count	Relative count (%)	Absolute count	Relative count (%)	in (%)
0-6 years	31.866	6,6	114.852	6,1	0,5
6-10 years	17.028	3,5	66.002	3,5	0,0
11-15 years	21.815	4,5	78.504	4,2	0,3
16-21 years	31.090	6,4	103.622	5,5	0,9
22-45 years	153.338	31,6	678.628	36,1	-4,5
46-65 years	120.708	24,9	497.665	26,5	-1,6
65+ years	109.143	22,5	341.724	18,2	4,3
Total	484.988	100,0	1.880.997	100,0	
* Hamburg's total population including people with private health insurance					
** from the reference data 'Hamburg Total'					

With this in view, while the results of the validation will surely provide insight into how well the model depicts reality, they should be regarded as an approximate evaluation rather than an absolute one because the underlying health insurance data does not capture the whole picture either.

Having clarified this, I do not wish to undermine the results of the model validation with the available insurance data. On the contrary, using data from three different health insurance funds to evaluate the model is a major benefit I want to put an emphasis on. As pointed out in the interviews with Prof Dr Busch and PD Dr Augustin, some health insurance funds lack a diverse client profile, which is why using data from more than one source increases the credibility of the results.

With that said, I am going to begin with the actual model validation. As already noted, my first step was to scale up the insurance data because it did not cover the entire population. To that end, I applied a weight to each unique group of individuals depending on their age, gender, and the status index of their statistical area of residence. This weight equals the ratio of the observed population belonging to a given group (e.g., males, aged 18-24, living in an area with high social status) to the corresponding population count in the insurance data.

I decided to use MAPE for illustrating the results of the external validation. It is by far the most intuitive metric and thus allows to quickly grasp how well the model illustrates different disease patterns. The errors for hypertension, heart failure, diabetes, cancer, and depression, classified by age and gender, are summarised in Table 41.

Table 41. MAPE for modelling disease patterns* classified by age and gender (own representation)

age	gender	hypertension	heart failure	diabetes	cancer	depression
18-44	female	6,2%	243%	630%	68,8%	9,2%
	male	100%	520,1%	828,9%	74,9%	14,8%
45-64	female	9,6%	150,8%	238,8%	68%	28,1%
	male	9,6%	133,3%	341%	66,7%	27%
65+	female	13%	28,7%	309,1%	68,9%	28,2%
	male	17,7%	12,5%	303,3%	66,3%	30,3%
18–44 (males and females)		53,1%	381,5%	729,4%	71,9%	12%
45-64 (males and females)		9,6%	142%	289,9%	67,3%	27,5%
65+ (males and females)		15,3%	20,6%	306,2%	67,6%	29,3%
females (all age groups)		9,6%	135,9%	469,5%	68,9%	18,7%
males (all age groups)		58,9%	266,3%	566,1%	70,6%	22,6%
Total		26%	181,4%	442%	68,9%	23%
* At the level of the seven status index classes						

According to the results, the model manages to simulate depression and hypertension relatively well as opposed to cancer, heart failure, and diabetes mellitus. Especially the latter two seem to be extremely poorly represented by the synthetic population. This is true for both genders as well as for all age groups. Looking at the absolute counts of the diseased people (Appendix, Table 45), both diabetes and heart failure affect much more individuals in the synthetic population than in the health insurance dataset. Nevertheless, comparing the total absolute counts of diseased individuals in the Morbidity Atlas, the synthetic population, and the

insurance dataset, suggests closer results between the Morbidity Atlas and the synthetic population, than the insurance dataset (Table 42). This is valid regardless of the status index of the statistical area of residence.

Table 42. Total counts of diseased people differentiated by age and gender according to different sources[25] (own representation)

Data source	Age	Gender	Hypertension	Heart failure	Diabetes	Depression	Cancer
Constraint data (Morbidity Atlas[26]; Cancer Registry)	18-64	Females	84.344	10.203	28.920	116.705	11.441
		Males	107.576	16.191	41.417	73.799	7.305
	65+	Females	125.935	35.562	42.608	45.930	15.633
		Males	94.050	27.383	41.962	19.052	16.663
Health insurance funds data	18-64	Females	94.410	3.884	7.207	138.641	42.985
		Males	99.494	6.041	8.234	83.270	27.220
	65+	Females	133.495	22.733	10.774	53.724	49.994
		Males	98.194	18.125	10.858	24.673	43.262
Synthetic population	18-64	Females	88.155	10.278	29.555	116.432	13.591
		Males	112.614	16.265	41.993	73.985	8.352
	65+	Females	124.828	34.588	42.405	43.274	17.366
		Males	93.191	26.525	41.212	17.649	18.054

Keeping in mind that the Morbidity Atlas served for constraining the synthetic population data at the level of the city quarters, these results are not surprising. The huge difference between the total count of individuals with heart failure and diabetes in the Morbidity Atlas and the health insurance dataset, however, is very much unexpected. While the data in the Morbidity Atlas is older (2011) and there is thus a 6-year-gap between both data sources, it is quite unlikely that so many ill individuals have been cured from those chronic conditions. Since the data in the Morbidity Atlas comes from all health insurance funds in Hamburg rather than from just three of them, the more likely explanation is that the larger proportion of individuals suffering from heart failure and diabetes are insured with health insurance funds different than AOK Rheinland/Hamburg, BKK Mobil Oil, and DAK-Gesundheit. This large discrepancy between the total counts in the Morbidity Atlas and the insurance data suggests that the latter is not that reliable source for validation. I therefore do not necessarily consider the model to have failed in illustrating the prevalence of heart failure and diabetes. Still, the simulated disease patterns cannot be verified using another data source, which puts their reliability into question.

With an average MAPE = 68,9%, the external validation with insurance data indicates that cancer is not as well represented by the synthetic population either. The number of individuals with any type of oncological disease in the insurance dataset is approximately three times larger than that in the Cancer Registry and the synthetic population (Table 42). Since I had to scale up the population available in the insurance dataset to carry out the external validation, the number of cancer patients in that 30%-sample of Hamburg's statutory insured population was scaled up as well. The most probable reason for this difference is hence a disproportionate

[25] Morbidity Atlas, Cancer Registry, Health insurance funds, synthetic population

[26] Prevalence percentages were used instead of absolute population counts and applied to population data from 31.12.2017 to constrain the model, because the absolute population counts in the Morbidity Atlas refer to 2011.

count of cancer patients being insured in the three health insurance funds that delivered the data. Against this background, the synthetic population does not necessarily fail in illustrating patterns of oncological disease. The available external data for validation, however, does not allow verifying how reliable the modelled data is.

In contrast, hypertension is relatively well simulated, especially if the population group of males aged between 18-44 years is excluded (Table 41). For some reason, the algorithm has allocated more individuals with hypertension to this category than necessary. The average error for males is 58,9% as opposed to just 9,6% for females. The MAPE for all population groups is hence 26%, with the largest proportion of error resulting from the poor fit for younger males.

Depression is represented by the synthetic population relatively well, whereby the results are more satisfying for individuals younger than 45. The best represented population group is that of females aged between 18 and 44 years (MAPE = 9,2%), whereas the worst is that of males older than 65 years (MAPE = 30,3%). With three times larger error, this is quite a difference. Still, the overall error of 23% ranks depression as the disease with the best model fit of all five.

Looking at the results depicted in Table 41, there is another observation worth mentioning – females are better represented by the synthetic population than males for all disease types. As for age, no such conclusion can be drawn because the trend varies depending on the illness. In terms of status index, however, there are certain variations in MAPE (Table 43). In other words, regarding different types of disease, the synthetic population matches the observed population from the health insurance dataset better in statistical areas with a certain social status, than in others. For instance, the error for simulating hypertension slightly increases (+ 4,3%) with decreasing social status. The same pattern is evident for depression, where the error increases even more (+10,7%). In the case of heart failure and diabetes, on the other hand, the reverse trend is exhibited – the error decreases with decreasing social status. Observed in absolute terms, the change seems significant (-47,9 % for heart failure and -62,6% for diabetes). Given the scale of the error for these two diseases, however, it is more advisable to track the change in error in relative terms – it decreases by 22,7% for heart failure and by 13% for diabetes. Cancer is the only disease that does not exhibit significant variation in MAPE based on status index.

Table 43. MAPE for simulating disease patterns differentiated by status index class (own representation)

status index	hypertension	heart failure	diabetes	cancer	depression
high	24%	210,7%	479,7%	69,9%	16,1%
average 1	24%	197,9%	474%	69,2%	19,6%
average 2	24,7%	187,2%	449,4%	68,9%	22,3%
average 3	26%	172,6%	420,4%	68,6%	23,4%
average 4	27,2%	175,5%	426,9%	68,4%	25,5%
low	27,8%	162,8%	425,6%	69,1%	27,1%
very low	28,3%	163,2%	417,1%	68,5%	26,8%

In summary, the available health insurance data did not prove to be a completely reliable source for model evaluation – especially in regard to heart failure, diabetes, and cancer. The reason for this assumption is that the differences in the total counts of diseased individuals in the insurance dataset and the datasets used for constraining the model are inexplicably large. In fact, the findings imply that the insured patients may not be a representative sample of

Hamburg's population – at least not in terms of these types of disease. Still, this outcome alone cannot serve as solid foundation for declaring the synthetic population a reliable illustration of the prevalence of diabetes, heart failure, and cancer. Instead, it simply suggests that these types of disease cannot be validated. Hypertension and depression, on the other hand, appeared to be relatively well simulated – especially for certain population groups.

Overall, external validation with both sample survey and health insurance data delivered mixed results. Neither of the two external data sources could be assessed as reliable representation of reality due to factors such as population coverage and population diversity. However, since there were no other sources available, I preferred to use them and carry out the model evaluation while keeping in mind their limitations.

The external validation with sample survey data suggested that the modelled health behaviour regarding smoking and sporting activity depicts reality very accurately. Obesity was also relatively well represented despite the unconvincing results of the logistic regression test about the extent to which the constraint variables manage to explain its variance. Out of the chronic illnesses, cancer was by far most accurately illustrated by the model. Depression and hypertension were relatively well, albeit far from perfectly simulated.

The results from the external validation with health insurance data were overall less encouraging. They only allowed verifying the five chronic illnesses available as constraints at the city quarter level – hypertension, heart failure, diabetes mellitus, depression, and cancer. Out of those, only hypertension and depression were relatively well simulated by the model, whereby this was not applicable for all population groups. It was not possible to evaluate the goodness of fit for heart failure, diabetes, and cancer because the absolute counts of diseased population in the available insurance dataset and in the health-related constraint datasets (Morbidity Atlas and Cancer Registry) differed greatly. The latter suggested an existing bias in the population insured with the three health insurance funds whose data I used for model validation. Since the health-related data used for constraining the model originates from all statutory health insurance funds rather than just three of them, it should be counted as the more reliable data source.

Against this background, I do not consider the external validation as failed despite what may appear as a poor result at the first glance. Rather than that, the findings should serve as evidence of the overall complexity of model validation with external data sources. The interpretation of the results must always be coupled with considerations regarding the known (and the suspected) limitations of the external data at hand. With these final words on the process of model validation, the section dedicated to generating synthetic population is concluded.

8. A CONTRIBUTION TO SETTING UP CITYWIDE HEALTH MONITORING SYSTEMS

In this chapter, I am going to introduce a couple of application examples for the generated small-scale health data. Thus, I will address the practical use of the model and demonstrate how it could contribute to setting up citywide health monitoring systems.

The greatest advantage of the model is that it provides disaggregated population data. Instead of total population counts for certain categories, individuals with attribute characteristics are available for each statistical area. Moreover, the generated synthetic population is not merely a population sample, but a complete representation of Hamburg's total population. Besides the socio-demographic and health attributes, each individual is assigned a spatial reference about the city quarter and the statistical area of residence.

This data format allows a much broader spectrum of possibilities for further in-depth analysis. For instance, it can aid in identifying areas, which are especially problematic in terms of health, social status, and characteristics of the living environment. Thanks to data availability, such strategy can be implemented without much effort for the city of Hamburg. To that end, geodata about the spatial distribution of noise pollution, public green space, social infrastructure, etc. and socio-economic data from the Social Monitoring can be used. The generated small-scale health data therefore represents the last missing ingredient in the triad 'environment – social status – health'.

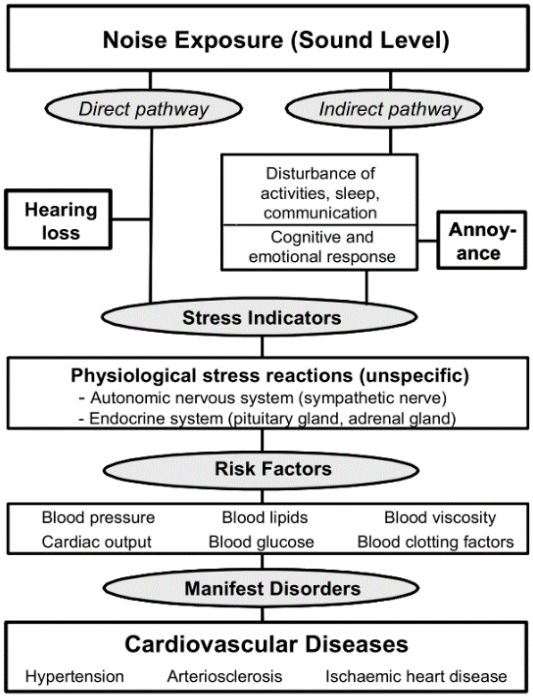
The validation of this data, carried out in the previous chapter, did not necessarily deliver encouraging results. Nonetheless, there were some inherent flaws to the used external data, which must be considered. The survey data originated from a relatively small and slightly biased sample covering only six statistical areas. The health insurance data, on the other hand, covered the entire city but only around 30% of the population with statutory health insurance. Because of this partial population coverage, the insurance data was weighted according to the population distribution in terms of age, gender, and social status of the statistical area of residence. The results showed significant differences between the weighted population counts related to heart failure, diabetes mellitus, and cancer in the insurance data and the corresponding population counts in the Morbidity Atlas. The latter contains data from all statutory health insurance funds based in Hamburg. This outcome hence indicated that the individuals insured with the three funds, which provided the data used for model evaluation, are not necessarily representative for Hamburg's total population – at least not in terms of those illnesses.

With this in view, the model is *not* flawed despite what the results of the validation partially suggest. Instead, it should be regarded as initial *database* suitable for identifying potentially problematic areas, but insufficient for issuing final statements about the health situation in them. To clarify the nature and extent of the suspected problems, further analyses are highly recommended.

All in all, this is the gain from the model – it can serve as citywide small-scale health warning system – highlighting neighbourhoods in need of timely health-promoting, or other measures. Combining the generated synthetic population with socio-economic data and geodata about the living environment can become a gold mine for urban health researchers. In this regard, the next sections are going to introduce two examples for possible data applications.

8.1. Application Example I: Identifying Hotspots of Hypertensive Individuals Exposed to Excessive Noise from Road and Air Traffic

Figure 13. Noise exposure reaction scheme (Source: Babisch 2002)



Continuous exposure to traffic noise can have a negative influence on the cardiovascular system (e.g., Zeeb et al. 2017; Chang et al. 2015; Dratva et al. 2012; World Health Organization Regional Office for Europe 2011; Umweltbundesamt 2006). In this context, Figure 13 illustrates the chain of reactions triggered by noise exposure. There are two pathways of reaction – a direct one manifested as hearing loss, and an indirect one, which can be the disturbance of various activities, including sleep and communication, or a cognitive and emotional response. The indirect pathway of reaction generally causes annoyance. Both the direct and the indirect pathways are considered stress indicators leading to (unspecific) physiological stress reactions of the autonomic nervous system and/or the endocrine system. These reactions are risk factors for the normal functioning of blood pressure, the blood lipids and blood glucose levels, etc. Eventually, these risk factors may cause the manifestation of disorders such as chronic cardiovascular diseases including hypertension, arteriosclerosis, and ischaemic heart disease.

