OPEN VS MICRO

comparing different agent populations and their impacts

Assessing the differences between populations in epidemiological Agent-Based Models

H.M. Bijlard

Master thesis submitted to Delft University of Technology in partial fulfilment of the requirements for the degree of MASTER OF SCIENCE in Engineering & Policy Analysis Faculty of Technology, Policy and Management

November 2022

Open vs Micro; comparing different agent populations and their impacts

Assessing the differences between populations in epidemiological agent-based models

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of
MASTER OF SCIENCE
in Engineering & Policy Analysis Faculty of Technology, Policy and Management

by

H.M.Bijlard

Student Name	Student Number
Hidde Bijlard	5418631

To be defended in public on November 17th, 2022

First supervisor:	Prof.Dr. T.C. Comes
Second supervisor:	Prof. Dr. Ir. J.H.Kwakkel
Advisor:	M.Sirenko
External supervisor:	Prof. Dr. F.P.Pijpers
External supervisor:	Dr. B.Braaksma
External supervisor:	A. Mitriaieva
Project Duration:	April, 2022 - November, 2022
Faculty:	Faculty of Technology, Policy and Management, Delft

Style: TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

A word of thank you

So after months of work, here I am, writing a thank you word for my thesis. This project is, by far, the most time-consuming, daunting, and rewarding experience of the Engineering & Policy Analysis program and it has taught me a lot, about data, science, data science, and myself.

I would like to express my gratitude. First and foremost to my family for their unconditional love and support. To my supervisors, for their guidance, feedback, and understanding. Especially to Mikhail, whose weekly meetings and coworking sessions I will cherish. To the EPA community, with too many names to mention. To Auriane, for our coffee breaks and being a rubber duck. To Ninarosa, for being the best. And finally to the CBS, for allowing me access to their data and connecting me to their people.

At the moment of writing this, my future is uncertain. Fortunately for me, if there is one thing the EPA program has taught me, it is dealing with uncertainty.

H.M.Bijlard
The Hague, November 2022

Executive Summary

Communicative diseases have put a burden on societal well-being throughout history. The global advent of the COVID-19 virus made the world painstakingly aware of the impact a pandemic can have on all levels of society. To combat the spread and impact of a pathogen, policymakers turn to specialized tools to craft policy. One of these tools is mathematical modeling and simulation. By using the agent-based modeling paradigm, it is possible to evaluate the impact of policy measures on the viral spread and healthcare burden. An agent-based model always contains a set of agents. Agents are the digital representations of a studied entity, for example, a person, but agents can also be an organization or a vehicle. One of the characteristics of agent-based modeling is the reliance on data to create a valid agent population. However, data on the level of individual people, also known as microdata, can be difficult to acquire and work with due to GDPR regulations. Aggregated population data, also known as open data, may serve as a substitute for microdata. Both open and microdata may be used to synthesize a population of agents, this process is known as synthetic population generation.

Although both of these forms of data can be used to synthesize an agent population, it is currently understudied how the results of the agent population synthesis process are affected by the type of used population data. This research project focuses on this gap in knowledge and systematically analyzes the differences between an open-data-based and a microdata-based agent population. The populations that are synthesized are part of the HERoS model. The HERoS model is an epidemiological agent-based model that maps the spread of COVID-19 in the city of The Hague. The agents in the HERoS model are digital individuals, each characterized by a set of attributes. These attributes define the behavior of an agent within the model. Examples of agent attributes in the HERoS model are age, household size, and social role. Two agent populations for the HERoS model are synthesized, one using a sample-free algorithm that uses open data, and one using an algorithm that converts microdata to an agent population. The differences between the agent populations are quantified by introducing two new metrics, the modified Freeman-Tukey statistic, and agent matching. It was found that the agent populations show similar characteristics on a city level, but show significant differences at the neighborhood level. Furthermore, it is shown that the quality of an agent population is both dependent on the number of attributes that are synthesized, as well as the type of attributes.

Moreover, the relationship between the quality of an agent population and the outcome a model produces is also understudied. This relationship is investigated by using the open-data-based and microdata-based agent populations as input for the HERoS model. The model outcomes are then compared on a global and neighborhood level, using phase comparison and the quantification of the precision of the model. It is found that the effect of using open data instead of microdata is multifaceted. The model outcome curves show similar characteristics, albeit they differ in a numerical sense. Using open-data increases the number of hospitalizations, ICU occupants, and deaths in the HERoS model. Furthermore, the precision of the model decreases when open data is used. Finally, it observed that neighborhoods that significantly differ in input population also significantly differ in model outcomes, showing the importance of the quality of an agent population for smaller resolutions. Conclusively, this research project shows that open data can serve as a viable alternative to microdata when it comes to synthesizing agent populations for epidemiological agent-based models.

Contents

List of Figures

List of Tables

1 Introduction

1.1. The Pandemic Problem

With over 625 million cases and over 6.5 million deaths reported globally in October 2022, the Coronavirus (COVID-19) has held the world in its tantalizing grip since its advent in 2019 (WHO, 2022). The impact of COVID-19 goes beyond health, as it has also led to increased domestic and racial violence, looting, and other manifestations of civil unrest (Bradbury-Jones & Isham, 2020; Eisner & Nivette, 2020), as well as major economic impact (Buheji et al., 2020; Kaye et al., 2021). In those 2.5 years, great progress has been made in the understanding and mitigation of the disease, leading to the development of vaccines by multiple pharmaceutical companies (Lopez Bernal et al., 2021).

Unfortunately, health crises caused by communicable pathogens are not one-off events, while historic examples, such as the Black Death or the Spanish Flu, spring to mind (Huremović, 2019). Additionally, the probability of pandemics occurring, and their corresponding severity, are increasing due to recent changes in global mobility, urbanization, population growth, antibiotic resistance, and human-animal relations (Dodds, 2019; Hughes et al., 2010). With COVID-19 acting as a catalyst for scientific research and data availability (Mercatelli, Holding, & Giorgi, 2021), it becomes possible, and pressing, to further develop the tools available that aid in reducing pandemic impact.

1.2. Using Models to combat Epidemics

One of the available tools is **epidemiological modeling**. According to Kretzschmar (2020), epidemiological modeling has played a vital role in the shaping of public health policy regarding pandemics in the past two decades. In essence, classical epidemiological modeling is based on differential equations, where individuals move from the Susceptible state to the Infected state, to the Recovered State. Unfortunately, differential equation methods are unable to capture the complexity of disease spread through social networks and the adaptation of behavior due to disease prevalence (Epstein, 2009). **agent-based models (ABMs)**, however, can capture these complex mechanics. To add to that, the heterogeneity of the population, the unequal distribution of disease burden in both spatial and social contexts, and the interconnectedness between human behavior and disease spread make ABMs a suitable tool to model pandemics (Dekker, Coffeng, Pijpers, Panja, & de Vlas, 2022; Epstein, 2009; Sirenko, Yap, Sarva, Verbraeck, & Comes, 2020).

ABMs create an artificial society, in which individuals (commonly referred to as agents) live their day-to-day lives (Epstein, 2009). By tracking these agents and their health status, the spread of a disease can be precisely captured on both individual and system levels. Furthermore, the effect of the change in human behavior due to the disease can be modeled (e.g. working from home instead of in an office). Hence, an ABM can serve as an exploratory tool to test and develop health policies. It is important to note that ABMs do not serve as **prediction** tools, but rather help in answering what-if questions. Such as, what is the effect of school closures on the stress of the healthcare system?

1.3. Generating Synthetic Populations

Central within ABMs is the **behavior** of individuals within the simulation. For instance, children in the simulation go to school, adults go to work, and the elderly may perform charity work. This example demonstrates that age plays an important role in the whereabouts of an individual during the day. Naturally, there are more behavior-determining attributes, such as the social status or the housing location of a person. In general, the behavior of agents is strongly determined by their attributes (Chapuis & Taillandier, 2019). Hence, it is of vital importance that the attributes of agents accurately capture the essence of the entities they represent. The process of the initialization and assignment of attributes to agents is known as **synthetic population generation**. The resulting agent population is a partial representation of reality, given that only the attributes of interest are synthesized for each agent in the simulation model. According to Chapuis and Taillandier (2019), the goal of synthetic population generation is to minimize the gap between the actual and synthetic populations. The smaller the gap between the actual and synthetic population, the more realistic a model becomes (Harland, Heppenstall, Smith, & Birkin, 2012).

For the process of synthetic population generation, data on the studied population is necessary to accurately capture the properties of the population. One form of such population data is **microdata**, which contains detailed information on the level of individuals, companies, and, addresses (Statistics Netherlands, 2022b). Microdata is **linkable**, meaning that it can be traced back to respective individuals, companies, or addresses. The linkability makes microdata subject to the General Data Protection Regulation (GDPR), a European regulation requiring any organization to safeguard personal data and uphold individual privacy rights when it concerns European data subjects (H. Li, Yu, & He, 2019). As a consequence, if one wants to work with microdata, one must comply with the GDPR and use the proper infrastructure to do so.

One of the organizations that uses, maintains, protects, and updates microdata is Statistics Netherlands ('Centraal Bureau voor de Statistiek' or CBS) in the Netherlands. CBS allows researchers to access microdata under strict guidelines, operating with the principle of safeguarding privacy, as well as preventing the disclosure of subjects. A synthetic population constructed from microdata is the population that most closely matches the studied population, as the data is sourced from the actual subjects on the same aggregation level.

Another form of data that can be used for synthetic population generation is **open data**, defined as; data that is freely available for use in a machine-readable format (Statistics Netherlands, 2022c). Open data is data on a higher aggregation level (e.g. neighborhood, city, or national level). Unlike microdata, aggregated data is not linkable, which means it is not subject to the GDPR, easier to obtain, and easier to work with. Often, statistical bureaus provide interfaces that allow one to download data and use it without restrictions. The open data that is shared from data holders is different per organization. Some agencies provide third parties with Public Use Micro Samples (PUMS), a sample of microdata that does not contain identifying properties, whereas others only provide aggregate data. Modelers need to select the appropriate method to synthesize an agent population, based on the type of data available, the number of agents, and the type of population (Yameogo, Gastineau, Hankach, & Vandanjon, 2021). From here on, methods that use PUMS for synthetic population generation are referred to as **sampled methods**, whereas methods that do not are referred to as **sample-free methods**. Unfortunately, regardless of the method used, a synthetic population based on open data is less representative of the studied population, as there is information loss due to aggregation when the open data is created.

Agent populations based on microdata better represent the studied population than agent populations based on open data. This must be, as microdata contains information on the studied population on an individual level, whereas open-data-based populations approximate the data on an individual level. However, it is unclear what the significance of the gap between the populations is. Studies show that researchers use the data that is available to create an agent population (Dekker et al., 2022; Gargiulo, Ternes, Huet, & Deffuant, 2010), but the impact of this choice is understudied. Although there have been studies comparing sample-free to sampled agent population synthesis methods (Lenormand & Deffuant, 2012; Wickramasinghe, 2019), there has not yet been a study that compared agent populations resulting from sample-free population generation methods to agent populations based on micro-

data. Furthermore, the currently existing body of literature and research does not reveal what effect the quality of the agent population has on the outcomes of the ABM. There are studies that investigate the sensitivity of a model to the model's parameters (Edeling et al., 2021; Nossent, Elsen, & Bauwens, 2011), but no special attention has been paid to the agent populations. Aiming to scientifically contribute to this topic, this research project focuses on these two gaps in knowledge and provides a foundation for further research on the subject.

1.4. The Research Questions

With the goal to fill the identified knowledge gap, the following main research question is posed:
What is the effect of using a sample-free, open-data-based agent population over a microdata-based agent population as input for a large-scale epidemiological agent-based model?
In order to answer the main question, the following sub-questions are posed:

1. What sample-free synthetic population generation techniques are available to generate a two-layered population for an epidemiological ABM?
2. What methods are available to quantify the difference between an open-data-based synthetic population and a microdata-based population?
3. What methods are available to analyze the sensitivity of a large-scale ABM to different inputs?
4. What are the statistical differences between an open-data-based agent population and a microdata-based agent population?
5. How does the agent population affect the model outcomes of the HERoS model?
6. What are the implications of using microdata over open data?
7. What is the added value of microdata over open data?

1.5. The Approach

The questions are answered using a variety of methods. To start with, a literature review is performed to review the state-of-the-art of sample-free population generation, and the methods available to compare them. The literature review allows for answering the first three sub-questions. The findings of the literature review will then be incorporated into a case study. For the case study, the HERoS model is used, which is a large-scale, epidemiological ABM, that simulates the spread of COVID-19 through the city of The Hague. The synthetic populations used in the case study are synthesized using two different data sources. Namely, microdata, made accessible by the CBS, and open data, sourced from Den Haag in Cijfers and the Statline databank. The open data is subjected to a synthetic population algorithm, which produces a set of agents. The resulting agent populations are statistically compared to one another to find the differences, and to assess the quality of the synthetic data algorithm. Next, the agent populations are used as input for the HERoS model, and the resulting model outcomes are compared. The case study enables the answering of questions 4 and 5. Questions 6 and 7 are answered by reflecting upon the experiences of the researcher during the execution of this research project.

1.6. The Structure of this Report

The structure of this report is as follows. Section 2 contains the literature review of the state-of-the-art and describes the knowledge gap observed in the literature. Section 3 then describes the methods and case study used in detail. Afterward, section 4 contains the results, as well as the implications and discussion of these results. Section 5 focuses on the limitations, implications, and recommendations for further research. To conclude the research, the research questions are answered and a conclusion is presented in section 6.

2

Literature Review

2.1. Introduction

The literature review presents the relevant literature on the topic of synthetic agent populations and their impact on model outcomes. The quantification of the model outcome sensitivity to its input is known as **Global Sensitivity Analysis** (Saltelli et al., 2008). The literature review starts off with a broad introduction to the topic and investigates why modeling is a useful tool to combat pandemics. After that, the concept of epidemiological agent-based models (ABMs) is elaborated upon, and examples of such models are analyzed. Next, agent population synthesis and its associated difficulties are discussed. Then, the connection between an agent population and the model outcomes is refined. The literature review identifies a gap in scientific knowledge and concludes by posing research questions.

2.2. Modeling to combat epidemics

One of the prominent threats to public health in all parts of the world are communicable pathogens. Mathematical modeling plays a vital role in shaping public health policy response to pathogens (Kretzschmar, 2020). By integrating scientific information in a consistent framework, it becomes possible to generate insight and analyze complex problems, such as a pandemic outbreak. Insights from mathematical modeling can range from predicting future developments of disease and guiding vaccination strategies, to estimating the impact of case isolation and lockdown measures (Kretzschmar, 2020; Van Kerkhove & Ferguson, 2012). Although the belief that mathematical modeling is important for the combat, mitigation, and study of pandemic diseases is widespread (Bedson et al., 2021; Heesterbeek et al., 2015; Kretzschmar, 2020), Knight et al. (2016) reports that mathematical modeling is still under-utilized in public health due to models being too complex, too assumption dependent, or insufficiently communicated between public health decision-makers and modelers. Metcalf, Edmunds, and Lessler (2015) reported on six challenges that modelers run into on the interface between modeling and public health: 1) Communication on the limits of modeling, 2) Maintaining the value of models over an extended period of time, 3) The challenge in dealing with 'black swan' events, 4) The integration of modeling in the policy-making process, 5) Economic evaluation of models, and 6) No cycle of policy and modeling generating insights for one another. The limits and strengths of epidemiological modeling are further elaborated on by Siegenfeld, Taleb, and Bar-Yam (2020). In the context of COVID-19, they present a set of guidelines that modelers, policymakers, and the general public need to be aware of when evaluating models. The first one is that a model is only useful if it accurately describes the large-scale behavior of the system of interest. Secondly, the usefulness of models does not only stem from their ability to predict the burden of disease, but also, more importantly, from their ability to assess the impact of interventions. Thirdly, the exponential nature of pandemics causes small inaccuracies in predictions to have large impacts, given enough time. This also applies to small differences in the implementation of interventions, stressing the importance of uncertainty and the dangers of a false sense of certainty for practitioners. Finally, they recognize there is a fundamental difference between academically relevant research and policy-relevant analysis, which resides in their assumptions. The former can handle wrong assumptions, as it will increase our understanding of the problem at hand, whereas the latter must validate assumptions and consistently think about real-world consequences during the modeling

process.

It is widely accepted that modeling plays a major role in combating pandemic diseases. However, it is very important to understand the strengths and limitations of models when using them to shape public health policy. These aspects of a model are, naturally, also dependent on the type of model and the underlying paradigms, which will be further elaborated on in the next sections.

2.3. Epidemiological agent-based models

The spread, course, and outcomes of pandemics are strongly determined by social and behavioral aspects of a population (Bedson et al., 2021; Dekker et al., 2022). ABMs are an effective tool for modeling these aspects, which makes them ideally suited for epidemiological modeling and simulation (Epstein, 2009). Not only can ABMs model the spread of disease through a population, they can also serve as a tool for testing the efficiency of policy measures against the spread. This has not gone unnoticed, as ABMs have contributed significantly to resolving urgent public health policy problems over the past 20 years (Ferguson et al., 2006; Longini Jr et al., 2007).

One of the most influential papers on the battle against COVID-19 is *"Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand"* by Ferguson et al. (2020). NPIs are public health measures that can be taken to minimize the spread of pathogens by limiting the number of contacts between individuals and thereby reducing transmission. NPIs are especially useful when there is no vaccine (yet) available. Examples of NPIs would be social distancing, home isolation, and school closures. In their influential paper, the authors use a previously published ABM (Ferguson et al., 2006; Halloran et al., 2008), adapted to COVID-19 specifically in the US and UK. This model will further be referred to as the **Imperial college model**. There are two fundamental NPI strategies: the first one is mitigation, in which public health institutions try to slow down the transmission of a pathogen and protect those that are the most vulnerable, reducing the stress on the healthcare system. The second strategy is suppression, where the pathogen is suppressed as much as possible to prevent it from spreading at all. According to the authors, the latter is the preferred option for COVID-19, since the former still results in hundreds of thousands of deaths and severe overburdening of various health systems.

Generally, an ABM needs an agent population that matches the population that it is studying as closely as possible (Chapuis & Taillandier, 2019). This agent population is commonly referred to as a *synthetic population*. The synthetic population of the Imperial college model was constructed based on population density, household size, age structure, school, workplace, commuting distance, and air travel data. Additionally, data is also needed to model the transmission of the pathogen properly and to model the schedules the agents have. Even with all these sources of data, assumptions still had to be made with respect to disease dynamics. Examples of this would be that $R_0 = 2.4$, and that individual infectiousness is described by a gamma distribution with parameters $\mu = 1$ and $\alpha = 0.25$. Given that this model was developed back when little was known about COVID-19, these assumptions are reasonable. Now that COVID-19 has been studied more extensively, modelers may require fewer assumptions to model the disease. The Imperial College model serves as an example of a characteristic of agent-based modeling, namely their data intensity. Following the words of Bedson et al. (2021), ABMs can be valuable for the design of behavioral interventions, when accompanied by sufficient, **individual-level** data for model parametrization and fitting. The key here is individual-level data, as this type of data, especially in health, can be problematic to obtain due to confidentiality issues (Reiter & Kinney, 2011; Rushton et al., 2006; Zandbergen, 2014).

Not only are privacy and confidentiality an issue regarding the data requirements of an ABM, but also data can be of poor quality or absent (Fefferman, O'Neil, & Naumova, 2005; Venkatramanan et al., 2018). Although data issues can be problematic in virtually any aspect of the ABM modeling process, one thing all ABMs have in common is that they need an agent population. Since the characteristics of an agent strongly determine an agent's decision-making and activity pattern, it is of great importance that the characteristics of an agent population match those of the studied population (Chapuis & Taillandier, 2019). The creation of a representative agent population, known as *synthetic population*

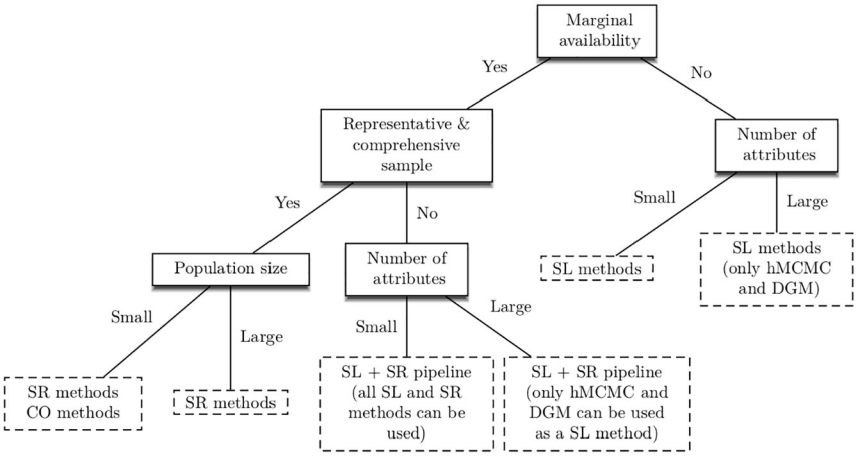

Figure 2.1: Decision tree to select the most appropriate technique for two-layered population synthesis, from Yameogo, Gastineau, Hankach, and Vandanjon (2021)

generation, can be difficult due to data being hard to collect or not representative of the studied population (Chapuis & Taillandier, 2019). Over the past 20 years, multiple algorithms and methods have been developed to overcome such data issues. In the following sections, these methods will be presented and discussed.

2.4. Generating synthetic agent populations

There is a wide range of synthetic population generation techniques presented in the literature. In a review presented by Chapuis and Taillandier (2019), these methods can be divided into two different families: Synthetic Reconstruction (SR) and Combinatorial Optimization (CO). However, as mentioned in reviews by Yameogo et al. (2021) and Sun, Erath, and Cai (2018), a third family of synthetic generation techniques exists, called Statistical Learning (SL). SR and CO techniques have been studied and applied for decades, whereas SL techniques are relatively new, and made possible due to advances in statistical learning techniques.

An ideal source of data for synthetic population generation is census data of a geographical region (e.g. a city or country). Census data is often collected by national statistical bureaus, like the CBS in The Netherlands or the U.S. Census Bureau. Because the information contained in a survey is confidential, complete census data is rarely made public. Even publishing a set of marginal distributions can be problematic at times, due to privacy concerns. To prevent disclosure, survey data is often made available to the public in two, independent ways. The first one is through Public Use Micro Samples (PUMS), containing a small sample of the whole population (typically 1-5%) with disclosure protection. This type of dataset is often called D1. The second type of public data, D2, is aggregated marginal distributions of a population, usually presented in the form of cross-tabulations. However, the policies employed by census data owners are different per country and region. For instance, some countries do not share D1 data (e.g. Germany and The Netherlands), whereas others (e.g. Great Britain and the USA) do. Consequently, modelers always need to select an appropriate technique based on the data available. This issue is recognized by Yameogo et al. (2021), where they present a decision tree to select the most appropriate technique for synthetic population generation.

A decision tree for synthesizing a two-layered population is shown in figure 2.1. A two-layered population is a population in which there is a link between different levels of the population (e.g. an individual and a household). This is an important distinction because the most common synthetic population generation technique, Iterative Proportional Fitting, is not able to generate a multilayered population. In the context of epidemiological modeling, a two-layered population is also required, because in-household transmission is an important source of a viral spread (Chao, Halloran, Obenchain, & Longini Jr, 2010; Lei, Xu, Xiao, Wu, & Shu, 2020; W. Li et al., 2020). Furthermore, a two-layered synthetic population

is also required when an individual agent can base its decisions on household characteristics (e.g. an agent using a car for transportation if there is one available on the household level).

