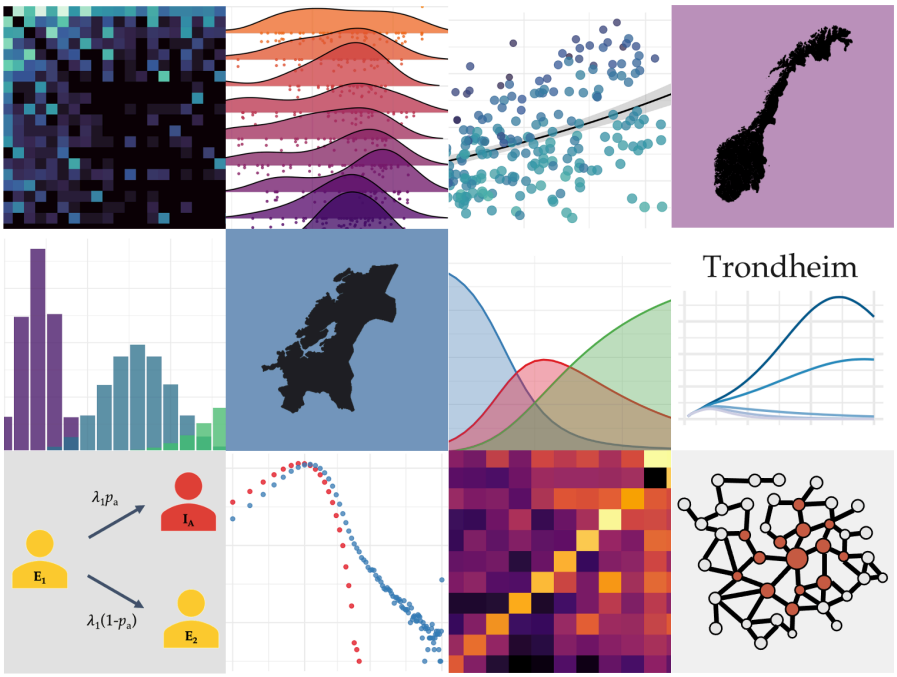
Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Helge Bergo

Agent-Based Modelling of SARS-CoV-2 Spread in a National Municipality Network

Master's thesis in Industrial Chemistry and Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: André Voigt
June 2021



NTNU
Norwegian University of
Science and Technology

Helge Bergo

Agent-Based Modelling of SARS-CoV-2 Spread in a National Municipality Network

Master's thesis in Industrial Chemistry and Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: André Voigt
June 2021

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Abstract

Since early 2020, the SARS-CoV-2 pandemic has upended daily life throughout the world. The virus has claimed 3.7 million lives, and over 174 million cases have been confirmed worldwide. Computational tools like agent-based models can help obtain a better understanding of how a pathogen like SARS-CoV-2 spreads and help both the public and decision-makers return more quickly to normality.

This Master Thesis presents a modelling framework for simulating Covid-19 spread in Norway, written in Python. The model is agent-based and implements a complex, scalable municipality network. The network structure is based on empirical data from Statistics Norway, and commuter data between municipalities is implemented.

Two regions in Norway were simulated. A smaller-scale Trøndelag region, and a complete national model with all municipalities in Norway. It was found that the average reproduction number varied significantly based on model input and population demographics. The most significant factors determining the reproduction number in a municipality was population size, population density, and the fraction of outgoing commuters.

Several model parameters are tunable and can be changed easily to facilitate different forms of analysis. Changes in the different parameters were simulated to evaluate the effect of disease characteristics, population demographics and network structure.

This project lays a foundation for more realistic and large-scale Covid-19 simulations of Norway, as well as a flexible agent-based model for different diseases and regions.

SAMMENDRAG

Helt siden starten av 2020 har SARS-CoV-2 snudd opp ned på hverdagen i hele verden. Viruset har tatt 3,7 millioner liv, og over 174 millioner har fått påvist smitte globalt. Datamodellering i form av agent-baserte modeller kan hjelpe oss å bedre forstå hvordan patogener som SARS-CoV-2 spres, og hjelpe befolkningen og beslutningstakere i å returnere til normalitet.

Denne masteroppgaven presenterer et modelleringsrammeverk for å simulere Covid-19-spredning i Norge, skrevet i Python. Modellen er agent-basert og implementerer et komplekst, skalerbart kommune-nettverk. Nettverksstrukturen er basert på empirisk data fra Statistisk Sentralbyrå, og pendlerdata mellom kommuner er implementert.

To regioner i Norge har blitt simulert. En mindre skala Trøndelag-region, og en komplett nasjonal modell med alle kommuner i Norge. Resultatene viser at det gjennomsnittlige reproduksjonstallet varierer markant basert på modell-verdier oppgitt av brukeren, i tillegg til demografisk befolkningsdata. De viktigste faktorene som avgjør reproduksjonstallet i en kommune var befolkningsstørrelse, befolkningstetthet og andel utgående pendlere.

Flere modell-parametere er regulerbare, og kan lett bli endret for å fasilitere ulike former for analyse. Endringer i ulike parametere var simulert for å evaluere effekten av sykdomstrekk, befolknings-demografi og nettverk-struktur.

Dette prosjektet legger et grunnlag for mer realistiske og stor-skala Covid-19-simuleringer i Norge, i tillegg til en fleksibel agent-baserte modell for ulike sykdommer og regioner.

PREFACE

"Hint, hint, no one has quoted me in their thesis yet."

– Eivind Almaas
Professor, NTNU

The last year can be summarised in one word: *Covid-19*. It is responsible for daily media coverage, countless conversations, and making sure everyone and their grandma knows what a "reproduction number" means. When everything closed down last year, I got the opportunity to take an even deeper dive into this topic and write my specialisation project and master thesis on Covid-19 for the Systems Biology group on NTNU. This not only gave me the chance to explore some of my now favourite topics, but also the ability to brag about my extensive knowledge of Covid-19 simulations at every dinner party for the rest of my life.

This thesis concludes my Master's degree in Chemical Engineering and Biotechnology at the Norwegian University of Science and Technology. It has been 5–6 delightful years brim-full of exciting courses, interesting people and the beautiful student city of Trondheim.

I would first and foremost like to thank my supervisor Eivind Almaas, for guidance and advice on Covid-19 and writing, and the opportunity to be a part of this amazing research group. My sincere gratitude to André Voigt, my co-supervisor and advisor throughout these last six months, for interesting and educational weekly meetings filled with funny derailments, professional and not.

To the reader: Beware! There is an extensive amount of figures in this thesis! But it could have been much, much worse. After stumbling upon a piece of advice from the great Claus O. Wilke, I learned that you should have a maximum of three to six figures per story when writing. I have tried my best to uphold this advice.

Ever since my Introduction to Information Technology course back in 2015, I have had a curiosity for programming and data visualisations, which have finally crystallised into what you are about to read.[1] It has been a joy to learn Python for simulations and take a deep dive back into R for the accompanying visualisations.

There are two particular student organisations I would like to extend my gratitude to after all these years in Trondheim. Revolve NTNU pushed my limits further than ever before, and my concept of work capacity got redefined again and again. Building a racecar while being a student has been highly educational,

[1] or dare I say experience?

and I am forever grateful for all the competent and intelligent students I got to know in Revolve.

The second is the chemistry union at Gløshaugen, Høiskolens Chemikerforening. I have never been bored for a single day in Trondheim, and I cannot count how many cups of coffee, beers, exciting discussions or hilarious parties I have attended with other chemistry students. Thank you!

Finally, a big thanks to my friends and family for always being supportive, especially the last few months. Last but not least, my partner Martine for her support. It has been many evenings with my nose deep in the computer "just checking out" something new. Thank you for being there for me.

HELGE BERGO

TRONDHEIM, JUNE 14TH, 2021

"NTNU: YOUR FIVE WORST YEARS, OR YOUR SIX BEST"

– inspired by KTH proverb

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ABM	Agent-Based Model
Covid-19	Coronavirus Disease 2019
HSØ	Helse Sør-Øst (South Eastern Norway Regional Health Authority)
HUNT	Helseundersøkelsen i Trøndelag (The Trøndelag Health Study)
IBM	Individual-Based Model
NPHI	Norwegian Public Health Institute (FHI)
R-number	Basic Reproduction Number
SARS-CoV-2	Severe Acute Respiratory Syndrome CoronaVirus 2
SEIR	Susceptible - Exposed - Infected - Recovered
SIR	Susceptible - Infected - Recovered
SIS	Susceptible - Infected - Susceptible
SSB	Statistics Norway (Statistisk Sentralbyrå)

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The Covid-19 pandemic has been ongoing for close to one and a half years, and the consequences have been devastating. As of June 2021, 174 million cases have been reported worldwide, and over 3.7 million have died[1]. The societal costs have been disastrous, with significant impacts on the economy, public health, and the daily lives of billions of people[2, 3, 4]. While vaccinations are well underway in many countries, significant restrictions on daily life continue. Mask use, travel restrictions and social distancing are still commonplace. Knowledge and insight into virus characteristics and disease dynamics have never been more needed. Creating models to understand how pathogens like SARS-CoV-2 spread throughout cities, countries, and the entire world is of major importance in assessing the current situation and the possible paths going forward.

Throughout history, the human race has always been susceptible to different pathogens, including virus and bacteria. Infectious diseases have been with us for a long time, and several of them reach epidemic or pandemic potential.[5] Despite decades of interventions and surveillance, the seasonal influenza virus cause epidemics throughout the world every single year, as the evolution of viruses continue[7].The seasonal influenza causes an estimated 610 000 life-years lost and 10 billion dollars in the United States alone[6]. In Norway, it is estimated that influenza kill nearly 1000 people yearly[9]. WHO has long warned about coming pandemics[10], and the fear is that the next pandemic may surpass previous pandemics like the “Spanish Flu” which had a death toll of over 15 million[11], or the “Black Death” which killed 75-200 million people[12]. A more recent example is Influenza A, caused by an H1N1 virus, similarly to the “Spanish Flu”. In 2009, the so-called “Swine Flu” quickly spread from Mexico and the United States to hundreds of countries. It is not clear how many were infected by the virus, but some estimates give the order of several tens of millions of cases[13].

The understanding of pathogens continues to increase for every research paper made, but there is still more to learn. Extensive knowledge of how a virus

like SARS-CoV-2 spreads in a modern and interconnected world remains a significant challenge. Substantial amounts of information is needed to realistically model disease spread. Simple models help to give insight into the processes of a pandemic[14, 15], but to stop the next pandemic, more advanced tools are needed. Information on everything from how individuals behave in their daily activities, to larger patterns like travelling within and between cities, countries and continents might be critical to understand the full picture. Fields like psychology, economy and politics need to be incorporated to create the most realistic models of pandemics possible. There is still a gap between "hard" mathematical models and "softer" psychological models[16, 17]. An example of this behaviour is how individuals in a population could spread disease through daily activities like travelling, commuting to work, or meeting a neighbour in their local store. Investigating such behaviour might lead to insights helpful in getting the current pandemic under control, as well as future pandemics.

This project aims to create a national agent-based model of SARS-CoV-2 spread in Norway by extending the modelling framework created at NTNU in the spring of 2020, to investigate the effect of commuter travelling and population demographics on the disease spread of Covid-19.

This aim is divided into three objectives: The first objective is to explore previously gathered empirical data from Statistics Norway to lay the foundation for a national agent-based model. The second objective is to extend the existing agent-based model into a national scale model by implementing commuter data between municipalities. The third and final objective is to investigate the effect of different parameter values on the modelling framework, to test the model stochasticity and influence of parameters and population demographics on the reproduction number of Covid-19.

CHAPTER 2
THEORY

This chapter describes relevant theory used throughout the project. It is divided into two main parts: The first part presents epidemic modelling and describes three main model categories: compartmental models, network models and agent-based models. The second part describes the statistical theory, going through important descriptive statistics and the probability distributions used throughout this project.

Parts of the following chapter are based on material presented in my project report *"Agent-Based Modelling of SARS-CoV-2 Spread in a Public Transport System"*[18]. An overview of the relevant sections can be found in Appendix A.1.

2.1 Epidemic Modelling

Epidemic modelling describes how infectious diseases spread throughout a population, using mathematical and computational tools[19]. This section is reused from [18], with minor changes.