Scientific evidence about the effects of road and air traffic on the risk of ischaemic heart disease and hypertension is ample. The number of studies related to rail traffic noise, however, is limited (World Health Organization Regional Office for Europe 2011, p.xv). For the purposes of this application example, I am therefore going to focus on air and road traffic exclusively.

The risk of hypertension from road traffic noise has been found to increase by '*1.38 (95% CI 1.06– 1.80) per 5-dB(A) in the 24-hour noise level (L24h ≈ 40–70 dB(A))*' (World Health Organization Regional Office for Europe 2011, p.40). Adjusting for air pollution does not affect the odds ratio for road traffic noise regarding the prevalence of hypertension (ibid.). Another study found that the effects of exposure to noise traffic at home were greater for people, who were also exposed to high noise levels at work (Umweltbundesamt 2006, p.12). Furthermore, epidemiological studies suggest that continuous, rather than occasional exposure to excessive levels of road and air traffic noise bears higher risk of cardiovascular disease, including hypertension and myocardial infarction (World Health Organization Regional Office for Europe 2011, p.33). Regarding the question what levels of noise are considered harmful, '*the German road traffic noise study (response rate 60%) carried out in Bonn suggested a relative risk for hypertension of 1.5 for subjects who lived in areas where the traffic noise level exceeded Lday = 65 dB(A)). This finding was significant*' (Umweltbundesamt 2006, p.26).

Dratva et al. (2012) carried out a stratified analysis by chronic disease status, which yielded larger effect estimates related to systolic blood pressure (SBP) and diastolic blood pressure

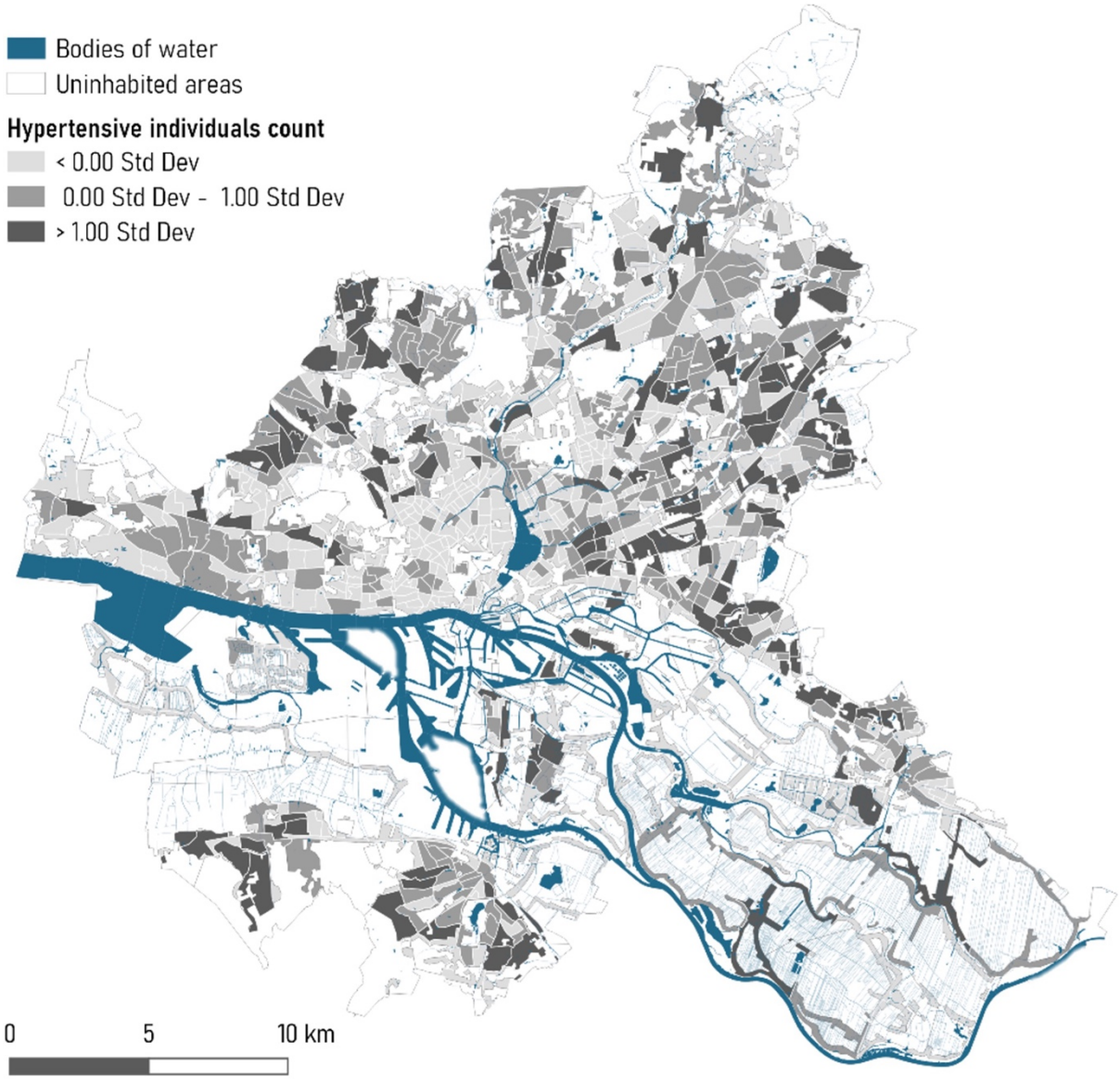
(DBP) in participants reporting physician-diagnosed hypertension, diabetes, and cardiovascular disease (p.53). Chang et al. (2015) also found that hypertensive adults are more susceptible to noise exposure, with a greater effect on ambulatory systolic blood pressure. There is hence a larger need for protecting this specific subpopulation from continuous exposure to excessive traffic noise levels.

Against this background, the generated small-scale health data can serve for identifying statistical areas with high concentration of hypertensive people. With the use of available geodata about road and air traffic noise, hotspots of hypertensive individuals exposed to excessive noise levels can be detected. Thus, timely noise-reduction measures can be implemented where they would have the most significant impact on protecting human health. This is one possible application of the generated synthetic population that I am going to address in the following paragraphs.

To single out statistical areas with a concentration of hypertensive individuals, I created a choropleth map where the colour of the area darkens as the number of people suffering from hypertension increases (Map 1). I chose to observe absolute rather than relative counts of hypertensive individuals to avoid the population size of the given statistical area influencing the selection. For instance, a statistical area with a total population of 500 people could end up classified as having high concentration even if the number of hypertensive individuals is just 200 because they would represent more than a third of its population. Rather than that, my aim was to identify statistical areas where the count of people suffering from hypertension is above the average at this spatial scale regardless of how populous the area is.

Against this background, there are three colour categories in Map 1 – from light, through medium, to dark grey. Each of them refers to the count of hypertensive individuals, whereby the classification is based on the standard deviation (Std Dev). The first category encompasses all statistical areas, where the count of hypertensive individuals is below the average of 493 people (< 0.00 Std Dev). The second category encompasses statistical areas with more than the average number of people suffering from hypertension but less than the average + 1 standard deviation, that is, between 493 and 720 individuals (0.00 Std Dev – 1.00 Std Dev). The third category encompasses statistical areas, where the number of hypertensive individuals is higher than the average + 1 standard deviation (> 1.00 Std Dev). These areas have more than 720 people suffering from high blood pressure.

Map 1. Distribution of hypertensive individuals at the level of the statistical areas (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021)

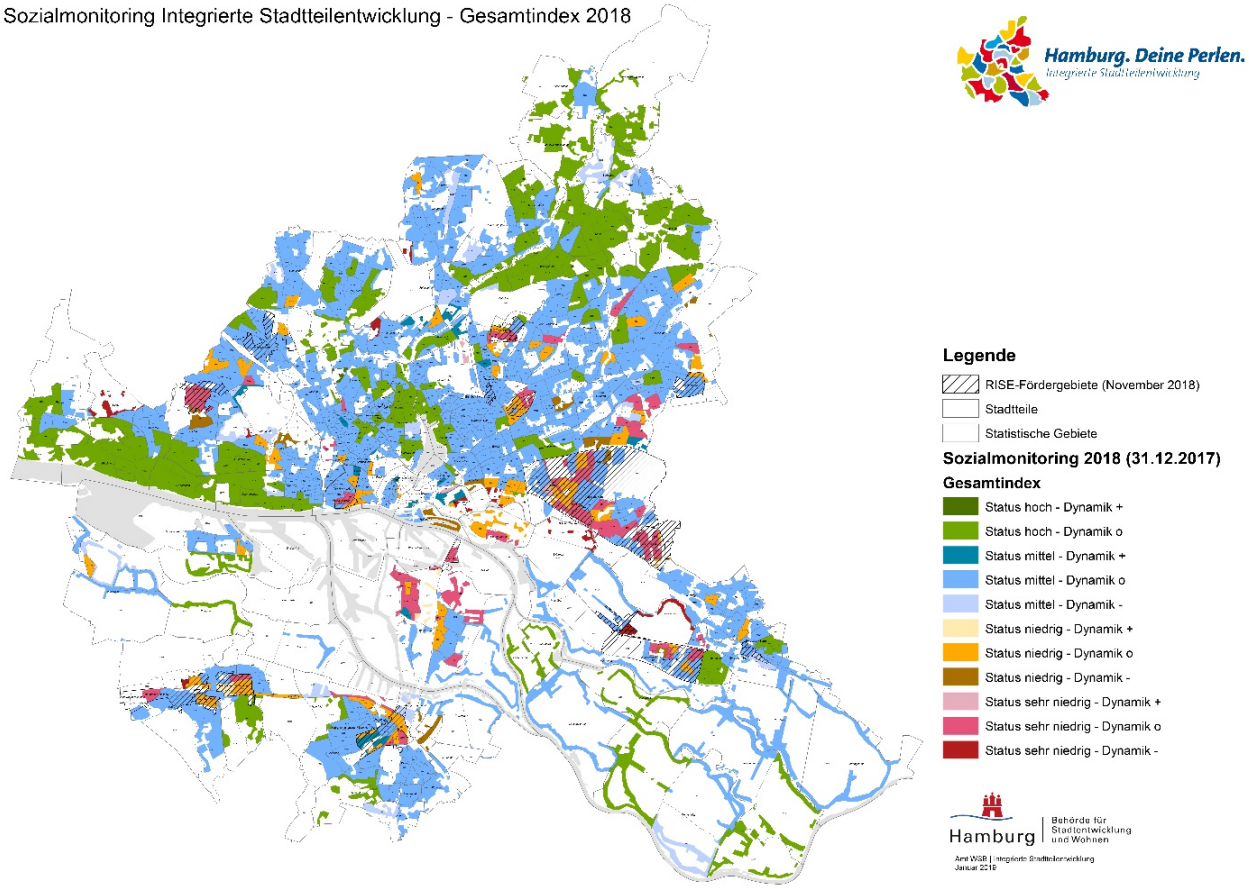


Map 1 indicates a certain spatial pattern in the distribution of people suffering from hypertension throughout the city. Statistical areas with more such individuals are mainly located in its periphery, east of the Alster Lake as well as south of the Elbe River. In contrast, statistical areas in the city centre and its vicinity exhibit lower concentrations of hypertensive individuals. Looking at the spatial distribution of the statistical areas in terms of their social status, as defined in the Social Monitoring 2018[27] (Map 2), there appeared to be a connection with the identified pattern for hypertension. To establish whether there were actual indications for such a relationship, I carried out Pearson's Correlation Test. For this purpose, I chose to look at the

[27] Social Monitoring 2018 is taken as reference instead of the most current one because the socio-demographic data that served for constraining the synthetic population was used for computing the status index for 2018 as well (time reference: 31.12.2017).

proportion rather than the absolute count of hypertensive individuals in order to control for population size.

Map 2. Social Monitoring Hamburg 2018 (Source: Behörde für Stadtentwicklung und Wohnen 2018)



The correlation between the percentage of hypertensive inhabitants and the standardised status index was weak, but what is more interesting is that it was negative (r = -0.211). This implies that statistical areas with higher social status have larger proportion of hypertensive population[28]. Controlling for age, r = -0.017 for those aged 18-64 years, and r = 0.294 for older individuals. Age is generally a risk factor for hypertension, and it is therefore interesting to observe that the proportion of elderly people suffering from this chronic disease is larger in deprived neighbourhoods. For younger individuals, there is no such trend. At the same time, when looking at the entire population without accounting for age, the relationship is reverse. Hence, other factors different than age, which I could not control for here, lead to a larger overall proportion of hypertensive population in more affluent urban neighbourhoods. There may be some relationship between these factors and social status, such as the latter influencing people's diet or lifestyle in a way that indirectly triggers hypertension. This, however, is merely an assumption, which is not based on actual numbers and figures.

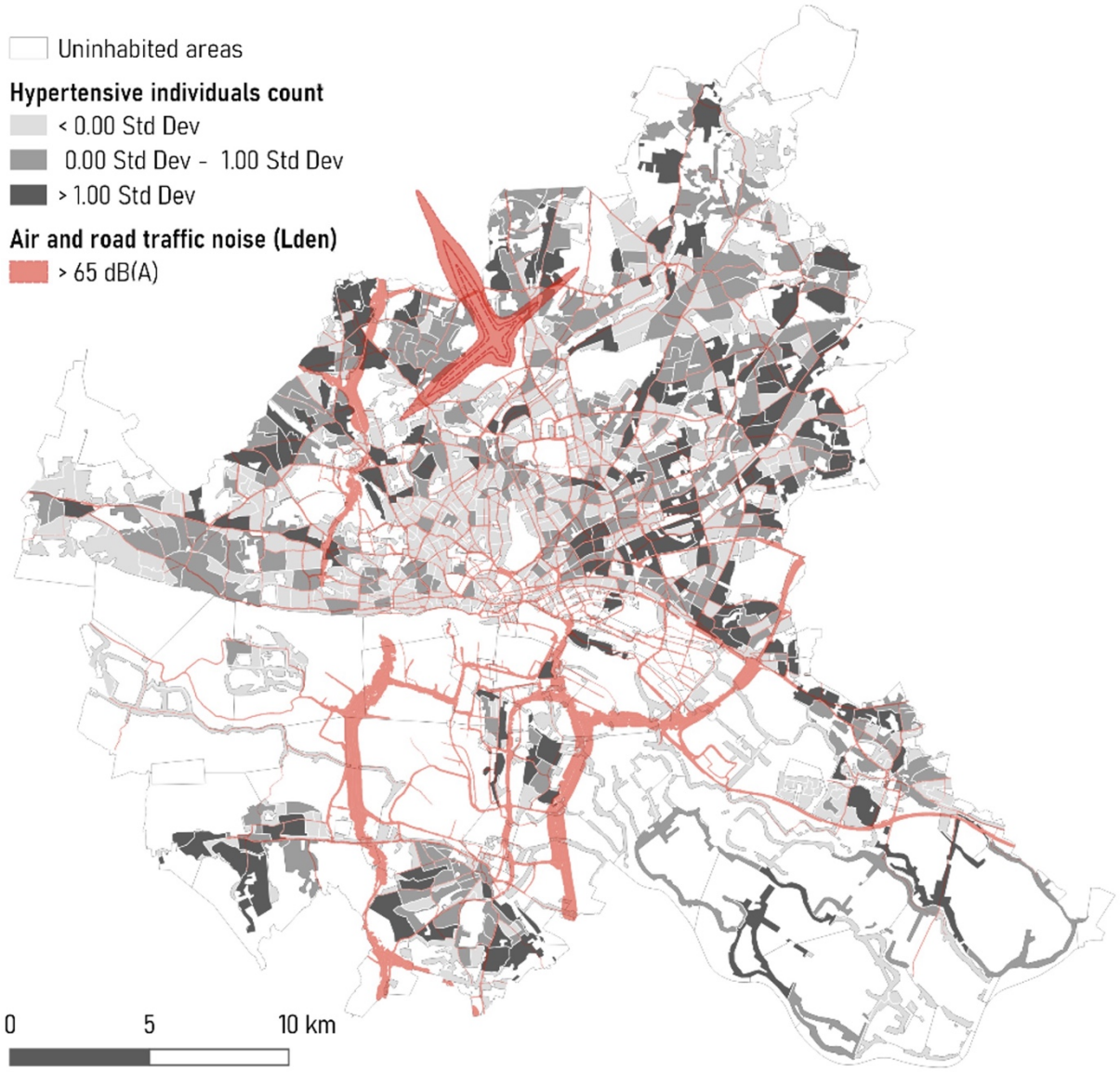
While r is weak both for the total, and for the elderly subpopulation, it still indicates a certain relationship between the proportion of hypertensive population and the social status of the

[28] Statistical areas with 'high', 'average1', and 'average2' status index have negative standardised status sums due to their proportions of unemployed population, proportions of population with migration background, etc. being below the city's average.

statistical area. Nonetheless, the difference in the direction of the correlation does not allow including social status as additional criterium for filtering statistical areas in need of timely noise-protection measures due to a high concentration of hypertensive individuals.