Although the decision tree by Yameogo et al. (2021), is fairly comprehensive and useful, it suffers from a couple of drawbacks. The first one is that it is explicitly designed for transport modeling, therefore it is unclear whether this decision tool can be used for epidemiological modeling. Since transport networks play an important role in disease spread (J. Li, Xiang, & He, 2021), it may be useful with some adaptation, but the authors do not reflect on this. The second drawback is the inconsistency of the quantification of nodes in the decision tree. Some deciding factors are quantified (e.g. a representative sample contains at least 5% of the studied population), whereas others are not. For instance, the number of attributes that need to be synthesized is either large or small, but there is no definition of a 'large' number of attributes. The population size is also not clearly defined. The authors report that CO methods are unfit for synthesizing large populations because they are computationally complex. The reviewed CO methods produced fewer than 60,000 agents and 30,000 households (Hafezi & Habib, 2014; Ma & Srinivasan, 2015). The reviewed methods that are able to synthesize large populations created millions or more agents (Fournier, Christofa, Akkinepally, & Azevedo, 2021; Mueller & Axhausen, 2011). A small and a large population size differ by multiple orders of magnitude and therefore it is unclear which technique is most appropriate to create 600,000 agents for example. The third, and most important shortcoming, is the assumption that a sample of microdata must be available. As we have concluded earlier, the availability of D1 (or PUMS) data is not guaranteed.

2.5. Sample-free synthetic population generation techniques

Contrary to the techniques covered in the previous section, there are also agent population synthesis techniques available that do not require a sample of the population. However, they seem to be less prevalent in literature than in other classes of agent synthesis techniques. In the following section, sample-free techniques, and the metrics to determine their quality, are discussed.

In a paper by P.-j. Ye, Wang, Chen, Lin, and Wang (2016), the authors provide an overview of sample-free methods for generic agent population synthesis techniques. To keep this literature review concise, the inner mechanisms of these algorithms will not be extensively elaborated upon. Instead, the up-and downsides of these algorithms are discussed and details are provided when needed. The first one is by Gargiulo et al. (2010), where the authors composed a sample-free algorithm because data on the micro-scale was not available for their purposes. The aggregate statistics of the resulting synthesized population are in agreement with the observed data according to the authors. However, they provide no metrics and seem to base their findings on visuals. This was later investigated by Lenormand and Deffuant (2012), who compare the sample-free algorithm with a sample-based approach and find that the sample-free method better fits the observed distribution, making use of the Proportion of Good Prediction (PGP) and χ^2 metrics. They are defined as follows:

$$PGP = 1 - \frac{1}{2}\frac{\sum_{k=1}^{p}|O_k - E_k|}{\sum_{k=1}^{p} O_k}$$

and

$$\chi^2 = \frac{\sum_{k=1}^{p}(O_k - E_k)^2}{\sum_{k=1}^{p} O_k}$$

, where p is the number of households of each household type, and O and E are the observed and expected values respectively. Note that the Observed population also was a generated population instead of an actual one.

Moreover, the authors discuss other shortcomings of their method. The first one is that the algorithm is strictly correlated to the data structure. The algorithm was tested in two French municipalities, which largely consist of nuclear families. A nuclear family is a family that consists of two parents and a couple of kids, or a subset thereof (e.g. a couple without kids or a single-person household). In other regions, however, the algorithm may struggle with synthesizing more complex family structures (e.g. grandparents living with their children). To add to that, the algorithm is tested by using a limited number of attributes for both households and individuals. The households in the study were characterized

by the household type, and the individuals had an age assigned to them, it is unclear how well this algorithm is able to handle synthesizing more complex individuals. The second sample-free algorithm is by Barthelemy and Toint (2013), where the total population of Belgium (approximately 10.6 million people) is synthesized. In their evaluation, they use the Absolute Percentage Difference (APD) and Freeman-Tukey Goodness-of-Fit Test, defined as follows:

$$APD_{T,T'}(u_1, ..., u_p) = \left| \frac{T'[u_1]...[u_p] - T[u_1]...[u_p]}{T'[u_1]...[u_p]} \right|$$

and

$$FT(T, T') = 4 \sum_i (\sqrt{T_i} - \sqrt{T_i'})^2$$

. Where T and T' are the generated and estimated populations. Once again, they tested against an estimated population instead of an actual population. Their findings are quite promising, as they are able to generate a synthetic population that is close to their estimates. Additionally, compared to a sampled method by Guo and Bhat (2007), the sample-free method was better able to handle data inconsistencies and produced a better fit. One downside of the proposed algorithm is that it is not able to generate multi-layered populations, which can be quite limiting if such a population is required.

The third algorithm that is discussed is by Wickramasinghe (2019). In this technique, a generic, application-independent, sample-free method is proposed, as well as a heuristic specification framework. This method was then applied to two case studies and compared to a sampled IPF approach by X. Ye, Konduri, Pendyala, Sana, and Waddell (2009). Based on the FT metric, the proposed method outperforms the IPF approach. Furthermore, it is able to generate layered households. Once again, the number of synthesized attributes for both the households and individuals was fairly limited and it is unclear how well this algorithm performs when a more complex population is needed.

All the discussed algorithms above, as well as others that exist in the literature (e.g (Huynh, Barthelemy, & Perez, 2016; P. Ye, Hu, Yuan, & Wang, 2017)), share some similarities, but also differ in some important aspects. Let us start with the similarities.

- They do not require a sample for agent population synthesis
- They are evaluated using the Freeman-Tukey metric
- They perform well compared to sampled methods
- They generate realistic, but not very complex households
- They use census data as input

And the differences:

- Some methods can synthesize layered populations, whereas others cannot
- The census data used as input sometimes contains marginals or contingency tables, other times not
- Some are able to generate large agent populations in reasonable computing time, whereas others are not tested

Since there is no data available on the disaggregate level, it is difficult to verify whether or not individuals in the studied population are in any way mirrored in the synthetic population. However, Wickramasinghe (2019) was able to gain access to microdata for comparison and reported promising results. Furthermore, the usage of the Freeman-Tukey statistic is quite consistent among various papers. This metric is able to quantify the goodness-of-fit and is able to handle 0-valued cells in a contingency table, contrary to the classic χ^2 statistic. However, it is unclear how well this statistic scales to more complex individuals, as the number of cells in a contingency table increases rapidly with the number of attributes considered.

In conclusion, sample-free methods are a viable alternative to sampled methods when it comes to synthetic population generation. However, there are still some significant challenges that need to be investigated. First of all, most of the methods in the literature have not been tested against the

actual population they try the emulate. Furthermore, it is unclear how well these methods are able to synthesize more complex layered populations, and how it affects their performance (both computational and goodness-of-fit wise). Another finding was the consequent usage of the Freeman-Tukey metric to assess algorithmic performance, however, it is also unclear how well this metric scales to larger populations.

2.6. Sensitivity Analysis of ABM models

In the previous sections, techniques for creating a synthetic population for ABMs are discussed. The goal of these population synthesis techniques is to create an agent population that resembles the actual population as well as possible. Intuitively, it makes sense that a synthetic population that more closely matches the real population would result in more accurate model outcomes. However, as we have seen, a more realistic population also comes at a price. As mentioned before, using the most accurate data possible (i.e. the actual population at the individual level) is cumbersome due to privacy and confidentiality issues. Because the relationship between the quality of an agent population and the quality of the model outcomes has not been investigated yet, it is unreasonable to assume that it is absolutely necessary to work with the most accurate data necessary to create the most accurate model outcomes. In the continuation of this literature review, this statement is further investigated by assessing the relationship between a model input (the agent population), and model output (the model outcomes).

The difficulty of properly analyzing the relationship between input and output (also known as sensitivity analysis) in the context of an ABM is highlighted by Ten Broeke, Van Voorn, and Ligtenberg (2016). This difficulty is due to multiple factors:

- The existence of multiple levels in the model (e.g. the environment and the agents in the model).
- Nonlinear interactions in the model, which suggest nonlinearity in the output as well.
- The existence of emergent properties.

In spite of these difficulties, they discuss two methods to assess the relationship between in- and output. The first one is One-factor-at-a-time (OFAT), in which one parameter changes while the other input parameters remain fixed. The result of OFAT is an understanding of model outcome sensitivity to all parameters individually and the identification of tipping points. For an accurate estimate of the sensitivity, the change in a single parameter needs to be sufficiently small, and a large number of runs is needed to account for stochastic effects if there are any present in the model. Therefore, OFAT can be difficult to execute if there is a large number of input parameters and/or lengthy model runtimes. Furthermore, OFAT only considers the impact of singular parameters, whereas interaction effects between parameters may be important.

The second method is Global Sensitivity Analysis (GSA), in which the connection between model in- and output is investigated by linking sets of input parameter values to model outcomes. This can be performed through either fitting a regression model (Thiele, Kurth, & Grimm, 2014) or through SOBOL analysis (Sobol′, 2001). The result of these global analysis methods is a set of sensitivity indices. A first-order index is a measure of the sensitivity of the model output to that particular input. In contrast to OFAT, second or higher-order indices can also be computed, allowing the modeler to gain insight into interaction effects between parameters. The estimated sensitivity indices are calculated based on a Monte-Carlo sample, and this sample is bootstrapped to compute the accuracy of the estimates. If the indices are normally distributed, the estimate for the index is reliable. This is one of the important assumptions of a SOBOL analysis. Another important assumption is that the evaluated model parameters are independent of each other. This can be problematic when evaluating an agent population, as the attributes of agents are often dependent on each other (for instance age and income level).

Although Ten Broeke et al. (2016) studies the difference between OFAT and GSA and recommends the usage of OFAT, the literature seems divided on which methods are appropriate to use for ABMs. For instance, Saltelli et al. (2019) and Steinmann, Wang, van Voorn, and Kwakkel (2020) explicitly mention that OFAT methods are not appropriate for investigating highly non-linear models, such as ABMs. When looking at epidemiological models, we see that Edeling et al. (2021) has performed a sensitivity

analysis using SOBOL on the Imperial college model (Ferguson et al., 2020), discussed earlier. The study presents interesting findings regarding the difficulty of performing a proper sensitivity analysis on large-scale epidemiological ABMs. One of the choices the modelers make is to include 19 out of a possible 940 parameters. The reason: it is impossible to acquire an accurate, data-informed value for all of the model's parameters. Following a process that involves expert knowledge and dimension-adaptive sampling, they end up with a set of 19 parameters that the model is most sensitive to. Although not explicitly mentioned, computational complexity may also play a role here. According to Nossent et al. (2011), the number of necessary model evaluations to compute the first- and total SOBOL indices, under the assumption that the input parameters are independent is given by $n * (p + 2)$, where n is the number of samples taken and p is the number of parameters. Since the strength of SOBOL is its ability to provide insight into the interaction between parameters, it is recommended to calculate second-order indices at the bare minimum. For the second order, the number of model evaluations is given by $n * (2p + 2)$. Following Saltelli (2002), if models have a runtime in the tenths of minutes to a day range and $n = 1000$, quantitative methods, such as SOBOL, are not applicable. For an environmental model with 26 input parameters, used in Nossent et al. (2011), a n value of 12,000 was used. As a reminder, this is all under the assumption of independent inputs. Saltelli et al. (2008) note that calculating SOBOL indices for dependent samples requires many more samples. It is therefore safe to conclude that a SOBOL sensitivity analysis is not suitable for a complex ABM with many parameters. One possible solution is to do a SOBOL analysis with a subset of parameters, such as in Edeling et al. (2021).

Campolongo, Saltelli, and Cariboni (2011) note that screening methods should be used to evaluate models that have a high number of input parameters and/or long computation time. The best practice is the usage of the elementary effects method for screening, which results in a ranking of input parameters based on the model's sensitivity to them. After the screening, quantitative methods can be applied to analyze combination effects. However, there is a nonzero chance that significant interaction effects with or between excluded parameters occur.

As mentioned earlier, the application of the SOBOL method relies on the assumption that the inputs to a model are independent. Kucherenko, Tarantola, and Annoni (2012) proposed a method that is able to calculate first and total order indices when inputs are dependent, requiring $n * (2p + 2)$ model evaluations, making this method computationally even more expensive than a regular SOBOL analysis. Naturally, this hinders the applicability of this method to large-scale ABMs. Although an analysis could be performed on a subset of inputs, the results will suffer from the same drawbacks as using a selection procedure when performing SOBOL. Furthermore, there have not been any studies that analyze a large-scale epidemiological ABM with this method.

In conclusion, various authors claim that the quality of an agent population is very important. This claim can be studied through global sensitivity analysis. Various quantitative methods, such as OFAT or SOBOL, are available to quantify model sensitivity. However, these methods suffer when the to-be-evaluated model is complex in terms of runtime and number of input parameters. Furthermore, the interaction effects and dependence between input parameters cannot be neglected. Other methods, such as the one proposed by Kucherenko et al. (2012), are able to handle dependence between variables, but these methods are even more computationally complex. In the context of large-scale epidemiological ABMs, such as the Imperial College model, performing a proper sensitivity analysis is non-trivial but nonetheless possible when using proper selection methods. However, the selection procedure itself can hinder the validity and usefulness of the results.

2.7. Knowledge gap

Based on the reviewed literature, two gaps in scientific knowledge can be identified. The first one is connected to the usage of sample-free synthetic population generation algorithms in ABMs. Sample-free methods are used when there is no data on the individual level available, or the usage of such data is not desirable. This makes it impossible to assess the quality of a generated population against the actual population. This study focuses on this gap in knowledge by investigating the differences in agent populations based on open and microdata.

The second gap in knowledge pertains to the relationship between the in and output of a large-scale epidemiological ABM. Currently, only the sensitivity of models to their policy and epidemiological parameters has been investigated, but no attention has been paid to the impact of an agent population on the output. This relationship between in- and output is further investigated in the continuation of this research project.

2.8. Research Questions

Based on the literature study and the identified gaps in knowledge, the following main research question is asked.

What is the effect of using a sample-free, open-data-based agent population over a microdata-based agent population as input for a large-scale epidemiological agent-based Model?

To support the main question, the following sub-questions are postulated:

1. What sample-free synthetic population generation techniques are available to generate a two-layered population for an epidemiological ABM?
2. What methods are available to quantify the difference between an open-data-based synthetic population and a microdata-based population?
3. What methods are available to analyze the sensitivity of a large-scale ABM to different inputs?
4. What are the statistical differences between an open-data-based agent population and a microdata-based agent population?
5. How does the agent population affect the model outcomes of the HERoS model?
6. What are the implications of using microdata over open data?
7. What is the added value of microdata over open data?

By answering all these subquestions, the main research questions can be answered. Questions 1, 2, and 3 have been answered by means of a literature study, and questions 4,5,6, and answered through the use of a case study. This case study and all the associated methods are discussed in the next chapter.

3

Methods

3.1. Introduction

This section presents an overview of the methods that are used in this research project. First, the HERoS model is introduced, which is used as a case study to assess the differences between agent populations based on different data sources. Then, methods of collecting population data are presented. Within the HERoS model, a synthetic population generation algorithm is used to create an agent population based on open data. This algorithm is explained in section 3.5.

Then, the used synthetic data algorithm is explained, as well as the data collection and preparation. Finally, the metrics that are used to quantify the population differences are presented.

3.2. Case study

The *Health Emergency Response in Interconnected Systems* (HERoS) model is used to assess the differences between populations and their resulting influence on model outcomes. The HERoS model is an ABM that simulates an artificial city and its inhabitants (further referred to as agents). The agents perform **activities** (e.g. shopping, working, going to school) at **locations** (e.g. working locations, school, home). The activities an agent performs are dependent on the **attributes** of the agent (e.g. social group, age). Next to their attributes, an agent also has an **epidemiological state**, indicating the state an agent is in with regards to COVID-19 (e.g. susceptible, exposed, hospitalized), and whether or not a person can infect others with COVID-19. By tracing infectious individuals, the spread and healthcare burden of COVID-19 can exactly be monitored. In this case study, the artificial city that is simulated is The Hague, due to the data available on the neighborhood level and the work previously done by Sirenko et al. (2020). For a more detailed description of the HERoS model, please refer to appendix A.

3.3. Agents in the HERoS model

The agents in the HERoS model are digital representations of individuals living in The Hague. An agent is composed of multiple attributes. An overview of these attributes is presented in table 3.1, as well as a description of the attributes and the aggregation level of the attributes (i.e. whether the value is the same for the whole household). An overview of the semantics of the possible values of the attributes is provided in appendix C. The double vertical line denotes the agent attributes that are assigned to the agents in the same manner for both the open-data and microdata agents.

3.4. Data collection

The HERoS model requires an agent population to run. This study focuses on the differences between agent populations constructed from two different forms of data: micro- and open-data. In this section, the collection of the data necessary for synthetic population generation is presented. The section starts with an explanation of the sourcing of open data, after which the sourcing of microdata is described. Please note that the translations in this section are made by the author.

Table 3.1: Overview of the attributes of an agent in the HERoS model

Attribute	Description	Aggregation level
person_id	Agent identification number	Individual
household_id	Household identification number	Household
household_structure	Household structure	Household
household_size	Number of people in household	Household
nb_children	Number of children in household	Household
income	Income of household	Household
age	Agent age	Individual
family_role	Role of agent in household	Individual
social_role	Social role of agent	Individual
neighborhood_code	ID of the neighborhood in which the agent lives	Household
district_code	ID of the district in which the agent lives	Household
age_group	Age group of agent	Individual
home_id	Identification of building in which household is located	Household
workplace_id	ID of agent workplace/school/college/university	Individual
building_geometry	Geometry of the building in which the agent lives	Household
geometry	Point that locates the building in which the agent lives	Household

3.4.1. Open-data collection

For the city of The Hague, demographic data is publicly available on the neighborhood level through Den Haag in Cijfers (Municipality of The Hague, 2022). The number of people per neighborhood, the number of households, the distribution of the household structures, the ages of household heads, the number of unemployed people, and the incomes of households are extracted from this source. The data is further supplemented by CBS data on the mother's age when her first child is born (Statistics Netherlands, 2019).

3.4.2. Microdata collection

The microdata that is used in this study is sourced from the CBS's System of Societal Statistic Datasets (SSB), which contains a wealth of data on virtually every aspect of The Netherlands (Bakker, Van Rooijen, & Van Toor, 2014). The data located in the SSB is mostly administrative data, characterized by its high level of completeness and detail. The individual-level data in the SSB can be linked together using linkage keys. The attributes of interest are linked together in a single data file, using the RIN number (a pseudorandom identification number used internally at CBS) as the merge key. Since the open data agent population is based on open data from 2019, the microdata was sourced from the same year. Initially, the data set contained >17.5 million records, but this was reduced to 550k by sub-setting based on the housing location of the agents (the housing location being The Hague).

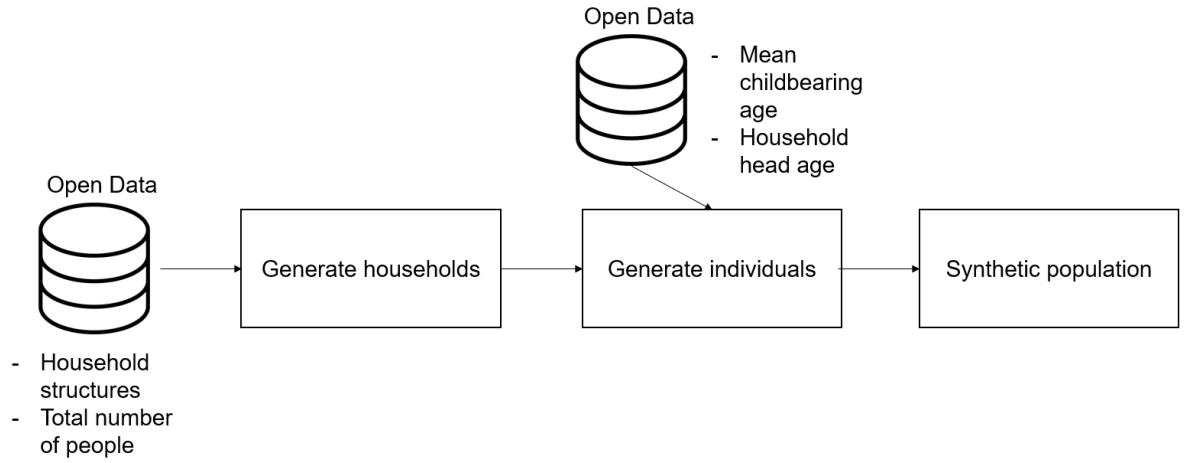

Figure 3.1: Overview of sample-free population generation algorithm used in HERoS model

3.5. Open-data-based population generation

The collected open data is used to synthesize an agent population for The Hague. This section explains this process in detail. Figure 3.1 presents a high-level overview of the mechanisms of the sample-free population algorithm. Each of the steps is elaborated on in the following sections.

3.5.1. Generating households

The generation of households is the first step in creating a synthetic agent population for The Hague. The household structures in the Den Haag in Cijfers dataset can be of four types:

1. Single-person household
2. Household w/o children
3. Household with children
4. Single-parent household

For each of these household types, the number of households per neighborhood is present in the open data. For example, the Leyenburg neighborhood contains 3,705 single-person households, 1,937 households w/o children, 1,497 households with children, and 777 single-parent households. Based on the data, a table is created which contains all the households in The Hague, where each row is a household. For each of these households, the number of people living in them needs to be determined. For childless households, this is straightforward (e.g. a single-person household contains one person). However, for households with children, the number of children needs to be determined. The number of children in a household is determined using a zero-truncated Poisson distribution, with a maximum of 5 children for households with children, and 4 for a single-parent household. Naturally, this limits the household size to a maximum of 7 agents (2 parents + 5 children).

3.5.2. Generating individuals

After the synthesis of households, the households are filled up with agents. First, the heads of households are synthesized and assigned attributes. When a person is not in a single-person household, the other household members are synthesized based on the partner's or parent's age. The income of a household is determined based on open data on income from Den Haag in Cijfers. The family_role is determined based on the position of the family member in the household. The social_role is based on the age of the agent (see figure 3.2). A selection of the agents is unemployed, based on the open data on unemployment.

3.6. Microdata-based population generation

This section focuses on the usage of microdata to synthesize an agent population that is compatible with the HERoS model. The synthesis of an agent population from microdata requires the conversion

of categories in the microdata to HERoS categories. To provide an example, a ”6” in the microdata type of household variable denotes a single-parent household, whereas in HERoS this category is denoted with a ”4”. The process of mapping is straightforward and did not present any issues. However, not every category in the microdata can be used in the HERoS model, because the model does not define behavior for every category of agents (i.e. there is no behavior defined for a recipient of benefits due to disability/disease). These cases are handled by assigning the individuals in these categories into categories that do have defined behavior. It is important to note that these categories of agents cannot be omitted because the total number of agents in the micro- and synthetic data must be very close to one another. Else, the model outcomes may vary too much for comparison (e.g. the number of infections will be lower if the density of agents in the model is lower). The steps taken to convert microdata to an agent population usable in the HERoS model are stated here, as well as the assumptions that are made during the data processing.

3.6.1. Handling missing data

In the microdata, there are agents present for which one or more of the variables are undefined or missing. Since the HERoS model must have complete data, these cases need to be handled. The number of micro agents with missing values is less than 0.05% of the total, hence the choice was made to remove these agents from the analysis. The resulting number of micro agents is 538 thousand, whereas the synthetic data algorithm produced 554 thousand agents. To keep the number of agents the same for both agent populations, households are randomly removed from the open-data-based agent populations until they have the same number of agents.

3.6.2. Family role

The first variable which is modified is the family_role variable. In the microdata, this category is known as PLHH, whereas it is known as family_role in the HeroS model. Table 3.2 shows the mapping of

Table 3.2: Mapping of PLHH to family_role

PLHH	family_role
1	7
2	1
3	2 if INPOSHHK == 2, else 3
4	2 if INPOSHHK == 2, else 3
5	4 if INPOSHHK == 2, else 5
6	4 if INPOSHHK == 2, else 5
7	6
8	2
9	1
10	1

microdata variables to HERoS variables for the family_role attribute. The definitions for these variables can be found in appendix C, in tables C.1 and C.2. There are a number of things to note here. The first one is the mapping of PLHH categories 3-6. Other than the microdata, the HERoS data distinguishes between the head and the partner of a household. However, this information is not present in the PLHH variable. To resolve this issue, the INPOSHH microdata variable is used. INPOSHH gives the position

of a person within the household. If this variable is equal to 2, it means that the person is classified as hoofdkostwinner met partner (main provider with a partner). It is assumed that the main provider is also the head of the household. Hence, if a person is known to be the main provider, this person is assigned to be the head of the household in the HERoS model definition. Furthermore, categories 8, 9, and 10 do not conform one-to-one to one of the categories of the HERoS model. These categories are assigned to the categories that most closely resemble their original category. For categories 8,9, and 10 these are 2, 1, and 1 respectively. These categorical assumptions affected 3.8% of the total number of micro agents.