The use of mathematical models in epidemiology is not a new invention, and they have been in use long before the invention of computers. Daniel Bernoulli investigated the effect of vaccination of smallpox virus using mathematical models already in 1766[19]. Lowell Reed and Wade Hampton Frost developed a mathematical model of disease spread in the 1920s, later to be known as the Reed-Frost model[20]. In 1927, Kermack and McKendrick laid much of the groundwork of modern mathematical epidemiological models in *"A Contribution to the Mathematical Theory of Epidemics"*[21].

During the last few decades, an increase in computing power have made the use of more complex network models and agent-based models more viable. In complex network models, every single individual in a given area is simulated from the bottom up. Epidemic models can be categorised into three main direc-

tions: Compartmental models, network models, and agent-based models, and the rest of this chapter will go through the three categories[14]. For an illustration of the categorisation of the different model types, see Figure 2.1.1.

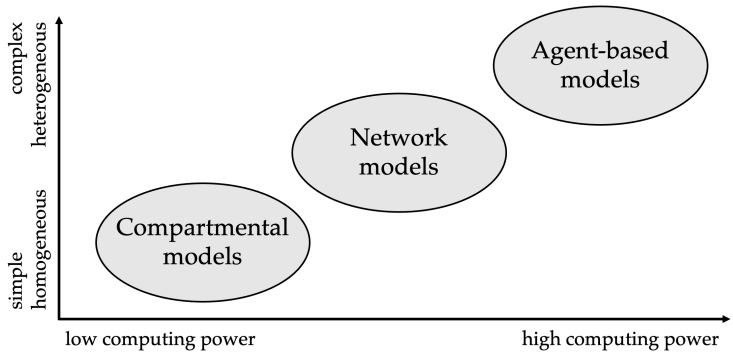


Figure 2.1.1: Epidemic models, categorised after complexity and computing power. One way to categorise epidemic models, where the three different types are arranged after complexity and computing power demand. Adapted from [14].

2.1.1 Compartmental Models

Compartmental models are simplified epidemic models with a focus on the macroscopic processes of an infectious disease. They can give important insight into theoretical aspects of an epidemic, such as epidemic threshold and size, without needing much computing power or high fidelity data.[14]

Most compartmental models categorise a population into different compartments, based on the Reed-Frost model, typically a variant of the SIR model[19, 22]. Here, the population is divided into three different states based on the health status of the individuals modelled[23]. All individuals are either susceptible (S), infected (I), or recovered[1] (R) from a given disease. They can move from one state to the next but only be in one state at a time. An illustration of the process is shown in Figure 2.1.2. Here individuals move from the susceptible to the infected state following the rate of infection parameter β, and from infected to recovered after a certain time with rate parameter γ.

Compartmental models using SIR dynamics are often based on differential equations with parameters β and γ controlling the flow of individuals from one state to the other for each time step, see Equations (2.1.1) to (2.1.3)[22].

[1] Or dead, which is practically the same from the perspective of disease transmission.

Figure 2.1.2: Illustration of the three epidemiological states in the SIR model.. The arrows indicate possible transitions, with β as the rate of infection, and γ determining the recovery rate.

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \qquad S(0) = S_0 \geq 0, \tag{2.1.1}$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \qquad I(0) = I_0 \geq 0, \tag{2.1.2}$$

$$\frac{dR}{dt} = \gamma I, \qquad R(0) = R_0 \geq 0 \tag{2.1.3}$$

Here $S(t)$, $I(t)$ and $R(t)$ will be the number of individuals in the different states at each time step t, for a population with size N. The transmission coefficient β describes the transmission rate between two individuals, and γ is the parameter deciding the length of disease, with $1/\gamma$ being the average infectious period. One assumption used in SIR models is the *homogeneous mixing hypothesis*, where it is assumed that the entire population is mixed, and everyone can, in theory, get the disease from an infected individual at any time.

These differential equations have analytical solutions and are therefore easily calculated. A SIR model with parameters $\beta = 0.5$ and $\gamma = 0.1$ is simulated and plotted in Figure 2.1.3.

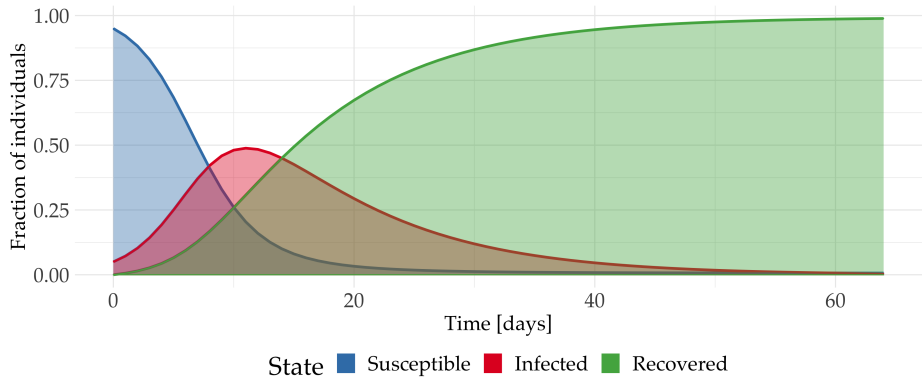


Figure 2.1.3: Example of a SIR model with $\beta = 0.5$ and $\gamma = 0.1$.

While the SIR model makes several simplifying assumptions about a disease, it can still give much insight into the processes driving disease spread in a population. Examples include the epidemic threshold, doubling time, peak infection rate, maximum theoretical infection rate, and the fraction of a population vaccinated to combat a particular disease. In practice, however, more complex variants of the SIR model are often used. The SIS model describes diseases where infected individuals do not acquire immunity after a successful recovery, and the SEIR model gives the possibility of having an exposed period where individuals are infected but not sick[22].

For the coronavirus SARS-CoV-2, while research is still ongoing, evidence suggests a high amount of infected individuals carry the disease while asymptomatic, so models incorporate two different exposed states. An example of this is the meta-population model from the Norwegian Institute of Public Health (NIPH), using a variant of the SEIR model as shown in Figure 2.1.4[24]. In this model, after a patient is exposed (E_1), they have a chance p_a to become asymptomatic (I_A). These infected may spread the disease further, but display very mild or no symptoms. Asymptomatic carriers is one of the reasons Covid-19 has spread so quickly, and this dynamic is therefore critical to include in models.

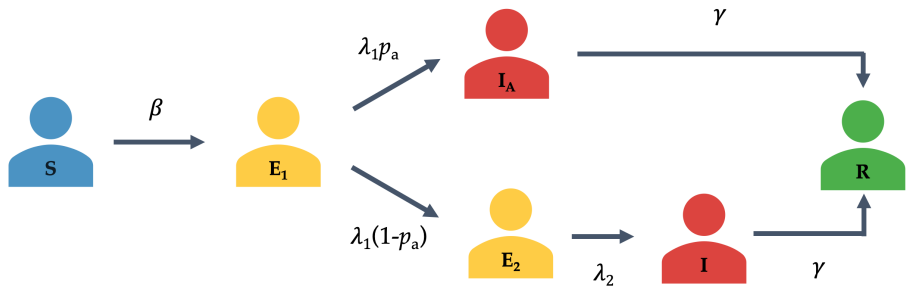


Figure 2.1.4: SEIR type model describing SARS-CoV-2, where exposed individuals have the probability p_a to become infectious and asymptomatic (I_A). Inspired by [24].

Reproduction number

Another key insight coming from the compartmental models is the basic reproduction number R_0. This number estimates how many individuals, on average, are infected by each infected person. In short, how many individuals does the average infected spread the disease to? For a compartmental model with SIR dynamics as described in Equations (2.1.1) to (2.1.3), the basic reproduction number is given by

$$R_0 = \frac{\beta}{\gamma}. \tag{2.1.4}$$

However, the basic reproduction is a theoretical value and often difficult if not impossible to estimate for a given disease. It is therefore usually given as a range, for example as 12-18 for measles[25] and 0.9-2.1 for seasonal influenza[26]. SARS-CoV-2 is estimated to have an R_0 somewhere between 3.3-5.7, without any restrictions[27]. In addition, the basic reproduction number often assumes no immunity in the population and no restrictions on movement and disease spread. This leads to the effective reproduction number R_e in practice, which is calculated for a given time period in a given population. R_e helps measure the effectiveness of different countermeasures to contain a given disease but is often overused or simplified, especially in the media[28].

To summarise, compartmental models focus on the disease transmission on a macroscopic level on a population and usually include several simplifying assumptions, like the homogeneous mixing hypothesis. They are theoretical but are easy to understand and require low computing power. Mathematical models provide key insight into the processes driving a viral disease and should not be disregarded, even though they cannot describe populations in detail.[14]

2.1.2 Network Models

The next category of models used in epidemic modelling is network models. These can be considered an intermediate step between simple compartmental models and more detailed agent-based models.[14] Instead of mathematical equations, populations are described as networks with individuals represented by nodes and contacts between individuals represented by the links. When using networks to represent epidemic systems, network theory can be used to calculate infectivity as a function of node degree. The connectivity of nodes can give valuable information about the epidemic properties.[29]

One of the main limitations of compartmental models is the assumption of homogeneous mixing. The possibility that everyone can meet and infect everyone else in a given population is rarely realistic, and this is where network models come into play. By representing populations through networks, a much more realistic view of contact networks can be given. More importantly, network models can capture heterogeneity during a disease outbreak. Where compartmental models are based on averages and a few parameters determining the properties of every individual, a network model can instead model differences in human behaviour through parameters like node connectivity and position in the network. Network models also incorporate the impact of network topology on the epidemic spread, incorporating properties like small-world and scale-free networks.[14, 29]

An illustration of disease spread on a fictional network is shown in Figure 2.1.5, to visualise how a possible disease might spread on a simple contact network as time goes by. Here it can be seen that new nodes can only be infected through links from other infected nodes, and transmission across the network cannot happen randomly.

Network models have seen more and more use in the past few years, and the

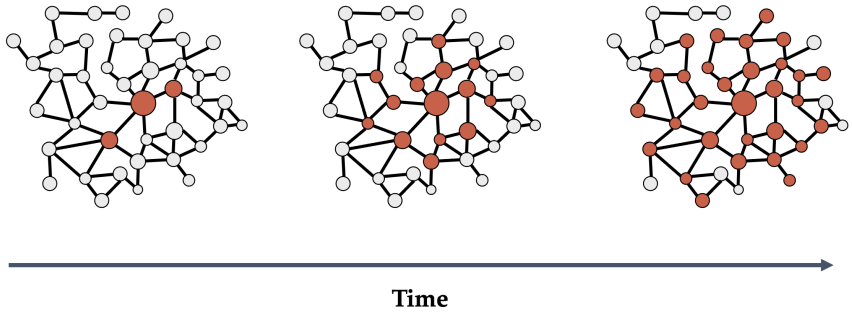


Figure 2.1.5: Illustration of a disease spreading on a simple, fictional network. Nodes coloured red are infected and may spread the disease to the grey, susceptible nodes.

applications in the field of epidemiology are many. Examples include simulating disease outbreaks in realistic urban networks[30], vaccination strategies in weighted networks[31], and cost-effective outbreak detection[32]. The types of networks used in the models can be described in many ways, from simple static networks to adaptive temporal networks that use feedback loops during the simulation, changing both links and weights throughout a simulated pandemic.

The advantages of network models compared to compartmental models are therefore many. They can capture heterogeneous contact patterns between individuals, as well as differences in population structure and topography. However, they are unable to capture the full complexity of factors interacting in a real-life pandemic situation and are lacking in representing daily human activities. Examples include daily commuting, differences in mobility and age, and consequences of interventions during a pandemic. For these levels of details to be possible, we turn to agent-based models, which can be seen as a further extension of the network models.

2.1.3 Agent-Based Models

Agent-based models (ABMs), also often called individual-based models, are complex, bottom-up simulation models that can give a more detailed description of real-life systems. By designing systems from the bottom-up with detailed descriptions of agents in a population, both the heterogeneous and stochastic nature of epidemics can be captured much better than in compartmental and network models. A higher granularity in the data gives rise to interactions between individuals on the micro-scale, leading to the emergence of macro behaviour in the entire system. Agent-based models have seen a large increase in popularity over the last few decades, partly thanks to the availability of more powerful computers and large data sets with demographic and environmental data.[14, 17]

Agents in the model can, in principle, be anything. In the last few years, applications for ABMs have been found in fields ranging from economics, ecology, social science and, of course, epidemic modelling[17, 33]. Examples include rumour spreading on Twitter[34] and simulating the entire public transport system of Zurich[35]. In epidemic modelling, the agents usually represent humans, but there are examples where parts of the environment might act as agents as well[36]. Agents are initialised with different properties like age, gender, occupation and geographic location. They move around in a simulated environment for each simulation step, meeting other agents and interacting with them. An example might be a simulation of a small city, where inhabitants move around in patterns resembling daily commuting, going to the store and visiting family. For each time step, it is logged whom they meet and where. If a set number of individuals are infected at the start of the simulation, and every agent they encounter might be infected by a probability p, realistic disease spread in a community might be simulated.