The distribution of the hypertensive population shown in Map 1, overlayed with noise from air and road traffic (Lden) is illustrated in Map 3. Only levels ≥ 65 dB(A) were considered as this is the benchmark regarded as potentially harmful for human health (Umweltbundesamt 2006, p.26).

Map 3. Distribution of hypertensive individuals overlayed with air and road traffic noise (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021; Behörde für Umwelt, Klima, Energie und Agrarwirtschaft 2017)

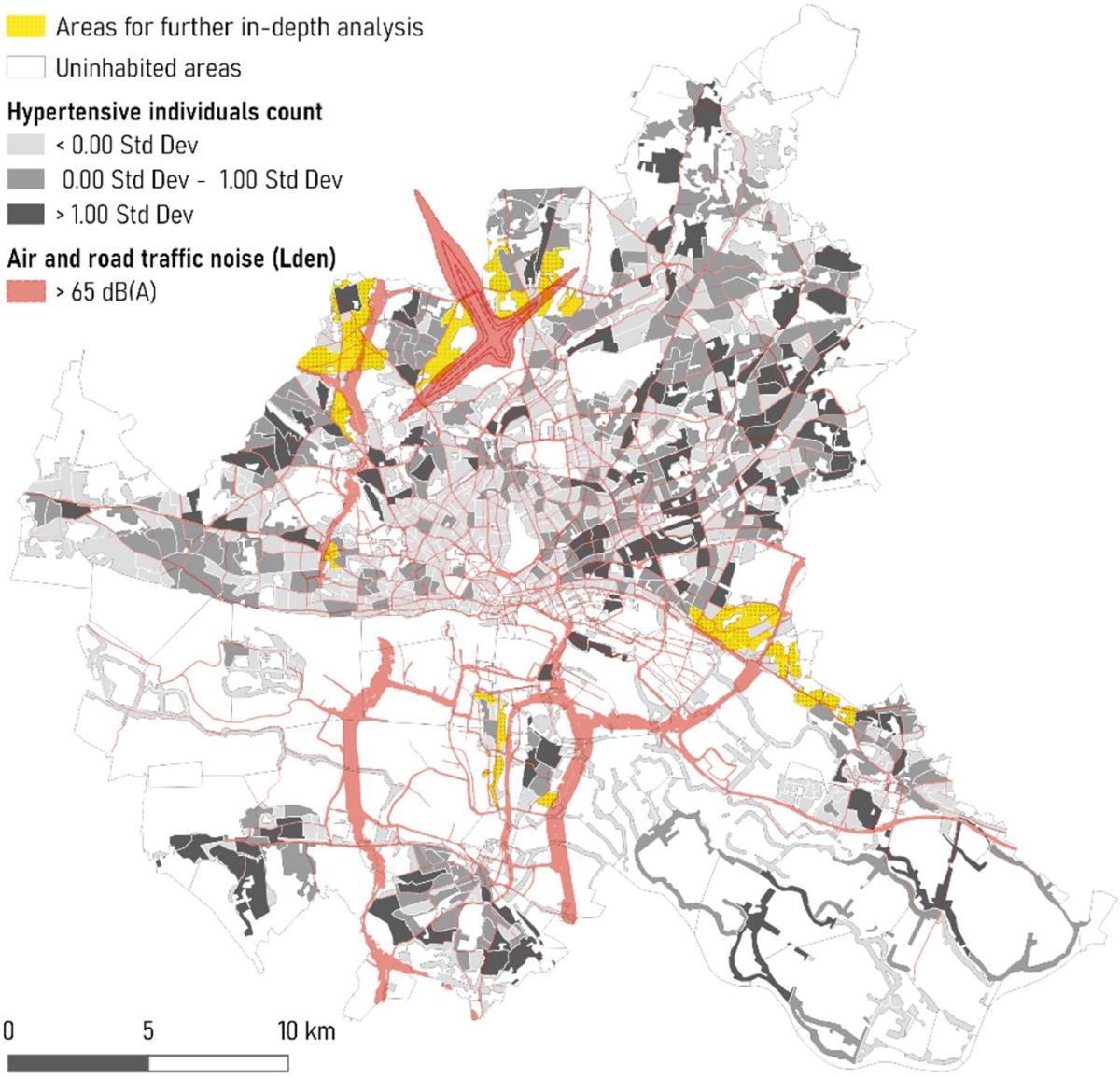


Almost all statistical areas located in the immediate vicinity of the airport exhibit an above average count of hypertensive individuals. West of the airport there is the A7 highway, an additional noise emitter. Especially the statistical areas on the west side of the highway are place of residence for many people suffering from hypertension. These neighbourhoods immediately draw the attention in terms of necessity to protect individuals with high blood pressure from traffic noise. Additionally, there are several statistical areas in Wilhelmsburg, such

as Kirchdorf-Süd located west of the A1 highway, and Elbinselquartier west of Bundesstraße 75, possibly requiring attention. Finally, the statistical areas in Billstedt and Mümmelmannsberg located right around the intersection of A1 and Bundesstraße 5 have large number of hypertensive residents exposed to potentially harmful levels of noise.

The benefit of the generated citywide individual health data is that it allows identifying areas, which may require timely health-promoting or protective measures, within the large urban realm. While modelled data cannot completely substitute *real* patient data, it can aid in sharpening the focus at the small scale and thus point the magnifying glass to those spatial units, which appear to be potentially problematic. Based on the modelled small-scale health data regarding hypertension, the above-mentioned statistical areas are highlighted in Map 4 and suggested for further in-depth analysis.

Map 4. Areas suggested for further in-depth analysis (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021; Behörde für Umwelt, Klima, Energie und Agrarwirtschaft 2017)



The areas were selected based on two main criteria: above average count of hypertensive residents and location near a large noise emitter, such as a highway, airport, or main federal

road (German: Bundesstraße). Naturally, each area must be explored in more detail in terms of building structure, existing natural noise barriers between the emitters and the buildings (such as trees), proportion of the population living in the area affected by the emitted noise, etc. Each of these aspects may have an ameliorating or, on the contrary – an aggravating effect, worth considering. The main use of the modelled health data therefore consists in providing a foundation for further analysis.

Nevertheless, there are certain limitations of the available data, which I would like to address. First, the noise geodata originates from estimations rather than actual measurements. Hence, both data sources – about noise, and about hypertension – are based on modelling approaches. The necessity to carry out further analyses of the local situation before starting to develop any conceptual measures aimed at reducing noise is thus crucial.

Second, only main roads are included in the noise geodata. Naturally, it was not intended for the noise estimations to encompass the entire road network as this would have led to an extremely high computation load. This inherent 'flaw' of the data may thus be leading to the underrepresentation of the real noise situation. As a result, it may not be possible to identify some of the areas with large numbers of hypertensive individuals exposed to excessive noise levels. Nevertheless, this risk must be relatively small as only roads with smaller traffic volume were excluded from the estimations.

Third, the exact place of residence of the hypertensive individuals within the statistical areas is unknown. It is not unlikely that they are not the ones living in the immediate vicinity of the noise emitters. Moreover, even if they do live right on an arterial road, for instance, their apartment may be looking in the opposite direction. They would hence be much less, if at all, affected by the high noise levels.

Finally, only the exposure at home is accounted for as there is no readily available data about where each individual works. While such information can theoretically be generated using agent-based modelling, this approach goes beyond the scope of this dissertation. Since scientific evidence points to higher vulnerability of individuals exposed to occupational noise in addition to being exposed to high noise levels at home (Umweltbundesamt 2006, p.12), neglecting this dimension results in presenting only part of the whole picture.

In summary, the generated synthetic population can be integrated with socio-economic data (e.g., from Hamburg's Social Monitoring), as well as with geodata about various health-relevant aspects of the living environment. In the previous paragraphs, I introduced an example of what such an analysis can look like. The available data allowed to identify potentially problematic neighbourhoods within the spatial realm of the city in terms of noise and its negative effect on the more vulnerable, hypertensive population. The adopted approach enabled determining where in the city are there potential hotspots of individuals suffering from a specific chronic disease exposed to characteristics of the living environment harmful for their health.

The findings should be viewed as a compass direction rather than an actual warning alert. For the latter, additional in-depth analyses must be conducted. The main advantage of the generated synthetic population is that it allows setting filters at the small scale. Thus, the danger of masking heterogeneity because of looking at aggregated population counts available for much

larger spatial units is avoided. Ultimately, this approach can save time and financial resources thanks to the tools it provides for more accurate citywide diagnoses.

In the context of the ongoing COVID-19 pandemic, making the right decisions and doing this quickly can save lives. The next application example is thus going to illustrate how the generated synthetic population can be used for identifying vulnerable population groups at risk of developing severe symptoms of the novel coronavirus at the small urban scale.

8.2. Application Example II: Identifying Spatial Concentrations of Vulnerable Populations within the Context of the COVID-19 Pandemic

In December 2019, Wuhan, a city in the Chinese province of Hubei, faced the outbreak of a novel coronavirus – SARS-CoV-2[29], which caused a fast-spreading respiratory disease, affecting millions of people worldwide. It all started with the emergence of clusters of pneumonia of unknown origin, whereby human infection is assumed to have occurred sometime between the beginning of October and mid-December 2019. On the last day of 2019, the World Health Organisation (WHO) announced the arrival of a novel coronavirus. The first case outside of China was reported in Thailand on January 13th, 2020, followed by cases in more than 20 other countries in South and Southeast Asia, Europe, the USA, and Canada. On February 11th, the new disease was officially named 'COVID-19'. By mid-March, Europe started to struggle with local epidemic outbreaks. At the same time over 170 countries worldwide were hit by the new virus, which led to the WHO declaring it a pandemic.

Next to SARS-CoV-2, there are six other human coronaviruses (as of September 2021). The Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome coronavirus (MERS-CoV) are known for their high mortality but are very rare and do not circulate for long time periods. The other four, HCoV-229E, HCoV-OC43, HCoV-HKU1, and HCoV-NL63, on the other hand, are very common and thus mutate frequently causing mostly colds and diarrhoea (van Damme et al. 2020, p.7). Genetically, SARS-CoV-2 is closely related to SARS-CoV – the first deadly human coronavirus, which posed a pandemic threat when it occurred in 2002 in Hong Kong. Still, although SARS-CoV had alarmingly high case fatality rate (CFR) of 9,7%, it disappeared quickly after the implementation of strict public health mitigation measures. SARS-CoV-2, albeit not as deadly, is far more transmissible. Compared to other coronaviruses and influenza, it has the longest incubation period of four to twelve days. On top of that, there is no interval between the onset of the first symptoms and the maximum infectivity, which significantly contributes to its high rate of transmission and impedes containment efforts (Petersen et al. 2020, p.e238). According to van Damme et al. (2020), the '*transmission dynamics of SARS-CoV-2 can be compared with influenza*', but unlike influenza, this is a new pathogen (p.5).

There are various factors affecting transmission: population density (e.g., people per household, indoor space per person), age structure (i.e., proportion of the elderly and children), forms of religious and social events, socially accepted mode of greeting (e.g., kissing, hugging, shaking hands, etc.), frequency of hand washing, availability of ventilation and air conditioning. Geographical factors such as climate, urbanisation rate, air traffic intensity, and population

[29] Severe acute respiratory syndrome coronavirus 2

movements also play a role in accelerating the transmission of the virus (van Damme et al. 2020, p.4).

In the case of SARS-CoV-2, the demand for hospitalization and intensive care is considerably higher compared to the 2009 influenza pandemic '*because of the subset of patients who develop acute respiratory distress syndrome*' (Petersen et al. 2020, p.e240). Mortality is '*strongly skewed towards people older than 70 years, dissimilar to the 1918 and 2009 influenza pandemics*' (Petersen et al. 2020, p.e238). Other factors, which strongly influence CFR are male gender, comorbidities, BMI, and the '*adequacy of supporting treatment, mainly oxygen therapy*' (van Damme et al. 2020, p.11).

Against this background, COVID-19 – the infectious disease caused by SARS-CoV-2, poses a worldwide challenge to national public health systems. Political decisions must be taken fast and yet remain open and flexible for timely changes depending on the current situation – be it skyrocketing surge of new cases or constantly dropping infection rates.

Responses to the pandemic can so far generally be divided into so-called 'coping strategies' and 'collective strategies'. Coping strategies encompass actions, which people and families take to protect themselves from infection and/or to combat the onset of symptoms. Collective strategies, on the other hand, are either voluntary or mandatory actions designed by local authorities and intended for the general public. While coping strategies are mainly about increasing personal hand hygiene, following respiratory etiquette, keeping distance, and wearing a face mask, collective strategies may include more drastic measures such as introducing mass masking, closing down schools, kindergartens, places of worship, cancelling cultural events, temporarily terminating public transport services, limiting national and international travel and even imposing complete lockdowns (van Damme et al. 2020, pp.7–8).

At the beginning of the pandemic, some epidemiologists (mostly in UK, Sweden, and the Netherlands) recommended to try and build herd immunity instead of imposing draconian measures aimed to contain the spread of SARS-CoV-2. However, the intensity of transmission quickly led to more and more countries imposing some form of lockdown '*ranging from very strict ('Chinese, Wuhan style'), over intermediary ('French/Italian/New York City style' and 'Hong Kong style'), to relaxed ('Swedish style'), or piecemeal*' (van Damme et al. 2020, p.8). Whether lockdowns turn out to be effective depends on a variety of factors including what stage of the local epidemic they are introduced at and what their scope is. The willingness of the population to adhere to the imposed measures, their trust in the government, and the degree of enforcement by the public authorities are of critical importance for making these decisions (ibid.).

Against this background, certain population groups carry heavier burden than others – be it in financial terms because of the imposed lockdowns, in terms of their physical health and mental wellbeing, etc. Since there are various manifestations of *vulnerability* resulting from COVID-19, it is important to define the use of this term for the purposes of the application example.