3.6.3. Household structure

The second variable where mapping is necessary to account for differences in structure is the household_structure variable. This category is named TYPP in microdata.

Table 3.3: Mapping of TYPHH to household_structure

TYPHH	household_structure
1	1
2	2
3	2
4	3
5	3
6	4
7	2
8	1

The mapping, as can be seen in table 3.3, is quite straightforward (see tables C.3 and C.4 for definitions). The only notable issues are the TYPHH categories ”Other household” and ”Institutional household”, since these categories cannot be handled by the HERoS model. The assumption was made to assign these categories to ”Household w/o children” and ”Single person household” respectively. This assumption is based on the most frequent household composition of these households, in terms of the number of people contained in that household in the microdata. This assumption affected 2.3% of the total number of households.

3.6.4. Social role

The third variable is the social role variable. This variable is known as SECM within microdata and social_role in the HERoS model.

Table 3.4 shows the mapping of SECM to social roles (see tables C.5 and C.6). Employees, directors/major stakeholders, and entrepreneurs are placed in the worker category. The receivers of both social and unemployment benefits are assumed to behave like unemployed people looking for work (affected 6.1% of agents). The people in SECM category ”Other without income” are also assumed to be unemployed people looking for employment, this affected 6.7% of agents. Receivers of benefits that are not able to work (Other and due to disease/disability) are categorized as pensioners (it is assumed here that they do not visit workplaces and behave like pensioners), this affected 4.8% of agents. People that are categorized as too young for school/school-going/studying are assigned to categories 1-6 based on their age, regardless of having income or not. The assignment of the social role for students follows the procedure, depicted in figure 3.2.

Space refers to a binary that is implemented to account for the limited capacity of colleges and

Table 3.4: Mapping of SECM to social_role

SECM	social_role
11	7
12	7
13	7
14	7
15	7
21	9
22	9
23	8
24	8
25	8
26	1-6
32	1-6
32	9

universities within The Hague. The number of college and university students in the Hague is 8527 and 9841 respectively (source HERoS model source code). If the colleges and universities are full, the agent is assumed to be working instead.

3.6.5. Income

Following the definition by the CBS, the incomes of micro agents are assigned based on their quintile position (Statistics Netherlands, 2022a). The people in the bottom two quintiles are considered low income, quintiles 3 and 4 are considered medium income, and quintile 5 is high income. Since the income percentile is known for each agent, only the mapping of the percentile to quintiles is needed. The agents with unknown incomes are left as is.

3.7. Determining other agent attributes

This section further describes the agent population synthesis process. This part of the process is the same for both the microdata-based and open-data-based agent populations. Each of the following sections focuses on a particular agent attribute and the synthesis thereof.

3.7.1. Age Groups

The agents are assigned age groups based on the brackets in table 3.5.

3.7.2. Assigning home IDs

Each household is assigned a home_id. The home_id is the key that links a household to a home location in the Hague. This home location determines where the agent sleeps and performs their at-home activities. Naturally, all the agents within a household get assigned the same home_id. Since the

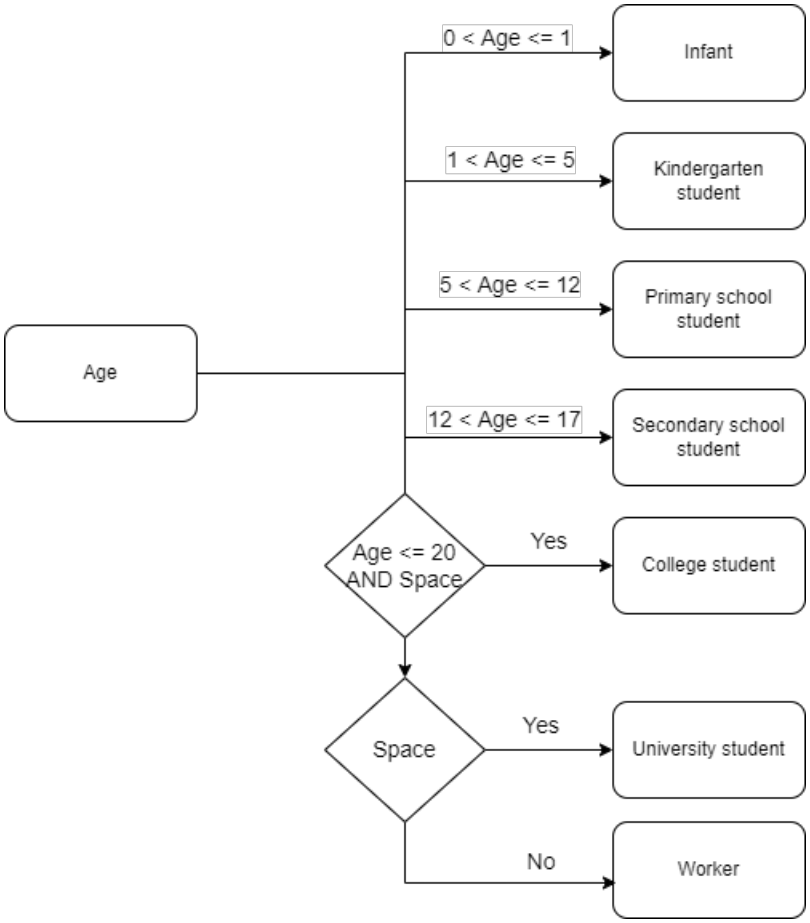

Figure 3.2: Flowchart of assigning the social role attribute to agents

home_id refers to a location, multiple households can have the same home_id, but this does not mean the households live together. It simply means that the households are housed in the same building but in different sub-locations. The assignment of a home_id to location is as follows:

Step 1. Determine the average number of sub-locations per location. The average is calculated by dividing the number of households in a neighborhood by the number of housing locations in a neighborhood.

Step 2. Sample the number of sub-locations per housing location from a normal distribution where μ is the average number of households per housing location and σ is 1.

Step 3. Round the number of sub-locations per location

Step 4. Reassign zero or negative sub-location values. Should a housing location have 0 or a negative number of sub-locations, the number of sub-locations is corrected to one.

Step 5. Check if there are enough sub-locations to house all households. If not, add one sub-location to n housing locations, where n is the difference in sub-locations and the number of families that need allocation.

Step 6. Assign home_id to households. This process is done randomly on the neighborhood level.

This algorithm is repeated for every neighborhood to account for the differences in housing characteristics per neighborhood.

3.7.3. Assigning workplaces

Now that the agents have a home_id, they can be allocated a workplace_id. The workplace_id is an umbrella term for a workplace or school. The workplace_id determines where an agent goes if this agent is scheduled to work or study. It is assumed that all of the people that live in The Hague also

Table 3.5: Age brackets for agents in the HERoS model

Age Range	Age Group
0-4	1
5-14	2
14-19	3
20-44	4
45-64	5
65-79	6
80+	7

work in The Hague.
At first, essential retail workers are assigned a workplace_id. Essential workers are the agents who go to work, regardless of whether a lockdown policy is in place. It is assumed that essential retail workers are paid less, and are selected from income groups 'low' and 'medium' with a ratio of 25 and 75 percent respectively. The number of essential retail workers that need to be allocated is calculated from the total number of needed employees for retail stores, which amounts to 1143 in The Hague.

Next, essential workers that work in the Police, at a Fire station, or in a Hospital are assigned a workplace_id. These essential workers are sampled from all income categories in a 25-50-25 percent distribution. In total, there were 5622 workplaces in this category that needed to be filled.

Finally, all the other workers are assigned a workplace. For each agent with the social_role of Worker, a random workplace within the city is assigned. Each workplace has an a priori assigned worker capacity that cannot be exceeded. The distributions of the number of employees per location type can be found in appendix B. The number of works that are subject to this assignment is equal to $n - number\ of\ essential\ workers$, where n is the total number of working agents.

Once every worker has a workplace_id, the final step for the workers is allocating the weekend workers. Workers in the categories of Retail, BarRestaurant, FoodBeverage, Recreation, Supermarket, Pharmacy, Police, Hospital, and FireStation continue to work on the weekends. These workers have been reassigned a social role, that enables the model to continue the operation of these places on the weekends.

3.7.4. Assigning school IDs

In a similar vein to the workers, all the school-going agents in the model need to be assigned a workplace_id as well. The number of students per school location is limited, following table 3.6.

Students that are in Kindergarten, Primary or Secondary School go to a school that is within a proximity of 2500 meters. The distance between a student and a school is computed by first dividing the city of The Hague into 500x500 grid cells and then calculating a haversine distance between the grid cells. Then, for each student, a set of school locations within this distance is filtered and a random one with leftover capacity is chosen. Should there be no suitable school location available, either due to limited capacity or no nearby schools, a random school is chosen. A heatmap of the density of students and locations per location type can be found in appendix B.
For the students in the categories College and University, the workplace_id is chosen randomly from the available locations. If the capacity of an entire category is exceeded (e.g. there are no University locations available at all), a student looks for a study place in the other category. University students

Table 3.6: Maximum student capacity per student location type

Location type	Maximum capacity
Kindergarten	125
Primary School	275
Secondary School	425
College	700
University	1555

will use a College as a workplace, and vice versa. If the capacity of both is exceeded, the students are reassigned to be unemployed. However, this was not the case for either population.

3.8. Operationalizing the difference between agent populations

After the synthesis of the open-data-based synthetic agent population and the microdata-based agent population, they are compared using quantitative methods. This section focuses on two techniques for doing so: the modified Freeman-Tukey test and agent matching.

3.8.1. Freeman-Tukey test

As came forward during the literature review part of this research project, the Freeman-Tukey (FT) test statistic is a measure that can be used to assess the goodness-of-fit of an observed against an expected population. The formula for calculating the test statistic in literature is as follows:

$$FT(O_i, E_i) = 4\sum_i (\sqrt{O_i} - \sqrt{E_i})^2$$

where O_i is the observed quantity, E_i is the expected quantity, and i refers to the number of cells in the contingency table. The advantage of using the Freeman-Tukey test statistic is that is able to handle 0 values in contingency table cells, contrary to the χ^2 statistic.

In this project, a modified version of the FT statistic is used, which is as follows:

$$FT(O_i, E_i) = \frac{\sum_i (\sqrt{O_i} - \sqrt{E_i})^2}{max\ error}$$

where

$$max\ error = 2\sum_i O_i$$

This leads to a test statistic in $[0, 1]$, where 0 is the best fit and 1 is the worst fit possible. The modified version removed the constant 4 in front of the sum sign and divided by the maximum error. Usually, the Freeman-Tukey statistic is used to determine a p-value, so that a null hypothesis can be tested. However, that is not of interest here, as it is known that the synthesized distributions for variables will not exactly match the actual distributions. Furthermore, due to the large number of agents in this project (>500,000), the null hypothesis will always be rejected if there is not a near-perfect match. Instead, it is interesting to know how **close** the two distributions are. The value of the metric is dependent on the size of the tested population, which makes it impossible to compare the metric for different population sizes. To account for this, the value of the metric is divided by the max error, which is given by $2\sum_i O_i$.

Another advantage of this metric is the scaling to more complex populations. The generated sample-free populations described in the literature are limited in complexity. The populations considered in this research project contain 19 attributes each, out of which 8 are considered during the analysis. The

other 11 are not directly comparable, as they are identification numbers or produced in the same manner for both populations.

The modified Freeman-Tukey statistic also has some drawbacks. The first one is that every error made by the synthetic data algorithm is weighted equally. For instance, the misclassification of the age group of an agent may have less impact on the model outcomes than the income group of an agent. Furthermore, a perfect fit does not necessarily mean a perfect population. Why is this the case? Because the microdata contains errors and logical inconsistencies as well. For instance, the income of some people in The Hague is not known. A perfect goodness-of-fit implies that the synthetic population also emulates the errors in the microdata, which is not desired. As there is no open-source implementation available of the modified FT statistic, it was implemented using the Python programming language. The source code will be made available on https://github.com/HMBijlard once the research project is finished.

3.8.2. Matching synthetic agents to real agents

Agent matching is another quantitative metric that is used to evaluate the quality of the agent population, which is easier to interpret than the modified Freeman-Tukey statistic. The problem with many statistical tests is that they provide information on the goodness of fit of a single attribute, but they do not provide insight into the overall quality of the agents. This can be partially solved by investigating conditional probabilities, as they give insight into the relationship between two (or more) attributes. The total amount of probability distributions (conditional or not) one can look at is given by

$$n + n \sum_{r=1}^{n-1} \binom{n-1}{r}$$

where n is the number of attributes for an agent. For the synthetic population in the HERoS model, $n = 7$, which means there are 448 density functions that can be compared. Since the number of distribution functions increases rapidly with the number of attributes, the scalability of classical goodness-of-fit methods to more complex populations is questionable. Therefore, another comparison metric is used, which is explained in the following section.

The quality of a synthesized agent can also be assessed with so-called counterparts. An agent has a counterpart if an agent with the same combination of attributes exists in the other population. One can deem a synthesized agent of good quality if this agent has a counterpart.

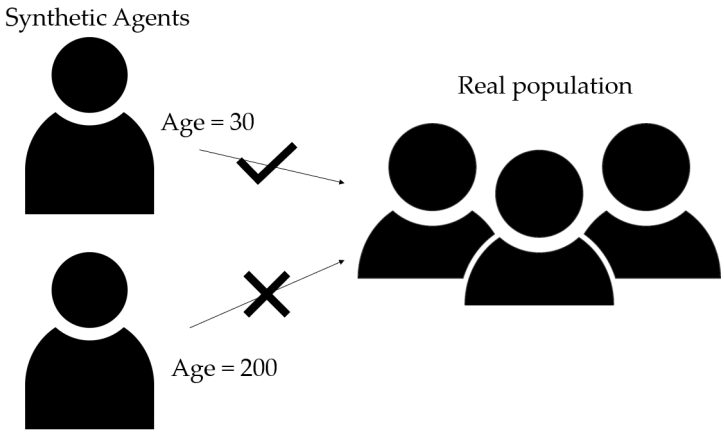

Figure 3.3: Graphical representation of a counterpart

This idea is graphically represented in figure 3.3, where 2 synthesized agents are on the left side. To assess whether these agents are of good quality, the microdata-based population is assessed to check whether the synthetic agents have counterparts. In this case, there is an agent in the real population with an age of 30, whereas there is no agent with an age of 200. Hence, the first agent has a counterpart, and the second agent does not. This method can be extended to include multiple attributes, the more

attributes selected, the closer the synthetic agents must be to the real population to have a counterpart. The percentage of agents that have a counterpart is a metric that indicates how closely the synthetic population resembles the real population. The agent matching method is implemented in the Python programming language, with the source code available on https://github.com/HMBijlard.

3.9. Experimental setup

After synthesizing the population and analyzing the differences between them, the agent populations are used as input for the HERoS model. In this section, the hardware setup of running the model, and the parameter specifications for the different model runs are given.

3.9.1. Hardware setup

As has been mentioned numerous times, the microdata used in this study is very privacy sensitive and can only be accessed in the secure CBS environment. The hardware of the standard Virtual Machine in the SEC environment is not sufficient to run the HERoS model properly. This problem is alleviated by using the Research Environment of the CBS. The Research Environment is a 16-core Linux machine with 512 GB of RAM. Using the Research Environment ensures that the microdata stays safe within the protected environment. The outcomes of the model are analyzed using the same computing resources. Once the analysis is done, non-identifying information can be exported out of the secure environment for publication.

3.9.2. Parameter setup

Table 3.7 contains the values of the model parameters that are used to run the HERoS model within the CBS environment. To account for the stochasticity, the HERoS model is run with 10 different seeds for both agent populations. Furthermore, the model is run without implementing policies. This is done to keep the number of factors that can influence the model outcomes minimal. All input parameters, apart from the random seed, are kept constant for all the simulation runs, similar to an OFAT analysis.

3.10. Model outcomes analysis

After simulating the model, the model outcomes are analyzed to gain insight into the sensitivity of the HERoS model outcomes to the different agent populations. First, a general introduction is given to the model outcomes of the HERoS model. Then, the methods that are used to analyze the model outcomes are explained.

3.10.1. Model Outcomes

The HERoS model produces a series of 8 outcomes, quantifying the spread and healthcare burden of COVID-19 in the city of The Hague. An example of the outcomes can be seen in figures 3.4. **Susceptible** refers to the number of agents that are susceptible to the virus, and **Recovered** refers to the number of agents that have been infected and since recovered. **Exposed** indicates the agents that have been in the proximity of an agent that was infectious at their time of contact. Both the **Infected-Asymptomatic** and **Infected-Symptomatic** outcomes are the agents that have contacted the virus, the former does not show symptoms, whereas the latter does. Infected agents go into the **Hospitalized** state if they are in the hospital. The **ICU** outcome specifies the number of agents in the Intensive Care Unit (ICU), and finally, the **Dead** outcome indicates the number of dead agents.

The model outcomes for both the synthetic and micro agent populations are compared both visually and numerically. The differences in model outcomes are assessed in terms of model**precision**. Based on ISO standard 5725-6, precision refers to the statistical variance between results (International Organization for Standardization, 1994).

3.10.2. Model precision

The precision of a model refers to the amount of statistical variance in the outcomes. Less variance in the outcomes is desired, as this narrows down the range of possible values and makes it easier to interpret the model outcomes. The statistical variance in the HERoS model is caused by the inherent randomness in the model (e.g. an agent moving from the exposed to the infected phase is probabilistic). This is accounted for by simulating the model 10 times for each agent population, using fixed random

Table 3.7: Experimental setup

Model parameter	Value
Run Length	150 days
Run Seed	[100-109]
Initial infected agents	100
Minimum age infected at t=0	0
Maximum age infected at t=0	100
Start policy1	-1 (no policy)
Start policy2	-1 (no policy)
Days between infection and early contagious period minimum	24 hours
Days between infection and early contagious period mode	24 hours
Days between infection and early contagious period maximum	48 hours
Incubation period minimum	2 days
Incubation period mode	4 days
Incubation period max	8 days
Period middle to late stage minimum	7 days
Period middle to late stage mode	11 days
Period middle to late stage maximum	14 days
Period late to recover stage minimum	14 days
Period late to recover stage mode	17.5 days
Period late to recover stage maximum	21 days
Infection probability	0.541495899185502
Asymptomatic fraction	0.5
Contagious factor	8 m^2

seeds. The precision of the HERoS model is then quantified by computing the mean difference between the minimum and maximum value of that outcome for each timestep.

3.10.3. Phase comparison

The goal of the HERoS model is to guide public decision-making under conditions of deep uncertainty (HERoS, 2020). Due to the strain of COVID-19 on ICUs, one of the major indicators for policymaking is

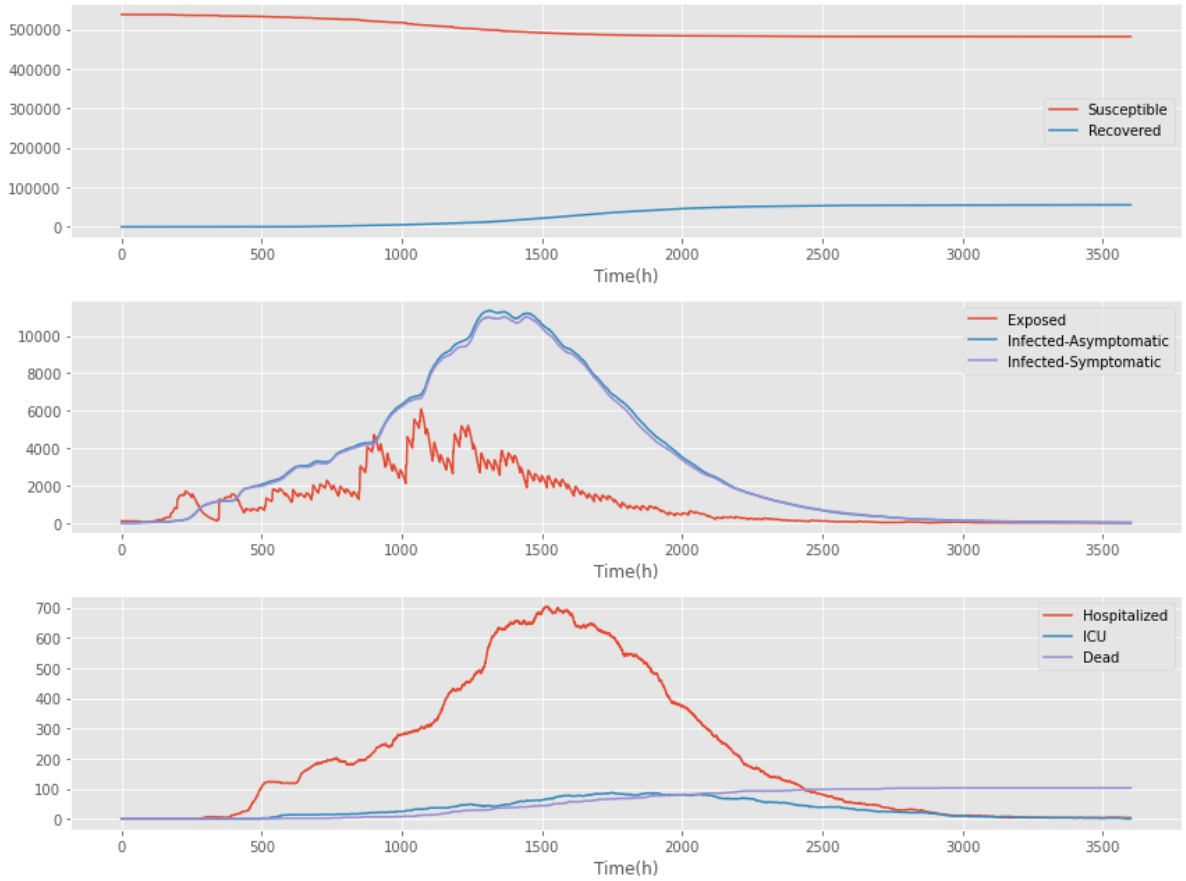

Figure 3.4: Example model outcomes of a single run

the ICU occupation (Haas, de Lange, van Dijk, & van Delden, 2020; Rijksoverheid, 2022). The HERoS model estimates this indicator, which is why it is important to quantify the impact of the microdata on the number of ICU admissions. The impact is investigated by dividing the ICU admissions curve into four different phases. These phases are defined in table 3.8.

Table 3.8: ICU occupancy phases

Phase	Phase definition
I	Phase where the number of ICU admissions is below $PI\ END$
II	Phase between $PI\ END$ and peak of ICU admissions
III	Phase where the number of ICU admissions is between peak and $PIII\ END$
IV	Phase where the number of ICU admissions is below $PIII\ END$

The threshold values for when a phase ends and a new one starts ($PI\ END$ and $PIII\ END$) can be chosen arbitrarily by the modeler. The phases pinpoint the periods with different characteristics. Phase I is the time at which the effect of COVID-19 on the ICU occupation is relatively low. Phase II is the period in which the ICU occupation increases until the peak of ICU admissions in that model run. Phase III is the phase in which the ICU occupancy is declining, but still above the threshold level. Phase IV is the phase in which the ICU occupancy is stable again. Note that the HERoS model does not account for the number of available ICU beds and assumes that there is always an ICU bed available

if necessary. Figure 3.5 graphically represents the phases on an example of an ICU occupancy curve.