A central aspect of ABMs is the element of stochastic processes[14]. Daily contact patterns and infection chance between individuals are often simulated as random stochastic variables drawn from a probability distribution. This means that each simulation run will produce different results, and conclusions are usually drawn after averaging several runs. This average gives more realistic results, as random processes often drive real-life human encounters, but comes at the cost of increased computing time, as simulations need to be run dozens or hundreds of times. This stochasticity also means that minor differences in input variables, for instance, the amount of initially infected agents, might lead to a significant difference in output, often termed as the butterfly effect[37]. An example of this effect in real life happened in South Korea, where the now infamous "Patient 31" spread Covid-19 to several clusters, leading to thousands of new cases[38].

As agent-based models have increased in complexity and scope during the last decade, a focus on the challenges of the modelling framework has led to a deeper understanding of the mechanics driving both pandemics and ABMs. Interdependent behaviour leads to agents responding to their environment in the short run, but in the long run, environments respond to the accumulation of agents choices or behaviour[39]. For complex ABMs with both a spatial and temporal dimension, the chosen granularity and level of detail might have considerable implications for the conclusions drawn. Evidence suggests decreasing spatial resolution leads to an increase in the speed and intensity of the epidemic while decreasing temporal resolution does the opposite[40]. In addition, there are issues with turning qualitative information from empirical research studies into quantitative data in an ABM[41]. This is especially important in epidemiological models, where the "human factor" often plays a significant role in the outcome of a pandemic. An example is the effect of differences in the degree of compliance to government policy during a pandemic. It has been estimated that this uncertainty might be lower than the built-in stochastic uncertainty in the models themselves[42], but this does not necessarily make it any easier to

incorporate these softer parameters into the models.

While agent-based models are complex and computationally demanding, they provide valuable information informing policymakers, giving a more detailed representation of reality than simpler network models and compartmental models. As data availability and computational power will continue to increase in the future, there is no reason to believe that ABMs will decrease in popularity and use[14].

2.2 Statistics

Since most agent-based models are driven by random processes and are stochastic, the use of statistics is vital for analysing and describing the results of the models after several runs. This section will go through the most important statistical measures and distributions used throughout this project. The theory presented is based on Walpole's *"Probability & Statistics for Engineers and Scientists"*[43]. This section is obtained from [18], with modifications.

2.2.1 Descriptive Statistics

Descriptive statistics is the process of summarising and describing data sets using different measures and statistics. These commonly fall into measures of centrality and measures of variability.

The first and possibly most used measure of centrality is the *sample mean* of a population, which describes the average of n observations, see Equation (2.2.1).

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2.2.1}$$

The sample mean sees widespread use, but for observations with substantial deviations from the mean, it is affected by extreme outliers. Typical examples are heavy-tailed distributions where one large measurement might skew the mean by a lot. Another centrality measure often used is, therefore, the *median*, which describes the middle value of a sorted dataset or the average of the two middle values for an even-numbered data set. An advantage of the median is that it is not affected by a few outliers if the rest of the data set is relatively normally distributed. Another important measure of centrality is the *mode*, which is the value found most frequently in a given data set.

For the description of measures of variability, the most commonly used for observational data is the *sample standard deviation*, which is a measurement of the variation or dispersion of a data set, see Equation (2.2.2).

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \tag{2.2.2}$$

Other measures of variability include the *range* of a set of values, where the difference between the maximum and minimum value is calculated, and the *kurtosis* and *skewness* of a distribution. These last two are often practical when dealing with non-normally distributed data, to compare a given heavy-tailed distribution to a normal one.

2.2.2 Probability Distributions

Another important tool when analysing and working with data sampled from experiments and observational studies is probability distributions. These can be thought of as the mathematical functions providing a sample space value, given a certain probability for different values. Probability distributions can be categorised into discrete and continuous distributions.

Binomial Distribution

Many experiments and complex real-life processes can be simplified into a binary yes-no response. Is a person infected with a disease or not? Is the output of this function higher than a given threshold? Is a product working or not? For these experiments or measurements, with several independent, random trials with a binary response, discrete probability distributions come into play. An example is the binomial distribution, which models the number of successes in a sample of size n, given a probability p. An example from epidemiological simulating is drawing k neighbours an infected individual transmits a disease to, given the probability of transmission p. A single binomial is called a Bernoulli trial, and the probability of getting x successes after n independent Bernoulli trials can be calculated by the binomial probability function, see Equation (2.2.3).

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, ..., n \tag{2.2.3}$$

An example of three different binomial distributions, with different values for p and n is shown in Figure 2.2.1.

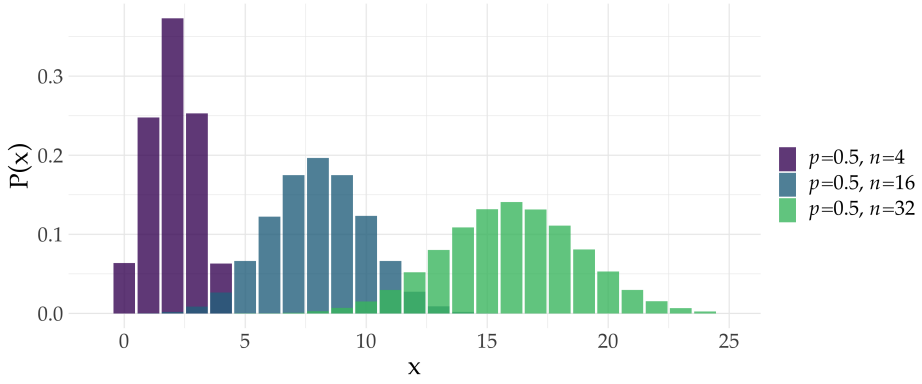


Figure 2.2.1: Binomial distributions with different parameters.

Poisson Distribution

Another practical discrete probability distribution is the Poisson distribution, used for expressing the probability of a given number of events in a given numeric interval. For example, how often a new customer appears in a queuing system or the length of phone calls. A surprisingly high number of everyday processes can be approximated well with a Poisson distribution[23]. It is also easy to work with mathematically and uses only a single fixed parameter λ, equal to both the mean and variance. The probability mass function is shown in Equation (2.2.4).

$$p(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ..., n \tag{2.2.4}$$

Three different Poisson distributions are plotted in Figure 2.2.2, with different λ values, and therefore different means and variances.

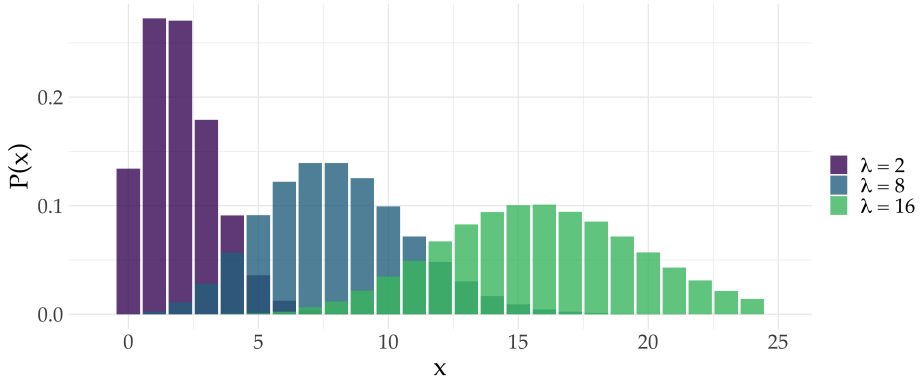


Figure 2.2.2: Poisson distributions with different parameters.

Normal Distribution

While discrete probability distributions are practical and useful for several purposes, many processes are not discrete but produce values on a continuous range instead. The normal distribution, or Gaussian, is the most used probability distribution and is well known for its recognisable bell-curved shape. The probability function is shown in Equation (2.2.5).

$$n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in (-\infty, \infty) \tag{2.2.5}$$

The normal distribution has several important properties. Only two parameters describe it, the mean value μ and variance σ. The mean value is also the median and mode, and the distribution is symmetric around the mean. In addition, for the *standard normal distribution*, where $\mu = 0$ and $\sigma = 1$, the sum of the area under the curve equals 1. A normal distribution well approximates many physical processes as the number of samples increases. A plot of three different normal distributions with different mean values and variances is shown in Figure 2.2.3.

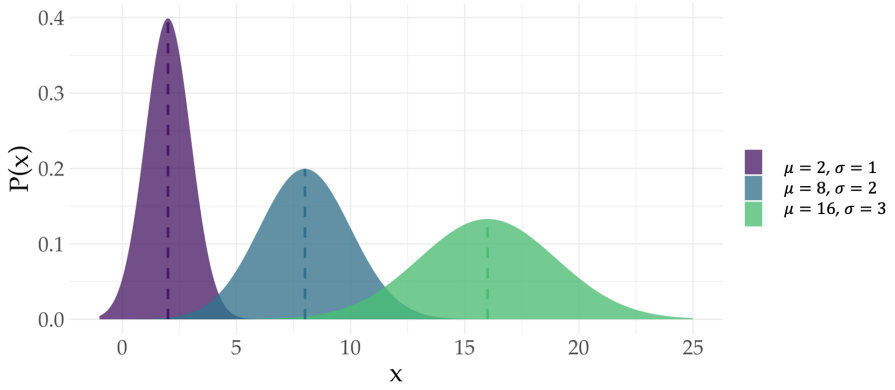


Figure 2.2.3: Normal distribution with different values of μ and σ. The dashed lines show μ.

CHAPTER 3

METHOD

The following chapter describes the methodology and work done in this project. The main bulk of the work consisted of incorporating the municipality network into the already existing agent-based model by implementing data from Statistics Norway (SSB) and running simulations on different input and parameter values.

The chapter is structured into four main parts, starting with the software used in the project. Following is a description of the agent-based model, including network structure, the epidemiologic dynamics driving the disease and intervention measures. The third part describes how the population data was used to integrate the municipality network in the model. Part four describes the modelling framework, how the algorithm works, as well as code structure and default model parameters and output.

3.1 Software

The software used in the project is Python for the model code and R for the data analysis and visualisations. Data were processed in digital labs at HUNT Cloud, Norwegian University of Science and Technology, Trondheim, Norway.

3.1.1 Python

The agent-based model and most supporting functions and scripts are all written in Python[44]. Python is a popular open-source programming language available for most operating systems.

Python is a high-level programming language, meaning it is easy to write expressive and readable code. While it is not one of the fastest languages available, its usability and clear syntax make it an excellent choice for projects and models

like this one. The original model was written in Python, and this was continued for this project.

Most of the model is written in the Python Standard Library, with some extra modules used. A table showing the installed modules and their respective versions is presented in Appendix B.1.

Python can be downloaded from `www.python.org`.

3.1.2 R

For most of the data analysis and visualisations, R was used in RStudio. R is an open-source language used for statistical computing and graphics, and RStudio is an integrated development environment for R.

For the analysis of the data from the agent-based model, the libraries from `tidyverse` were used. These offer a common, underlying design philosophy designed for data science and makes working with large datasets simple and intuitive. For the visualisations, `ggplot2`, part of the `tidyverse`, was used.

These tools, in combination, make for an effective and clean working environment, perfect for exploring and visualising the different types of output data from the model.

A table showing the installed libraries and their respective versions is presented in Appendix B.2.

R can be downloaded from `www.r-project.org`, and RStudio from `www.rstudio.com`.

3.2 The Agent-Based Model

The agent-based model used in this project is an extension of the NTNU Covid-19 model developed by Voigt *et al.* in the spring of 2020. The model framework is described in [45]. In addition, more documentation about the model and the NTNU Covid-19 Modelling Taskforce can be found here: `www.ntnu.edu/biotechnology/ntnu-covid-19`.

As most of the underlying logic and structure of the model builds on the NTNU Covid-19 model, this section will describe the mechanisms of the computational modelling network. Most of it is therefore based on [45].