8.2.1. Defining vulnerability in the context of COVID-19

Over the past year, COVID-19 has touched the life of every one of us. Nevertheless, the exact implications vary significantly depending on individual characteristics including age, health status, socioeconomic standing, working conditions, and living situation. In the context of the novel coronavirus, '*deep-rooted inequalities* […] *can lead to the pandemic having a disproportional*

impact on groups that were already in a situation of greater vulnerability' (United Nations 2020, p.3). Having an unstable job and irregular income, living in poor housing conditions, being socially isolated, perceiving one's own health as poor, and/or struggling from a mental health condition puts people at particular risk of experiencing a more severe impact of the coronavirus pandemic (OECD 2020, p.3). The World Health Organisation (2020) identifies the following population groups as especially vulnerable in an urban setting: those living in informal settings, homeless people, refugees and migrants, the elderly – especially those living in isolation, those with underlying medical conditions, socially marginalised groups as well as those at risk of violence during lockdown (p.5). Where in the city one lives is not of lesser importance. Data from London and New York suggests that per-capita infection rates are higher in socially deprived neighbourhoods where the average household size is larger compared to more affluent urban areas (United Nations 2020, p.9).

Still, depending on the individual situation, the specific challenges vulnerable population groups must face may differ. Elderly people, for instance, are mainly concerned with direct effects on their physical health – serious complications may result from underlying medical conditions, or their general health may worsen even after they have successfully healed from the virus. Furthermore, strong confinement measures aimed at protecting them from an infection can have an adverse side effect as they significantly restrain their independence and limit their social interactions. For those living alone or in long-term care, social isolation is especially difficult to bear. Last, but not least, the imposed lockdown measures (may) cause the disruption of necessary routine check-ups for those suffering from a chronic disease (OECD 2020, p.18).

Women constitute another particularly vulnerable population group. Single parenthood and having a lower or irregular income is more often the case for women than men. Moreover, prolonged times of social isolation have led to increasing numbers of domestic violence affecting women's personal safety (OECD 2020, pp.18–19).

Home confinement is expected to have a long-lasting adverse effect on children's mental health, thus making them another vulnerable population group in the context of the COVID-19 pandemic. Especially in the case of families living in overcrowded accommodations, the lack of personal space often leads to frustration. Boredom and the prolonged loss of contact with friends and classmates can manifest as additional stressors and thus negatively influence the emotional well-being of children and adolescence (OECD 2020, p.19).

In view of these multifaceted implications that COVID-19 can have for different population groups, I want to narrow down the definition for *vulnerable populations* for the purpose of this application example as *individuals at higher risk for developing severe symptoms of COVID-19*.

What constitutes *severe symptoms* of COVID-19? The WHO defines severe symptoms as '*severe acute respiratory illness (fever and at least one sign/symptom of respiratory disease, e.g. cough, shortness of breath; AND requiring hospitalization)*' (Clark et al. 2020, p.e1003). Robert Koch-Institute (2020a) outlined several risk factors, which often cause the development of severe symptoms of COVID-19 and thus possibly lead to complications. These include age[30],

[30] the risk is constantly rising from 50 years onwards (Robert Koch-Institute 2020b)

gender[31], smoking, obesity, coronary heart disease, hypertension, chronic pulmonary disease, chronic liver disease, chronic kidney disease, diabetes mellitus, cancer, and weakened immune system due to chronic illness and/or the regular intake of specific medications such as Cortison. Clark et al. (2020) also point out that '*older individuals, males, and those with underlying conditions such as cardiovascular disease and diabetes are at increased risk of severe COVID-19 and death*' (p. e1004). Asthma, on the other hand, is relatively common and only moderate to severe cases are considered a risk factor according to US guidelines (ibid. e1005). Data from UK suggests that COVID-19-related death is associated with: being male, older age, deprivation, diabetes, severe asthma, and various other medical conditions (Williamson et al. 2020, p.430).

8.2.2. Identifying vulnerable populations at the small urban scale

In an urban setting, there are diverse subpopulations. Hence, their individual characteristics make up a complex, heterogeneous spatial reality. While spatially aggregated data does provide some information about existing (e.g., socio-demographic) patterns, information about the interconnectedness of multiple risk factors is still missing at the small scale.

For the governments to be able to design place-specific responses taking into account existing spatial disparities of the pandemic impact, '*disaggregated mapping of COVID-19 vulnerability* […] *within cities is critical*' (United Nations 2020, p.12). This would facilitate a more efficient outlining of strategies aimed at reducing the transmission of the virus in vulnerable groups. The WHO defines this approach as 'shielding', i.e., introducing a combination of '*measures to protect vulnerable persons at increased risk of severe disease from COVID-19* […] *or increased risk of infection*' (Clark et al. 2020, p.e1004). This strategy can bring direct as well as indirect benefits. On the one hand, it can reduce mortality in susceptible population groups. On the other, it can facilitate the mitigation of a surge in demand for hospital beds (ibid.). Furthermore, identifying vulnerable population groups within cities can foster reaching out to them directly or targeting local actors (e.g., chronic disease support groups) for a more efficient dissemination of information (World Health Organization 2020, p.15). At the same time, an optimised allocation of measures and resources could be facilitated. For instance, neighbourhoods with a concentration of people at risk of developing severe symptoms can be prioritised in the process of vaccination. Allocation of pop-up testing sites for COVID-19 and/or pop-up flu shot clinics within the city could also be more effective if based on such knowledge.

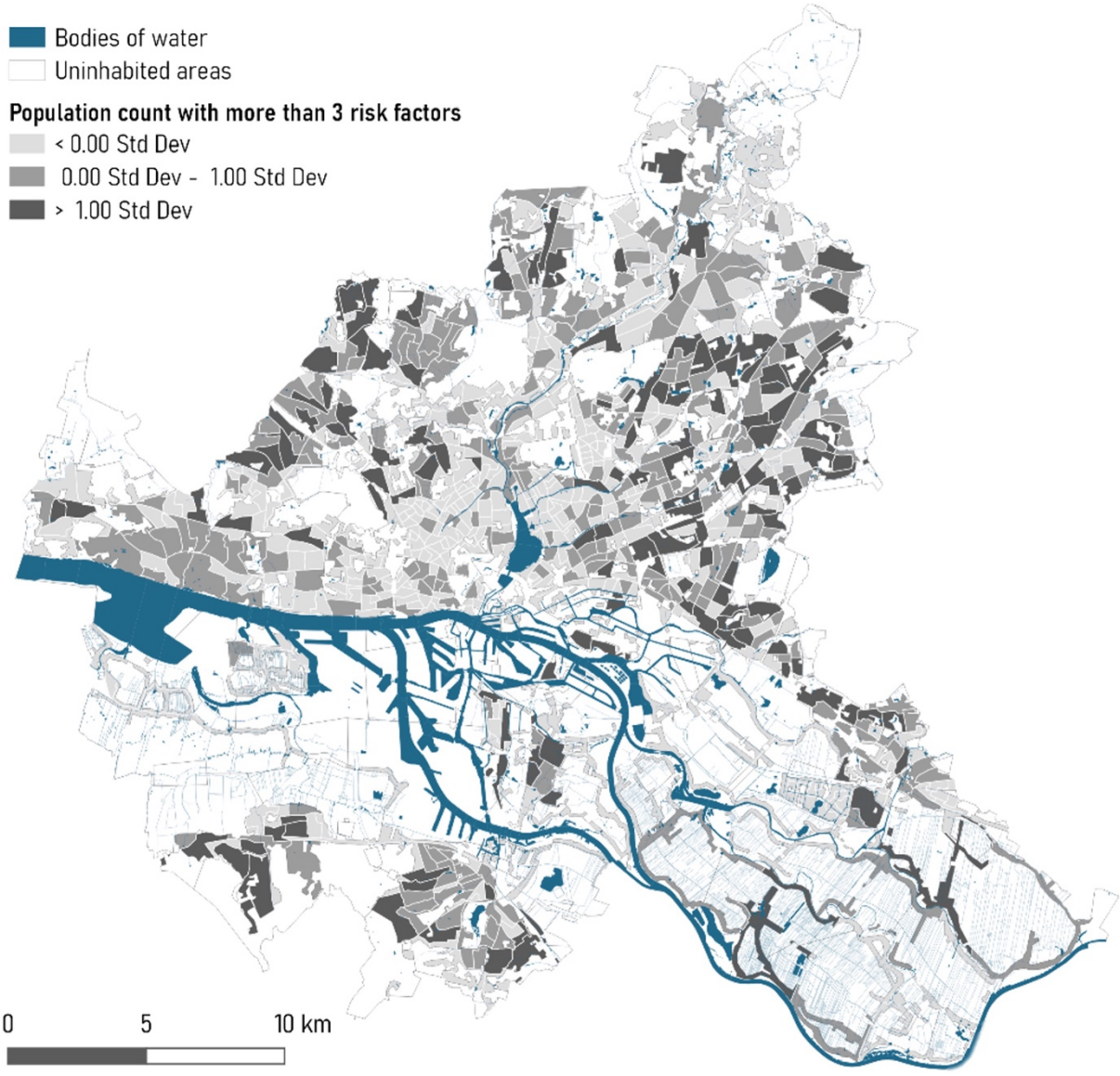
As of June 2020, the USA have introduced a 'COVID Local Risk Index' which *'estimates city- and neighbourhood-level risk of COVID infection and illness severity based on social and economic factors and the distribution of age, race/ethnicity and underlying health outcomes in the community'* (City Health Dashboard 2020). While this index ultimately allows mapping population vulnerability at the neighbourhood level, it is composed using aggregated data. In contrast, the generated synthetic population allows much more flexibility in terms of defining and localising vulnerable population groups within the city depending on any number of combinations of different risk factors.

[31] males are slightly more vulnerable than females according to frequency of hospital admission (Clark et al. 2020, p.e1007)

As highlighted in the previous paragraphs, the main risk factors for developing severe symptoms of COVID-19 are age, gender, chronic medical conditions such as hypertension, diabetes mellitus, cancer, respiratory, liver, and kidney diseases, obesity, and smoking. Several of those are available as individual attributes in the synthetic population. Although the external validation did not manage to completely verify all of them, the generated small-scale data can still lay a foundation for further analyses.

According to the Robert Koch-Institute (2020a), the risk for the elderly people with underlying medical conditions to develop a severe reaction to the novel coronavirus leading to hospital admission rises with each chronic disease. With this in view, Map 5 illustrates spatial concentrations of individuals with more than three risk factors, including age of more than 65 years, male gender, hypertension, heart failure, diabetes mellitus, cancer, smoking, and obesity.

Map 5. Individuals with more than 3 risk factors for developing severe symptoms of COVID-19 (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021)



Age and gender alone constitute two risk factors. However, since they are universally distributed personal characteristics, they cannot lead to the formation of spatial concentrations. Looking at individuals with more than three risk factors therefore ensures accounting for (at least)

two other specific conditions and thus increases the odds of identifying hotspots of vulnerable populations.

Like in the previous application example, I classified the statistical areas into three groups using standard deviation. Areas, where the total count of individuals with more than three risk factors is below the average of 170 people, fall into the first category (< 0.00 Std Dev). Areas, where the corresponding count is between 170 and 250 people, fall within the second category (0.00 – 1.00 Std Dev). Finally, areas with more than 250 such individuals belong to the third category (> 1.00 Std Dev). The dark grey areas in Map 5 thus represent the highest spatial concentrations of populations at greater risk for developing severe symptoms of COVID-19.

The exact operationalisation of vulnerability in terms of risk factors count, weighting of certain comorbidities, and the like, is not the primary goal here. As the virus is mutating, risk factors will be subject to change. The modelled individual health data offers the much-needed flexibility in this regard. It can be used to illustrate the small-scale distribution of one or several specific diseases in combination – both in absolute and relative terms. Visualisations such as Map 5 can thus be used as tools facilitating the navigation of the complex vaccination process.

In the City of Toronto, for instance, residents living in so-called 'hot spot postal codes' have a priority for getting a vaccine: '*Hot spot areas are neighbourhoods identified by the Province of Ontario with ongoing and historic high rates of COVID-19 transmission, hospitalization and death. These areas were also identified by the province to have populations with higher risk factors including racialization, income, quality of housing, immigration status and education attainment*' (City of Toronto 2021).

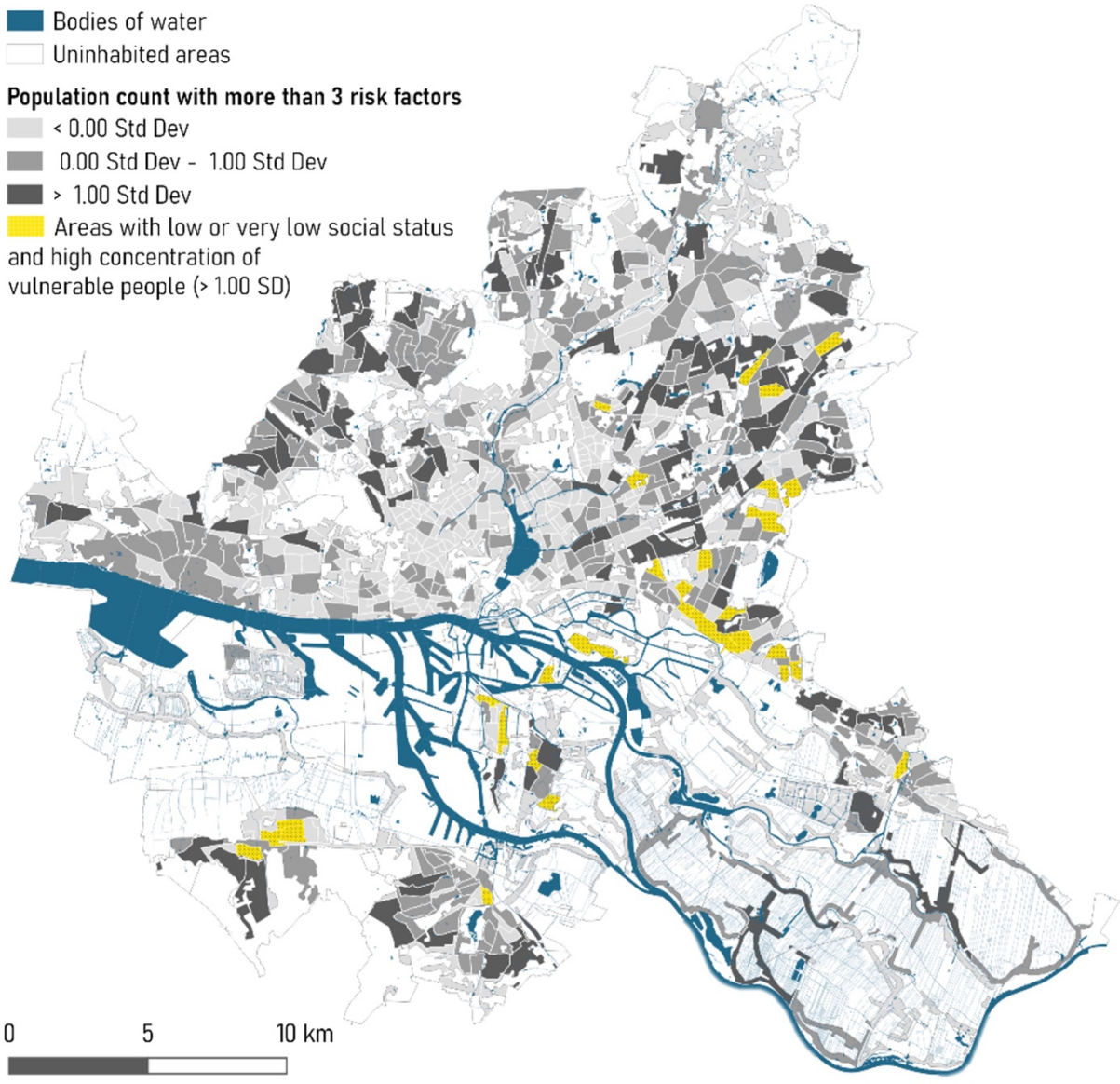
The Canadian approach is an example of using small-scale spatial data to prioritise the vaccination of certain population groups. While the city of Toronto has set the focus on the socio-economic dimension of vulnerability, this does not mean that the physiological dimension is irrelevant. It is much more likely that there was simply no small-scale health data available. The generated synthetic population can thus be used to build upon the Canadian approach and thus reveal an even broader spectrum of individual factors increasing vulnerability.