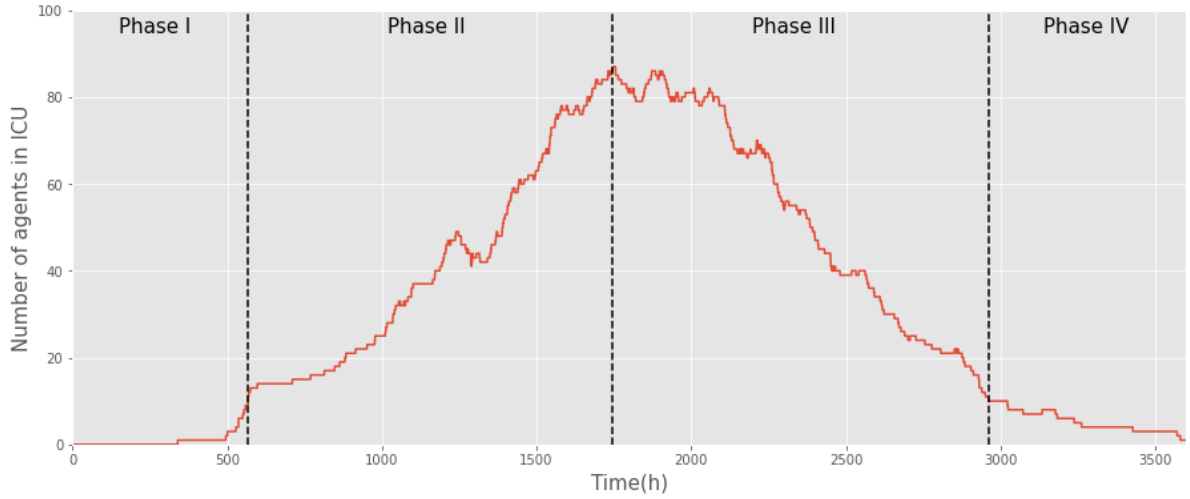

Figure 3.5: Visual representation of Phases, with $PI\ END = 10$ and $PIII\ END = 10$

After defining the phases, the different ICU occupancy curves for both the open- and microdata agent populations are compared. The comparison is made by computing the differences between the properties of each phase, table 3.9 denotes the properties of interest. The slope of the curve is the mean increase of ICU occupation per time unit, calculated using $Slope = \frac{\Delta ICU occupation}{\Delta Time}$

Table 3.9: Phase properties

Phase	Interesting properties
I	Duration
II	Duration, Slope
III	Duration, Slope
IV	Duration

4 Results

4.1. Introduction

The Results section focuses on the results obtained by applying the methods from the Methods section. The results are presented in the form of a table or graph. After presenting the result, the findings and implications are discussed. The results section is divided into two parts, the first part focuses on the comparison between a generated sample-free agent population and a microdata-based agent population. The second section focuses on the effect these two different populations have on the model outcomes of the HERoS model.

4.2. Population comparison

This section focuses on the quantitative comparison between the open-data and microdata agent populations. First, the differences are visualized for the whole population, then the populations are compared at the neighborhood level. The final sections focus on the quantification of the population differences, using the modified Freeman-Tukey and agent-matching metrics.

4.2.1. General population comparison

Figure 4.1 shows the differences between the populations on the whole population level. Three agent populations are plotted, two of which are constructed from open data (individuals0 and individuals8) and the microdata-based agent population. The x-axis shows the different categories for each attribute, and the y-axis the density. Since the two open-data-based populations have significant overlap, the plot shows no significant difference between the two.

The first observation is the fit between populations for the household structures, which is great because of the overlapping densities for all agent populations. This can be explained quite easily, as the open data contained exact information on the distribution of different households for the city of The Hague. However, there is still some error. This is due to the categorical mapping of variables. The microdata contains more information on households than open data, since the CBS uses eight categories for household types, whereas Den Haag in Cijfers uses four. By mapping institutional households as Type 1 households, some information is lost. The fit of the household size attribute is considerably worse than the one of the household structure. One of the assumptions of the open data algorithm is that households can consist of maximally seven people (two parents + five children). However, the microdata shows that this assumption does not always hold. However, it is not really clear what type of agents live in these large households. The open-data algorithm is only able to synthesize households that consist of one or two parents and their children. However, other types of households, such as student households or elderly homes are not present. Furthermore, one can see that the number of children per household does not follow the assumed truncated Poisson distribution. Instead, there are not enough one or two-children households, and too many three to five-children households. For the age variable, one can see three significant differences. The first difference is the peak of people between the age of 20 and 25 in the open-data-based population, the second observation is the undersampling of people between the age of 25 and 42, and the third observation is a large number of

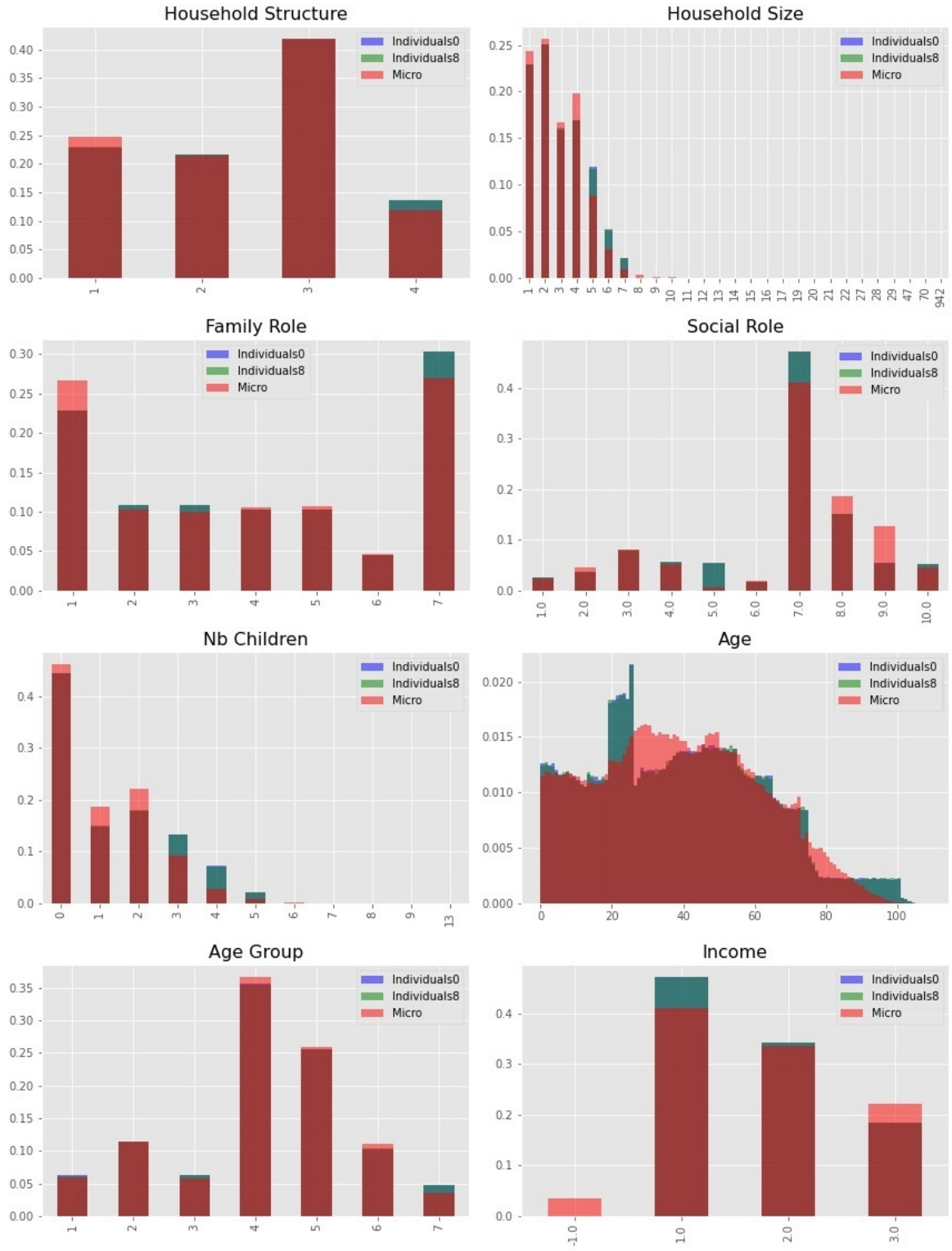

Figure 4.1: Histograms of agent attributes in populations based on micro and open data

agents over the age of 90 in the open-data-based population. However, if the people are put in age groups, the differences seem less apparent, as the densities of agents in group 4 (20-44), group 5 (45-64), and group 7 (80+) are similar. This shows that it is more difficult to synthesize an agent population on a more detailed level. Finally, for the income attribute, one can see there exists a difference in the number of categories, which is due to imperfections in the microdata. Whenever the income data is incomplete, the income of an individual is noted as -1. The synthetic data algorithm does not suffer from missing data, so every agent has an income. All differences between the micro and synthetic agent populations mentioned above serve as an example for a more general description of why differences between attributes occur. The following list identifies these reasons.

- **Mapping of micro to model data.** To be able to both compare the populations one-to-one and run the HERoS model with microdata, the microdata needs to be mapped to a HERoS-compatible format (For details, see section 3.6). The mapping of categories is absolutely necessary, as simply removing agents in categories that do not fit the definition of the HERoS model leads to fewer agents in the model. This is a problem, as the spread of pathogens such as COVID-19 is strongly affected by the density of the population. The mapping of categories is based on assumptions. For instance, it is assumed that members of an institutional household behave as if they are in a single-person household. However, a member of an institutional household may be an elderly person in a home for the elderly. As one has seen during COVID, elderly home residents are at major risk for hospitalization due to COVID-19 (Lu et al., 2021). Later on, the impact of such assumptions on model outcomes will be investigated. However, this finding also has implications for the design of models, because the designer of the model may make different choices based on the available data. In this case, the choice was made to implement four different types of households, because the available open data only contained four categories. However, if the modeler had access to microdata containing eight categories, the modeler may have opted to explicitly model all categories, or perhaps group together households that presumably are similar in behavior (e.g. married and unmarried couples). This finding implies that the design of the model is dependent on availability, as well as the level of detail in the data. Hence, if a modeler would have access to microdata before starting the modeling process, the resulting model may look very different, and therefore may produce different results.
- **Limitations of the open-data-based algorithm.** Some of the differences in the density for certain attributes can be assigned to the mechanisms of the synthetic data algorithm. The synthetic data algorithm produces agents above the age of 80 uniformly, which leads to too many 100-year-olds and too few 80-year-olds. Naturally, this can lead to an overestimation of the number of deaths/hospitalizations, as the older a person, the more susceptible to severe implications of pathogens. The open-data-based algorithm is also limited due to the assumptions made during the design of the algorithm. As mentioned before, assumptions are difficult to verify without a sample of the actual population. However, sample-free methods may be improved if the designer would have access to a sample/full population to verify assumptions.
- **Errors in microdata.** Finally, mismatches in attributes can be caused by incompleteness in microdata. As mentioned in the methods section, a perfect match between the micro and synthetic data is not desired, as the errors in microdata must also be synthesized if this is the case. This problem is a form of overfitting. To account for this problem, one can filter out values that are not consistent with definitions (e.g. single-person households that contain multiple people) and replace them with more appropriate values based on conditional probabilities. This form of processing microdata is outside of the scope of this study.

Based on visual inspection, one can see that the open-data-based generation algorithm generates a population that shows similar patterns to the microdata-based agent population. The differences between the populations can be explained, using three different reasons. This comparison was made on the highest level, but for an ABM, where spatial heterogeneity plays a major role, it is interesting to see how well the populations match on a smaller spatial scale. The differences between the agent populations on a smaller spatial scale are investigated in the next section.

4.2.2. Spatial comparison

The open data for The Hague consists of 109 neighborhoods, whereas the microdata consisted of 111 neighborhoods. The two extra neighborhoods that have been disregarded from this analysis. Figure 4.2 shows the differences per neighborhood for each agent attribute. The δ on the y-axis is the difference in the number of agents after subtracting the open-data-based attribute frequencies from the microdata attribute frequencies. The assumption here is that the microdata-based population is considered ground truth.

One can see that the differences per attribute of the agents differ in magnitude per neighborhood. The reasons for why these discrepancies happen are discussed in section 4.2.1. It is interesting to see that fit on a lower spatial level can greatly differ from the fit on a higher level (see figure 4.1 or for the city-level differences), for example, the number of agents in the single-person household structure. This goes on to show that a good fit on a global level does not necessarily implicate a good fit on a lower spatial level, and can be problematic if the model outcomes are very sensitive to the spatial heterogeneity of the agent populations.

Furthermore, there is a significant number of outliers, which shows that for certain neighborhoods, the difference between the two agent populations in that neighborhood is much larger than for the same attribute in other neighborhoods. Examples are the number of agents in social roles 7 and 9 (working and unemployed agents), the number of agents in a household without children, and the number of agents in family role 1 (single person). These differences may be attributed to two reasons. The first one is that the mechanisms of the open-data algorithm are not suited to predict these neighborhoods that are significantly different from the average neighborhood. For example, a neighborhood that is very dense with students or young professionals will have a high number of agents in the younger age groups. Applying the generic rules of the open-data algorithm population results in a population that does not show this discrepancy in age. The other reason could be that the data is lacking. For instance, one can see that there is a neighborhood in which over 1500 people do not have an income registered in the microdata. However, it is unclear what caused this lack of data. It could for instance be that the neighborhood houses many new people, who have not submitted their tax statements yet, or that the income of people in certain neighborhoods is unreported. Unfortunately, it is impossible to tell which one of these reasons is responsible for the discrepancies between the agent populations on the neighborhood level. In other words, is the poorness of fit for certain neighborhoods caused by the data, or the population generation algorithm? To alleviate this problem, it is recommended to involve other sources of information in the synthetic population synthesis process, in the form of expert knowledge. An expert on the studied population is able to tell whether the neighborhoods that are classified as outliers in particular attributes are truly different in comparison to other neighborhoods, or that the data is misrepresenting the population.

The analysis of the differences between the agent populations on the neighborhood level shows that the fit per attribute can differ greatly. This can be problematic if the outcomes of the model in which the populations are used are sensitive to spatial homogeneity, Furthermore, reasons for the differences between the agent populations are hypothesized, however, it is impossible to test these hypotheses without incorporating other sources of information. Therefore, it is recommended to involve expert opinion in the synthetic population generation process to improve the quality of the resulting agent population on a smaller spatial scale. Now that the differences between the agent populations have been visualized, they are subjected to quantitative methods, namely the modified Freeman-Tukey goodness of fit test and agent matching metrics.

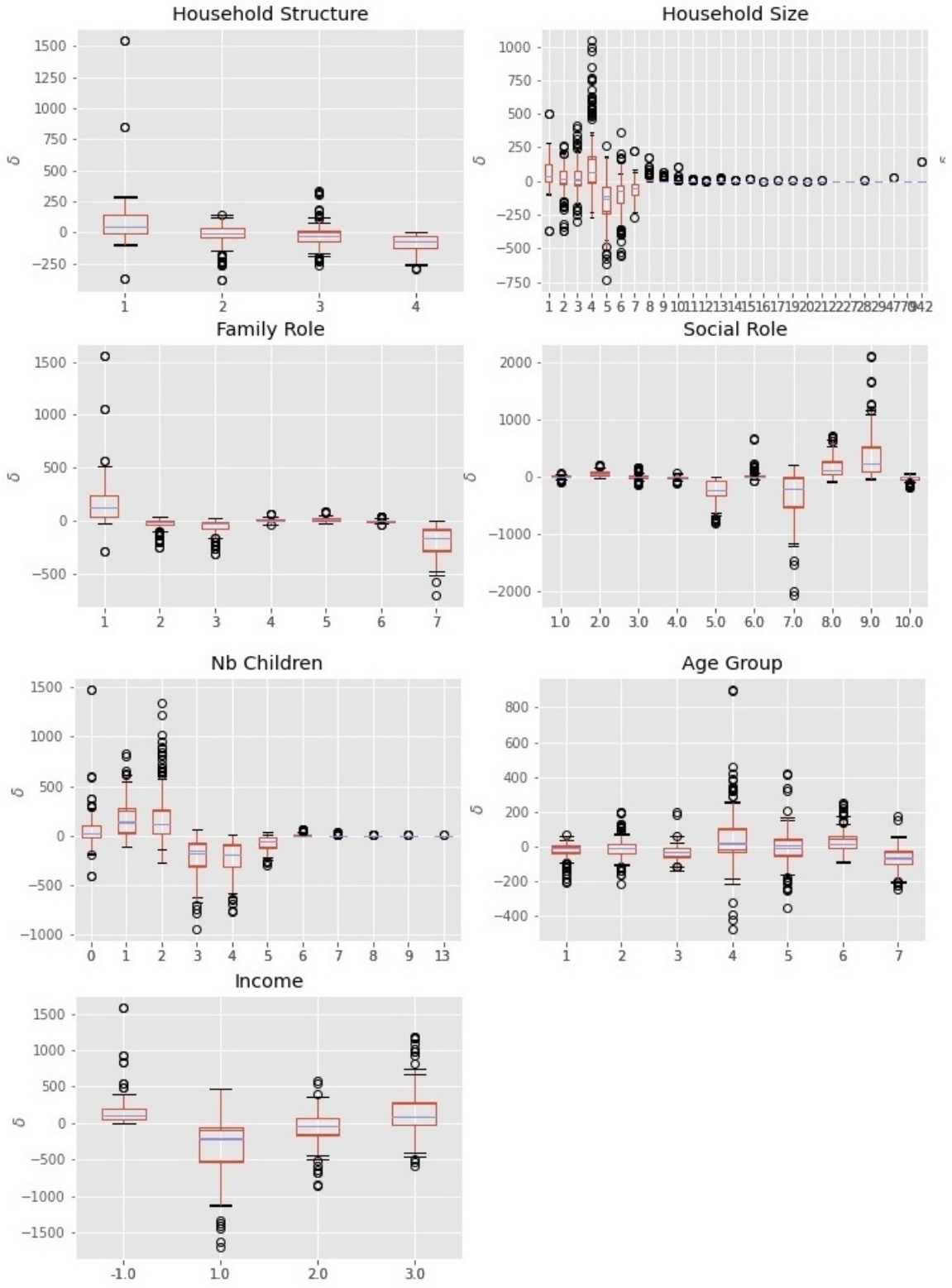

Figure 4.2: Differences in agent attributes on neighborhood level

4.2.3. Freeman-Tukey

The modified Freeman-Tukey statistic is able to capture all the complexity of comparing the agent populations in a single metric. The closer the FT value is to 0, the better the fit is between the population. Each point in figure 4.3 is a combination of attributes, all of which have been tested. One can see that the goodness-of-fit between the populations worsens as the number of synthesized attributes increases, showing that the more complex the agent populations are, the more difficult it is to achieve a proper fit. Secondly, the similarity of fit between the two agent populations suggests that the metric is not sensitive to small differences in the synthetic populations, albeit more populations should be tested to verify the robustness of this observation.

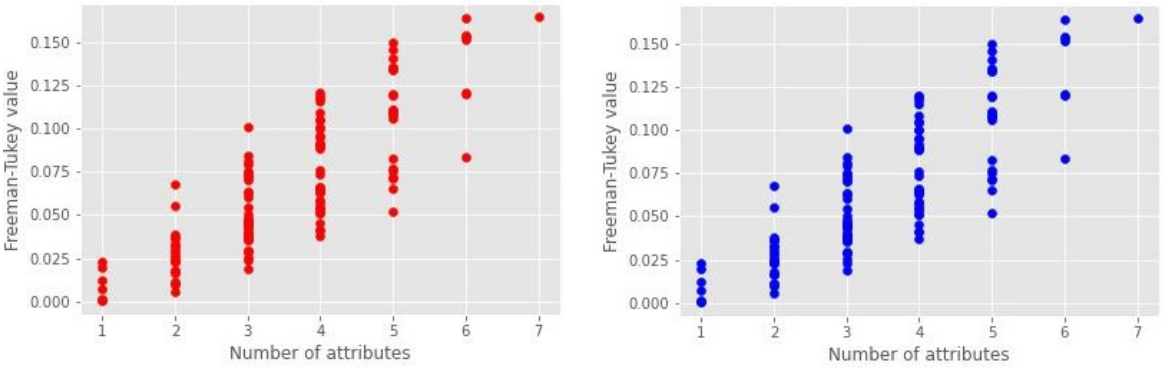

Figure 4.3: Modified Freeman-Tukey statistic for individuals0 (left), and individuals8 (right)

Interestingly enough, the modified Freeman-Tukey statistic also shows that the goodness-of-fit is dependent on the attributes taken into consideration. Hence, one can also use this metric to identify the attributes, or combinations thereof, that are the easiest to reproduce from open data. For both populations, the attribute with the best fit is the household structure, and the worst fit attribute is the social role. For an overview of the attribute combinations that produce the best and worst fits, see appendix B. Interestingly enough, agents that have 6 attributes may have a better fit than agents with 3 attributes, given that the right attributes are chosen. The implication here is that the **type** of attributes of the generated agents is just as important as the **quantity** of attributes. Therefore, the complexity of an agent (both in terms of kind and number of attributes) needs to be carefully considered when designing and implementing an ABM.

Overall, the modified Freeman-Tukey statistic is able to quantify the goodness-of-fit between populations, regardless of the complexity and size of the population. It is recommended for further research to compare different synthetic population algorithms using this metric and to report on the shortcomings of the metric. One of the shortcomings that can be noted here is the fact that access to microdata is necessary to compute the modified Freeman-Tukey statistic. Now that the modified Freeman-Tukey statistic is computed, the agent matching method is calculated for the agent populations on a neighborhood level to look into the spatial differences in goodness-of-fit.

4.2.4. Matching agents

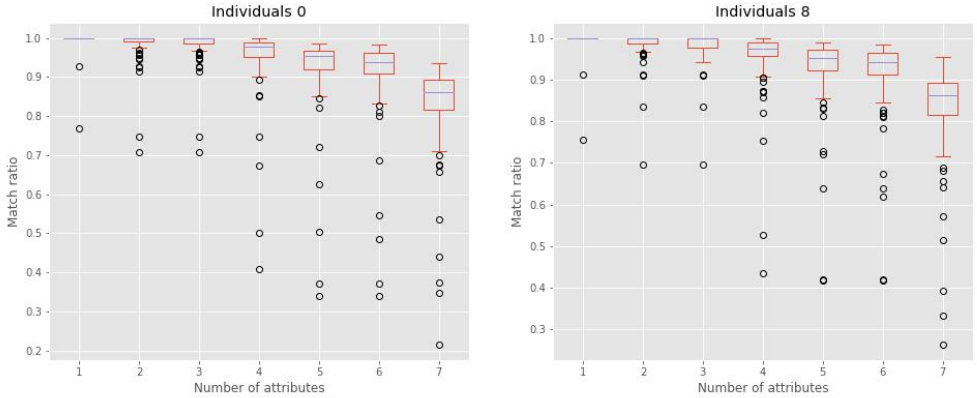

Figure 4.4: Results of matching agents in each neighborhood for individuals0 (left), and individuals8 (right)

The agent matching method computes the fraction of agents in the open-data-based population that have a counterpart in the microdata-based population (refer to section 3.8.2 for details). The results of the agent matching method are shown in figure 4.4, and also show that the more attributes are considered, the worse the fit to the micro population is. However, the difference in goodness-of-fit varies greatly per neighborhood. In the average neighborhood, 86% of the synthetic agents have a microdata counterpart when considering all attributes, however, in a particular neighborhood, this number is closer to 20%. Due to the ease of computation and interpretability, the agent matching method shows to be capable of quantifying the difference between agent populations.

4.2.5. Summary of population comparison results

In conclusion, it is clear that there are significant differences between the open-data-based and microdata-based agent populations. Additionally, the differences vary in magnitude per attribute and the spatial resolution that is considered. Furthermore, shortcomings of the open-data algorithm may be resolved when involving expert opinion in the synthesis process. The results of this section also show that both the type and number of agent attributes determine how good the fit between the agent populations is, which implies that the modeler should carefully think about the attributes of agents. The next step in this research project is to quantify the effect of population differences on model outcomes.

4.3. Model outcomes comparison

This section focuses on the influence of the different agent populations on the model outcomes. First, the impact of the input agent population on global model outcomes is investigated. After that, the effect of the agent population on model outcomes at the neighborhood level is investigated.