3.2.1 Network structure

The model is an agent-based (also known as an individual-based) complex network model, consisting of different network structures to simulate demographic data and realistic human behaviour on a municipality level. The model has nine different layers consisting of a varying number of cliques. The nodes (or agents) are created from population demographics and placed into one or several groups in different layers. The groups are designed as k-cliques, where all clique members are in contact with each other and can meet and interact daily. The exception to this is the generic contact layer, more on this later.

Nodes

Each node in the simulation has the following attributes: age, domicile, layer memberships, disease state, and activity. The activity represents the maximum number of daily contacts. Young persons below age 20 and elderly over 80 years follow a normal distribution with parameters μ and σ. The remaining population follows a combination of the normal distribution and a power-law distribution with parameter γ to capture a more significant heterogeneity in the contact patterns.

The formula is shown in Equation (3.2.1), where act_1 represents young and old people, and act_2 the remaining population. The actual parameters used are presented in Table 3.2.2.

$$f(x;\mu,\sigma,\gamma) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, & \text{for } act_1 \\ \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} + x^{\gamma}, & \text{for } act_2 \end{cases} \tag{3.2.1}$$

The activity is set at the start of the simulation by a random draw for each node. Every simulation day, the daily number of contacts is drawn from a uniform distribution ranging from 1 to the maximum number of daily contacts, C_D. The activity also has a maximum hard limit of 100, so any values above this are set to 100.

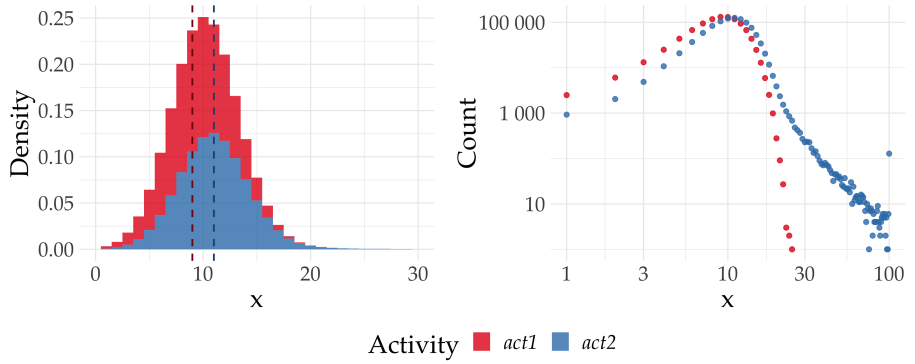


Figure 3.2.1: Distribution of the activity types. A million values drawn for each activity type. The vertical lines mark the median values which is 9 for $act1$ and 11 for $act2$. The right plot shows the number of each value in logarithmic scale.

A million values drawn from each distribution is shown in Figure 3.2.1. As can be seen, the two different functions produce relatively similar distributions, but the second type can create much higher activity values. On the right-hand side, a log-log-plot shows the number of times large values are drawn. For a million values drawn, there will be several with very high values. There are 165 individuals with an activity over 90 in this plot and 638 with an activity over 50.

These individuals have the potential to be what is often termed "super-spreaders" in the media.[1]

Layers

All nodes are present in the household layer and generic contact network, and one other layer. For example, a person of age 32 might be present in a household clique representing its family, one work clique together with its colleagues, and the generic contact network, representing the daily contact pattern.

The assignment of individuals to the layers are described below:

1. **Household:** The household layers consists of separate households with size and age distributions from municipality data.
2. **Day-care:** The number of day-care facilities is based on demographic data. For households with multiple children of day-care age, these children are placed in the same day-care.
3. **Schools:** The school layers are separated into three different layers: primary, secondary and high school. The size of each school is based on demographic data. Class structure is not included. For households with multiple school-age children, these are placed in the same schools for primary and secondary schools, while high-school students are randomly assigned.
4. **Nursing homes:** Both numbers of nursing homes and population sizes are based on demographic data.
5. **Work:** The number of companies and the number of workers are based on demographic data. This layer represents spread between co-workers, so for professions with extensive exposure to the general public, customer contact is represented in the generic contact layer.
6. **Hospital:** Sick persons are removed from their household or nursing home, as well as their work or school layer. The hospital layer also incorporates the possibility to be placed in an intensive care unit (ICU), but this is not modelled as a separate layer.
7. **Generic contact network:** This layer represents the daily contact patterns of a person and uses the activity and daily contact number as described previously.

All the layers except the generic contact network and hospital layer can be seen as static networks created during the model initialisation. However, nodes are inactive or not present in a layer depending on their disease or quarantine

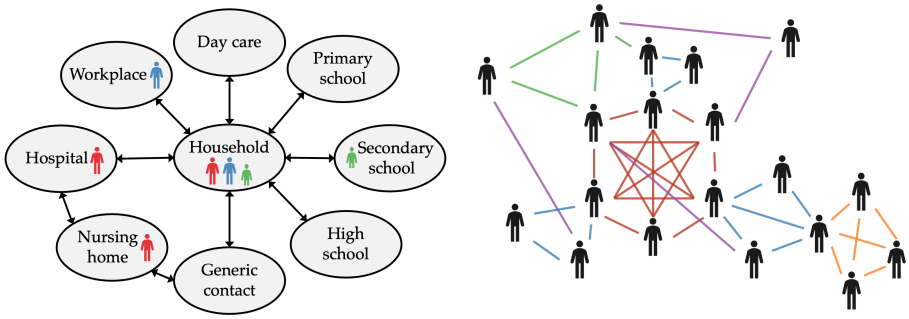
[1] The literature often distinguish between "super-spreaders" and "super-spreader events", which is two quite different things.

status. Suppose an infected person manifests symptoms or is confirmed Covid-19 positive, or an asymptomatic person tests positive. In that case, they self-quarantine from activities in all layers except their domicile (household or nursing home). Infected individuals have a set chance of becoming sick enough to be moved to the hospital layer based on age-related risk. They are either moved back to their standard layers when cured or removed from the simulation when dead.

All individuals are present in the generic contact network, which is a random time-dependent scale-free network meant to capture the heterogeneity in daily contact patterns. This network is generated daily for the entire municipality, and the number of contacts for each person varies each day. The generic contact network represents chance encounters between individuals, like meeting random residents in your city or municipality in the store or during commuting to work.

For individuals with professions like teachers or health care workers, their workplace is in a school or nursing home, meaning that they do not belong to a regular work-layer but are present in a clique in one of the other layers.

Figure 3.2.2 shows an illustration of the layers in the agent-based model, with an example of a family in the left plot and an illustration of a small social network in the right plot.



(a) Layer structure example **(b)** Contact network example

Figure 3.2.2: Illustration of the layer structure in the ABM. (a) Possible layer affiliations for an example family of 3 persons. Named circles show available layers that a person can be member of. All household members are also part of the generic contact layer. **(b)** Example of a contact network between individuals caused by shared group membership in different layers: household (blue), primary school (red), day care (orange), workplace (green) and generic (purple). Adapted from [46].

The different layers have different infection probabilities to simulate different behaviour in different settings, meaning that the probability of infection when an infected meets another person depends on which layer the contact happens. The infection probability is higher in the household layer than in the work layer. It is assumed that family members have more contact and spend more time together in a smaller area than someone does in their workplace. All clique members in each layer have the same constant probability of infecting other members of the

clique, except for children which have a reduced infectiousness to 30%, and a lower presymptomatic infection rate. Infection probability for each layer is listed in Table 3.2.1.

Table 3.2.1: Infection chances for the different layers in the model. All percentages are used in a Bernoulli function for each possible contact between a susceptible and an infected individual. All values are fitted to Norwegian clinical data.

Layer type	Infectiousness
Household	30%
Nursing Home	20%
Generic Contact Network	0.75%
Day Care	0.015%
Work	0.015%
High School	0.015%
Secondary School	0.015%
Primary School	0.005%

3.2.2 Epidemiologic Dynamics

The dynamics of the disease spread on the layers and cliques follow an SEIR-type dynamic, as described in more detail in Section 2.1.1. All individuals in the model have a given state, ranging from healthy to dead or recovered. The complete list of states is the following: Susceptible (S), Exposed (E), Infected asymptomatic (Ia), Infected pre-symptomatic (Ip), Infected symptomatic (Is), Hospitalised (H), Intensive Care Unit (ICU), Recovered (R), or Dead (D). Figure 3.2.3 shows the different states and their possible transitions and parameters.

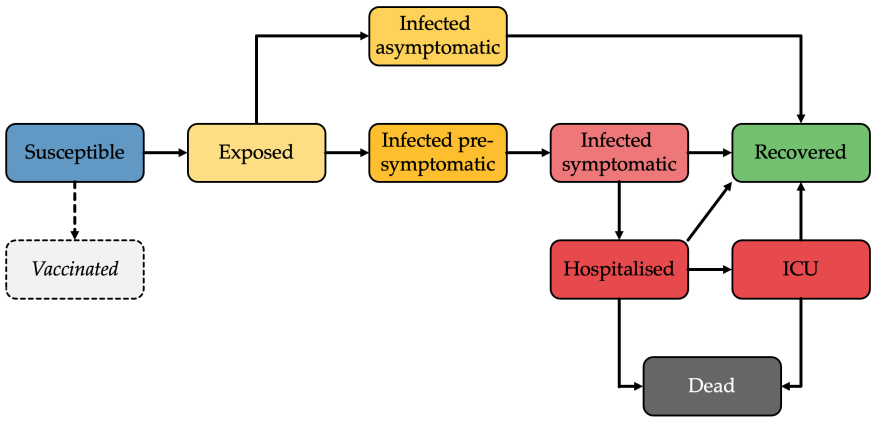


Figure 3.2.3: SEIR disease dynamics used in the model. The different state changes are shown. Note that the vaccinated state is illustrated, but is not included in the model itself.

After a susceptible person is infected in the contact network, the individual's disease course follows the SEIR dynamics, based on empirical data for Covid-19. The waiting times between each state is determined from a stochastic process based on the model parameters. These are shown in Table 3.2.2. The probability for different transition states vary based on age groups and are shown in Table 3.2.3.

Table 3.2.2: Parameters used for the Covid-19 disease dynamics. Symbols corresponds to Figure 3.2.3. Source HSØ indicates data from South-Eastern Norway Regional Health Authority (Helse Sør-Øst), and comes from email correspondence in March-April 2020. *Adjusted for reduced incubation time.

Parameter	Symbol	Value	Unit	Function	Source
Probability of infection	β	-	-	Network effect	-
Incubation time	λ_E	1	Days	Fixed	NHPI [24]*
Pre-symptomatic duration	λ_{IpS}	5	Days	Poisson	NHPI [24]
Symptomatic time before recovery	λ_{IsR}	5	Days	Poisson	HSØ
Asymptomatic time before recovery	λ_{IaR}	8	Days	Poisson	NHPI [24]*
Symptomatic time before hospitalisation	λ_{IsH}	6	Days	Poisson	HSØ
Symptomatic time in nursing home before death	λ_{ND}	10	Days	Poisson	HSØ
Hospital time, before recovery (no ICU)	λ_{HR}	8	Days	Poisson	HSØ
Hospital time, before ICU	λ_{HI}	4	Days	Poisson	HSØ
ICU time, before recovery	λ_{IR}	12	Days	Poisson	HSØ
ICU time, before death	λ_{ID}	12	Days	Poisson	HSØ
Exposed developing symptoms	P_I	50	%	Bernoulli	NHPI [24]
Hospitalised needing ICU	P_{HI}	30	%	Bernoulli	NHPI [24]
Not developing immunity	P_{RS}	0	%	Bernoulli	NHPI [24]
Mean, daily contacts	μ	10	-	Gaussian	Model fit
Variance, daily contacts	σ	3	-	Gaussian	Model fit
Exponent, daily contacts	γ	-0.5	-	Power law	Model fit

Table 3.2.3: Age stratified parameters. P_{IsH} represents the chance of a symptomatic becoming hospitalised. P_{HD} is the chance of a symptomatic patient dying in the hospital, and P_{ND} is the chance of symptomatic dying outside hospital. All parameters use Bernoulli functions. Source: Verity *et. al* [47].
*Nursing home residents only, adjusted to Norwegian hospital death rates.