To add the socio-economic dimension of vulnerability following the example of Toronto, the social status of the statistical areas can be considered as well. Map 6 thus highlights the previously identified statistical areas with concentrations of vulnerable populations, which additionally have either low or very low social status according to the recent Social Monitoring (Behörde für Stadtentwicklung und Wohnen 2020).

All selected areas are located either east of the Alster Lake or south of the Elbe River, thus confirming already familiar patterns of social inequality. In this context, recent evidence from Cologne points out to higher infection rates in densely populated, socially deprived districts (German: *Stadtbezirke*). During the third wave of COVID-19 infections, there were significant differences in the 7-day incidence across city districts, ranging from 0 to 700 (Deutschlandfunk 2021). Johannes Nießen, the Head of the Public Health Department in Cologne, stated that deprived city quarters were especially affected at the time being. In this regard, the Public Health Department established a local testing centre in one of the severely affected areas, where there is also a scarcity of general practitioners - Cologne-Chorweiler. The next step was to open a vaccination centre (ibid.). At the same time, *mobile*, or *pop-up* vaccination centres

can be set up in areas where the concentrations of vulnerable populations are higher at a certain point in time – due to virus mutations related to specific risk factors, for instance. The latter can serve as examples for protecting the most vulnerable population groups by improving their access to testing and vaccination, thus adopting a location-oriented approach.

Map 6. Areas with low social status and high concentration of individuals at increased risk for developing severe symptoms of COVID-19 (own representation, geodata source: Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg 2021; social status source: Behörde für Stadtentwicklung und Wohnen 2018)



There are many possibilities for using the generated synthetic population to illustrate patterns of vulnerability – be it solely physiological (as in the case of comorbidities), behavioural (e.g., smoking), socio-economic (social status of the statistical area of residence), or all the above. The introduced examples of visualising vulnerability at the scale of the statistical areas represent merely a couple of many possible approaches that can prove to be useful for developing future strategies for protecting the most vulnerable population groups. The bottom line is that

the generated small-scale data provides a multitude of opportunities. It fills a gap. It can serve as a starting point. It can lay a foundation for more detailed explorations.

In general, the interviewed experts (see Chapter 5. 'Modelling Health Data on a Small Urban Scale from the Perspective of German Public Health Researchers') were rather reserved about how reliable the modelled data would be and what kind of impact the knowledge of the spatial distribution of vulnerable people at the small scale may have on the further unfolding of the pandemic. Nonetheless, there are examples from abroad about small-scale data being used to identify spatial hotspots of vulnerable individuals and thus support local authorities in their efforts to save lives.

While COVID-19 is the most recent example of a pandemic we must deal with in a densely urbanised world, it is unlikely to be the last one. I therefore doubt that having spatially fine-grained health data can ever be obsolete, or even a downside. We cannot know for sure what exactly is coming our way and thus having as detailed data as possible can only be a plus. Epidemics always have a spatial *ingredient* to them. Monitoring how viruses are spreading is vital. Many researchers are currently dedicated to the development of models predicting the rise and scope of infections, including their spatial and temporal manifestation. In this context, knowing where the most vulnerable people live can be an additional asset for combating the current and future pandemics. This is exactly the contribution that the generated spatial microsimulation model can make.

9. SUMMARY AND DISCUSSION

In the previous pages, I explored a spatial microsimulation approach to modelling individual health data on a small urban scale. To that end, the city of Hamburg served as case study. The entire data modelling process included several main stages. First, I examined the spatial structure of Hamburg to determine which level of administrative division is going to be most suitable as *data holder*. Then, I singled out available data sources for the purpose of population synthesis. These included a micro dataset from a national representative survey and several geographic datasets containing socio-demographic and health data aggregated at different spatial scales. Next, I carried out logistic regression tests to determine to what extent the identified constraint variables can explain the variance of the selected health-related target variables. I was thus able to compile the final list of constraint and target variables. Then, I used deterministic iterative proportional fitting to design an algorithm for constraining the synthetic population at two spatial scales – the city quarters and the statistical areas. Thus, I generated a dataset containing synthetic individuals with health-related attributes for each statistical area. To estimate the coherency of the model, I carried out an internal validation using various metrics such as TAE, RE, RMSE, and MAPE. For the evaluation of the model's goodness of fit, I used two external data sources – sample data from a survey conducted in six statistical areas with differing social status, and insurance data aggregated at the level of the status index classes from three health insurance funds in Hamburg – AOK Rheinland/Hamburg, BKK Mobil Oil, and DAK-Gesundheit.

Overall, the model validation pointed to differences in the goodness of fit depending on type of disease. The model exhibited adequate fit for hypertension and depression both according to the survey and the health insurance data. The sample survey data suggested extraordinary model fit for cancer as opposed to the health insurance data, which delivered less encouraging results. Diabetes mellitus appeared to be poorly modelled both according to the survey and the insurance data. The simulated spatial distribution of heart failure was evaluated as relatively satisfying with the survey data. According to the insurance data, however, the model fit for this variable was completely inadequate.

The survey data allowed verifying health behaviour variables as well. The model fit for smoking and sporting activity was nearly perfect. Obesity and overweight were also well represented by the synthetic population. In contrast, impairment in daily activities due to illness and suffering from a chronic medical condition were rather poorly illustrated.

Notwithstanding, the external data used for model evaluation had some inherent flaws, such as insufficient population diversity and limited geographic coverage. These limitations were explored in detail in Chapter 7.2. 'External validation results'. For this reason, I am interpreting the validation results from a comparison rather than an actual evaluation perspective. Since there was no other readily available data, this approach proved to be the better alternative to skipping external validation altogether.

Following the model validation, I introduced two application examples for the generated small-scale health data. First, I demonstrated the opportunities for visualising the spatial distribution of chronic disease and its possible interactions with factors of the living environment. To that end, I examined the small-scale distribution of hypertensive individuals and their exposure to

noise from road and air traffic. This example illustrated the main purpose of modelling individual health data at the small urban scale – it allows revealing effects of the living environment taking place in a highly localised manner. When health data is limited to absolute or relative disease counts at some level of spatial division encompassing tens of thousands of inhabitants, such potential patterns generally remain undetected.

The second application example demonstrated how the synthetic population can be used to identify local hotspots of vulnerable individuals in the context of the COVID-19 pandemic. The novel coronavirus confronts us with an ever-changing flow of information about the risk factors for developing severe symptoms leading to hospital admission and, possibly, death. Chronic diseases including hypertension, diabetes mellitus, and cancer, as well as socio-demographic factors such as old age and male gender were identified as such from the very beginning. The generated synthetic population thus allows detecting spatial concentrations of individuals subject to several risk factors. Naturally, the spectrum of risk factors may change over time, but my aim was not to exhaust all imaginable possibilities. Instead, I strived to demonstrate the overall advantages of the explored approach to modelling individual health data. Depending on data availability, the introduced spatial microsimulation method can generate an even broader palette of individual health-related attributes, thus offering more possibilities for analysis.

Next to the health-related risk factors, scientific evidence about the impact of social deprivation of urban neighbourhoods on COVID-19 infection rates is mounting. I thus integrated data available from the recent Social Monitoring to identify statistical areas with low or very low social status on the one hand, and high concentration of inhabitants with multiple risk factors for developing severe symptoms on the other. Similar approaches are already adopted in Canada to navigate more efficiently the complex vaccination process. The added value of the generated synthetic population is that it offers individual health data next to the already available socio-economic data classifying neighbourhoods in terms of social status. It thus builds upon the methodology applied in the city of Toronto for prioritising the vaccination of population groups living in deprived areas.

It was beyond the scope of this dissertation to delve deeper into the topic of hypertension and noise exposure, or that of COVID-19 and population vulnerability. Exhausting the spectrum of possibilities for specific measures for health promotion and disease prevention after the additional in-depth analysis of the local situation was never defined as objective of this dissertation. Instead, my goal was to contribute to the scientific dialogue about available options for obtaining individual health data on a small urban scale, and to demonstrate its advantages over aggregated health data.

Getting access to health data is a sensitive topic. There are solid arguments in favour of its protection. Looking at Germany, and more specifically – Hamburg, the procedure for acquiring data from health insurance funds is extremely complex. To obtain such data at the neighbourhood level is not at all possible – at least not for now. As each health insurance fund manages the data of its own insurants, there is currently no centralised point of contact for data acquisition. This poses an additional challenge for researchers, who strive to quickly gain insight into ongoing health processes and their manifestation within the spatial realm of the city. At the same time, health has, now more than ever, become a topic everyone is interested in. To

combat the COVID-19 pandemic, researchers should be making use of all tools available for getting as much and as detailed information as possible. Settling for spatially aggregated health data – especially when the latter is delivered at the scale of areas inhabited by tens of thousands of individuals – is no longer an option.

The appreciation for small-scale data is growing – especially in light of the current health crisis. In North America, socio-demographic and socio-economic data available at the neighbourhood scale is increasingly used to facilitate decision-making. One of its applications, in USA, is to estimate the local risk of infection and its potential severity. Prioritisation of vulnerable populations for vaccination is an example of how such small-scale data is used in Canada.

This type of small-scale data is already available in Hamburg, but it is not made use of – at least not as targeted. Hamburg also has the advantage of readily available health data aggregated at the level of the city quarters. The city has therefore a lot of potential in terms of utilising available data sources. The generated spatial microsimulation model offers a possible solution. It is the product of a well-established modelling technique, which has been implemented for decades in countries such as the UK, Australia, and the USA. It opens a whole new spectrum of possibilities for analysis.

Implementing the proposed method to generate individual health data for the entire city of Hamburg is a huge opportunity. The biggest asset of the introduced approach is the two-tier modelling strategy, which allowed constraining the synthetic population using available health data aggregated at the level of the city quarters. This contributed significantly to achieving a more reliable distribution of disease prevalence at the lower spatial scale of the statistical areas. Another essential advantage of the proposed spatial microsimulation approach is that the model can be updated with more current or more detailed data anytime. It should thus not be regarded as final product subject to no further changes but as a starting base that can be continuously upgraded and improved. At its current state, the model still represents a huge step forward because it fills the gap of not having any idea what kind of disease patterns unfold below the surface of the city quarters – as heterogenous as they are. The generated synthetic population allows combining individual health data with available social, economic, and geographic data to put all the pieces of the puzzle together. It provides direction and encourages further analysis.

The conducted interviews showed that some public health researchers in Germany are still sceptical about modelling health data mainly due to the possibility of generating unreliable results, not reaching to any relevant conclusions, or simply confirming what has already been known. Nonetheless, constant technological advances in the field of data modelling, coupled with the ever-increasing computational power of computers make it hard to find excuses *not* to use such techniques for research purposes. There being a potential danger to develop a model, which cannot fully do justice to reality with all its diverse characteristics should not be a reason not to try out such methods. The proposed spatial microsimulation model offers something extremely valuable and not readily available from another source – individual health data georeferenced at the scale of Hamburg's neighbourhoods and covering its entire population.

10. CONCLUSION AND OUTLOOK

Obtaining health data on a small urban scale – one corresponding to neighbourhoods – is a highly challenging endeavour due to data protection regulations. Against this background, the contribution of this dissertation was providing an alternative solution for the exploration of health-related patterns within cities. A spatial microsimulation method for generating georeferenced individual health data was introduced, whereby each step of the modelling process was described in detail. Using Hamburg as case study allowed going beyond the theory of spatial microsimulation and compiling an entirely new dataset with synthetic individuals – each of them with health-related attributes and spatial reference about the statistical area of residence. Thus, the first one of the defined research questions – 'How can health-related data be generated at a small urban scale?' – was answered.

The synthetic population was used to demonstrate how individual health data available at the scale of urban neighbourhoods can serve to unmask spatial interactions between environmental and socioeconomic factors on the one hand, and the prevalence of chronic disease on the other. To that end, spatial concentrations of hypertensive individuals exposed to excessive noise from road and air traffic were explored and detected within the spatial realm of Hamburg. The social status of the respective neighbourhoods was also considered in order to assess its contribution to the vulnerability of the inhabitants. With that, the second research question – 'How can spatial interactions between environmental and/or socioeconomic factors and the prevalence of chronic disease be made evident using the modelled data?' – was addressed.

The benefit of using the generated synthetic population in light of the current COVID-19 pandemic was also brought to attention. Knowing which are the risk factors for developing severe symptoms possibly requiring hospital admission, hotspots of vulnerable individuals subject to more than three such factors were identified at the neighbourhood scale. Evidence from the USA and Canada supports the use of similar approaches to prepare for rising infections and navigate the vaccination process more efficiently. The last research question – 'How can the generated small-scale health data facilitate the efforts of public health officials to combat the novel coronavirus?' – was thus answered as well.

There are only a few necessary requirements for transferring the proposed spatial microsimulation approach to other cities. First, an established level of spatial division for data collection purposes must be in place. Second, there must be at least two datasets available – one geographic dataset containing aggregated socio-demographic data for each small area, and one individual level dataset, typically a representative survey, containing health-related information. Finally, some type of external data is necessary for model evaluation. This can be a small survey encompassing several spatial units or aggregated health-data at a larger spatial scale compatible with the one used for the modelling purposes. There are no strict guidelines to follow. Nonetheless, it is advisable to use different data sources for the model evaluation if such are available. Thus, the final 'verdict' about how well the model manages to depict reality is going to be more reliable.

In conclusion, I believe the dissertation managed to hold its promise. The proposed spatial microsimulation approach presents a huge opportunity to fill the existing gap of missing health data at the urban neighbourhood level and thus offers a whole new spectrum of possibilities for analysis. With that, researchers in the field of public health and urban planning could find a

common ground for more intensive cooperation so that cities, not only in Germany or Europe, but also all over the world, could continuously increase their knowledge about existing interactions between disease prevalence and factors of the living environment and thus become healthier places.