4.3.1. General outcome comparison

Figure 4.5 shows how the HERoS model outcomes change by altering the input population. For visualization purposes, the outcomes in the figure are a subset of all outcomes, please refer to appendix B for a complete overview. By changing the input agent populations, one can see that the results differ over an average of 10 runs for each population, indicating that the HERoS model is sensitive to the input agent population. Furthermore, the number of hospitalized and dead agents is over 15% lower when using microdata instead of open data. Even though fewer deaths are often desirable, the number here does not say anything in itself. Only when comparing the number of deaths to the ground truth, the difference can be considered desirable or not. Because the model is sensitive to the agent population, the quality of the model can be affected by the quality of the agent population. In the following section, the effect of the agent population on the precision and accuracy of the model is quantified.

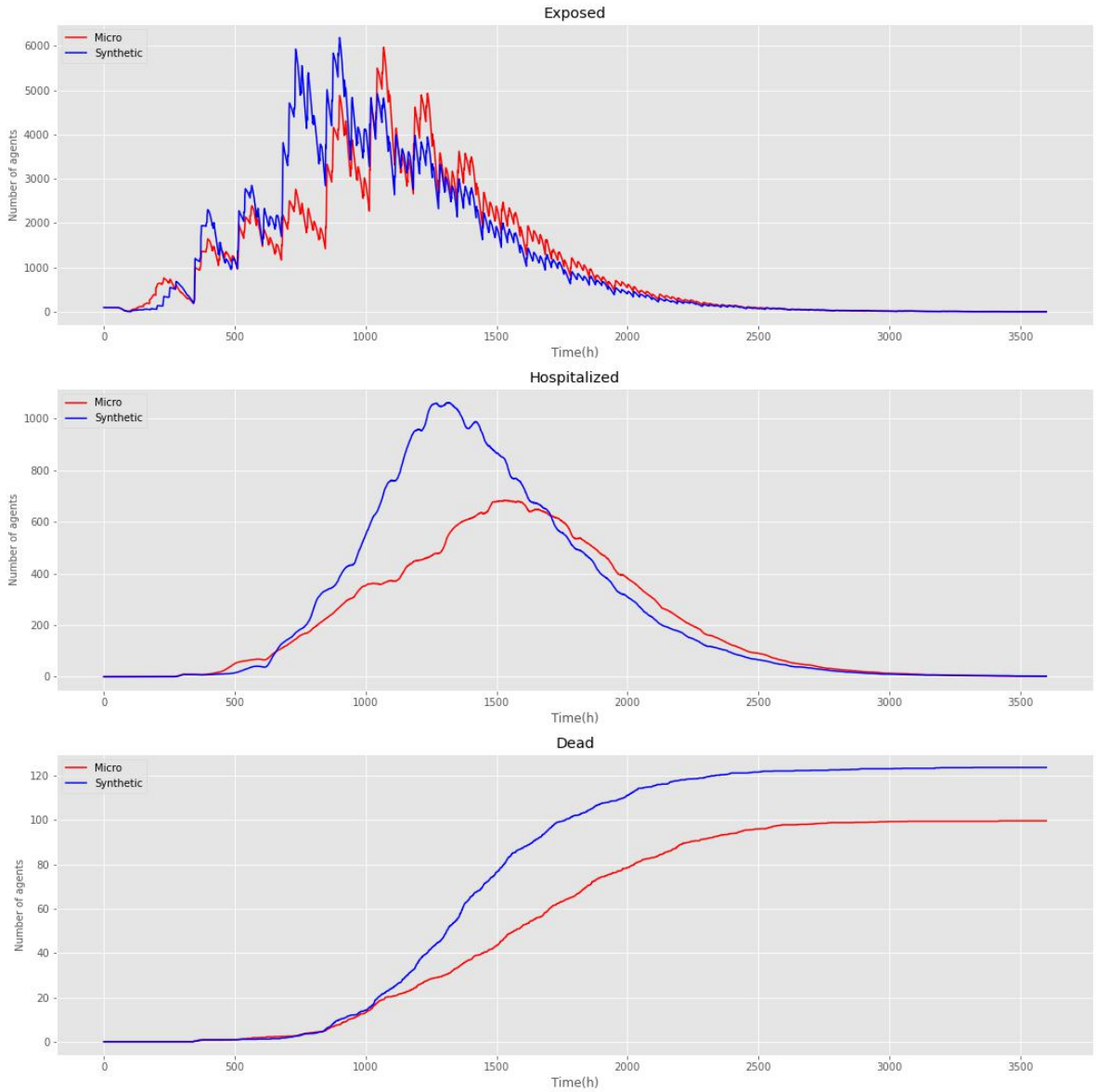

Figure 4.5: Mean model outcomes for 10 replications of HERoS model

4.3.2. The effect of agent populations on model precision

Figure 4.6 shows a subset of the model outcomes, with the bandwidths of the model outcomes. The bandwidths are determined by the minimum and maximum values over the 10 replications, and the black line is the mean over all replications. One can see that the bandwidths are smaller when using microdata instead of synthetic data. In the top right of the figure, the mean distance between the minimum and maximum values is given. In line with the visuals, one can see that the mean distance is consistently smaller when using microdata instead of open data. Since the bandwidth of the model outcomes decreases, the precision of the HERoS model increases when microdata is used. Hence, one of the advantages of using microdata over open data is the fact that the model outcomes are more precise. An advantage of higher precision is that it is easier to build a policy based on the model, as there is less uncertainty involved and there are fewer scenarios that need to be accounted for. However, an extremely precise model should always be met with caution due to the danger of a false sense of certainty. Now that it is clear that the model is more precise when using microdata, the effect of the agent population on the shape of the model outcomes is examined.

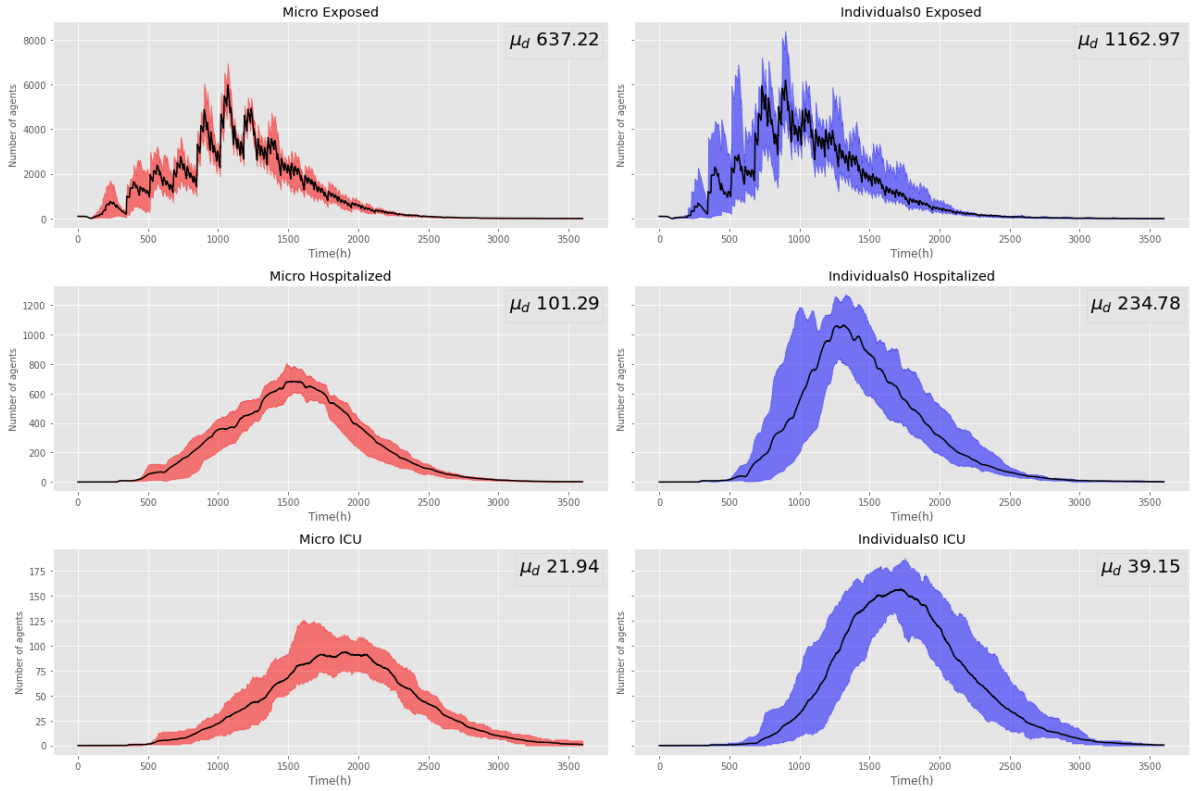

Figure 4.6: Model outcomes for 10 replications, with bandwidths and the mean distance between the minimum and maximum value

4.3.3. Phase comparison

Table 4.1: Mean differences in phase characteristics for 10 model replications

	I	II	III	IV
Duration	123.5	-305.5	66	116
Slope	NaN	0.084	-0.043	NaN

Figure 4.7 shows the phases I-IV, averaged over 10 replications. One can see that the uptake of ICU patients takes place in lower volumes and at a lower pace when using microdata. The values of the phase characteristics are in table 4.1, where one can see that the wave of COVID-19 when using synthetic data, starts earlier, peaks higher, and declines faster. Although the speed of viral spread is higher in the runs with an open-data-based synthetic population, the total number of recovered people is similar, which means the total contagiousness is not affected. Despite a similar number of agents being infected by the virus, the number of deaths is significantly higher in the synthetic data runs. An overestimation of the modeled number of deaths caused by COVID-19 can have many adversarial implications for policymakers who base their decision on this model and on model-based decision-making in general. It is difficult to establish a cause-effect relation between the population and the higher number of predicted deaths when using synthetic data, because multiple agent attributes may contribute to the change in model outcomes. One possible hypothesis to explain the higher number of deaths when using an open-data-based agent population is the fact that very elderly (>90) agents are more frequent in the synthetic data. However, an investigation of the model outcomes shows that this is not the case. Because of time constraints, it was not feasible to further investigate this effect.

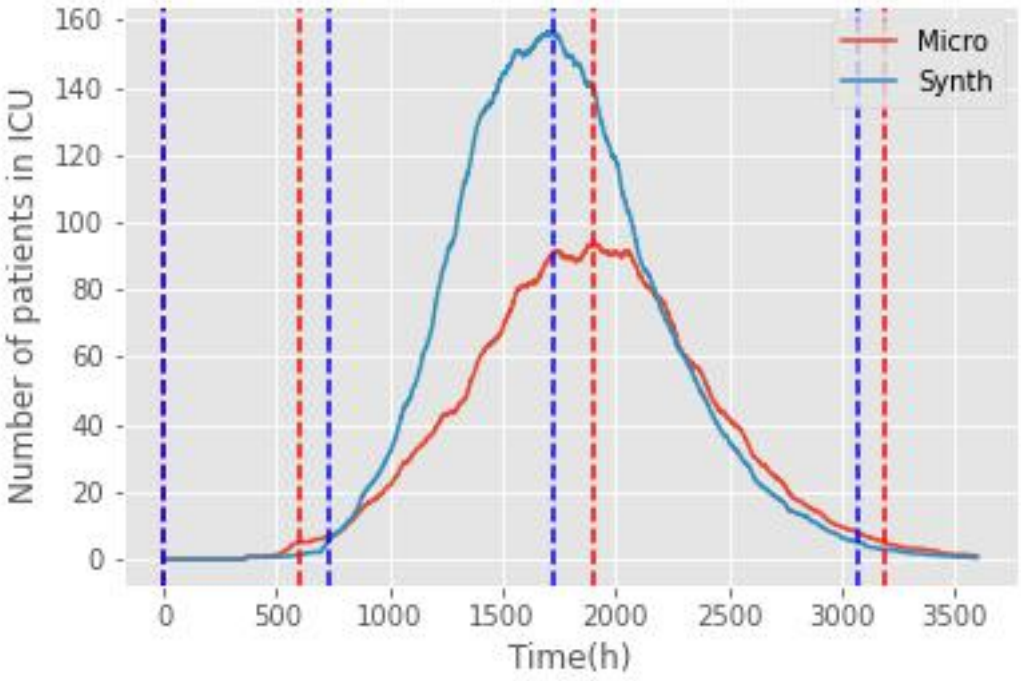

Figure 4.7: ICU phases, with PI = 5 and PIII = 5

4.3.4. Summary of the effects of agent population on model outcomes

Based on visually comparing the model outcomes, it becomes clear that the model is sensitive to the input population. Using a microdata-based agent population slows down the spread of the virus, but does not decrease the total number of infected agents. Despite a roughly equal number of infected agents, the number of predicted deaths is lower when using microdata. Furthermore, the precision of the model increases when using microdata, which is generally positive but should be met with caution. Similar to the previous section, the next section focuses on the effect of the agent population on the model outcomes at the neighborhood level.

4.3.5. Spatial variations in model outcomes

This section focuses on the spatial differences in model outcomes per neighborhood. Neighborhoods with outlying outcomes are investigated to further understand the impact of model inputs on model outcomes.

Figure 4.8: Difference in ICU patients at t=1008 per 1000 agents for 10 model replications

Figure 4.8 shows the difference in the number of ICU patients per neighborhood on a map of The Hague. The time t=1008 was chosen because the peak of ICU occupancy was at this moment in simulation time. The neighborhood where the difference in ICU occupancy per 1000 inhabitants was equal to 0 is grayed out. Please note that the difference here is calculated by subtracting the synthetic outcomes from the micro outcomes. A positive value means the number of ICU uptakes per neighborhood was higher when the model is run using micro agents. One can see here that the outcomes in ICU patients per neighborhood for most neighborhoods are similar. This implies that the dynamics of the virus per neighborhood are comparable regardless of the input agent population. By using Fisher-Jenks natural breaks optimization to create the color gradient, one particular neighborhood stands out. This neighborhood is the Binckhorst, depicted in dark blue in figure 4.8. The Binckhorst neighborhood is further investigated later in this chapter.

Besides visualizing the differences in the ICU occupancy at the peak of ICU load, the differences in the number of agent deaths per neighborhood can be seen in figure 4.9. The red line shows where the difference in the number of deaths per neighborhood is equal to 0. The black lines are drawn at distance d from the red line and visualize an artificial bound for which the difference between the model outcomes is not significant. In this figure, $d = 0.3$. Most neighborhoods are located in the lower left quadrant, with relatively few deaths per 1000 inhabitants, regardless of the input agent population. Since there are more points above the red line than below, using synthetic data tends to increase the number of deaths per neighborhood, which is also observed when looking at non-spatial model outcomes. The two data points that deviate the furthest from the red fit line warrant further interest. These neighborhoods, where the difference between the model outcomes is the largest, are worth investigating in order to increase the understanding of the HERoS model and its sensitivities to agent populations. The top left refers to the neighborhood Vlietzoom-Oost, where there were no deaths in the micro population, but 1.8 deaths per 1000 synthetic agents, the most of any synthetic neighborhood. Why is the difference so large? The explanation here is quite simple. The Vlietzoom-Oost neighborhood consists of only 97 agents. Due to the law of small numbers, it is unreasonable to draw any conclusions from this particular neighborhood. Next to Vlietzoom-Oost, the other major outlier is the Binckhorst, which was also identified as an outlier in the previous section. The following paragraph further investigates the Binckhorst, and describes the connection between the difference in model outcomes and the agent population.

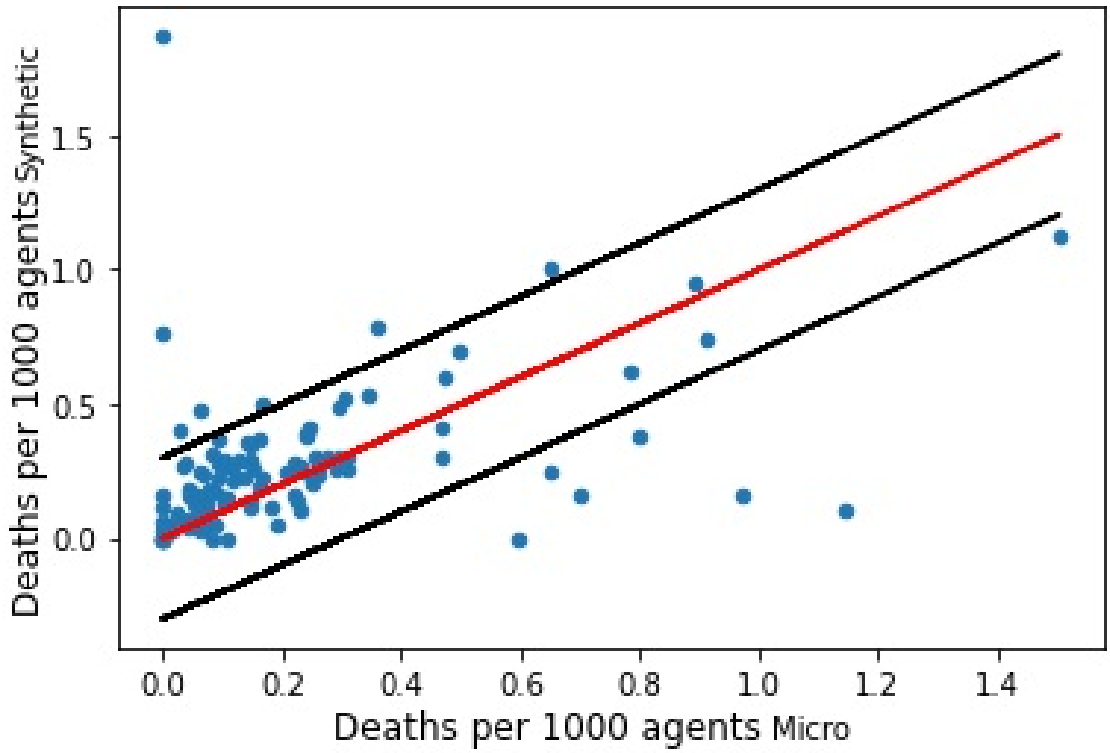

Figure 4.9: Mean neighborhood deaths per 1000 agents for 10 replications, with $d = 0.3$

4.3.6. Investigating the Binckhorst neighborhood

In figure 4.10 the difference between the open-data-based and microdata-based agent populations can be seen. The figure clearly shows major differences between the population. Based on the microdata, the Binckhorst is a neighborhood that is characterized by large households, many single-people households, many unemployed people, and a lack of data in the income category. However, when observing the synthetic population, these characteristics are not represented at all. Since the model is clearly sensitive to the agent population, it is not surprising that the results for the Binckhorst are substantially different when using the different populations as input. It is hypothesized that it is due to the differences in household sizes and unemployment. Because many more people live in a single household, the COVID-19 virus spreads more easily to other household members as they are in close proximity to one another. Furthermore, most of these agents are unemployed and do not leave the house often. If agents spend more time together in close proximity, it is of no surprise that the total number of infections, ICU uptakes, and deaths is higher when using microdata. Although the results of this hypothesis are in line with findings in the literature (Allen, Darlington, Hughes, & Bellis, 2022), the pathways through which they occur are different. This raises questions on how well the unemployed social group is modeled. Hence, it is recommended to further investigate the behavior of unemployed agents. It could be that unemployed people in the Binckhorst do work but in an unregistered manner. Additionally, it is recommended to further look into the behavior of agents in larger households. This is not an issue when using synthetic data, where the complexity of a household is limited to a maximum of 7 people, but does play a role when microdata is used. Other models deal with this by implementing home isolation behavior, where agents stay isolated within their own homes. On a more global level, using more detailed data for agent-based models may require substantially more effort by modelers, as the detailed data must be handled properly and appropriate behavior must be constructed for the detailed agents. The trade-off between complexity and realistically modeling complex behavior is complicated, and modelers should always be aware of this issue when using more detailed data. Finally, investigating outliers in the model outcomes leads to an additional understanding of the sensitivity of the HERoS model to input agent populations, as well as the mechanics of the model.

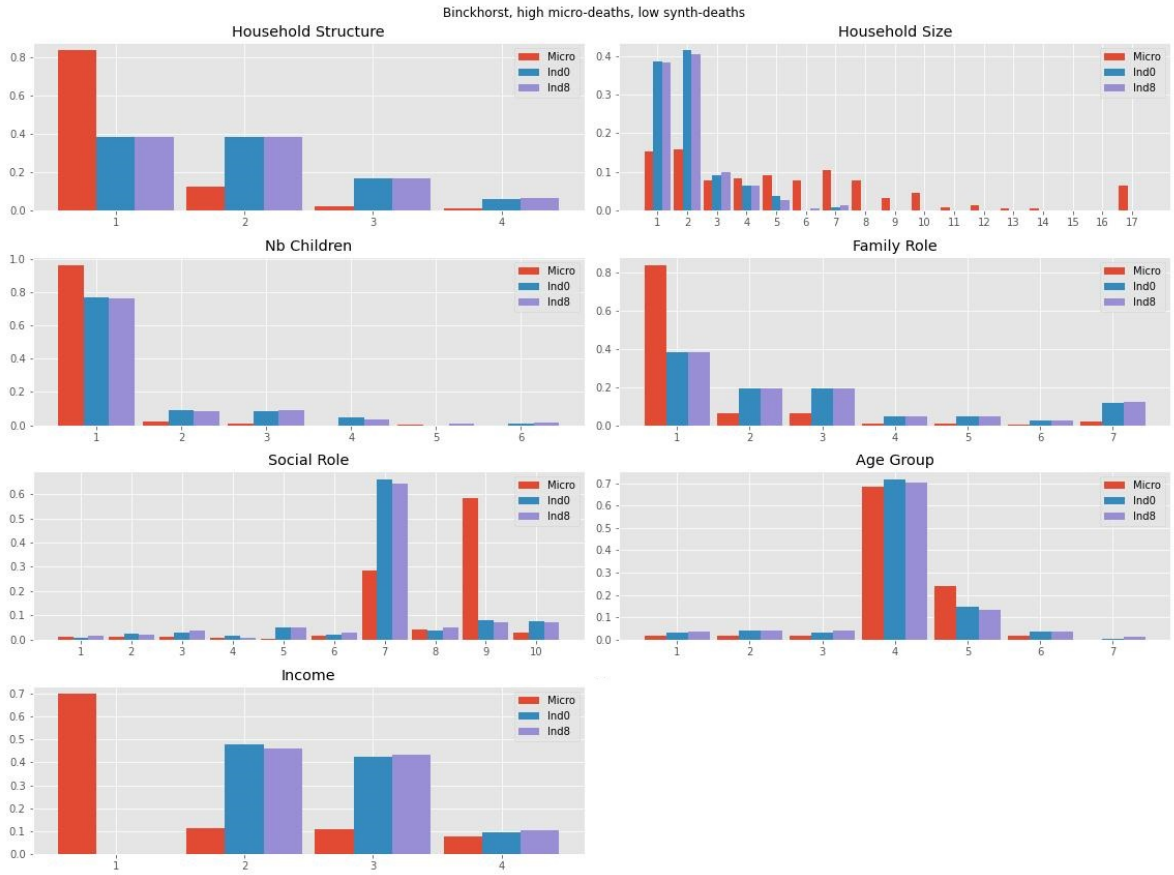

Figure 4.10: Population differences between synthetic and micro agents for the Binckhorst neighborhood

By investigating the spatial differences in model outcomes, some valuable insights have been obtained. The first one is that the spatial differences in model outcomes per neighborhood do not differ substantially for most neighborhoods. Nonetheless, the differences between model outcomes in certain neighborhoods were substantial. By investigating these outliers, it was found that the populations of the open-data-based and microdata-based agent populations of the Binckhorst substantially differentiated in terms of household size and unemployment. It is hypothesized that these combine to set a causal chain in effect that increases the number of deaths in the neighborhood. Even though the final results are confirmed by literature, as unemployed people tend to be more at risk for severe health consequences of COVID-19, it is doubtful the model represents this phenomenon adequately. Therefore, it is recommended to further look into the behavior of unemployed agents and agents in large households.