Age group	P_{IsH}	P_{HD}	P_{ND}
$0-9$ years	0%	1.61e-3%	0%
$10-19$ years	0.048%	6.95e-3%	0%
$20-29$ years	1.04%	3.09e-2%	0%
$30-39$ years	3.43%	8.44e-2%	0%
$40-49$ years	4.25%	0.161%	0%
$50-59$ years	8.16%	0.595%	0%
$60-69$ years	11.8%	1.93%	0%
$70-79$ years	16.6%	4.28%	26*%
$80+$ years	18.4%	7.80%	42*%

During a simulation, each infected individual keeps track of its current and next state. In addition, the date of the last and subsequent state of change is tracked. These four data points are what the model checks for each day and updates infected individuals accordingly. When the day of the following state change occurs, the new state is determined from a stochastic draw, with the duration determined according to a Poisson-distributed random variable, λ, plus 1, to avoid the possibility of two state changes in one day.

To illustrate, a person in state Ip will develop symptoms (state Is) with a chance P_I, and the duration of the coming state (Is to R) will be determined by a Poisson draw of $p(\lambda_{IsR} + 1)$.

3.2.3 Reproduction number

There are several different ways to calculate the reproduction number in an ABM. For this model, the daily average reproduction number is calculated by first counting through the secondary infections of all recovered individuals. Second, the daily R-number is calculated by taking the average number of secondary infections caused each day.

This method makes for some stochasticity in the determined value of R between consecutive days. There is a significant increase in the first few days and an artificial drop in calculated R numbers for the last few days since individuals have to be recovered to count in the average reproduction number, biasing individuals with shorter illnesses. After a simulation, the average reproduction number is calculated after filtering out the first and last simulation days to counteract this bias. For a more detailed discussion on this topic, more information is found in the original paper and supporting material [46].

Another consequence worth mentioning is that the average daily reproduction number for the municipality model is calculated both for the entire region and separately for each municipality. The reproduction number for a given

municipality will include the number of secondary infections each inhabitant cause, regardless of in which municipality they are caused. This may create a delayed increase in the reproduction number for smaller municipalities infected by commuters from other municipalities.

In addition, many small municipalities will have several days or even entire simulations with zero secondary infections, on average. In the calculations, these are counted as nans, or not-a-number, instead of 0. Otherwise, the average number would be much lower for that municipality. This definition means that the average reproduction number for a municipality represents the average number of secondary infections when there are, in fact, secondary infections.

3.2.4 Intervention measures

One of the main goals of creating an agent-based model instead of a simpler compartmental model is to simulate different intervention measures. Examples include school closure, social distancing, testing and vaccinations.

Lock-down

The first intervention strategy is the lock-down of one or more layers in the model. Lock-down is implemented differently depending on layer type:

1. **Day care:** The layer is completely disabled.
2. **Schools:** Secondary and high schools are disabled completely. Primary schools can separately shut down for grades 1.-4. and grades 5.-7., which means that the younger kids can go to school while the older stay at home, for example.
3. **Work:** A fraction of cliques in the work layer is closed, representing workplaces where working from home is possible.
4. **Generic contact:** A shut-down in the contact layer is implemented by decreasing the infection probability in the layer, simulating fewer contacts and increased social distancing.
5. **Household and nursing homes:** These layers are never disabled.

Testing and Quarantine

The model implements a testing regime to find and quarantine individuals with Covid-19 that do not display symptoms. Testing is done by returning a positive test if the individual is asymptomatic or pre-symptomatic. One of the goals of the original model was to investigate different testing strategies and how these affect the reproduction number.

When an individual test positive, they are put into quarantine. Quarantine is modelled by disabling the workplace, school and generic contact layers for this individual. The same happens for individuals who self-quarantine after

they display symptoms of Covid-19. In practice, quarantined nodes are only present in their domicile (household or nursing home) or hospital if they turn sick enough.

Testing and quarantine strategies have not been a focus during this project. However, the functionality is implemented in the model.

3.3 Implementing a Municipality Network

The previous section described the underlying logic of the agent-based model and how the nodes and layers make up the complex network structure. In addition, base parameters and possible intervention measures were described.

This section will go through the methodology behind the national agent-based model, which has been the main focus of this thesis. Whereas the original model worked for a single municipality, it was primarily used for simulating Oslo or Trondheim and had no commuting or travel incorporated. This project has focused on implementing national data into the model to simulate the entirety of Norway simultaneously.

The first part describes how the model was turned into an object-oriented program. The second part goes into detail on how the population data of Norway was implemented and used in the commuter algorithm. Finally, a brief description of the municipality network is given.

3.3.1 Making the Model Object-Oriented

The original model was written in base Python, using a list- and dictionary-based approach to simulate the nodes and layers. A municipality consisted of all nodes present, with a dictionary-structure representing the nine different layers, with references to nodes in the different cliques in each layer. While this is made for effective simulations, a more intuitive and object-oriented approach is a worthy trade-off, as code is read many more times than written[48].

Therefore, in the process of understanding the code base, the code was turned into an object-oriented model, using classes and a hierarchical approach to represent the different object types. The nodes were turned into objects instead of dictionaries, and the layer structure was made into a hierarchy of classes.

Each layer is a class object that contains a list of cliques. Each clique is a class containing the nodes present, in addition to specific clique methods, like pooled testing and quarantine functions. This structure makes working with a relatively large object structure more straightforward, like looping and iterating through the object hierarchy is simpler and more intuitive.

A schematic of the class structure is shown in Figure 3.3.1, and a simple explanation is given in Table 3.3.1.

There are several advantages of object-oriented programming. Some strengths include hiding implementation details in lower-level classes, so more time can be spent ensuring the overall model structure and logic are functioning as intended and moving similar methods and classes to separate files and folders. This

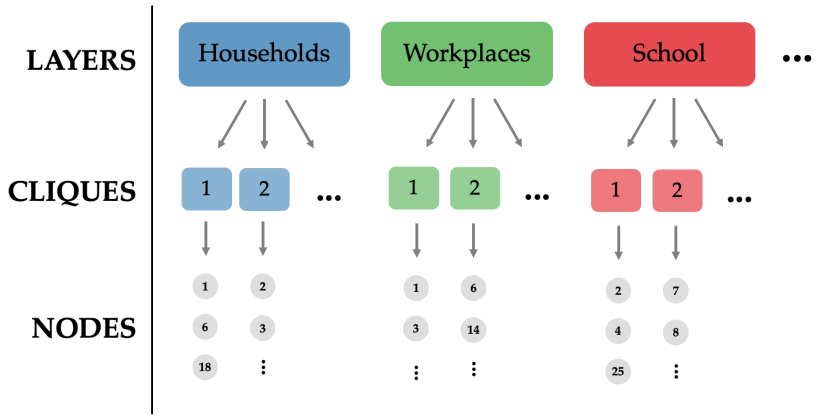


Figure 3.3.1: Illustration of the object hierarchy.

implementation has reduced debugging time and made new features easy to implement and test.

Table 3.3.1: An overview of the different class types in the model.

Class	Description
Person	Represents individuals. Contains attributes like id-number, age, disease state. Implements state change and testing functions.
Commuter	Subclass of the Person class, with commuter information like home and destination municipality.
Clique	A class containing persons. Implements pooled test and quarantine functions.
Layer	A container class containing different cliques.
Municipality	A container class containing all the layers and nodes for each municipality.

3.3.2 Population Data

High-resolution demographic data is essential to get realistic and representative agents when implementing population data. When creating the original model, a script to download data from Statistics Norway (SSB) was made to create the national data for the network structure. The actual datasets used are described in Appendix B.3.

The national data is divided into the different municipalities in Norway, each municipality represented by two text files. One contains the id and age for each

individual, and the other the social network used to create the layers. These two pieces of information create all the agents and put them into the correct cliques and layers. An example is shown in Table 3.3.2.

Table 3.3.2: Example of the two types of individual data tables. The left table simply contains all individuals in a municipality with id and age. The right table includes all cliques in a municipality, with reference to each clique's individuals, with varying size and age distributions.

Id	Age
1	65
2	24
3	11
4	56
5	7
...	...

Clique type	Node ids
Household	1, 13, 14, 23
Household	2, 4, 5
Workplace	4, 17, 22, 143, 178, 201, 202, 203
PrimarySchool	3, 28, 29, 30, 67, 68, 69
NursingHome	6, 71, 72, 73, 88, 155, 156, 157
...	...

Note that the dataset represents demographic data but not an entirely correct recreation of actual population data. For example, the data for Trondheim will include the correct number of schools and workplaces. However, the actual individuals going to each different clique or household will be based on random draws from the age and household distributions. The original article discusses this aspect in more detail[49].

For this model, the creation of nodes from the same dataset will be deterministic, and node 1 will always be present in the same household and workplace. If one were to create the municipal data from SSB data again, however, this would change. For this thesis, the age of nodes and the network structure are unchanged between runs. However, parameters like activity and random contacts change for every simulation[2].

3.3.3 Commuter Algorithm

To simulate spread between municipalities, information about commuters is critical. Commuters work or go to school in a different municipality than their household and regularly travel between two municipalities. These individuals are represented with a node id in the municipality data that points to a different municipality. This means that one or more nodes are references to nodes in other municipalities for a given workplace.

For example, a workplace in Trondheim might have a commuter coming from Oppdal, and their node id would be `oppdal76`, meaning this node is node 76 in Oppdal, not node 76 in Trondheim.

The following considerations were taken into account when linking these nodes and cliques together: Commuter nodes have a home municipality, where they are present in their household, and a commute municipality, where they are

[2] And simulated disease states and deaths, naturally.

present in their workplace or school. In addition, commuter nodes are present in the generic contact layer of both municipalities, but with an activity of only 50% in each layer. These considerations are meaningful while still being computationally efficient.

During the initialisation of the municipalities in a simulation, a temporary commuter layer is created. All cliques with incoming commuters get a commuter node reference placed in a dummy clique in the commuter layer. After the municipalities are initiated, the incoming commuters are linked to their respective references in their home, and the dummy variables are deleted. This means that the workplace in Trondheim not only has a reference to a node in Oppdal but includes the actual node so that the node is present both in Oppdal and Trondheim.

However, for smaller simulations, for example, when simulating only a region or county, not all commuter node references will have an initiated home municipality. In this case, the node is still created but has a `missingHome` attribute and has a small, daily chance of infection in its home municipality, based on a given prevalence level. This implementation gives a realistic representation of workplaces with many commuters while limiting the simulation scope.

The number of commuters between municipalities and regions will be described in more detail in the subsequent sections.

3.3.4 Municipality Network

After population data is initiated and the commuters are placed in the right home and work municipalities, the municipality network is created. In practice, this is a list of municipalities, where each municipality is a class object containing a list of inhabitant nodes and the network layer structure.

The network is scalable, and by changing the list of municipalities, the region to be simulated changes. This list can be a single municipality like Trondheim, a region like Trøndelag county, or the entirety of Norway. For testing purposes, the municipalities in the county of Trøndelag have been used, for the most part, to be able to simulate a relatively large region with several municipalities and commuting, without too long computation times.

Intermittent Travelling

Intermittent travelling is not incorporated in the model. The same is true for leisure travel or holidays. This will be discussed in more detail in later chapters.

3.4 Modelling Framework

The model itself consists of several Python files that together make up the simulation framework. These were designed to reuse the code as much as possible and minimise code duplication by using object-oriented approaches and inheritance. In addition, several R scripts were used for the data analysis and visualisations.

This section will go through the modelling framework, starting with an overview of the main simulation algorithm. A description of the code structure will follow, with a sequence diagram of the pseudocode and file structure. After this, simulation protocol and model output will be described, explaining default parameter values and some example visualisations of the results.

All code used in the project, as well as raw data, is available in the following GitHub repository: `https://github.com/helgebergo/master-thesis`.

3.4.1 Algorithm

A flowchart of the main algorithm is shown in Figure 3.4.1, giving an overview of the four main simulation steps:

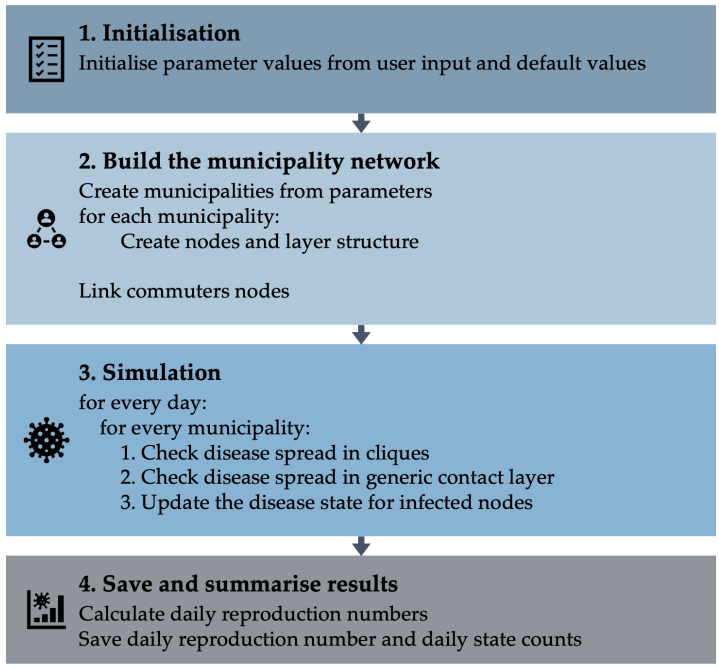


Figure 3.4.1: A flowchart illustrating the simulation algorithm.