11. APPENDIX

Table 44. Sample from the generated synthetic population (own representation)

city quarter	statistical area	age	sex	single household	employed	hypertension	heart failure	diabetes	cancer	depression	sport in the last 3 months	subjectively perceived health	impairment due to illness	chronic illness	smoking	Body Mass Index
Hausbruch	100002	50-54	f	no	yes	no	no	no	no	no	no	good/very good	no	no	no	overweight
Hausbruch	100002	50-54	f	yes	no	no	yes	no	no	no	yes	average/bad/ very bad	yes	yes	no	overweight
Hausbruch	100002	35-39	f	no	no	no	no	no	no	no	no	good/very good	no	no	no	normal weight
Hausbruch	100002	30-34	f	no	yes	no	no	no	no	no	yes	average/bad/ very bad	no	yes	no	normal weight
Hausbruch	100002	50-54	m	yes	yes	no	no	no	no	no	yes	good/very good	no	no	no	overweight
Hausbruch	100002	40-44	f	no	yes	no	no	no	no	no	no	good/very good	no	no	no	overweight
Hausbruch	100002	50-54	f	no	yes	no	no	no	no	no	no	good/very good	yes	no	no	normal weight
Hausbruch	100002	18-24	m	no	yes	no	no	no	no	no	yes	good/very good	yes	no	no	normal weight
Hausbruch	100002	60-64	m	no	yes	no	no	no	no	no	no	good/very good	no	no	no	overweight
Hausbruch	100002	30-34	m	no	no	no	no	no	no	no	yes	average/bad/ very bad	yes	yes	no	normal weight
Hausbruch	100002	35-39	f	no	no	no	no	yes	no	no	yes	average/bad/ very bad	no	no	no	overweight
Hausbruch	100002	40-44	f	no	no	no	no	no	no	no	no	average/bad/ very bad	yes	yes	no	obesity
Hausbruch	100002	55-59	f	no	no	no	no	no	no	no	no	good/very good	no	no	no	obesity
Hausbruch	100002	40-44	m	no	yes	no	no	no	no	no	yes	good/very good	no	no	no	normal weight
Hausbruch	100002	55-59	f	no	yes	no	no	no	no	no	yes	good/very good	yes	yes	no	normal weight
Hausbruch	100002	25-29	f	yes	yes	no	no	no	no	no	yes	good/very good	no	no	no	normal weight
Hausbruch	100002	50-54	f	no	yes	no	no	no	no	no	no	good/very good	yes	no	no	normal weight
Hausbruch	100002	50-54	f	no	no	no	no	no	no	yes	yes	average/bad/ very bad	yes	no	no	obesity
Hausbruch	100002	18-24	m	no	yes	no	no	no	no	no	yes	good/very good	yes	no	no	normal weight
Hausbruch	100002	25-29	m	no	yes	no	no	no	no	no	yes	average/bad/ very bad	yes	yes	no	normal weight
Hausbruch	100002	35-39	m	no	yes	no	no	no	no	no	yes	good/very good	no	no	no	overweight
Hausbruch	100002	40-44	f	no	no	no	no	no	no	no	no	average/bad/ very bad	yes	yes	no	obesity
Hausbruch	100002	35-39	f	no	yes	no	no	no	no	no	yes	good/very good	no	no	no	normal weight
Hausbruch	100002	55-59	f	no	yes	no	no	no	no	yes	yes	good/very good	no	no	no	normal weight

Table 45. External validation with insurance data: Absolute counts by age, gender, and status index (own representation)

Status Index	Age & Gender	Hyper. (insur.)	Hyper. (model)	Diab. (insur.)	Diab. (model)	Heart failure (insur.)	Heart failure (model)	Cancer (insur.)	Cancer (model)	Depr. (insur.)	Depr. (model)
high	18-44_m	17 928	35 914	1 117	10 263	560	3 497	7 416	1 855	31 953	36 806
	18-44_f	15 699	14 741	1 313	9 689	535	1 768	11 620	3 636	54 894	49 776
	45-64_m	11 730	11 299	798	4 133	694	1 803	3 635	1 015	5 431	4 938
	45-64_f	10 027	10 611	529	2 303	290	1 144	5 612	1 722	9 769	8 381
	over65_m	40 689	34 114	3 931	16 735	9 139	11 086	20 668	7 044	8 958	6 776
	over65_f	64 014	56 469	4 227	18 712	14 447	19 271	23 912	7 490	23 690	17 941
average1	18-44_m	17 980	36 026	1 105	10 312	559	3 497	7 412	1 863	32 142	36 908
	18-44_f	15 730	14 766	1 334	9 697	512	1 778	11 683	3 652	55 436	50 305
	45-64_m	14 651	13 634	1 017	5 493	858	2 220	3 504	1 148	7 605	6 101
	45-64_f	13 833	13 361	783	3 308	483	1 475	6 311	1 859	13 023	10 499
	over65_m	40 996	34 518	4 205	17 223	9 468	11 169	20 111	6 969	9 636	6 842
	over65_f	65 285	57 687	4 765	19 606	14 875	19 665	23 871	7 577	24 591	18 404
average2	18-44_m	18 013	36 178	1 111	10 333	570	3 497	7 443	1 870	32 314	37 039
	18-44_f	15 760	14 820	1 324	9 702	516	1 795	11 698	3 654	55 685	50 719
	45-64_m	14 756	13 741	1 207	5 897	889	2 361	3 546	1 173	8 777	6 355
	45-64_f	14 727	13 665	1 082	3 700	610	1 519	5 938	1 814	14 550	10 575
	over65_m	40 411	33 909	4 376	17 195	9 556	11 112	19 454	6 820	9 717	6 835
	over65_f	65 090	57 655	4 843	19 874	15 136	19 780	23 353	7 387	24 711	18 321
average3	18-44_m	17 980	36 020	1 112	10 288	561	3 497	7 440	1 888	32 231	36 945
	18-44_f	15 809	14 772	1 337	9 698	516	1 778	11 702	3 645	55 575	50 323
	45-64_m	11 748	10 487	1 120	4 350	849	1 818	2 638	871	7 208	5 122
	45-64_f	11 561	10 682	971	2 825	564	1 205	4 060	1 406	11 451	8 240
	over65_m	36 535	30 022	3 979	15 663	9 028	9 992	17 804	5 977	8 923	6 183
	over65_f	60 750	52 846	4 572	18 226	14 596	18 832	21 457	6 613	23 235	16 603
average4	18-44_m	18 014	35 989	1 112	10 310	567	3 497	7 399	1 862	32 259	36 949
	18-44_f	15 756	14 779	1 331	9 694	522	1 784	11 709	3 647	55 327	50 163
	45-64_m	10 748	9 456	1 144	4 337	841	1 766	2 252	765	7 035	4 639
	45-64_f	10 749	9 290	817	2 646	443	1 091	3 356	1 144	10 548	7 115
	over65_m	35 536	29 143	3 815	15 329	8 905	9 839	17 253	5 892	8 881	5 990
	over65_f	59 933	51 662	4 485	17 981	14 515	18 553	20 821	6 526	23 106	16 184
low	18-44_m	18 007	35 908	1 104	10 275	570	3 497	7 432	1 855	32 079	36 894
	18-44_f	15 752	14 774	1 327	9 682	519	1 783	11 693	3 630	54 993	49 925
	45-64_m	8 224	7 031	839	3 253	615	1 289	1 669	567	5 268	3 288
	45-64_f	8 220	7 173	707	2 076	441	784	2 431	793	7 974	5 043
	over65_m	33 043	26 516	3 486	14 187	8 669	9 222	16 169	5 252	8 262	5 552
	over65_f	56 572	48 469	4 169	16 870	14 167	17 847	19 513	5 928	21 416	14 774
very low	18-44_m	18 018	35 920	1 100	10 309	561	3 497	7 496	1 843	32 084	36 938
	18-44_f	15 743	14 780	1 328	9 679	521	1 801	11 667	3 643	54 952	50 016
	45-64_m	8 455	7 417	918	3 537	671	1 443	1 660	641	5 506	3 736
	45-64_f	8 854	7 348	866	2 283	484	817	2 427	784	8 710	5 382
	over65_m	32 978	26 251	3 626	14 240	8 804	9 180	16 277	5 245	8 420	5 617
	over65_f	55 977	47 758	4 233	16 716	14 461	17 657	19 241	5 832	21 767	14 612

Approval confirmations from the interviewees

Heiko Becher:

From: h.becher@uke.de
Sent: 28 April 2021 15:25
To: evgenia.yosifova@hcu-hamburg.de
Subject: Re: Freigabe Experteninterview für meine Dissertation

Liebe Frau Yosifova,

besten Dank, ich bin mit dem Text voll einverstanden.

Viel Erfolg weiterhin bei der Dissertation und viele Grüße

Enno Swart:

From: enno.swart@med.ovgu.de
Sent: 28 April 2021 16:28
To: evgenia.yosifova@hcu-hamburg.de
Subject: AW: Freigabe Experteninterview für meine Diss

Liebe Evgenia,

Sehr schön. Geht für mich so in Ordnung.

Liebe Grüße
Enno

Susanne Busch:

From: susanne.busch@haw-hamburg.de
Sent: 14 September 2021 21:05
To: evgenia.yosifova@hcu-hamburg.de
Subject: AW: Freigabe Experteninterview für meine Dissertation

Liebe Evgenia,

hier ist meine Zustimmung zu dem Wortlaut unten.

Alles Liebe und für den Endspurt alles Gute.
Susanne

Jobst Augustin:

From: jo.augustin@uke.de
Sent: 11 May 2021 08:44
To: evgenia.yosifova@hcu-hamburg.de
Subject: AW: Freigabe Experteninterview für meine Dissertation

Guten Morgen,

anbei nun das noch einmal gesichtete Dokument - sie können es so verwenden.

Beste Grüße
Jobst Augustin

12. LIST OF ABBREVIATIONS

ALKIS	Amtliches Liegenschaftskatasterinformationssystem (engl: Authoritative real estate cadastre information system)
AOK	Allgemeine Ortskrankenkasse (engl: General local health insurance fund)
BKK	Betriebskrankenkasse (engl: Company health insurance fund)
BMI	Body Mass Index
CFR	Case Fatality Rate
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
COVID	Coronavirus Disease
CPU	Central Processing Unit
DBP	Diastolic Blood Pressure
EEA	European Environmental Agency
EU	European Union
EUR	Euro
FUN	Function
GDPR	General Data Protection Regulation
GEDA	Gesundheit in Deutschland aktuell' (engl: 'Current health situation in Germany')
HAW	Hochschule für Angewandte Wissenschaften (engl: University of Applied Sciences)
HIPS	Health Information and Planning System
ICD	International Classification of Diseases
ID	Identification number
IPF	Iterative Proportional Fitting
ISMHSR	Institute of Social Medicine and Health Systems Research
IVDP	Institut für Versorgungsforschung in der Dermatologie und bei Pflegeberufen (engl: Institute of Health Care Research in Dermatology and Nursing)
MAPE	Mean Absolute Percentage Error
MERS	Middle East Respiratory Syndrome
NO_2	Nitrogen Dioxide
O_3	Ozone
OECD	Organisation for Economic Co-Operation and Development
PD	Privatdozent
PM_{10}	Particulate Matter 10
$PM_{2,5}$	Particulate Matter 2,5
QGIS	Quantum Geographic Information System
R^2	R squared
RE	Relative Error

RKI	Robert Koch-Institute
RMSE	Root Mean Squared Error
SARS	Severe Acute Respiratory Syndrome
SBP	Systolic Blood Pressure
SGB	Sozialgesetzbuch (engl: Social code)
SHI	Statutory Health Insurance
SPSS	Statistical Package for the Social Sciences
Std Dev	Standard Deviation
TAE	Total Absolute Error
TRS	Truncate, Replicate, Sample
UHI	Urban Heat Island
UK	United Kingdom
UKE	Universitätsklinikum Eppendorf (engl: University Medical Center Hamburg-Eppendorf)
USA	United States of America
USD	U.S. Dollar
WHO	World Health Organisation

13. FIGURES

14. FORMULAS

15. MAPS

16. TABLES

17. BIBLIOGRAPHY

Aether. (2017). Updated analysis of air pollution exposure in London: Report to Greater London Authority. [online]. Available from: https://www.london.gov.uk/sites/default/files/aether_updated_london_air_pollution_exposure_final_20-2-17.pdf.

Akademie für Raumentwicklung in der Leibniz-Gemeinschaft ed. (2021). *SARS-CoV-2-Pandemie: Was lernen wir daraus für die Raumentwicklung?* Hannover.

American Society for Reproductive Medicine. (2018). Smoking and infertility: a committee opinion. *Fertility and Sterility*, 110(4), pp.611–618.

An, R. (2020). Projecting the impact of the coronavirus disease-2019 pandemic on childhood obesity in the United States: A microsimulation model. *Journal of sport and health science*, 9(4), pp.302–312.

Arbuthnott, K.G. and Hajat, S. (2017). The health effects of hotter summers and heat waves in the population of the United Kingdom: a review of the evidence. *Environmental health : a global access science source*, 16(Suppl 1), p.119.

Ballas, D. (2004). Simulating trends in poverty and income inequality on the basis of 1991 and 2001 census data: a tale of two cities. *Area*, 36(2), pp.146–163.

Ballas, D. et al. (2007). Using SimBritain to Model the Geographical Impact of National Government Policies. *Geographical Analysis*, 39(1), pp.44–77.

Ballas, D., Clarke, G. and Wiemers, E. (2006). Spatial microsimulation for rural policy analysis in Ireland: The implications of CAP reforms for the national spatial strategy. *Journal of Rural Studies*, 22(3), pp.367–378.

Ballas, D. and Clarke, G.P. (2001). Modelling the Local Impacts of National Social Policies: A Spatial Microsimulation Approach. *Environment and Planning C: Government and Policy*, 19(4), pp.587–606.

Battaner-Moro, J., Barlow, C. and Wright, P. eds. (2010). *Social Deprivation and Accessibility to Quiet Areas in Southampton*.

Behörde für Stadtentwicklung und Wohnen. (2018). Sozialmonitoring Bericht 2018 - Karte Gesamtindex. [online]. Available from: https://www.hamburg.de/sozialmonitoring/11884576/sozialmonitoring-bericht-2018/.

Behörde für Stadtentwicklung und Wohnen. (2020). Sozialmonitoring-Bericht 2020: Leichter Trend zur Mitte statt Auseinanderdriften. [online]. Available from: https://www.hamburg.de/sozialmonitoring/14763076/sozialmonitoring-bericht-2020/.

Behörde für Umwelt, Klima, Energie und Agrarwirtschaft. (2017). Lärmkarten Hamburg (§47c BImSchG). [online]. Available from: https://metaver.de/trefferanzeige?docuuid=030A8F47-EBEF-4669-94FC-0299BB7D5C88.

Birkin, M. and Clarke, M. (1988). Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environment and Planning A: Economy and Space*, 20(12), pp.1645–1671.

Braun-Fahrländer, C. (2004). Die soziale Verteilung von Umweltbelastungen bei Kindern in der Schweiz. In G. Bolte & A. Mielck, eds. *Umweltgerechtigkeit: Die soziale Verteilung von Umweltbelastungen*. Gesundheitsforschung. Weinheim: Juventa-Verl., pp. 155–173.

Brosius, F. (2011). *SPSS 19*. 1. Auflage. Heidelberg; München; Landsberg; Frechen; Hamburg: mitp.

Brunt, H. et al. (2017). Air pollution, deprivation and health: understanding relationships to add value to local air quality management policy and practice in Wales, UK. *Journal of public health (Oxford, England)*, 39(3), pp.485–497.

Buchcik, J. et al. (2021). Gesundheitliche Situation in Quartieren mit unterschiedlicher sozialer Lage. In J. Westenhöfer et al., eds. *Gesunde Quartiere: Gesundheitsförderung und Prävention im städtischen Kontext*. München: oekom Verlag, pp. 53–74.

Campbell, M. and Ballas, D. (2013). A spatial microsimulation approach to economic policy analysis in Scotland. *Regional Science Policy & Practice*, 5(3), pp.263–288.

Campbell, M. and Ballas, D. (2016). SimAlba: A Spatial Microsimulation Approach to the Analysis of Health Inequalities. *Frontiers in public health*, 4, p.230.