4.4. Summary of results

In the previous sections the differences between the agent populations, as well as the impacts of the agent population on model outcomes are visualized and discussed. The first finding was that the sample-free algorithm was able to generate an agent population that is similar to the microdata-based agent population. However, there are differences, which can be attributed to three reasons, discussed in section 4.2.1. Secondly, it was found that data missing in the microdata set is not uniformly distributed over the neighborhoods. This lack of data, combined with the difficulty in verifying neighborhood populations, makes it recommended to involve expert opinion when working with sample-free synthetic data algorithms. Thirdly, it was ascertained that the modified Freeman-Tukey statistic is able to capture the goodness-of-fit in a sample-size independent single value, regardless of population complexity. The goodness-of-fit is strongly dependent on the number and combinations of attributes, confirmed by both

the FT and agent matching metrics. Agent matching also shows promise to capture the goodness-of-fit between populations, and also allows the modeler to filter synthetic outliers and hard-to-predict micro agents. Moreover, agent matching shows that the need for microdata to produce valid models is very much dependent on the type of model and its reliance on agents that possess a unique set of properties. In the model outcomes comparison, it was found that the HERoS model is sensitive to the changes in input population and that using an open-data-based agent population increases the viral spread speed, the number of hospitalizations, ICU occupants, and deaths. Additionally, the precision of the model is improved when using a microdata-based population. Similar to the populations of agents, the model outcomes did not differ substantially for most neighborhoods. However, one particular neighborhood stood out. This neighborhood, the Binckhorst, was further investigated to determine the cause of the differences between the outcomes for this neighborhood. It is presumably caused by a combination of different household sizes and unemployment rates. This investigation led to an increased understanding of the HERoS model.

5 Discussion

This chapter discusses the limitations and implications of this research project and provides recommendations for further research.

5.1. Limitations

As with any research, the implications of this study should be met with caution, as limitations are in place. This section focuses on these limitations and addresses why the results are nonetheless valid.

5.1.1. Accounting for stochasticity in agent populations

Wherever possible, all the tests have been carried out using two different open-data-based agent populations; individuals0 and individuals8. These populations are based on different random seeds of the synthetic population generation algorithm. It is currently unclear how much the findings of this project are influenced by the limited number of tested populations. However, a quick and dirty analysis of five different open-data-based populations suggests that the population differences between different seeds of the population are orders of magnitude smaller than the differences between the open-data-based and microdata-based populations. The findings of this project have shown that the model is sensitive to the agent population, yet no significantly different patterns in model outcomes have been observed. Therefore, it is expected that the influence of the synthetic population algorithm seed does not significantly alter the conclusions drawn in this research project.

5.1.2. Missing data

Within the microdata, data on the income of individuals is largely missing in some neighborhoods. One of these neighborhoods is the Binckhorst, where the model outcomes differed greatly for both agent populations. It is unclear how large the influence of missing income data is on these outcomes, and thus may affect the conclusions. One way to test this is to create income data for the neighborhoods using imputation and rerun the model, but that was not carried out due to time constraints.

5.1.3. Structural analysis of model outcomes

This research project focuses on the impact of agent populations on model outcomes. As was found in the literature review, the preferred method for doing this type of analysis is SOBOL. However, it was also found that SOBOL is expensive computationally, which made it impossible to carry out in the given time period of this project. As an alternative, an OFAT analysis was carried out, where all input parameters, except for the agent population, were kept constant. OFAT turns a blind eye to interaction effects between parameters, which limits the amount of information gained over the impact of agent populations. Furthermore, the OFAT analysis was also not completed fully, as only one parameter has been changed. The result is little extra understanding of the importance of the agent population, compared to other model parameters.

5.1.4. Generalization of findings

The ambitious goal of this research project is to investigate the difference between open- and microdata for synthesizing agent populations that are used in large-scale epidemiological ABMs. A case study in the form of the HERoS model was used to investigate, however, a single case study is not enough to draw conclusions for the entire field of epidemiological ABMs. Therefore, this study is only limited to drawing conclusions that are valid for the HERoS model but does provide a beginning for a more broad analysis where other models are investigated as well.

Despite these limitations, this study has progressed the knowledge of synthetic agent populations significantly. As with any research, there are still many unanswered questions, which can be answered through further research. These questions are discussed in the recommendations for further work sections. But first, the implications of this study are delved into.

5.2. Implications

In this section, the implications of this research are discussed. First, a recap of the most important implications from the results section is given, after which the most important implications of the project itself are elaborated upon.

5.2.1. Recap of findings

Based on the previous section, the most important findings in this research project are as follows:

- The sample-free generation algorithm employed by the HERoS model is able to construct a synthetic population that shows the same patterns as the microdata-based population on a global level. However, population differences at the neighborhood level can be significant.
- The modified Freeman-Tukey and agent matching metrics are able to capture agent population differences in a single, comparable value.
- The HERoS model is sensitive to the changes in agent populations and using a synthetic agent population increases the total number of hospitalizations, ICU occupants, and deaths.
- The precision of the HERoS model improves when using a microdata-based agent population.
- Significant differences in agent populations lead to significant differences in model outcomes.

5.2.2. Implications of findings

The most important finding of this research project is that sample-free synthetic population generation algorithms are able to create a population that exhibits the same patterns as the population that is created using the most detailed data available. Although very case-dependent, this means that microdata or public use micro samples are not a necessity when an agent population must be synthesized. Although, using microdata has advantages as well. The first one is that the precision of a model is improved when microdata is the data source for the agent population. Secondly, a sample-free population generation algorithm may be less suitable to create an agent population if agents on the edge of distributions play a major role. As seen in the comparison between the populations, the agent with relatively unique combinations of attributes are difficult to synthesize based on aggregates. Furthermore, modelers also need to be aware of the drawbacks of working with microdata. Getting access to microdata in the Netherlands is not trivial and requires substantial resources. Moreover, working with microdata happens in a secure environment, where the data going in and out of the system is closely monitored. This can slow down research and needs to be taken into consideration when searching for suitable data sources. Finally, a modeler should keep in mind that a model is only as good as its worst part. Hence, having a perfect agent population does not guarantee that the model is a better fit for its purpose, as a model consists of more submodels that the agent population. This implies that modelers need to be able to identify the parts of their model that negatively influence the model quality, and act accordingly.

5.2.3. Implications for CBS

The current CBS policy regarding microdata is to restrict access for anyone, except for scientific institutions that perform statistical research. Access can be granted through an extensive application procedure. Based on the findings in this research, this policy should be kept in place, as it has been

shown that open data can indeed be a viable alternative to microdata. Since the main priority of the CBS is to maintain privacy, it is recommended to continue investigating the implications of synthetic data on privacy. However, synthetic data generation techniques span wider than sample-free techniques, and this research only spans a specific application domain of synthetic data generation techniques.

5.3. Why do the results matter?

One of the sustainable development goals of the United Nations is to ”Ensure healthy lives and promote well-being for all at all ages” (United Nations, n.d.). The COVID-19 pandemic has had and is still having, a detrimental impact on this goal, through both direct and indirect pathways. The advent of technology has presented policymakers with new tools that can aid in mitigating a pandemic, and building resilience against them. Agent-based models are generic tools that can be used to tackle specific problems, which have proven their worth in aiding organizations to combat pathogen threats. Improving these tools directly contributes to the ability of policymakers to minimize the harm done by pathogens. One of the challenges in the synthesis of an agent-based model is the population. Since population data is often GDPR bound and thus difficult to acquire, it can prove challenging to build an agent population that is representative of the studied population. Sample-free algorithms provide a solution for this problem by only requiring aggregate data of the population, minimizing the risk of privacy issues. However, it had not been studied how well these synthetic populations are able to represent an actual population, and what impact the population quality has on the model outcomes. This research project focuses on this gap in knowledge and takes the first steps to fill this gap. By doing so, the research project contributes to the understanding of agent-based models and their relationship to the world. These contributions hope to aid modelers to create better decision-making tools, and policy-makers to make optimal use of them.

5.4. Relation to literature review

The literature review of this project identified two gaps in knowledge: The quality of sample-free synthetic generation algorithms and the impact of the population quality on ABM outcomes had not been studied yet. By performing a case study on the HERoS model, which employs a sample-free synthetic data algorithm to simulate the spread of COVID-19 in a city, these knowledge gaps are investigated. The findings of the research project extend the literature on this topic, but there is additional work needed to resolve them. Another extensively covered topic in the literature review is general sensitivity analysis and in particular the SOBOL method. Although SOBOL and its extensions are applicable in this case study, it was not executed due to time constraints. With this limitation in mind, it provides an opportunity for further research.

5.5. Recommendations for further research

This section provides opportunities for other researchers to further expand this research.

5.5.1. Generalization of findings

The literature review of this research project found two knowledge gaps, which both pertained to the usage of synthetic data in agent-based models. By investigating the HERoS model and the usage of open- and microdata therein, the first steps investigating the difference between micro- and open data are taken. Naturally, more research in this field is necessary to increase scientific understanding. By analyzing other models, methods, and their relationship to one another, a more complete overview can be created that can fill the knowledge gaps more completely. Since this research proposes comparable metrics that are applicable regardless of the synthetic population generation method, they can be used to assess the performance of different generation techniques.

5.5.2. Hypothesis testing

In the results section of this research, hypotheses are often proposed that serve as a possible explanation for the observed outcomes. However, these are hypotheses and require testing to either confirm or deny them. Due to time constraints, it was not possible to fully investigate these propositions. Fortunately, the main findings are not affected, as these are independent of the internal mechanisms of the HERoS model. It is recommended to test these hypotheses to further understand internal mechanisms,

which may improve the quality of the HERoS model and epidemiological models in general.

5.5.3. SOBOL analysis

One of the main methods investigated in the literature review is the SOBOL analysis, which allows for the quantification of the sensitivity of a model to individual and combinations of attributes. Though, as is described in the literature, SOBOL requires substantial computational power. Due to time and computational constraints, a SOBOL analysis has not been carried out in this study, but doing so allows for the ranking of agent attributes in terms of sensitivity. This ranking can be used to determine which attributes are the most important to get 'right'. If it turns out certain attributes may not matter for the model outcomes, the complex problem population generation algorithms face can be significantly simplified.

5.5.4. Investigating attribute interdependence

As mentioned in the literature review, the complexity of a SOBOL analysis is dependent on the interdependence of the attributes. A preliminary analysis using Spearman rank correlation suggests that there is a strong correlation between attributes. However, Spearman rank correlation is not a suitable method to investigate the relationship between nominal attributes. The first step in a SOBOL analysis should investigate this further, for example with Cramer's V metric. Furthermore, heavily correlated attributes may make population synthesis easier, as only one of the attributes needs to be known to accurately model the other one (e.g. if an agent lives in a single-person household, the household size must be one).

5.5.5. Importance of the agent population relative to other data sets

Most ABMs, including the HERoS model, rely on multiple datasets to realistically model the phenomena they study. This research project is concerned with the agent population part of the model. Naturally, this is not the only part of the model that requires data. Other examples would be the activity schedules of the agents or their travel patterns. The quality of each of these datasets affects the usability of the model. It would be interesting to investigate the importance of each of these datasets, and question which of these datasets bottlenecks the usability of the HERoS model.

5.5.6. Investigating privacy

The agent matching metric has shown that it is possible to create agents from open data that exactly match the properties of agents synthesized from microdata. However, this also raises concerns about privacy, as theoretically it is possible that a synthesized agent can be traced back to an actual person. In this study, this seems not to be an issue, as the cells in the contingency table are not small enough to trace back individuals, but it may be possible to trace back individuals when the synthetic dataset is combined with other datasets. It is recommended to study the implications of synthetic data on individual privacy.

5.5.7. Investigating societal cost of NPI's

Currently, the HERoS model limits itself to reporting back disease numbers in terms of infections, hospitalizations, ICU uptakes, and deaths. In HERoS deliverable 2.1, it was found that lockdowns are a robust and effective NPI to combat viral spread. However, it is also known that lockdowns significantly impact other aspects of health negatively. Models that only report back infection numbers may lead to policymaker tunnel vision to only minimize the number of deaths associated with a pathogen, instead of considering the total societal costs. It is therefore recommended to expand the HERoS model with model outcomes that serve as an indicator for societal cost, so that policymakers may able to minimize that instead of just infection rates. This relationship between the sources of data used and the societal costs then also warrants investigation.

5.5.8. Hybrid agent populations

Another interesting experiment would be to investigate agent populations that are partially based on microdata, and partially on open data. If in combination with for instance a SOBOL analysis, it turns out that open data is unable to properly synthesize the most important attributes of agents, it may be interesting to replace exactly these columns with values based on microdata. Since microdata contains

the full conditional probability $P(x|y)$, where y is the full set of the other agent attributes, it is possible to synthesize the missing attribute x. It is recommended to investigate the gain in population quality and the tradeoff between the quality and privacy of such populations.

5.5.9. Assessing differences between models that were designed with microdata in mind

The design of a model is dependent on the data that is available. In the HERoS model, the number of social roles is limited, whereas many more categories could be implemented if data is available. In the microdata, the SECM variable denotes at least 13 distinct categories of social classes. The additional detail in microdata allows modelers to model a larger variety of agents, each with their own behavior. The first question one should ask here is, does this improve the model in terms of quality or usability? It is currently unclear what the answer to this question is. Hence, it is proposed to build a model that is designed with microdata in mind but generates the same model outcomes. Only then, one is able to fully use the potential of microdata, and answer whether or not microdata is truly necessary for agent-based modeling.

5.6. Summary

The discussion section discussed the limitations, implications, and recommendations for further research. This study is limited due to the limited seed analysis, missing data, lack of structural analysis of the model outcomes, and the generalization of the findings. In spite of these limitations, the most important implication is that models do not require microdata to provide policymakers with valuable insights, but that microdata can be useful. Furthermore, an extensive list of topics for further research is presented. Although cliche, more research is simply needed to further fill the knowledge gap identified in the literature review.

6

Conclusion

6.1. Main conclusion

Epidemics have been shown to be a major threat to societal well-being throughout history. The still ongoing COVID-19 pandemic is a recent example that shows that the global population is vulnerable to communicable diseases. Fortunately, advancements in technology have allowed policymakers to turn to advanced decision-making tools that aid in mitigating the effects of a pandemic as effectively as possible. One of these tools are agent-based models (ABMs), a simulation paradigm that explicitly simulates the behavior of agents, whose interaction with each other results in system-level emergent behavior. An ABM is perfectly suited to study the dynamics of disease spread, due to the interconnectedness of human behavior and disease spread, the unequal distribution of health burden, and the heterogeneity of the population. In most ABMs, each agent represents an individual in the studied population, but agents can take any form. The set of agents is known as the **agent population**. Each agent in the agent populations has **attributes**, which determine their behavior. An example of an attribute is the age of an agent, which determines whether an agent goes to school, works, or is retired.

One of the characteristics of an ABM is the amount of data that is needed to model a valid ABM. For example, data is needed to capture the attributes of agents, but also to simulate their activity patterns and the environment in which they exist (e.g. a city). Obtaining timely and accurate data is difficult and presents a major obstacle to the advancement of ABMs. This research project zooms in on the data that is needed for the creation of an agent population. The process of the creation of a population is known as **synthetic population generation**. There are multiple forms of data that can be used for the synthesis of an agent population. The first one is microdata, which is traceable data on an individual level. Microdata is perfectly suited to synthesize an agent population, as the data can be directly translated into a model-usable form, and no assumptions need to be made about the population. However, working with microdata is difficult due to privacy concerns and GDPR regulations. Microdata is safeguarded by the data owners, and access to it is severely limited. To circumvent this problem, modelers turn to agent population synthesis techniques that are able to generate an agent population from other sources of data. **Sample-free algorithms** are a subfamily of population generation techniques that only require aggregate data on the population for population synthesis. The resulting agent population is of lower quality than an agent population created from microdata, due to the translation that is necessary to transform aggregate to individual-level data.

Before this study, the difference in quality between the agent populations had not been studied yet. Furthermore, the impact of the different agent populations on the model outcomes also had not been studied. These two gaps in knowledge led to the following main research question:
What is the effect of using a sample-free, open-data-based agent population over a microdata-based agent population as input for a large-scale epidemiological agent-based model?
To answer this question, the problem was investigated by making use of a case study, in the form of the HERoS model. The HERoS model is an agent-based model that models the spread of COVID-19 through the city of The Hague. The HERoS model employs a sample-free algorithm to synthesize an

agent population. Using microdata, made available by CBS, another agent population was synthesized, and the differences between the populations were investigated. The agent populations were then used as input for the HERoS model, and their corresponding outcomes were compared. It is concluded that the effect of using open data instead of microdata is multifaceted. First of all, it is shown that the agent populations differ in a variety of attributes. These differences become larger when evaluating the differences between populations on a smaller spatial scale (in this case neighborhood level). Since the HERoS model is sensitive to the input agent population, the model outcomes change when another population is used as input. If an open-data-based agent population is used, the number of hospitalization, ICU occupants, and Deaths is higher than when a micro-data-based agent population is used. Furthermore, the precision (i.e. the statistical variance) of the model is higher when using microdata instead of open data. With the limitations of this study in mind, namely the lack of seed analysis for the agent population generation, missing data, and the lack of a structural method to analyze model outcomes, these findings imply that the usage of microdata is not necessary in the case of the HERoS model. In this particular case, the benefits (higher precision), do not outweigh the drawbacks (getting access to the data, working in a secure environment, and cost of data). However, there are still many open-ended questions pertaining to the usage of open data in epidemiological ABMs. The continuation of this chapter focuses on answering the subquestions.

6.2. Subquestions

6.2.1. What sample-free synthetic population generation techniques are available to generate a two-layered population for an epidemiological ABM?

The literature review presented a multitude of sample-free synthetic population generation techniques. Examples are the techniques by Gargiulo et al. (2010), Barthelemy and Toint (2013), and Wickramasinghe (2019). In this project, the technique presented in Sirenko et al. (2020) is used to synthesize agent populations.

6.2.2. What methods are available to quantify the difference between an open-data-based synthetic population and a microdata-based population?

The Freeman-Tukey test statistic is often used in the literature the quantify the difference between open-data- and microdata-based populations. Other usable metrics are the Proportion of Good Prediction, the χ^2 metric, and the Absolute Percentage Difference. This study modifies the Freeman-Tukey statistic to account for differences in population sizes and make the value of the statistic comparable. Furthermore, the agent matching metric is introduced and applied in this study, adding two new metrics to the collection.

6.2.3. What methods are available to analyze the sensitivity of a large-scale ABM to different inputs?

There is a multitude of methods available to analyze the sensitivity of a large-scale ABM. Amongst these are SOBOL and one-factor-at-a-time (OFAT) analysis. SOBOL is part of the global sensitivity analysis methods, which also take interaction effects between model parameters into account. For a more complete overview, please refer to Saltelli et al. (2008). In this project, only an OFAT analysis is performed due to time constraints but is recommended to perform a SOBOL analysis as well.

6.2.4. What are the statistical differences between an open-data-based agent population and a microdata-based agent population?

The differences between an open-data-based and microdata-based population present themselves at various levels. The most notable ones are of course the data and technique used to convert data to an agent population. In the case of the HERoS model agent populations, the differences between the populations differ per attribute. The household structure attribute is similar, whereas the household size attribute can vary greatly between the populations. Furthermore, if one zooms into a smaller spatial level, the differences between the populations are larger. Effects on a local level may play a role here, for instance, the underreporting of income data for the Binckhorst neighborhood in The Hague.

6.2.5. How does the agent population affect the model outcomes of the HERoS model?

It has been shown that the agent population affects the outcomes of the HERoS model. If an open-data-based agent population is used, the number of hospitalizations, ICU occupants, and Deaths is higher. In contrast with the general trend that using synthetic data leads to more deaths, the Binckhorst (characterized by a high unemployment rate and large households) shows the opposite. Finally, the precision of the model decreases when using an open-data-based agent population.

6.2.6. What are the implications of using microdata over open data?

If one uses microdata over open data, there are several implications that are important. Working with microdata requires the user to work in a secure environment, where the microdata cannot be exported from. In a secure environment, the privileges of the users are limited and the preferred software of the modeler may not be used there. Additionally, the publication of results based on microdata must first be checked to minimize the risk of the disclosure of subjects. To add to that, there are also financial costs involved when one wants to work with microdata, which is especially relevant for external researchers that would like to have access to the data in the CBSs SSB. Working with open data suffers none of these drawbacks, but there are limits to what one can do with open data.

6.2.7. What is the added value of microdata over open data?

Microdata adds value in multiple ways. The first one is the fact that having access to microdata allows the modeler to check their assumptions and validate their findings with another source of data. Without microdata, this remains guesswork. The microdata also aids in further refining open-data-based synthetic population algorithms, as it allows the designer of the algorithm to spot and eradicate bugs in the algorithm. Moreover, microdata reduces the amount of time it takes to create an agent population, as the disaggregation and synthesis steps in an open-data-based algorithm can be skipped. The creation of an agent population is rather straightforward, as it was only required to map categories and handle data inconsistencies.

These answers conclude this research project, which contributed valuable insights to the study of epidemiological ABMs, synthetic agent populations, open and microdata, and the HERoS model.