These steps are performed for every simulation, and the results vary based on user and default input. By changing the municipality list in the initialisation step,

the size of the simulation is changed, for example. A list of the default parameter values is explained later in Section 3.4.3.

3.4.2 Code Structure

The code is structured into several different files with different functions during the simulations. The main model file is `nationalModel.py`, which initialises the model and calls on the other functions as needed. For the simulations presented later in this chapter, `run.py` is the script that controls all parameter input and parallel simulations. Table 3.4.1 gives a summary of the main files and their functions during simulations.

Table 3.4.1: Overview of the different files in the model framework.

Filename	Description
`run`	Script file for running all simulations and saving output files.
`nationalModel`	Main functions, including setup and run functions.
`model`	Updated version of the original model code, adapted to work with municipalities and the object oriented approach.
`nationalModel-Functions`	Supporting functions to create the municipality network for a given region.
`modelFunctions`	Supporting functions for the original model, including counting, test pool creations, test rules and strategy.
`parameters`	Disease dynamics, virus parameters and important model parameters for the simulations.
`classes`	Includes the person class, in addition to the cliques, layers and municipality classes.
`modelUtilities`	Utilities including saving, printing and profiling functions.

The hierarchy of different function calls is not trivial, and a simplified illustration of the sequence diagram is shown in Figure 3.4.2.

This diagram shows an example simulation started from `run.py`, which orchestrates the entire simulation. The parameters are initialised after the type of simulation, based on the default values defined in Section 3.4.3. After this, the municipality network is created, and the model loops through every municipality every day, simulating disease spread in the different layers and cliques. After the simulations, the different result files, including the disease states and daily reproduction numbers, are saved to larger summary files.

There are two main ways to run the simulations. One is to initialise the model, run it until a given number of nodes are infected, and then simulate for a set number of days. This way is useful for recreating real-life scenarios, for instance, last year in Norway, when the country closed down at around 100 confirmed cases.

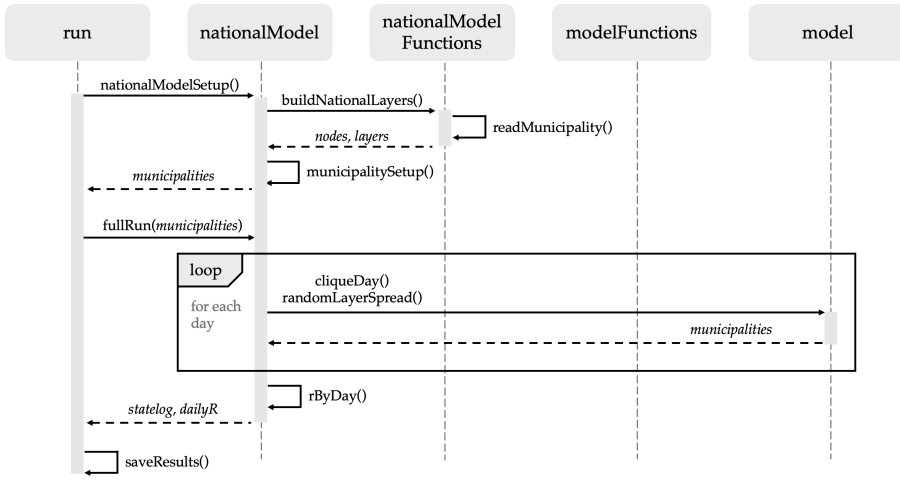


Figure 3.4.2: Simplified sequence diagram of the code. The `run` script orchestrates the simulations by first calling `nationalModel` and `nationalModelFunctions` to create the municipalities with nodes and layers. This is then used in the fullRun-function, which loops through every municipality for every day, and simulates disease spread in both cliques and generic contact layer. Finally, the results are summarised in text files for further analysis and visualisations.

The second way is to run the entire model for a given number of days without any threshold values defined. This method makes for more straightforward simulations that may be more comparable between runs but display less stochasticity in the initial values. It is also more challenging to simulate containment measures or strategies implemented after a given condition.

The second version is used in this project, though both versions are incorporated in the model.

3.4.3 Model Parameters

When doing simulations, the model is fed a set of user input, combined with the default values to create the parameters used in each simulation. The main parameters used in the simulations are shown in Table 3.4.2, with a short description of what each parameter does, together with the default value.

3.4.4 Model Output

The model output consists of two different types of text summaries, and these are further analysed in R to produce the visualisations.

Table 3.4.2: Main model parameters varied throughout the simulations. Short description and default values given.

Parameter	Description	Default
Days	Number of simulation days	60
Seed	Which municipalities should get initialised with infected	All
Region	Which region or list of municipalities to simulate	Trøndelag
Prevalence	Initial fraction of infected in each seed municipality	0.005
Mutation infectivity	Increase in base virus infectivity	0.0
Commuter fraction	Fraction of commuters simulated	1.0
Testing	If a test regime should be active or not	None
Strategy	The degree of general containment strategies implemented, like social distancing, hand washing, fewer contacts. Implemented as a factor that influences the base infection parameter.	0.25

Summary Files

The summary files are a log of the number of individual in each disease state and the average reproduction number, both saved for every municipality, for every day. Table 3.4.3 shows an example of these two tables.

Table 3.4.3: Example model output. The left table summarises all states' counts for every day, and the right table shows the calculated daily reproduction number. Note that the values displayed are arbitrary and only meant to illustrate the two different output types and trends.

Day	S	E	Ia	Ip	Is	R	H	ICU	D
...	...	...	...	...	...	...	...	...	...
55	188	39	56	25	33	42	3	0	0
56	186	40	61	27	37	50	4	1	0
57	184	44	67	27	41	60	4	2	0
58	181	47	73	29	45	70	5	2	1
...	...	...	...	...	...	...	...	...	...

Day	R
...	...
55	1.66
56	1.63
57	1.59
58	1.57
...	...

The tables are merged into larger tables containing additional info about simulation parameters, like the municipality, region, commuter fractions and differences in seeding municipality, mutations or containment strategies.

Visualisations

The combined data tables have been analysed in R, and several graphical representations have been created. To illustrate a typical simulation, five replications with the default parameters have been run, and two plots have been made. The first is a line chart with the different number of individuals in each state throughout the simulation, in Figure 3.4.3.

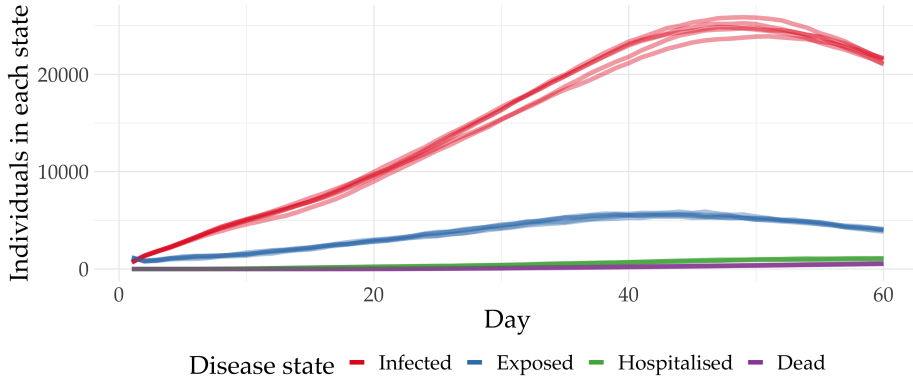


Figure 3.4.3: Daily state counts after 5 runs in Trøndelag. Total number of individuals in the entire region, for four of the disease states. The three different infected states have been aggregated into one.

Four of the states have been included, the number of infected, exposed, hospitalised and dead. The numbers show the total number of individuals in each state in the Trøndelag region.

The second plot, Figure 3.4.4, shows the average reproduction number in the region after the five replications.

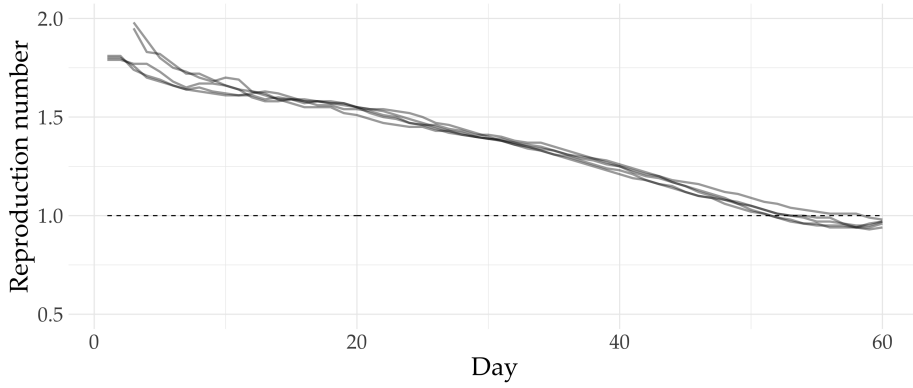


Figure 3.4.4: Daily reproduction number after 5 runs in Trøndelag. Average values for all infected in the entire region.

CHAPTER 4

RESULTS AND ANALYSIS

The previous chapter introduced the agent-based model logic and the modelling framework.

This chapter is structured into three parts. The first section explores the datasets used throughout the project, consisting of Trøndelag and Norway regional data. The second part goes through the simulation results of Trøndelag, where different parameter values have been explored. The third and final part will describe the results from a complete Norway simulation, with an analysis of crucial demographic parameters influencing the Covid-19 disease spread. Figure 4.0.1 gives an overview of the three parts of the chapter.

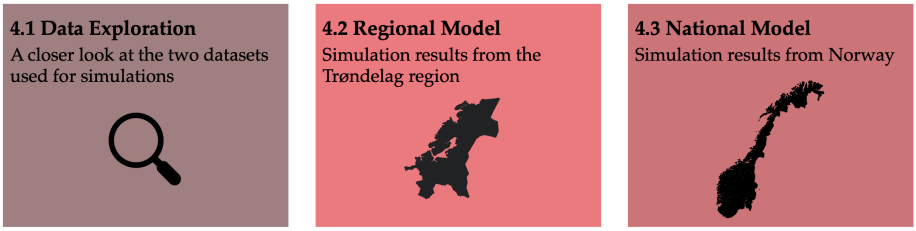


Figure 4.0.1: Overview of the three sections of the chapter.

4.1 Data Exploration

This section will go through essential aspects of the data sets used in the project and lay the groundwork for the simulation results to come later. The data sets were created a year ago by someone else, and it is crucial to explore the population and commuter data. There are two versions of the dataset that is used

throughout this thesis. The first consists of 27 municipalities in Trøndelag county, and the second is all counties in Norway, with over 300 municipalities simulated. The following subsections will first go through the Trøndelag data and then the complete Norway data set to give some insight before the results in Section 4.2.

It should be noted that the original municipality data for Norway was assembled only months after a significant municipality amalgamation, where several of the smallest municipalities in Norway were merged with others, moved to different counties, or changed names[50]. This means that several of the municipalities are non-functional in the model, either because of name changes or changes in the underlying data sets from Statistics Norway. A more thorough discussion on this topic is found in Section 5.2.2.

4.1.1 Trøndelag Region Data

Trøndelag county is a large region in the middle of Norway, with a population size of around 400 000. The central city and municipality is Trondheim, with 190 000 inhabitants. It is relatively representative of a "normal" Norwegian county, with a large, central city and several smaller cities and municipalities nearby. Trøndelag has a few municipalities with a population of 15-20 000 and several smaller ones. There is extensive commuting within the region, especially to and from Trondheim. Stjørdal has an airport, and trains and main roads are connecting Trøndelag to nearby counties, facilitating commuting to other large cities in the country.