Carter, H.E., Schofield, D. and Shrestha, R. (2017). The long-term productivity impacts of all cause premature mortality in Australia. *Australian and New Zealand journal of public health*, 41(2), pp.137–143.

Cassells, R., Miranti, R. and Harding, A. (2013). Building a Static Spatial Microsimulation Model: Data Preparation. In R. Tanton & K. L. Edwards, eds. *Spatial microsimulation: A reference guide for users*. Understanding population trends and processes. Dordrecht: Springer, pp. 9–16.

Cassels, S., Clark, S.J. and Morris, M. (2008). Mathematical models for HIV transmission dynamics: tools for social and behavioral science research. *Journal of acquired immune deficiency syndromes (1999)*, 47 Suppl 1, pp.34–9.

Chai, T. and Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp.1247–1250.

Chang, T.-Y. et al. (2015). Acute effects of noise exposure on 24-h ambulatory blood pressure in hypertensive adults. *Journal of hypertension*, 33(3), pp.507–14514.

Chen, X., Meaker, J.W. and Zhan, F.B. (2006). Agent-Based Modeling and Analysis of Hurricane Evacuation Procedures for the Florida Keys. *Natural Hazards*, 38(3), pp.321–338.

Chernick, H.A., Holmer, M.R. and Weinberg, D.H. (1987). Tax policy toward health insurance and the demand for medical services. *Journal of Health Economics*, 6(1), pp.1–25.

Chin, S.-F. et al. (2005). Spatial Microsimulation Using Synthetic Small-area Estimates of Income, Tax and Social Security Benefits. *Australasian Journal of Regional Studies, The*, 11(3), pp.303–335.

City Health Dashboard. (2020). COVID Local Risk Index. [online]. Available from: https://www.cityhealthdashboard.com/metric/1422.

City of Toronto. (2021). COVID-19: How to Get Vaccinated. [online]. Available from: https://www.toronto.ca/home/covid-19/covid-19-protect-yourself-others/covid-19-vaccines/covid-19-how-to-get-vaccinated/?accordion=vaccine-eligibility.

Clark, A. et al. (2020). Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. *The Lancet Global Health*, 8(8), pp.1003–1017.

Clarke, M. et al. (1985). A Strategic Planning Simulation Model of a District Health Service System: The In-patient Component and Results. In K. Davis et al., eds. *THIRD INTERNATIONAL CONFERENCE ON SYSTEM SCIENCE IN HEALTH CARE: Troisime*. Health Systems Research. [Place of publication not identified]: Springer, pp. 949–954.

Clarke, P.M. et al. (2004). A model to estimate the lifetime health outcomes of patients with type 2 diabetes: the United Kingdom Prospective Diabetes Study (UKPDS) Outcomes Model (UKPDS no. 68). *Diabetologia*, 47(10), pp.1747–59.

Dahlgren, G. and Whitehead, M. (1991). Policies and strategies to promote social equity in health. , p.70.

van Damme, W. et al. (2020). The COVID-19 pandemic: diverse contexts; different epidemics-how and why? *BMJ global health*, 5(7).

Deutschlandfunk. (2021). Corona-Hotspots in Köln - Gesundheitsamt: Armut und Wohnen sind maßgebliche Gründe. [online]. Available from: https://www.deutschlandfunk.de/corona-hotspots-in-koeln-gesundheitsamt-armut-und-wohnen.1769.de.html?dram:article_id=496369.

Diener, E. et al. (2017). If, Why, and When Subjective Well-Being Influences Health, and Future Needed Research. *Applied psychology. Health and well-being*, 9(2), pp.133–167.

Dragano, N. and Conte, A. (2020). „Health in All Policies“ und gesundheitliche Chancengleichheit: COVID-19 als Fallstudie. *Public Health Forum*, 28(3), pp.185–187.

Dratva, J. et al. (2012). Transportation noise and blood pressure in a population-based sample of adults. *Environmental health perspectives*, 120(1), pp.50–5.

Edwards, K.L. and Clarke, G. (2013). SimObesity: Combinatorial Optimisation (Deterministic) Model. In R. Tanton & K. L. Edwards, eds. *Spatial microsimulation: A reference guide for users*. Understanding population trends and processes. Dordrecht: Springer, pp. 69–86.

Edwards, K.L. and Clarke, G.P. (2009). The design and validation of a spatial microsimulation model of obesogenic environments for children in Leeds, UK: SimObesity. *Social science & medicine (1982)*, 69(7), pp.1127–34.

Erhart, M. et al. (2013). *Morbiditätsatlas Hamburg - Zusammenfassung: Gutachten zum kleinräumigen Versorgungsbedarf in Hamburg – erstellt durch das Zentralinstitut für die kassenärztliche Versorgung in Deutschland im Auftrag der Behörde für Gesundheit und Verbraucherschutz Hamburg*. Berlin: Zentralinstitut für die kassenärztliche Versorgung in Deutschland.

European Court of Auditors. (2018). Air pollution: Our health still insufficiently protected. [online]. Available from: https://www.eca.europa.eu/Lists/ECADocuments/SR18_23/SR_AIR_QUALITY_EN.pdf.

European Environmental Agency. (2017). Air quality in Europe — 2017 report. [online]. Available from: https://www.apren.pt/contents/publicationsothers/small-airquality2017-15-29-1.pdf.

European Environmental Agency. (2018). Unequal exposure and unequal impacts: social vulnerability to air pollution, noise and extreme temperatures in Europe. [online]. Available from: https://op.europa.eu/en/publication-detail/-/publication/affed40d-1554-11e9-81b4-01aa75ed71a1/language-en/format-PDF/source-85207538.

Evans, W.K. et al. (2013). Canadian Cancer Risk Management Model: evaluation of cancer control. *International journal of technology assessment in health care*, 29(2), pp.131–9.

Fecht, D. et al. (2015). Associations between air pollution and socioeconomic characteristics, ethnicity and age profile of neighbourhoods in England and the Netherlands. *Environmental pollution (Barking, Essex : 1987)*, 198, pp.201–10.

Federal Ministry for the Environment, Nature Conservation and Nuclear Safety. (2015). *Grün in der Stadt - Für eine lebenswerte Zukunft. Grünbuch Stadtgrün.* Bonn.

Fernandez Milan, B. and Creutzig, F. (2015). Reducing urban heat wave risk in the 21st century. *Current Opinion in Environmental Sustainability*, 14, pp.221–231.

Frohlich, K.L. (2013). Area Effects on Behaviour and Lifestyle: The Spatiality of Injustice. In C. Stock & A. Ellaway, eds. *Neighbourhood structure and health promotion*. New York, NY: Springer, pp. 39–60.

Gilbert, N. (2000). Models, Processes and Algorithms: Towards A Simulation Toolkit. In R. Suleiman, K. G. Troitzsch, & G. N. Gilbert, eds. *Tools and techniques for social science simulation*. Heidelberg: Physica-Verlag, pp. 3–16.

Gold, C. et al. (2012). *Aktiv werden für Gesundheit – Arbeitshilfen für Prävention und Gesundheitsförderung im Quartier // Aktiv werden für Gesundheit: Präventiv handeln: Ernährung - Bewegung - Stressbewältigung*. [3., aktualisierte Aufl.]. Berlin: Gesundheit Berlin-Brandenburg.

Hahad, O. et al. (2019). The Cardiovascular Effects of Noise. *Deutsches Arzteblatt international*, 116(14), pp.245–250.

Hamburg Cancer Registry. (2020). Occasion-related estimation of cancer prevalence. Years of diagnosis until 2018. Data version: 01.10.2020.

Harding, A. et al. (2004). *Assessing poverty and inequality at a detailed regional level: New advances in spatial microsimulation*. [online]. Available from: https://www.econstor.eu/handle/10419/63532.

Hoffmann, B., Robra, B.-P. and Swart, E. (2003). Soziale Ungleichheit und Strassenlärm im Wohnumfeld--eine Auswertung des Bundesgesundheitssurveys. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany))*, 65(6), pp.393–401.

Hollander, Y. and Liu, R. (2008). The principles of calibrating traffic microsimulation models. *Transportation*, 35(3), pp.347–362.

Horwich, T.B. and Fonarow, G.C. (2017). Prevention of Heart Failure. *JAMA cardiology*, 2(1), p.116.

Hynes, S. et al. (2009). Building a static farm level spatial microsimulation model for rural development and agricultural policy analysis in Ireland. *International Journal of Agricultural Resources, Governance and Ecology*, 8, pp.282–299.

Jacobs, E. et al. (2017). Healthcare costs of Type 2 diabetes in Germany. *Diabetic Medicine*, 34(6), pp.855–861.

Katzschner, A. and Bruse, M. (2012). Stadtklima und soziale Vunerabilität. In *Umweltgerechtigkeit. Chancengleichheit bei Umwelt und Gesundheit: Konzepte, Datenlage und Handlungsperspektiven*. Bern: Hans Huber, pp. 99–112.

Kavroudakis, D., Ballas, D. and Birkin, M. (2012). A Spatial Microsimulation Approach to the Analysis of Social and Spatial Inequalities in Higher Education Attainment. *Applied Spatial Analysis and Policy*, 3, pp.1–23.

Khreis, H., May, A.D. and Nieuwenhuijsen, M.J. (2017). Health impacts of urban transport policy measures: A guidance note for practice. *Journal of Transport & Health*, 6, pp.209–227.

Kohlhuber, M. et al. (2006). Social inequality in perceived environmental exposures in relation to housing conditions in Germany. *Environmental research*, 101(2), pp.246–55.

Kohlhuber, M. and Bolte, G. (2012). Einfluss von Umweltlärm auf Schlafqualität und Schlafstörungen und Auswirkungen auf die Gesundheit. *Somnologie - Schlafforschung und Schlafmedizin*, 16(1), pp.10–16.

Kohlhuber, M., Schenk, T. and Weiland, U. (2012). Verkehrsbezogene Luftschadstoffe und Lärm. In *Umweltgerechtigkeit. Chancengleichheit bei Umwelt und Gesundheit: Konzepte, Datenlage und Handlungsperspektiven*. Bern: Huber, pp. 87–98.

Konduri, K.C. et al. (2016). Enhanced Synthetic Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions. *Transportation Research Record: Journal of the Transportation Research Board*, 2563(1), pp.40–50.

Kongmuang, C. et al. (2006). *SimCrime: A Spatial Microsimulation Model for the Analysing of Crime in Leeds*. The School of Geography, University of Leeds. [online]. Available from: http://eprints.whiterose.ac.uk/4982/.

Kono, S. et al. (1983). A bio-demographic analysis of Japanese fertility via micro-simulation. *Jinko mondai kenkyu. [Journal of population problems]*, (168), pp.1–29.

Kosar, B. and Tomintz, M. (2014). simSALUD: A Web-based Spatial Microsimulation to Model the Health Status for Small Areas Using the Example of Smokers in Austria. In R. Vogler et al., eds. *Geospatial innovation for society: GI_Forum 2014 ; [Geoinformatics Forum held in Salzburg from July 1-4, 2014*. Berlin; Wien: Wichmann; Verl. der Österr. Akad. der Wiss. ÖAW, pp. 207–216.

Krauth, C. et al. (2014). Resource utilisation and costs of depressive patients in Germany: results from the primary care monitoring for depressive patients trial. *Depression research and treatment*, 2014, p.730891.

Kruize, H. and Bouwman, A.A. (2004). Environmental (in)equity in the Netherlands - A case study on the distribution of environmental quality in the Rijnmond region. [online]. Available from: http://hdl.handle.net/10029/8982.

Landesbetrieb Geoinformation und Vermessung (LGV) Hamburg. (2021). Geoportal Hamburg. [online]. Available from: https://geoportal-hamburg.de/geo-online/#.

Laussmann, D. et al. (2013). Soziale Ungleichheit von Lärmbelästigung und Straßenverkehrsbelastung: Ergebnisse der Studie zur Gesundheit Erwachsener in Deutschland (DEGS1). *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 56(5–6), pp.822–31.

Lay-Yee, R. and Cotterell, G. (2015). The Role of Microsimulation in the Development of Public Policy. In M. Janssen, M. A. Wimmer, & A. Deljoo, eds. *Policy practice and digital science: Integrating complex systems, social simulation and public administration in policy research*. Public Administration and Information Technology. Cham; s.l.: Springer International Publishing, pp. 305–320.

Lejeune, Z. et al. (2016). Housing quality as environmental inequality: the case of Wallonia, Belgium. *Journal of Housing and the Built Environment*, 31(3), pp.495–512.

Lesyuk, W., Kriza, C. and Kolominsky-Rabas, P. (2018). Cost-of-illness studies in heart failure: a systematic review 2004-2016. *BMC cardiovascular disorders*, 18(1).

Loll, B.-U. (1991). Statistische Gebiete als kleinräumige Gliederungseinheiten Hamburgs. *Hamburg in Zahlen*, (4), pp.92–100.

Lone, N.I. et al. (2021). Influence of socioeconomic deprivation on interventions and outcomes for patients admitted with COVID-19 to critical care units in Scotland: A national cohort study. *The Lancet Regional Health - Europe*, 1, p.100005.

Lovelace, R. and Ballas, D. (2013). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, 41, pp.1–11.

Lovelace, R., Ballas, D. and Watson, M. (2014). A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. *Journal of Transport Geography*, 34(1), pp.282–296.

Lovelace, R. and Dumont, M. (2016). *Spatial microsimulation with R*. Boca Raton; London; New York: CRC Press Taylor & Francis Group a Chapman & Hall Book.

Majstorovic, D. et al. (2015). Features and Added Value of Simulation Models Using Different Modelling Approaches Supporting Policy-Making: A Comparative Analysis. In M. Janssen, M. A. Wimmer, & A. Deljoo, eds. *Policy practice and digital science: Integrating complex systems, social simulation and public administration in policy research*. Public Administration and Information Technology. Cham; s.l.: Springer International Publishing, pp. 95–123.

McGeehin, M.A. and Mirabelli, M. (2001). The potential impacts of climate variability and change on temperature-related morbidity and mortality in the United States. *Environmental health perspectives*, 109 Suppl 2, pp.185–9.

Meijer, M. (2013). Neighbourhood Context and Mortality: An Overview. In C. Stock & A. Ellaway, eds. *Neighbourhood structure and health promotion*. New York, NY: Springer, pp. 11–37.

Mindermann, N. et al. (2021). GKV-Routinedaten und Einsatzdaten des Rettungsdienstes mit Quartiers- und Soziallagenbezug. In J. Westenhöfer et al., eds. *Gesundheitsförderung und Prävention im Setting Quartier*. Hamburg: oekom Verlag.

Monteiro, M. de F. and Sobral Filho, D.C. (2004). Physical exercise and blood pressure control. *Revista Brasileira de Medicina do Esporte*, 10(6), pp.513–516.

Moshammer, H., Petersen, E. and Silberschmidt, G. (2002). Ökologische und gesundheitliche Folgen der Mobilität. *Umwelt Medizin Gesellschaft*, 15(3), pp.242–248.

Muñoz, E. and Peters, I. (2014). Constructing an Urban Microsimulation Model to Assess the Influence of Demographics on Heat Consumption. *International Journal of Microsimulation*, 7(1), pp.127–157.

Mustafa, A.F.M. (1973). *Fecundability Differentials Among Acceptors and Non-Acceptors of Family Planning: A Simulation Experiment*. Duke, Chapel Hill: North Carolina State University, Department of Statistics.

Neuhauser, H., Kuhnert, R. and Born, S. (2017). 12-Month prevalence of hypertension in Germany. *Journal of Health Monitoring*, 2(1).