References

Allen, J., Darlington, O., Hughes, K., & Bellis, M. A. (2022). The public health impact of loneliness during the covid-19 pandemic. *BMC public health*, *22*(1), 1–9. doi:10.1186/s12889-022-14055-2

Bakker, B. F., Van Rooijen, J., & Van Toor, L. (2014). The system of social statistical datasets of statistics netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, *30*(4), 411–424. doi:10.3233/SJI-140803

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *47*(2), 266–279. doi:10.1287/trsc.1120.0408

Bedson, J., Skrip, L. A., Pedi, D., Abramowitz, S., Carter, S., Jalloh, M. F., … Chowell, G., et al. (2021). A review and agenda for integrated disease models including social and behavioural factors. *Nature human behaviour*, *5*(7), 834–846. doi:10.1038/s41562-021-01136-2

Bradbury-Jones, C., & Isham, L. (2020). The pandemic paradox: The consequences of covid-19 on domestic violence. *Journal of clinical nursing*, *29*, 13–14. doi:10.1111/jocn.15296

Buheji, M., da Costa Cunha, K., Beka, G., Mavric, B., De Souza, Y., da Costa Silva, S. S., … Yein, T. C. (2020). The extent of covid-19 pandemic socio-economic impact on global poverty. a global integrative multidisciplinary review. *American Journal of Economics*, *10*(4), 213–224. doi:10.5923/j.economics.20201004.02

Campolongo, F., Saltelli, A., & Cariboni, J. (2011). From screening to quantitative sensitivity analysis. a unified approach. *Computer Physics Communications*, *182*(4), 978–988. doi:10.1016/j.cpc.2010.12.039

Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini Jr, I. M. (2010). Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS computational biology*, *6*(1), e1000656. doi:10.1371/journal.pcbi.1000656

Chapuis, K., & Taillandier, P. (2019, September 27). A brief review of synthetic population generation practices in agent-based social simulation. Retrieved from https://www.researchgate.net/publication/335601121_A_brief_review_of_synthetic_population_generation_practices_in_agent-based_social_simulation

Dekker, M. M., Coffeng, L. E., Pijpers, F. P., Panja, D., & de Vlas, S. J. (2022). Reducing societal impacts of sars-cov-2 interventions through subnational implementation. *medRxiv*. doi:10.1101/2022.03.31.22273222. eprint: https://www.medrxiv.org/content/early/2022/03/31/2022.03.31.22273222.full.pdf

Dodds, W. (2019). Disease now and potential future pandemics. In *The world's worst problems* (pp. 31–44). doi:10.1007/978-3-030-30410-2_4

Edeling, W., Arabnejad, H., Sinclair, R., Suleimenova, D., Gopalakrishnan, K., Bosak, B., … Coveney, P. V. (2021). The impact of uncertainty on predictions of the covidsim epidemiological code. *Nature Computational Science*, *1*(2), 128–135. doi:10.1038/s43588-021-00028-9

Eisner, M., & Nivette, A. (2020). *Violence and the pandemic: Urgent questions for research*. Harry Frank Guggenheim Foundation. Retrieved April 6, 2022, from https://www.hfg.org/hfg_reports/violence-and-the-pandemic-urgent-questions-for-research/

Epstein, J. M. (2009). Modelling to contain pandemics. *Nature*, *460*(7256), 687–687. doi:10.1038/460687a

Fefferman, N. H., O'Neil, E. A., & Naumova, E. N. (2005). Confidentiality and confidence: Is data aggregation a means to achieve both? *Journal of public health policy*, *26*(4), 430–449. doi:10.1057/palgrave.jphp.3200029

Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*, *442*(7101), 448–452. doi:10.1038/nature04795

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., … Cuomo-Dannenburg, G., et al. (2020). Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. doi:10.25561/77482

Fournier, N., Christofa, E., Akkinepally, A. P., & Azevedo, C. L. (2021). Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, *48*(2), 1061–1087. doi:10.1007/s11116-020-10090-3

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An iterative approach for generating statistically realistic populations of households. *PloS one*, *5*(1), e8828. doi:10.1371/journal.pone.0008828

Guo, J. Y., & Bhat, C. R. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, *2014*(1), 92–101. doi:10.3141/2014-12

Haas, L. E., de Lange, D. W., van Dijk, D., & van Delden, J. J. (2020). Should we deny icu admission to the elderly? ethical considerations in times of covid-19. *Critical Care*, *24*(1), 1–3. doi:10.1186/s13054-020-03050-x

Hafezi, M. H., & Habib, M. A. (2014). Synthesizing population for microsimulation-based integrated transport models using atlantic canada micro-data. *Procedia Computer Science*, *37*, 410–415. doi:10.1016/j.procs.2014.08.061

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini Jr, I. M., Cummings, D. A., Lewis, B., … Germann, T. C., et al. (2008). Modeling targeted layered containment of an influenza pandemic in the united states. *Proceedings of the National Academy of Sciences*, *105*(12), 4639–4644. doi:10.1073/pnas.070684910

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. H. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, *15*(1). doi:10.18564/jasss.1909

Heesterbeek, H., Anderson, R. M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., … Funk, S., et al. (2015). Modeling infectious disease dynamics in the complex landscape of global health. *Science*, *347*(6227), aaa4339. doi:10.1126/science.aaa4339

HERoS. (2020, April). Project overview. Retrieved June 7, 2022, from https://www.heros-project.eu/

Hughes, J. M., Wilson, M. E., Pike, B. L., Saylors, K. E., Fair, J. N., LeBreton, M., … Wolfe, N. D. (2010). The origin and prevention of pandemics. *Clinical Infectious Diseases*, *50*(12), 1636–1640. doi:10.1086/652860

Huremović, D. (2019). Brief history of pandemics (pandemics throughout history). In *Psychiatry of pandemics* (pp. 7–35). doi:10.1007/978-3-030-15346-5_2

Huynh, N., Barthelemy, J., & Perez, P. (2016). A heuristic combinatorial optimisation approach to synthesising a population for agent based modelling purposes. *Journal of Artificial Societies and Social Simulation*, *19*(4), 11. doi:10.18564/jasss.3198

International Organization for Standardization. (1994). Iso 5725-6. Retrieved November 4, 2022, from https://www.iso.org/obp/ui/#iso:std:iso:5725:-6:en

Kaye, A. D., Okeagu, C. N., Pham, A. D., Silva, R. A., Hurley, J. J., Arron, B. L., … Gamble, J. W., et al. (2021). Economic impact of covid-19 pandemic on healthcare facilities and systems: International perspectives. *Best Practice & Research Clinical Anaesthesiology*, *35*(3), 293–306. doi:10.1016/j.bpa.2020.11.009

Knight, G. M., Dharan, N. J., Fox, G. J., Stennis, N., Zwerling, A., Khurana, R., & Dowdy, D. W. (2016). Bridging the gap between evidence and policy for infectious diseases: How models can aid public health decision-making. *International journal of infectious diseases*, *42*, 17–23. doi:10.1016/j.ijid.2015.10.024

Kretzschmar, M. (2020). Disease modeling for public health: Added value, challenges, and institutional constraints. *Journal of public health policy*, *41*(1), 39–51. doi:10.1057/s41271-019-00206-0

Kucherenko, S., Tarantola, S., & Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer physics communications*, *183*(4), 937–946. doi:10.1016/j.cpc.2011.12.020

Lei, H., Xu, X., Xiao, S., Wu, X., & Shu, Y. (2020). Household transmission of covid-19-a systematic review and meta-analysis. *Journal of Infection*, *81*(6), 979–997. doi:10.1016/j.jinf.2020.08.033

Lenormand, M., & Deffuant, G. (2012). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *arXiv preprint arXiv:1208.6403*. doi:10.18564/jasss.2319

Li, H., Yu, L., & He, W. (2019). The impact of gdpr on global technology development. *Journal of Global Information Technology Management*, *22*(1), 1–6. doi:10.1080/1097198X.2019.1569186

Li, J., Xiang, T., & He, L. (2021). Modeling epidemic spread in transportation networks: A review. *Journal of Traffic and Transportation Engineering (English Edition)*, *8*(2), 139–152. doi:10.1016/j.jtte.2020.10.003

Li, W., Zhang, B., Lu, J., Liu, S., Chang, Z., Peng, C., … Tao, K., et al. (2020). Characteristics of household transmission of covid-19. *Clinical Infectious Diseases*, *71*(8), 1943–1946. doi:10.1093/cid/ciaa450

Longini Jr, I. M., Halloran, M. E., Nizam, A., Yang, Y., Xu, S., Burke, D. S., … Epstein, J. M. (2007). Containing a large bioterrorist smallpox attack: A computer simulation approach. *International Journal of Infectious Diseases*, *11*(2), 98–108. doi:10.1016/j.ijid.2006.03.002

Lopez Bernal, J., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., … Ramsay, M. (2021). Effectiveness of covid-19 vaccines against the b.1.617.2 (delta) variant. *New England Journal of Medicine*, *385*(7), 585–594. doi:10.1056/NEJMoa2108891

Lu, Y., Jiao, Y., Graham, D. J., Wu, Y., Wang, J., Menis, M., … Izurieta, H. S. (2021). Risk Factors for COVID-19 Deaths Among Elderly Nursing Home Medicare Beneficiaries in the Prevaccine Period. *The Journal of Infectious Diseases*, *225*(4), 567–577. doi:10.1093/infdis/jiab515. eprint: https://academic.oup.com/jid/article-pdf/225/4/567/42508705/jiab515.pdf

Ma, L., & Srinivasan, S. (2015). Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, *30*(2), 135–150. doi:10.1111/mice.12085

Mercatelli, D., Holding, A. N., & Giorgi, F. M. (2021). Web tools to fight pandemics: The covid-19 experience. *Briefings in bioinformatics*, *22*(2), 690–700. doi:10.1093/bib/bbaa261

Metcalf, C. J. E., Edmunds, W., & Lessler, J. (2015). Six challenges in modelling for public health policy. *Epidemics*, *10*, 93–96. doi:10.1016/j.epidem.2014.08.008

Mueller, K., & Axhausen, K. W. (2011, September). *Hierarchical IPF: Generating a synthetic population for switzerland* (No. ersa11p305). European Regional Science Association. Publication Title: ERSA conference papers. Retrieved March 8, 2022, from https://www.researchgate.net/publication/254457473_Hierarchical_IPF_Generating_a_synthetic_population_for_Switzerland

Municipality of The Hague. (2022, October). Den haag in cijfers. Retrieved October 31, 2022, from https://denhaag.incijfers.nl/home

Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S.-m., Hayashi, K., … Akhmetzhanov, A. R., et al. (2020). Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). *International journal of infectious diseases*, *94*, 154–155. doi:10.1016/j.ijid.2020.03.020

Nossent, J., Elsen, P., & Bauwens, W. (2011). Sobol' sensitivity analysis of a complex environmental model. *Environmental Modelling Software*, *26*(12), 1515–1525. doi:10.1016/j.envsoft.2011.08.010

Reiter, J. P., & Kinney, S. K. (2011). Commentary: Sharing confidential data for research purposes: A primer. *Epidemiology*, *22*(5), 632–635. doi:10.1097/EDE.0b013e318225c44b

Rijksoverheid. (2022, August). Intensive care-opnames. Retrieved August 9, 2022, from https://coronadashboard.rijksoverheid.nl/landelijk/intensive-care-opnames

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (2006). Geocoding in cancer research: A review. *American journal of preventive medicine*, *30*(2), S16–S24. doi:10.1016/j.amepre.2005.09.011

Sahoo, B. K., & Sapra, B. K. (2020). A data driven epidemic model to analyse the lockdown effect and predict the course of covid-19 progress in india. *Chaos, Solitons & Fractals*, *139*, 110034. doi:10.1016/j.chaos.2020.110034

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. doi:10.1016/S0010-4655(02)00280-1

Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., … Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling Software*, *114*, 29–39. doi:10.1016/j.envsoft.2019.01.012

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., … Tarantola, S. (2008). *Global sensitivity analysis: The primer*. doi:10.1002/9780470725184

Siegenfeld, A. F., Taleb, N. N., & Bar-Yam, Y. (2020). What models can and cannot tell us about covid-19. *Proceedings of the National Academy of Sciences*, *117*(28), 16092–16095. doi:10.1073/pnas.2011542117

Siqueira, C. A. d. S., Freitas, Y. N. L. d., Cancela, M. d. C., Carvalho, M., Oliveras-Fabregas, A., & de Souza, D. L. B. (2020). The effect of lockdown on the outcomes of covid-19 in spain: An ecological study. *Plos one*, *15*(7), e0236779. doi:10.1371/journal.pone.0236779

Sirenko, M., Yap, J. R., Sarva, S., Verbraeck, A., & Comes, T. (2020, October 14). D2.1 – agent-based model & scenario analyses. Retrieved July 3, 2022, from https://www.heros-project.eu/output/deliverables/

Sobol′, I. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, *55*(1), 271–280. The Second IMACS Seminar on Monte Carlo Methods. doi:10.1016/S0378-4754(00)00270-6

Statistics Netherlands. (2019, May 1). Average age of first-time mothers up to 29.9 years. Retrieved October 31, 2022, from https://www.cbs.nl/en-gb/news/2019/19/average-age-of-first-time-mothers-up-to-29-9-years

Statistics Netherlands. (2022a). Inkomensgroepen. Retrieved July 18, 2022, from https://www.cbs.nl/nl-nl/dossier/dossier-verstedelijking/waar-zijn-de-inkomens-het-hoogst-/inkomensgroepen#:~:text=De%5C%20hoogste%5C%20inkomens%5C%20vormen%5C%20de,twee%5C%20bovenste%5C%2010%5C%25%5C%2Dgroepen.

Statistics Netherlands. (2022b, February). Microdata: Conducting your own research. Retrieved April 6, 2022, from https://www.cbs.nl/en-gb/our-services/customised-services-microdata/microdata-conducting-your-own-research

Statistics Netherlands. (2022c, February). Open data. Retrieved April 6, 2022, from https://www.cbs.nl/en-gb/our-services/open-data

Steinmann, P., Wang, J. R., van Voorn, G. A., & Kwakkel, J. H. (2020). Don't try to predict covid-19. if you must, use deep uncertainty methods. *Review of Artificial Societies and Social Simulation*, *17*. Retrieved from https://rofasss.org/2020/04/17/deep-uncertainty/

Sun, L., Erath, A., & Cai, M. (2018). A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, *114*, 199–212. doi:10.1016/j.trb.2018.06.002

Ten Broeke, G., Van Voorn, G., & Ligtenberg, A. (2016). Which sensitivity analysis method should i use for my agent-based model? *Journal of Artificial Societies and Social Simulation*, *19*(1), 5. doi:10.18564/jasss.2857

Thiele, J. C., Kurth, W., & Grimm, V. (2014). Facilitating parameter estimation and sensitivity analysis of agent-based models: A cookbook using netlogo and 'r'. *Journal of Artificial Societies and Social Simulation*, *17*(3), 11. doi:10.18564/jasss.2503

United Nations. (n.d.). Health - United Nations Sustainable Development. Retrieved October 5, 2022, from https://www.un.org/sustainabledevelopment/health/

Van Kerkhove, M. D., & Ferguson, N. M. (2012). Epidemic and intervention modelling: A scientific rationale for policy decisions? lessons from the 2009 influenza pandemic. *Bulletin of the World Health Organization*, *90*(4), 306–310. doi:10.2471/BLT.11.097949

Venkatramanan, S., Lewis, B., Chen, J., Higdon, D., Vullikanti, A., & Marathe, M. (2018). Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, *22*, 43–49. doi:10.1016/j.epidem.2017.02.010

WHO. (2022, October 26). WHO coronavirus (COVID-19) dashboard. Retrieved October 27, 2022, from https://covid19.who.int

Wickramasinghe, B. N. (2019). Application independent heuristic data merging methodology for sample-free agent population synthesis. *Journal of Artificial Societies and Social Simulation*, *22*(1). doi:10.18564/jasss.3844

Yameogo, B. F., Gastineau, P., Hankach, P., & Vandanjon, P.-O. (2021). Comparing methods for generating a two-layered synthetic population. *Transportation Research Record*, *2675*(1), 136–147. Place: Thousand Oaks Publisher: Sage Publications Inc WOS:000616951200011. doi:10.1177/0361198120964734

Ye, P.-j., Wang, X., Chen, C., Lin, Y.-t., & Wang, F.-y. (2016). Hybrid agent modeling in population simulation: Current approaches and future directions. *Journal of Artificial Societies and Social Simulation*, *19*(1), 12. doi:10.18564/jasss.2849

Ye, P., Hu, X., Yuan, Y., & Wang, F.-Y. (2017). Population synthesis based on joint distribution inference without disaggregate samples. *Journal of Artificial Societies and Social Simulation*, *20*(4). doi:10.18564/jasss.3533

Ye, X., Konduri, K., Pendyala, R. M., Sana, B., & Waddell, P. (2009, January 11). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th annual meeting of the transportation research board, washington, dc*. Retrieved from https://www.researchgate.net/profile/Karthik-Konduri/publication/228963837_Methodology_to_

match_distributions_of_both_household_and_person_attributes_in_generation_of_synthetic_populations/links/0912f509c17afbbb00000000/Methodology-to-match-distributions-of-both-household-and-person-attributes-in-generation-of-synthetic-populations.pdf

Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in medicine*, *2014*. doi:10.1155/2014/567049

A

Background Information

A.1. HERoS model

The HERoS model consists of agents performing activities in certain locations, guided by their personal schedules. The behavior of these agents can be altered through policies. Furthermore, the agents have an epidemiological state, indicating their relation to the epidemic disease that is under investigation. Structurally speaking, the HERoS model can be split up into 5 sub-models: Citizens, Locations, Activities, Epidemiological states, and Policies. These submodels and their relations are depicted in figure A.1.

A.1.1. Citizens

The agents in the HERoS model are the digital representation of citizens in the city of The Hague. These agents are synthesized through an open data algorithm which provides them with age, social role, family role, income category, and household structure. This algorithm is discussed in more detail in section 3.5. Alternatively, the agents in the HERoS model can be constructed from microdata. As mentioned throughout, this research focuses on the comparison between the two data sources. An agent has a set of attributes that define the behavior of an agent. These attributes are interdependent. For example, in the synthetic population, a 40-year-old agent is assigned the social role of Worker. Because this agent is a Worker, the agent gets assigned a workplace_id, which determines where this agent goes to work during the week.

A.1.2. Locations

The locations within The Hague are obtained through the use of Open Street Maps (OSM). OSM data consists of the location and points of interest within buildings. These points of interest have an amenity tag, which indicates the location type. Each location houses one or more sub-locations (e.g. a regular house has one sub-location, whereas an apartment building contains multiple (sub-locations)). Based on the location type and number of sub-locations, the surface area of a location is determined. The surface area of a location plays an important role, due to the fact that mixing of the crowd and spatial contacts are the primary sources of infection (see section A.1.4 for more information).

Table A.1 shows an overview of all the locations in the HERoS model. The locations are categorized in various categories and assigned a surface area. At the start of a day, an agent starts at their home and commutes to a location based on their upcoming activity. For instance, if an agent is scheduled to work, the agent commutes to their workplace and stays there for the allocated time. If the next activity is to go grocery shopping, the agent moves to the supermarket closest to their home. At the end of a day, the agents move to their home again. The allocation of agents to their respective home and work location is further explained in 3.7.2 and 3.7.3

A.1.3. Activities

As mentioned before, each agent has a personal schedule. Schedules are composed of the following activities: working, personal care, shopping, social activity, or travel. Based on an agent's social role, the schedules are different. Workers are expected to go to work, children go to school and the elderly

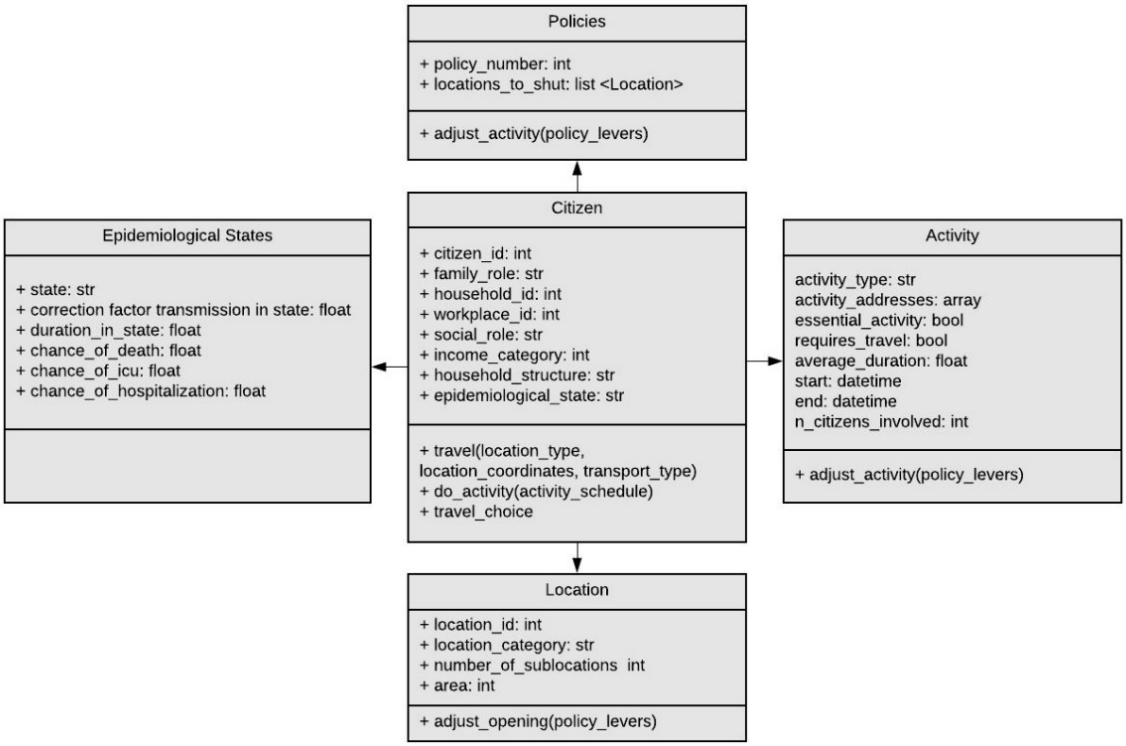

Figure A.1: HERoS UML diagram, from Sirenko, Yap, Sarva, Verbraeck, and Comes (2020)

are performing voluntary work. An activity has two components, the location, and duration of the activity. The duration per activity is either sampled from the distribution for that particular activity (e.g. voluntary work is drawn from a uniform (1,1.5) (hours) distribution) or, in the case of travel, dependent on the length of travel. When an agent has to travel, the agent picks a mode of transport, dependent on the length of travel. If the distance is smaller than 1km, the agent walks, if the distance is between 1-5 km, the agent cycles, and if the distance is more than 5km, the agent drives a car (if the agent is sufficient age). Based on the mode of transport, the travel time is calculated. The difference in the duration of activities ensures variety in the schedules and that not all agents travel between locations at the same time. The location at which an agent performs the activity is dependent on the activity type. Working is always at an agent's defined work location and personal care (e.g. sleeping or having breakfast) is done at home. Shopping works differently, as an agent can choose to go to either a supermarket, pharmacy, retail store, food and beverage store or a mall. The agent will then always choose the location that is closest to home. For social activity, agents can pick to go to a bar, park, or a recreational POI. The location for this activity is randomly chosen within a 2.5 km radius. The schedules from Monday to Thursday are identical, but they are different on Friday, Saturday, and Sunday. This is done to account for shifts in people's schedules on Friday and on the weekend by not working, performing more social activities, and staying relatively more at home. When an agent becomes infected, their schedule changes based on the severity of the infection. Symptomatic agents stay at home more, whereas an agent in the Hospital or ICU relocates their schedule to the hospital until either recovered or dead. The activities of agents can also be altered through policy. Should a lockdown be implemented through a policy, any location that is not a supermarket or pharmacy becomes unavailable and agents choose to stay at home instead.

A.1.4. Epidemiological model

Founded in scientific literature, the HERoS model employs a modified Susceptible-Exposed-Infected-Recovered (SEIR) model to mimic the effect of COVID-19 on the population. This SEIR model consists of two different parts, the transmission of the disease and the stages of infection when an agent gets infected.

Table A.1: Locations in the HERoS model

Location category	Location type	Sublocation Area (m^2)	No. of locations
Housing	Accommodation	100	84348
Occupation	Workplace	30	53458
Education	University	100	7
	College	100	13
	Secondary school	75	68
	Primary school	75	159
	Kindergarten	50	114
Essential Services	Hospital	75	13
	Fire station	100	6
	Police station	30	15
	Healthcare	100	277
	Pharmacy	75	63
	Supermarket	200	133
Shopping	Mall	100	9
	Retail	100	2014
	Food and Beverages	100	511
Social activity locations	Bars and restaurants	200	1652
	Recreation	250	244
	Parks	300000	21
	Religion	250	33

Transmission

The transmission of COVID-19 can happen through multiple pathways, but in the HERoS model, only Peer-to-Peer (P2P) transmission is considered. That is, if an agent is in the proximity of another agent that is infected, there is a chance that transmission of the disease occurs. The agent is then said to move from the Susceptible to the Exposed phase. The chance that this happens is dependent on four factors: the duration of exposure, the area of the location the agents are in, the infected agent's disease stage, and the number of infected people in that area. The formula for this is as follows:

$$chance\ of\ exposure = \sum_{n=1}^{N} \frac{C_f}{A} * C_s * t$$

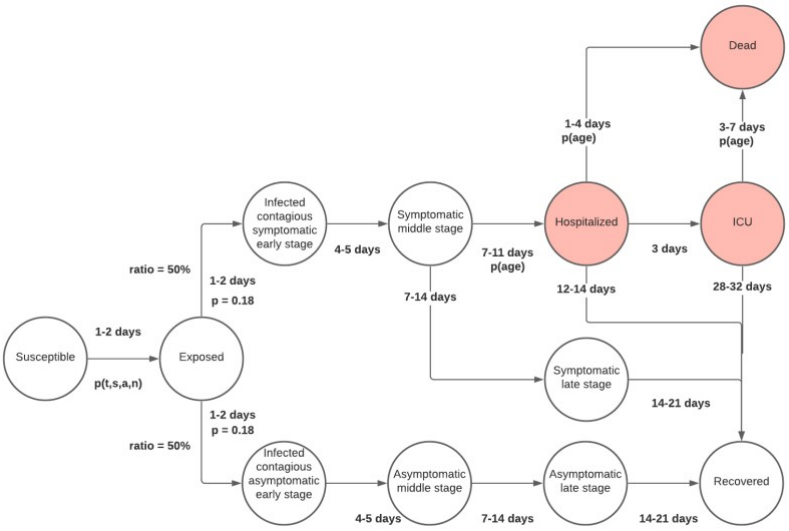

Figure A.2: Epidemiological state diagram

where n is the number of agents, C_f is the contagious factor, A is the shared area, C_s is the correction factor for the contagiousness in state s, and t is the duration of exposure. The contagious factor is the minimum area at which a one-hour exposure guarantees an agent to move to the exposed phase. If a shared area is larger, the chance of exposure should be lower as it is less likely that people have contact with infected people. The correction factor is there to account for the difference between contagiousness in different phases of the disease. A person is most contagious in the middle stage of the disease and least contagious in the final stages. Furthermore, the correction factor is dependent on whether or not a person is symptomatic or asymptomatic. An asymptomatic person is less likely to infect other people and therefore has a lower value for the correction factor.