There is a significant spread in population size between the municipalities, as illustrated in Figure 4.1.1.

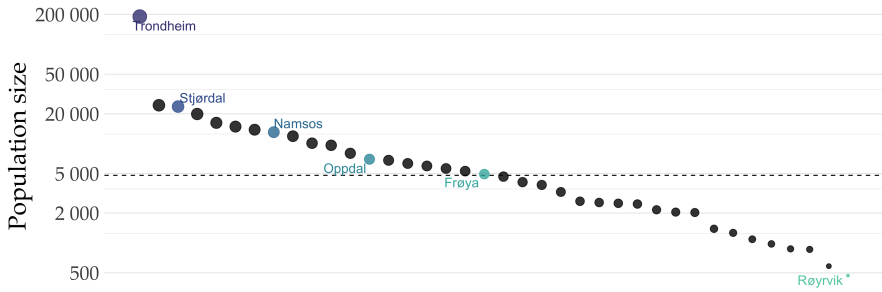


Figure 4.1.1: Population sizes of municipalities in Trøndelag. Ordered by size. The dashed line shows the median municipality in population size, which is Frøya. The vertical axis is in logarithmic scale, and the size of points follow population size.

The figure shows that Trondheim dwarfs the other municipalities in size and that there are several tiny municipalities with only a few thousand inhabitants. The median municipality is Frøya, with a population of 5000. The smallest municipality is Røyrvik, with only 500 inhabitants. In addition to these two, there are four more municipalities highlighted, and these will be used throughout the project. They have a wide range of population sizes, in addition to differences

in area and population density, the fraction of commuters, and mean age. An overview is shown in Figure 4.1.2. The actual numbers can be found in table form in Appendix C.1.1.

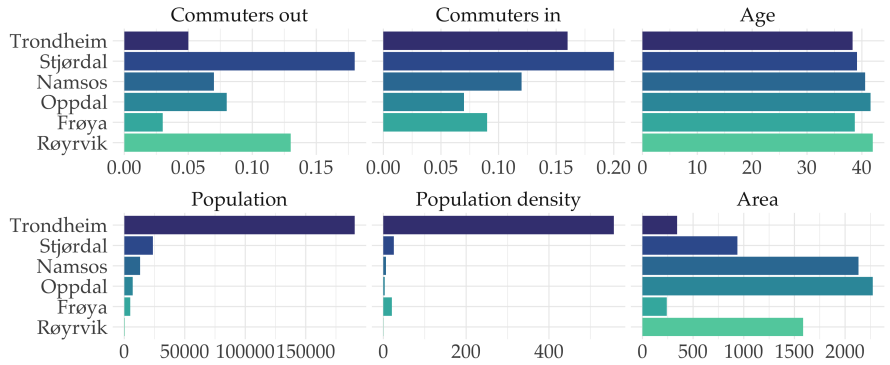


Figure 4.1.2: Overview of the six highlighted municipalities in Trøndelag.

As can be seen, there are some key differences other than the apparent population size. The commuter fraction is quite different between the municipalities, with Stjørdal and Røyrvik having a significant fraction of their population commuting to other municipalities. In contrast, Trondheim and Stjørdal have a large number of workers coming into the municipality. The age distribution is quite similar, but the smaller municipalities have an older population, on average. In addition, population density is very varied, with Trondheim knocking the other municipalities out of the ballpark.

Commuters in Trøndelag

Trøndelag has extensive commuting within the region, and several of the municipalities are typical commuter cities, where a substantial fraction of the population commutes daily to other cities for work. To investigate the differences between municipalities, Figure 4.1.3 was created. Here the fraction of incoming commuters by population size is plotted against the outgoing commuters. This figure shows the differences between the municipalities and the significant variance in the region.

Typical commuter municipalities are Skaun, Malvik and Melhus, while Trondheim and Namsos have a much larger fraction coming into the city than travelling out. The plot shows the fraction of commuters, not the actual amount, so remember that for larger municipalities like Stjørdal and Trondheim, there will be several thousand people travelling to and from work every single day.

However, while this gives insight into individual municipalities, even more, interesting is looking at how the different areas of the region are connected. Figure 4.1.4 shows the actual amount of commuters on the left-hand side and the correlation between the municipalities on the right-hand side.

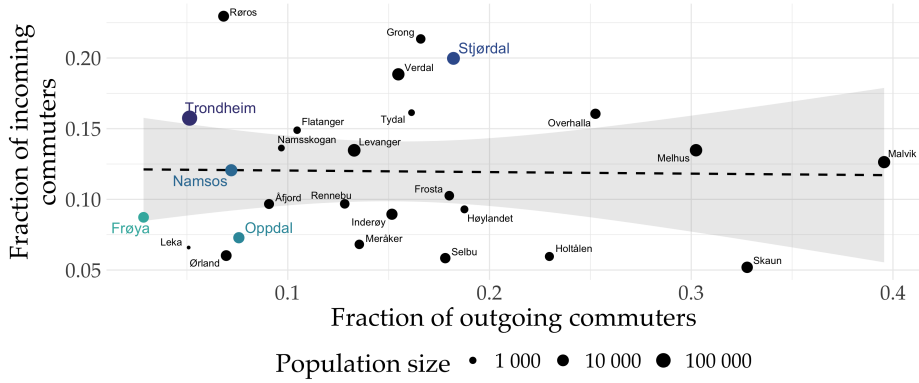
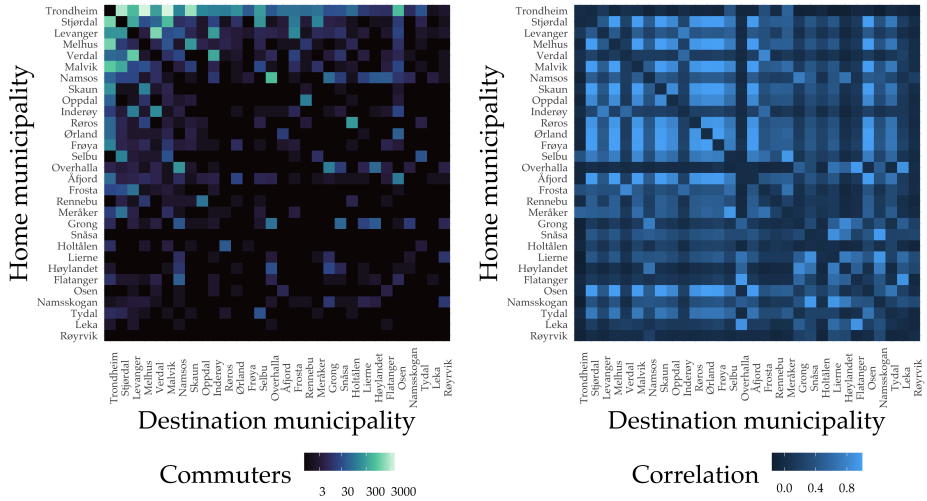


Figure 4.1.3: Commuter fractions for all municipalities in Trøndelag. Trend line shown as dotted line.



(a) Number of commuters.

(b) Commuter correlation.

Figure 4.1.4: Commuters and correlation between municipalities in Trøndelag. Sorted by descending population size.

This figure shows several points of interest. First of all, the majority of commuters travel to and from the largest municipalities. Second, the plot is very symmetric. Trondheim has several outgoing commuters to nearly every municipality and vice versa. The correlation plot on the right side shows some clustering, especially between the largest municipalities.

Throughout this project, the Trøndelag data will be used extensively, as this is a large region with a large population, but still small enough that simulations

are efficient and manageable.[1] However, some of the later simulations will incorporate the entire country so that the next section will describe some key features of the Norwegian country data.

4.1.2 Norway Data

The Norway data set consists of 5 million individuals, separated into 11 counties and 365 municipalities. The spread in population sizes is vast, as well as differences in population density and commuters. As can be seen in Figure 4.1.5, the differences in fractions of incoming and outgoing commuters are significant. This pattern is not unlike the distribution in Trøndelag, actually, and seems to point to the fractality of city sizes and counties. Oslo has few outgoing commuters, and a considerable fraction is coming in, especially when considering the population size of Oslo. The age distributions are pretty similar, but there is a slight trend for lower mean age in the largest counties.

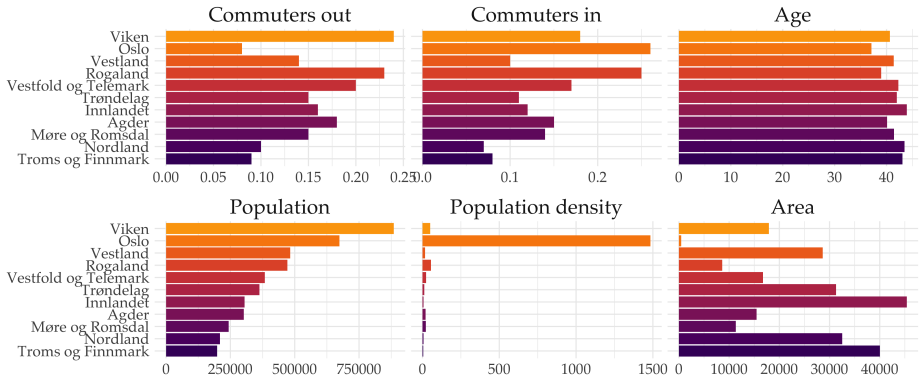


Figure 4.1.5: Overview of demographic distributions of counties in Norway. The upper three panels show the mean of each county, the lower three show the sums.

Figure 4.1.6 shows the spread of population sizes for all municipalities in Norway. As can be seen, the spread is large. Most of the municipalities in Norway fall somewhere between 1000 and 20 000 inhabitants, but there are quite a few in the 50 000 to 200 000 range, and some tiny ones with only a few hundred inhabitants.

Commuters in Norway

As the population distribution is varied, it is unsurprising that the number of commuters between municipalities and regions displays significant variations. Heatmaps of the number of commuters within and between counties are shown in Figure 4.1.7. A similar plot with commuters between all municipalities can be found in Appendix C.1.2.

[1] And the fact that this thesis is written in Trondheim by an NTNU student, after all.

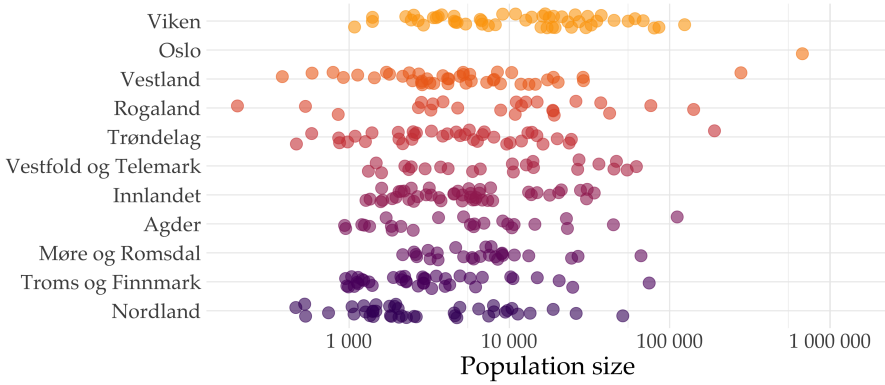
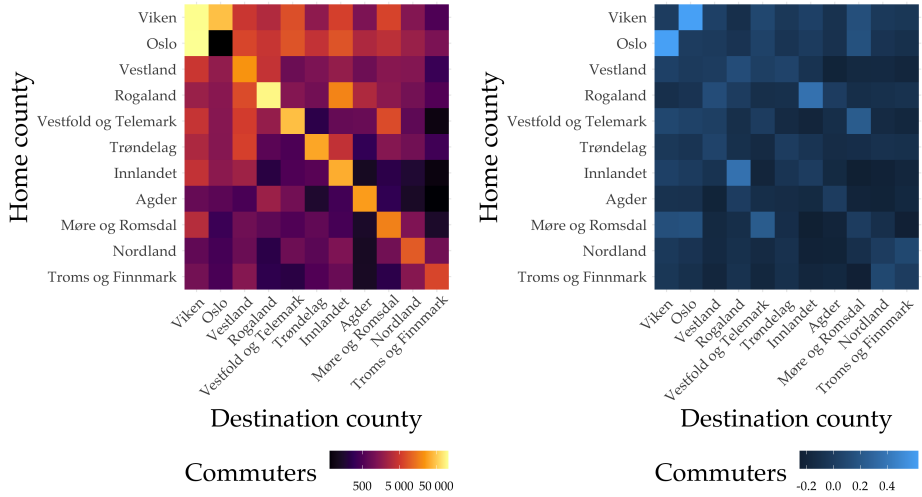


Figure 4.1.6: Distribution of population sizes in Norway. Horizontal axis in logarithmic scale. Counties sorted by decreasing total population size.