Noctor, E. and Dunne, F.P. (2015). Type 2 diabetes after gestational diabetes: The influence of changing diagnostic criteria. *World journal of diabetes*, 6(2), pp.234–44.

O'Donoghue, C., Morrissey, K. and Lennon, J. (2014). Spatial Microsimulation Modelling: a Review of Applications and Methodological Choices. *International Journal of Microsimulation*, 7(1), pp.26–75.

OECD. (2020). *COVID-19: Protecting people and societies*. OECD.

OECD. (2018). *Divided Cities: Understanding Intra-urban Inequalities*. OECD. [online]. Available from: https://www.oecd-ilibrary.org/urban-rural-and-regional-development/divided-cities_9789264300385-en [Accessed August 4, 2021].

Orcutt, G.H. (1957). A New Type of Socio-Economic System, May 1957. *Review of Economics and Statistics*, (39), pp.116–123.

Orcutt, G.H. et al. (1961). *Microanalysis of Socioeconomic Systems, A Simulation Study*. New York: Harper & Brothers.

Paavola, J. (2017). Health impacts of climate change and health and social inequalities in the UK. *Environmental health : a global access science source*, 16(Suppl 1), p.113.

Padilla, C.M. et al. (2016). City-Specific Spatiotemporal Infant and Neonatal Mortality Clusters: Links with Socioeconomic and Air Pollution Spatial Patterns in France. *International journal of environmental research and public health*, 13(6).

Parkin, D.M. (1985). A computer simulation model for the practical planning of cervical cancer screening programmes. *British Journal of Cancer*, 51(4), pp.551–568.

Petersen, E. et al. (2020). Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases*, 20(9), pp.238–244.

Plaisier, A.P. et al. (1990). ONCHOSIM: a model and computer simulation program for the transmission and control of onchocerciasis. *Computer Methods and Programs in Biomedicine*, 31(1), pp.43–56.

Pohlan, J., Pohl, T. and Selk, A. (2010). Sozialmonitoring im Rahmenprogramm Integrierte Stadtteilentwicklung (RISE).

Pohlan, J. and Strote, J. (2017). Cities under observation: Social monitoring in integrated neighbourhood development in Hamburg. *Procedia Computer Science*, 112, pp.2426–2434.

Popadiuk, C. et al. (2016). Using the Cancer Risk Management Model to evaluate the health and economic impacts of cytology compared with human papillomavirus DNA testing for primary cervical cancer screening in Canada. *Current oncology (Toronto, Ont.)*, 23(Suppl 1), pp.56–63.

R Core Team. (2020). R: A language and environment for statistical computing. [online]. Available from: https://www.R-project.org/.

Rahman, A. (2017). Estimating small area health-related characteristics of populations: a methodological review. *Geospatial health*, 12(1), p.495.

Rees, P., Martin, D. and Williamson, P. (2002). Census data resources in the United Kingdom. In P. Rees, D. Martin, & P. Williamson, eds. *The Census data system*. London: Wiley, pp. 1–24.

Remen, T. et al. (2018). Risk of lung cancer in relation to various metrics of smoking history: a case-control study in Montreal. *BMC cancer*, 18(1), p.1275.

Robert Koch-Institute ed. (2014). *Daten und Fakten: Ergebnisse der Studie »Gesundheit in Deutschland aktuell 2012«: Beiträge zur Gesundheitsberichterstattung des Bundes*. Berlin: RKI.

Robert Koch-Institute. (2011). Facts and Trends from Federal Health Reporting: Diabetes mellitus in Germany. *GBE Kompakt*, 2(3).

Robert Koch-Institute ed. (2015). *Gesundheit in Deutschland aktuell 2012. Public USE File.* Berlin: RKI.

Robert Koch-Institute. (2019). Gesundheitsmonitoring. [online]. Available from: https://www.rki.de/DE/Content/Gesundheitsmonitoring/gesundheitsmonitoring_node.html.

Robert Koch-Institute. (2020a). Informationen und Hilfestellungen für Personen mit einem höheren Risiko für einen schweren COVID-19-Krankheitsverlauf. [online]. Available from: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Risikogruppen.HTML.

Robert Koch-Institute. (2020b). Steckbrief zur Coronavirus-Krankheit-2019 (COVID-19). [online]. Available from: https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html#doc13776792bodyText14.

Roy, S.G. (1984). Demography of sterilization: Indian experience. *Janasamkhya*, 2(1), pp.51–65.

Saier, U. (2020). Basisinformationen zur Gesundheit in Hamburg. [online]. Available from: https://www.hamburg.de/contentblob/14867104/e26661d9b398dbb772b4503ebabae810/data/basisbericht.pdf.

Santow, M.G. (1978). A microsimulation of Yoruba fertility. *Mathematical Biosciences*, 42(1–2), pp.93–117.

Schlander, M., Hernandez-Villafuerte, K. and Thielscher, C. (2018). Kosten der Onkologie in Deutschland. *Forum*, 33(5), pp.330–337.

Schneider, U. and Kleindienst, J. (2016). Monetising the provision of informal long-term care by elderly people: estimates for European out-of-home caregivers based on the well-being valuation method. *Health & social care in the community*, 24(5), pp.81–91.

Schnur, O. (2008). Quartiersforschung im Überblick: Konzepte, Definitionen und aktuelle Perspektiven. In O. Schnur, ed. *Quartiersforschung: Zwischen Theorie und Praxis*. Quartiersforschung. Wiesbaden: VS Verlag für Sozialwissenschaften / GWV Fachverlage GmbH Wiesbaden, pp. 21–56.

Schofield, D.J. et al. (2017). A Brief, Global History of Microsimulation Models in Health: Past Applications, Lessons Learned and Future Directions. *International Journal of Microsimulation*, 11(1), pp.97–142.

Schulz, M., Romppel, M. and Grande, G. (2018). Built environment and health: a systematic review of studies in Germany. *Journal of public health (Oxford, England)*, 40(1), pp.8–15.

Seebaß, K. (2017). Who Is Feeling the Heat?: Vulnerabilities and Exposures to Heat Stress--Individual, Social, and Housing Explanations. *Nature and Culture*, 12(2), pp.137–161.

Sexton, K. (2014). Environmental Justice. In *Encyclopedia of Toxicology*. Elsevier, pp. 381–384. [online]. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123864543010629 [Accessed August 4, 2021].

Shrestha, R. et al. (2016). Environmental Health Related Socio-Spatial Inequalities: Identifying 'Hotspots' of Environmental Burdens and Social Vulnerability. *International journal of environmental research and public health*, 13(7).

Singh, P. et al. (2014). Dementia care: intersecting informal family care and formal care systems. *Journal of aging research*, 2014, p.486521.

Šlachtová, H. et al. (2016). Environmental and Socioeconomic Health Inequalities: a Review and an Example of the Industrial Ostrava Region. *Central European journal of public health*, 24 Suppl, pp.26–32.

Sozialbehörde. (2021). Indikatoren Gesundheitsberichterstattung: Gesundheitsbezogene Daten und Informationen. [online]. Available from: https://www.hamburg.de/indikatoren/.

Statistisches Amt für Hamburg und Schleswig-Holstein. (2017). Statistische Gebiete - Soziale Indikatoren 31.12.2016.

Statistisches Amt für Hamburg und Schleswig-Holstein. (2018). Statistische Gebiete - Soziale Indikatoren 31.12.2017.

Stellwagen, E. (2019). Forecasting 101: A Guide to Forecast Error Measurement Statistics and How to Use Them. [online]. Available from: https://www.forecastpro.com/Trends/forecasting101August2011.html.

Stephens, C. and Church, C. (2017). Environmental Justice and Health. In *International Encyclopedia of Public Health*. Elsevier, pp. 499–506. [online]. Available from: https://linkinghub.elsevier.com/retrieve/pii/B978012803678500134X [Accessed August 4, 2021].

Strauch, D. et al. (2004). Linking Transport and Land Use Planning: The Microscopic Dynamic Simulation Model ILUMASS. In *GeoDynamics*. CRC Press, pp. 319–336. [online]. Available from: https://www.taylorfrancis.com/chapters/edit/10.1201/9781420038101-25/linking-transport-land-use-planning-microscopic-dynamic-simulation-model-ilumass-dirk-strauch-rolf-

moeckel-michael-wegener-j%c3%bcrgen-gr%c3%a4fe-heike-m%c3%bchlhans-guido-rindsf%c3%bcser-klaus-beckmann.

Swamidass, P.M. (2000). Encyclopedia of Production and Manufacturing Management.

Tanton, R. (2014). A Review of Spatial Microsimulation Methods. *International Journal of Microsimulation*, 7(1), pp.4–25.

Tanton, R. (2011). Spatial microsimulation as a method for estimating different poverty rates in Australia. *Population, Space and Place*, 17(3), pp.222–235.

Tanton, R. and Edwards, K.L. (2013). Introduction to Spatial Microsimulation: History, Methods and Applications. In R. Tanton & K. L. Edwards, eds. *Spatial microsimulation: A reference guide for users*. Understanding population trends and processes. Dordrecht: Springer, pp. 3–8.

Tanton, R. and Vidyattama, Y. (2010). Pushing It To The Edge: Extending Generalised Regression As A Spatial Microsimulation Method. , 3(2), pp.23–33.

Techopedia. (2014). For Loop. *Techopedia*. [online]. Available from: https://www.techopedia.com/definition/19415/for-loop.

Tiwari, P. et al. (2015). Living Environment. In *India's Reluctant Urbanization*. London: Palgrave Macmillan UK, pp. 153–173. [online]. Available from: http://link.springer.com/10.1057/9781137339751_5 [Accessed August 4, 2021].

Tomintz, M.N., Clarke, G.P. and Rigby, J.E. (2008). The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. *Area*, 40(3), pp.341–353.

Umweltbundesamt ed. (2006). *Transportation Noise and Cardiovascular Risk: Review and Synthesis of Epidemiological Studies*. Berlin.

United Nations. (2020). Policy Brief: COVID-19 in an Urban World. , pp.1–30.

United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. [online]. Available from: https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E [Accessed July 15, 2021].

United Nations. (2018). World Urbanization Prospects: The 2018 Revision. Online Edition. [online]. Available from: https://population.un.org/wup/Download/ [Accessed July 15, 2021].

Upshaw, T.L. et al. (2021). Social determinants of COVID-19 incidence and outcomes: A rapid review M. Camacho-Rivera, ed. *PLOS ONE*, 16(3), p.e0248336.

Urban, A. et al. (2017). Impacts of the 2015 Heat Waves on Mortality in the Czech Republic-A Comparison with Previous Heat Waves. *International journal of environmental research and public health*, 14(12).

Vidyattama, Y., Tanton, R. and Biddle, N. (2013). Small Area Social Indicators for the Indigenous Population: Synthetic data methodology for creating small area estimates of Indigenous disadvantage.

Watts, H.W. (1991). Distinguished Fellow: An Appreciation of Guy Orcutt. *Journal of Economic Perspectives*, 5(1), pp.171–179.

Westenhöfer, J. et al. (2021). Gesundheitsförderung und Prävention im Setting Quartier: Hintergrund und Stand der Forschung. In *Gesunde Quartiere: Gesundheitsförderung und Prävention im städtischen Kontext*. Hamburg: oekom Verlag, pp. 31–46.

Weycker, D. et al. (2007). Cost-effectiveness of memantine in moderate-to-severe Alzheimer's disease patients receiving donepezil. *Current medical research and opinion*, 23(5), pp.1187–97.

White, M.P. et al. (2013). Would You Be Happier Living in a Greener Urban Area? A Fixed-Effects Analysis of Panel Data. *Psychological Science*, 24(6), pp.920–928.

WHO Regional Office for Europe. (2019). *Environmental health inequalities resource package. A tool for understanding and reducing inequalities in environmental risk*. Copenhagen. [online]. Available from: https://www.euro.who.int/__data/assets/pdf_file/0018/420543/WHO-EH-inequalities-resource-package.pdf [Accessed August 1, 2021].

Williamson, E.J. et al. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821), pp.430–436.

Williamson, P. (2013). An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation. In R. Tanton & K. L. Edwards, eds. *Spatial microsimulation: A reference guide for users*. Understanding population trends and processes. Dordrecht: Springer, pp. 19–47.

Willis, M., Asseburg, C. and He, J. (2013). Validation of economic and health outcomes simulation model of type 2 diabetes mellitus (ECHO-T2DM). *Journal of medical economics*, 16(8), pp.1007–21.

van Wissen, L. (2000). A micro-simulation model of firms: Applications of concepts of the demography of the firm. *Papers in Regional Science*, 79, p.111.

Wolf, T., Chuang, W.-C. and McGregor, G. (2015). On the Science-Policy Bridge: Do Spatial Heat Vulnerability Assessment Studies Influence Policy? *International journal of environmental research and public health*, 12(10), pp.13321–49.

Wolf, T. and McGregor, G. (2013). The development of a heat wave vulnerability index for London, United Kingdom. *Weather and Climate Extremes*, 1, pp.59–68.

Woock, K. and Busch, S. (2021). Sind vor dem Virus alle gleich? Gerechte Gesundheitsversorgung in der Krise. *Sozialer Fortschritt*, 70, pp.437–453.

World Health Organization. (2013). A global brief on Hypertension: Silent killer, global public health crisis. [online]. Available from: https://www.who.int/cardiovascular_diseases/publications/global_brief_hypertension/en/.

World Health Organization. (2019). Hypertension. [online]. Available from: https://www.who.int/news-room/fact-sheets/detail/hypertension.

World Health Organization. (2020). Strengthening Preparedness for COVID-19 in Cities and Urban Settings: Interim Guidance for Local Authorities.

World Health Organization Regional Office for Europe. (2011). *Burden of Disease from Environmental Noise: Quantification of Healthy Life Years Lost in Europe*. World Health Organization Regional Office for Europe, ed. Geneva: World Health Organization. [online]. Available from: https://www.who.int/quantifying_ehimpacts/publications/e94888.pdf?ua=1.

World Health Organization Regional Office for Europe. (2018). Environmental Noise Guidelines for the European Region. [online]. Available from: https://www.euro.who.int/__data/assets/pdf_file/0008/383921/noise-guidelines-eng.pdf.

World Health Organization Regional Office for Europe. (2013). Review of evidence on health aspects of air pollution – REVIHAAP Project. Technical Report. Edited by World Health Organization, Regional Office for Europe. Copenhagen.

Wuerzer, T. (2014). Urban Health. In A. C. Michalos, ed. *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer Netherlands, pp. 6835–6837. [online]. Available from: http://link.springer.com/10.1007/978-94-007-0753-5_3127 [Accessed July 28, 2021].

Wüstemann, H., Kalisch, D. and Kolbe, J. (2017). Access to urban green space and environmental inequalities in Germany. *Landscape and Urban Planning*, 164, pp.124–131.

Yett, D.E. et al. (1975). A microsimulation model of the health care system: the role of the hospital sector. *Applied Mathematics and Computation*, 1(2), pp.105–130.

Yosifova, E. (2021). Auswahl und Beschreibung der Untersuchungsgebiete. In J. Westenhöfer et al., eds. *Gesundheitsförderung und Prävention im Setting Quartier*. Hamburg: oekom Verlag, pp. 49–52.

Yosifova, E. and Pohlan, J. (2021). Umwelt- und Umgebungsmerkmale von Quartieren mit unterschiedlicher sozialer Lage. In *Gesunde Quartiere: Gesundheitsförderung und Prävention im städtischen Kontext*. München: oekom Verlag, pp. 75–104.

Zeeb, H. et al. (2017). Traffic noise and hypertension - results from a large case-control study. *Environmental research*, 157, pp.110–117.