Epidemiological States

When an agent is exposed to the disease, the epidemiological state of an agent changes from susceptible to exposed. Every agent starts at the beginning of the simulation as either susceptible or infected. Figure A.2 shows the various states and flow between them. When an agent is exposed, the agent can become either symptomatic, asymptomatic, or susceptible again. Following Ferguson et al., 2020 and Nishiura et al., 2020, up to 50% of the population can be asymptomatic. Hence, the split between symptomatic and asymptomatic is 50/50. After getting infected, an agent moves through the early, middle, and late stages of the disease. Since asymptomatic people are not burdened by the disease, it is assumed that they do not end up in the hospital or the ICU. Symptomatic infected agents can end up in the hospital once they reach the symptomatic middle stage. The probability of this happening is based on their age. Older people are more likely to suffer heavy symptoms and end up in the hospital due to Covid-19 (Ferguson et al., 2020). Once an agent is in the hospital, the agent can either recover, move to the ICU or die. The logic here is the same as mentioned earlier and is based on age. Once an agent has moved through the various disease stages, the final stage is either Recovered or Dead. A recovered agent is considered to be immune from the disease and cannot be infected or exposed again. The same applies to a Dead agent, because this agent is, well, dead.

A.1.5. Policies

The final sub-model in the HERoS model is the Policies sub-model. Policies allow the user of the model to implement policies that affect the behavior of citizens in the model, which in its turn leads to different macroscopic outcomes. A common non-pharmaceutical strategy to cope with pandemic spread is lockdowns. Though varying in degree of severity, lockdowns are an effective method to combat disease spread (Sahoo & Sapra, 2020; Siqueira et al., 2020). The HERoS model simulates a lockdown by making only a few select locations available for a visit. These are the essential services (Hospitals, Fire stations, Police stations, Healthcare locations, Pharmacies, and Supermarkets), and parks. By altering their schedules, agents will stay at home more often.

B

Supplementary figures

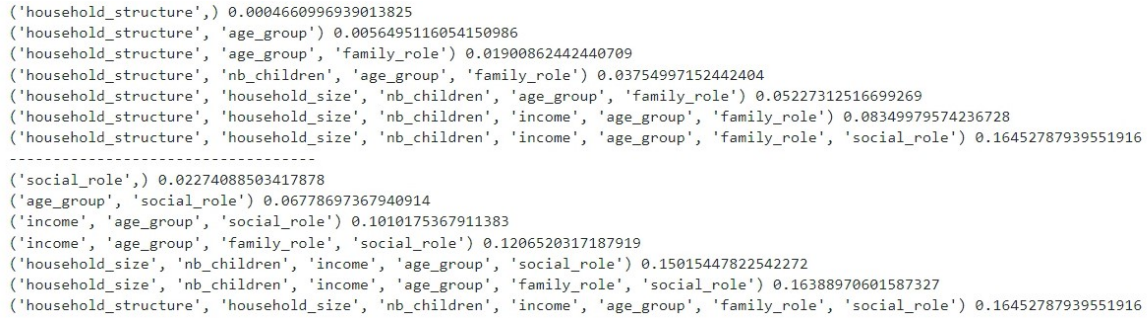

Figure B.1: Results of Freeman-Tukey statistic for individuals0, the top is the best fit per combination, the bottom is the worst fit per combination

```
('household_structure',) 0.0004624946881545793
('household_structure', 'age_group') 0.005626745014893123
('household_structure', 'age_group', 'family_role') 0.018875857069037747
('household_structure', 'nb_children', 'age_group', 'family_role') 0.037243826669875475
('household_structure', 'household_size', 'nb_children', 'age_group', 'family_role') 0.05196824272430777
('household_structure', 'household_size', 'nb_children', 'income', 'age_group', 'family_role') 0.08339174381203164
('household_structure', 'household_size', 'nb_children', 'income', 'age_group', 'family_role', 'social_role') 0.1643110987026201
-----------------------------------
('social_role',) 0.02263484805765482
('age_group', 'social_role') 0.06768258452092489
('income', 'age_group', 'social_role') 0.10069550348197932
('income', 'age_group', 'family_role', 'social_role') 0.12026477331499984
('household_size', 'nb_children', 'income', 'age_group', 'social_role') 0.1499036970673091
('household_size', 'nb_children', 'income', 'age_group', 'family_role', 'social_role') 0.1636735038710627
('household_structure', 'household_size', 'nb_children', 'income', 'age_group', 'family_role', 'social_role') 0.1643110987026201
```

Figure B.2: Results of Freeman-Tukey statistic for individuals8, the top is the best fit per combination, the bottom is the worst fit per combination

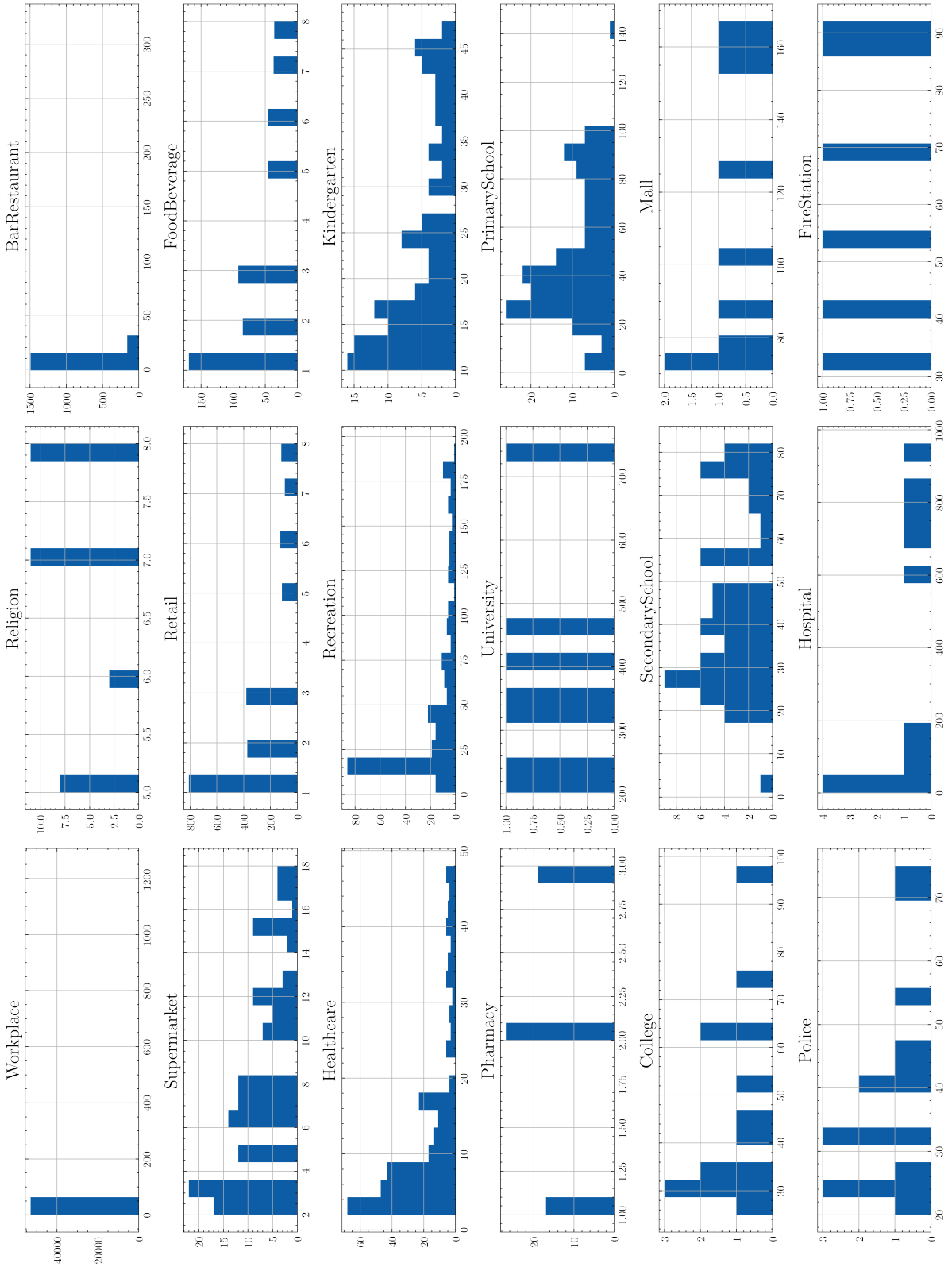

Figure B.3: Number of employees per location type

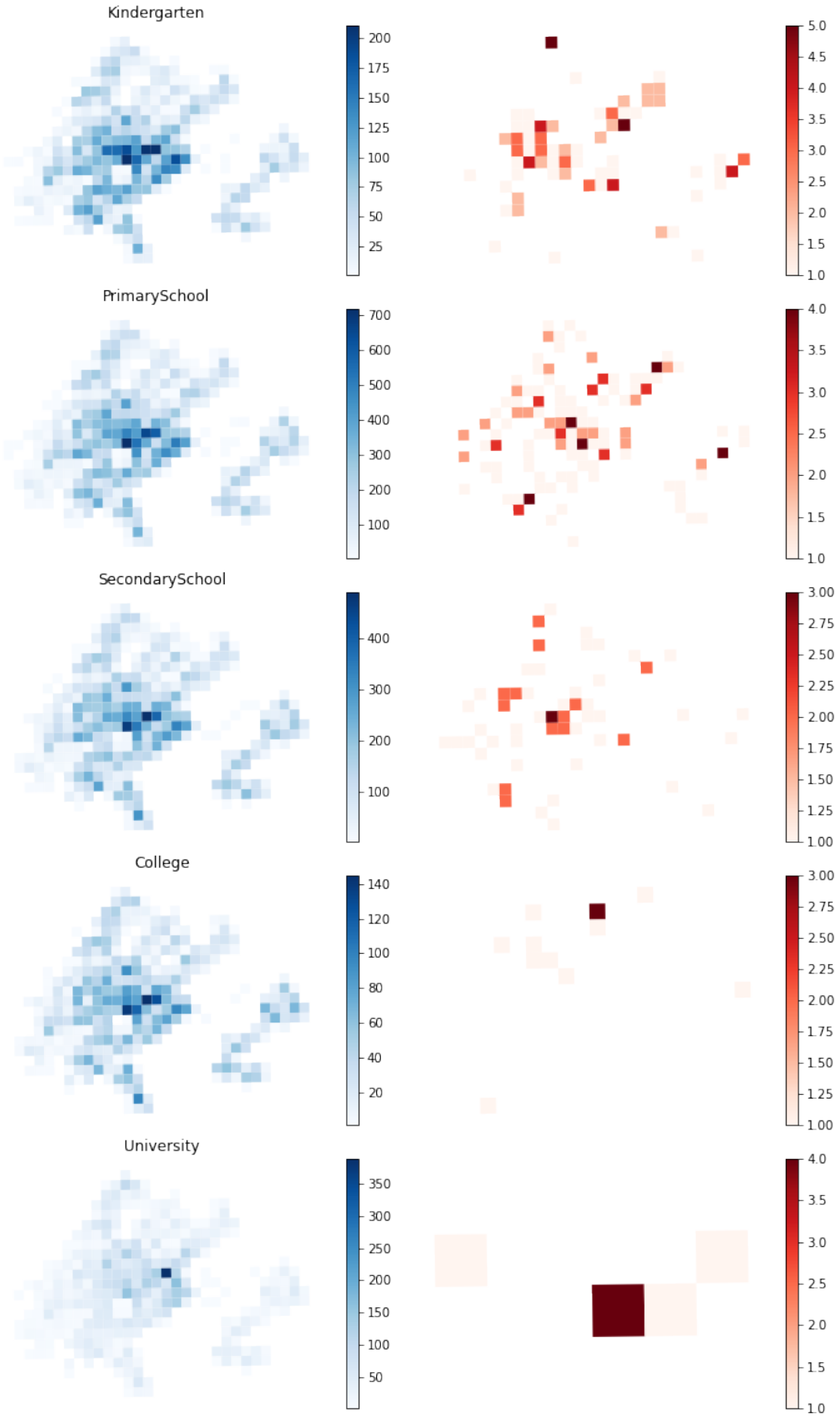

Figure B.4: Heatmap of students and student locations

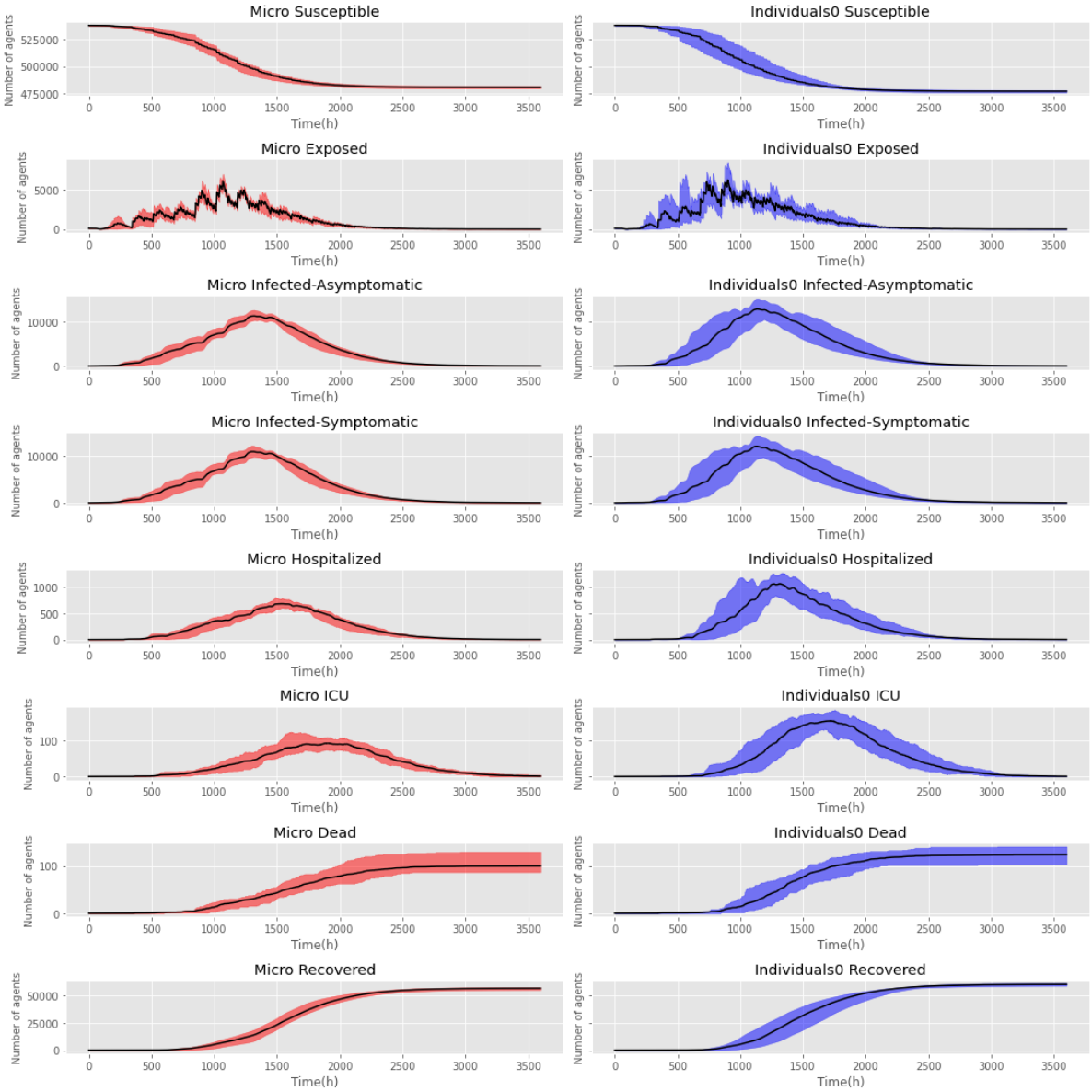

Figure B.5: Model outcomes for synthetic and microdata, including bandwidth

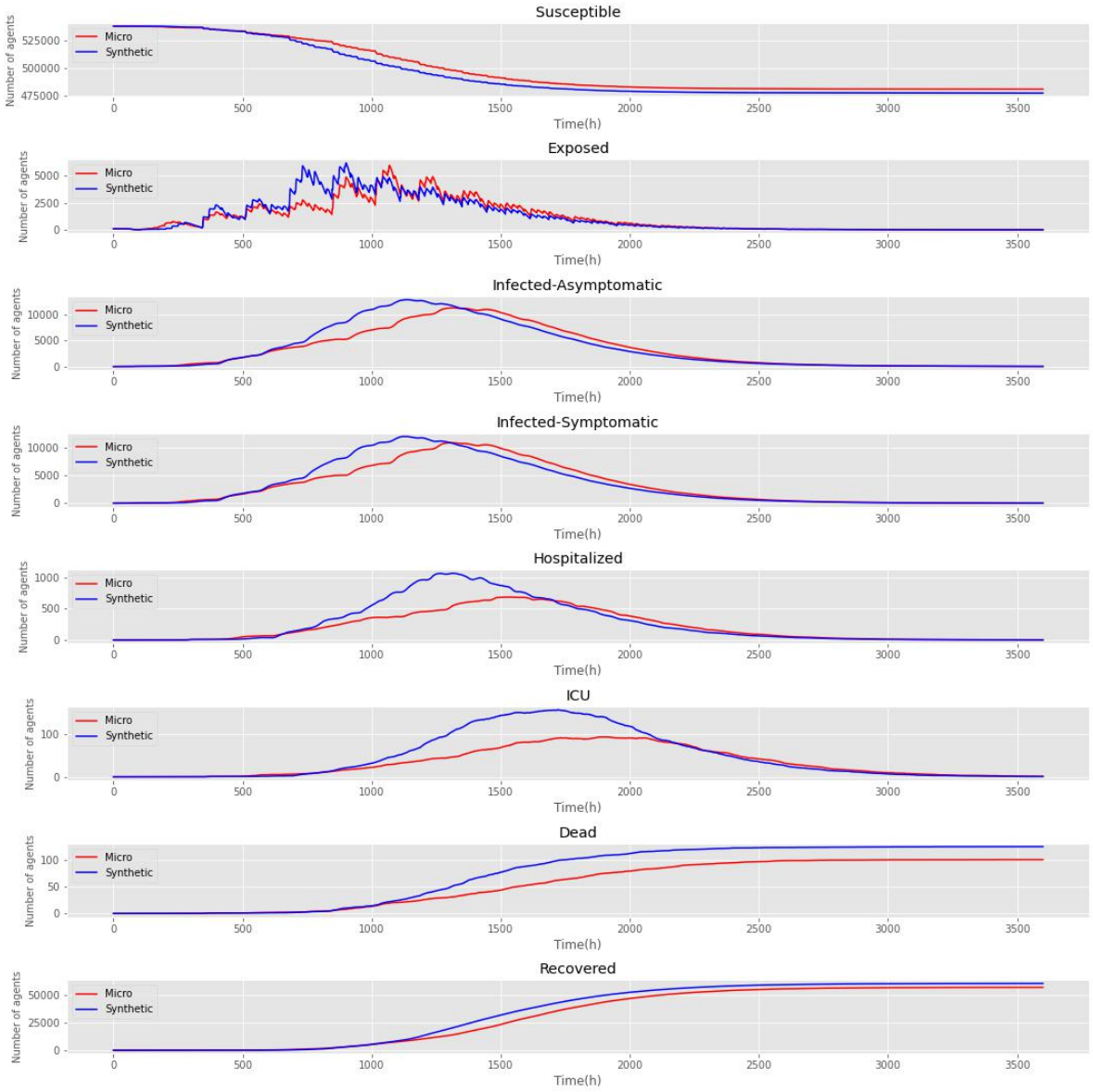

Figure B.6: Model outcomes for synthetic and microdata

C

Data variables

Table C.1: Microdata categories in PLHH variable

Category	Variable name	Variable translation
1	Thuiswonend kind	Child living at home
2	Alleenstaande	Single person
3	Partner in niet-gehuwd paar zonder kinderen	Childless, unmarried partner
4	Partner in gehuwd paar zonder kinderen	Childless, married partner
5	Partner in niet-gehuwd paar met kinderen	Unmarried partner with children
6	Partner in gehuwd paar met kinderen	Married partner with children
7	Ouder in eenouderhuishouden	Parter in single-parent household
8	Referentiepersoon in overig huishouden	Reference person in other household
9	Overig lid in huishouden	Member of other household
10	Lid van institutioneel huishouden	Member of institutional household

Table C.2: HERoS categories in family_role variable

Category	Name
1	Single person
2	head of household w/o children
3	partner in household w/o children
4	Household w/ children head
5	Household w/ children partner
6	Single parent
7	Child

Table C.3: Microdata categories in TYPHH variable

Category	Variable name	Variable translation
1	Eenpersoonshuishouden	Single-person household
2	Niet-gehuwd paar zonder kinderen	Umarried couple without children
3	Gehuwd paar zonder kinderen	Married couple without children
4	Niet-gehuwd paar met kinderen	Unmarried couple with children
5	Gehuwd paar met kinderen	Married couple with children
6	Eenouderhuishouden	Single-parent household
7	Overig huishouden	Other household
8	Institutioneel huishouden	Institutional household

Table C.4: HERoS categories in household_structure variable

Category	Name
1	Single person household
2	Household w/o children
3	Household with children
4	Single person household

Table C.5: Microdata categories in SECM variable

Category	Variable name	Variable translation
11	Werknemer	Employee
12	Directeur-Grootaandeelhouder	Director / Major stakeholder
13	Zelfstandig ondernemer	Entrepreneur
14	Overige zelfstandige	Other entrepreneurs
15	Meewerkend gezinslid	Co-working family member
21	Ontvanger werkloosheidsuitkering	Unemployment benefits recipient
22	Ontvanger bijstandsuitkering	Social benefits recipient
23	Ontvanger uitkering sociale voorz. overig	Other social benefits recipient
24	Ontvanger uitkering ziekte/AO	Recipient of benefits due to disease/disability
25	Ontvanger pensioenuitkering	Pensioner
26	Nog niet schoolg./schol./stud. met ink	Not yet going to school/school-going/ student with income
31	Nog niet schoolg./schol./stud. zonder ink	Not yet going to school/school-going/ student w/o income
32	Overig zonder inkomen	Other without income

Table C.6: HERoS categories in social_role variable

Category	Name
1	Infant
2	Kindergarten student
3	Primary school student
4	Secondary school student
5	College student
6	University student
7	Worker
8	Pensioner
9	Unemployed job seeker

D

Assumptions

- Referentiepersoon in overig huishouden can be assigned to the same family role as the Head of a household wo children
- Overig lid in huishouden can be assigned to the same family role as a Single person
- Lid van instiutioneel huishouden can be assigned to the same family role as a Single person
- Other households can be assigned to the same household structure as Household w/o children
- Institutional households can be assigned to the same household structure as Single person household
- Receivers of social benefits are assumed to behave like an unemployed person searching for work
- Receivers of unemployment benefits are assumed to behave like an unemployed person searching for work
- Receivers of benefits that are not able to work (Other and due to disease/disability) are assumed to behave like pensioners
- People in SECM category Other without income can be assigned to Unemployed job seeker
- The number of sub-locations per location is normally distributed, with a left cutoff at 1.
- Essential retail workers are in income groups low and medium, with a distribution of 25 and 75 percent respectively
- Other essential workers are from all income categories, with a distribution of 25/50/25 percent.
- All the people that live in The Hague also work in The Hague.