(a) Number of commuters.

(b) Commuter correlation.

Figure 4.1.7: Commuters and correlation between counties in Norway. The counties are sorted by decreasing population size.

What is most interesting in these two plots is perhaps the large degree of commuters travelling within counties or neighbouring counties. The diagonal in the left plot is by far the brightest, meaning that most commuters travel to other municipalities but in the same county. This trend is not surprising, as most commuters travel to nearby areas or cities instead of across the country, but its extent is worth noting. The number of commuters varies a lot, and the largest counties have tens of thousands of commuters travelling daily.

4.2 Regional Model

The following section will go through the simulation results from the Trøndelag region. First is a discussion of the average reproduction number for the different municipalities with the default parameter values. Following is a walk-through of the different parameters tested. An overview of the subsections in this part is illustrated in Figure 4.2.1.

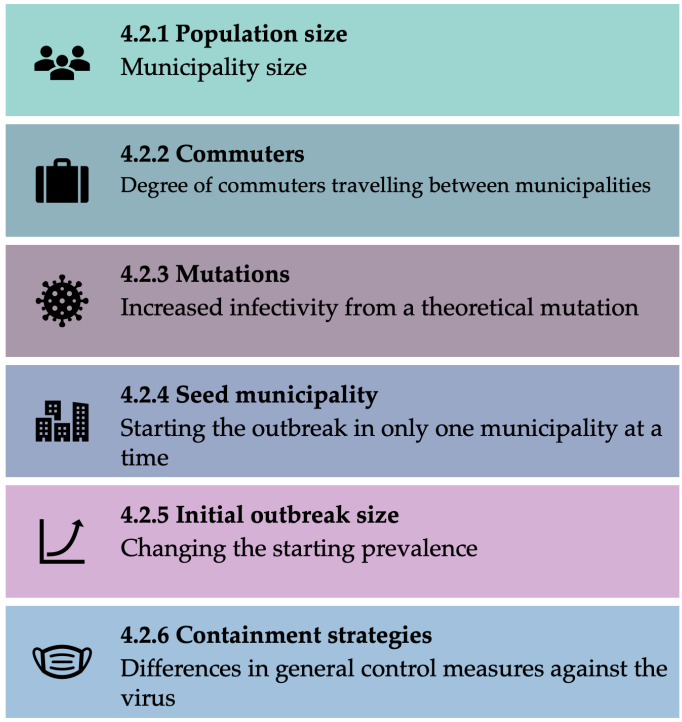


Figure 4.2.1: Overview of the subsections, with a short description of what was changed for each parameter.

The two visualisations types described in Section 3.4.4 will be used extensively throughout this section, but with a twist. A limit of the previous plots is that they only show the overall trend in the Trøndelag region. One of the main points of this project has been to simulate entire regions and look at differences between municipalities. The same plot as Figure 3.4.4 is recreated, but this time the reproduction number for every municipality is shown in a different facet of the plot. The plots are sorted after decreasing population size with Trondheim in the upper left. This sorting will be consistent throughout this chapter. In addition, the same six municipalities of interest as described in Section 4.1.1 are highlighted using the same colour scheme to facilitate easy comparisons between runs and parameter values throughout the section.

There are several interesting things to point out in this figure. The stochasticity

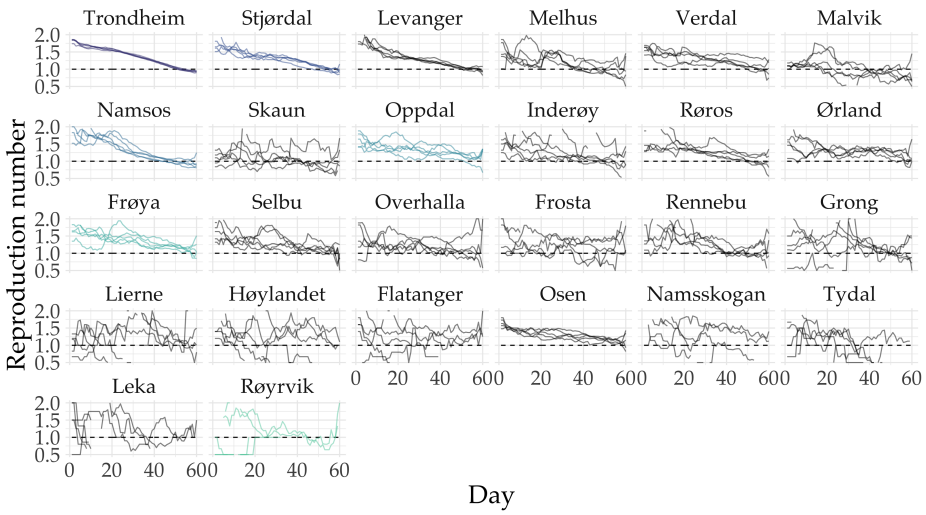


Figure 4.2.2: Daily reproduction number for 5 runs in Trøndelag, split to show each municipality. The municipalities are sorted by decreasing population size.

is significant, especially for the smallest municipalities. There are considerable differences both between and within municipalities. By comparing the two smallest municipalities, Røyrvik and Leka, with Trondheim and Stjørdal, the differences in fluctuations are pretty clear.

The overall trend is a high reproduction number in the first days, which gradually decreases throughout the simulation for most of the municipalities. The trend is most evident in the largest municipalities and more challenging to pin down in the smaller ones, like Røyrvik, for instance.

When it comes to the disease states, Figure 3.4.3 is also recreated with a split for each municipality, shown in Figure 4.2.3. The differences in numbers between municipalities differ on several orders of magnitude, but the trends are very similar. It can once again be seen that the stochasticity increases for the smaller municipalities and the differences in the number of infected are huge. Trondheim has close to 20 000 infected on day 50. In contrast, Frøya has around 300, and Røyrvik only has 20-30, meaning that 10% of the population in Trondheim is infected, compared to roughly 6% in Frøya and Røyrvik. The increased number of initial cases may cause a much larger number of infected in Trondheim. However, the most probable reason is that a larger city means a more extensive social network, with more infected interacting with susceptible individuals, increasing the risk of super spreader events and disease in large cliques.

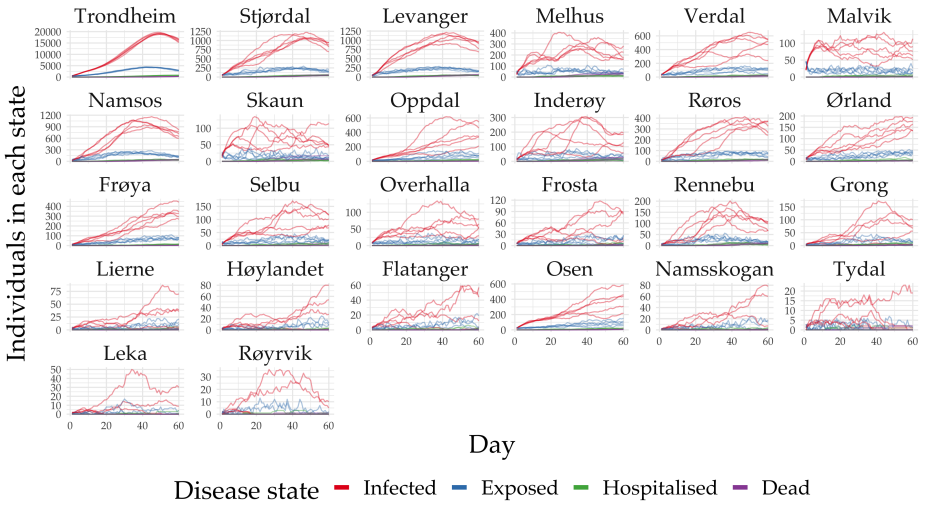


Figure 4.2.3: Daily states for 5 runs in Trøndelag. Note the limits on the y axis is different for each municipality, meaning that the numbers are not directly comparable, but shows the overall trends.

Mean Reproduction Number

The previous figures showed the aggregates of only five different runs. However, more runs are needed to compare different municipalities and investigate the degree of fluctuations in the model. Figure 4.2.5 shows the mean and standard deviation after 100 runs with default values.

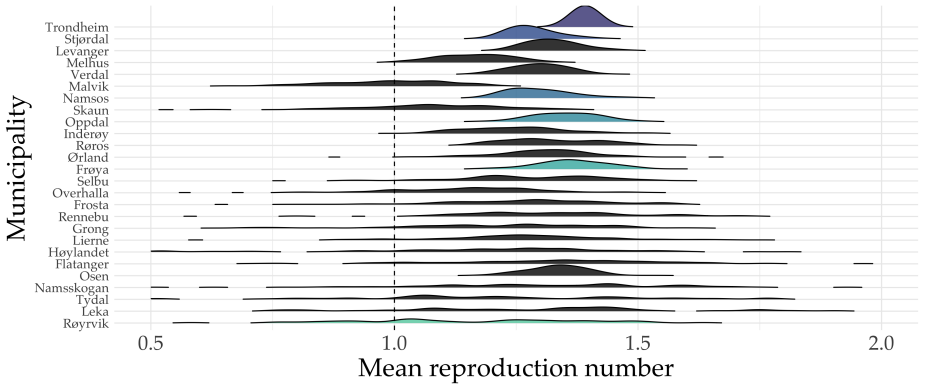


Figure 4.2.4: Density ridge plot of mean reproduction number for Trøndelag. Mean values after 100 runs. The municipalities are sorted after decreasing population size.

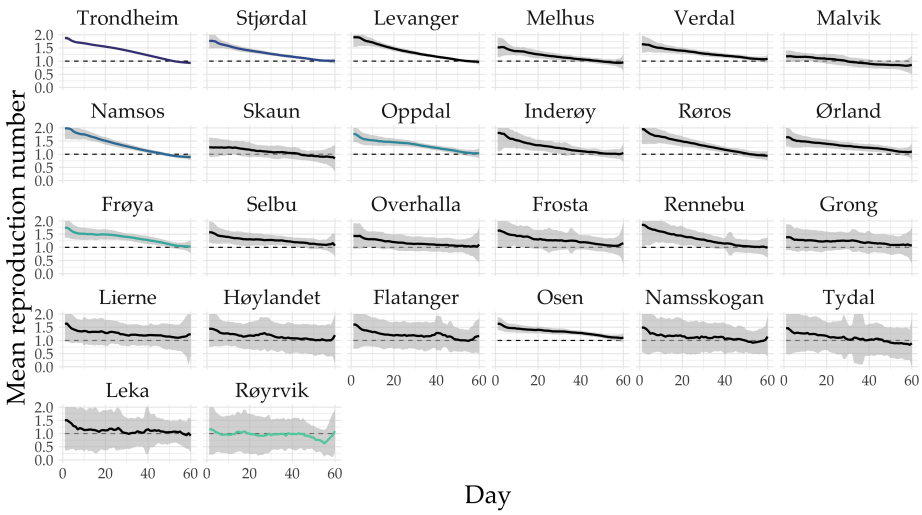


Figure 4.2.5: Mean daily reproduction numbers in Trøndelag. Values calculated after 100 runs. Standard deviation represented by the grey ribbons.

The plot shows significant differences between municipalities in the standard deviation, but the mean reproduction numbers do not vary much. Most municipalities start with an R-number of slightly below 2 and end up at around 1 after 60 days of running time. The variance increases steadily as population size decreases, with some notable exceptions, like Osen. A possible reason for this might be how the average reproduction number is calculated for each municipality. As described in Section 3.2.3, a simulation with no secondary infections in a given municipality is defined as not-a-number instead of zero.

The aggregated mean reproduction number for each municipality is shown in Figure 4.2.4, a density ridge plot showing the average of 100 runs after filtering out the first and last 15 days.

The density ridge plot has some advantages compared to a regular box plot, as it is easier to see the overall trends in reproduction numbers. The larger municipalities, like Trondheim and Stjørdal, display thinner density curves, meaning they have less variation between runs. On the other hand, the smallest municipalities show a much larger dispersion of values and range from 1 to 2 in R-number.

Another aspect of this type of visualisation is that the average of every run is displayed. Trondheim seems to have an average R-number of around 1.4 for every single run. In contrast, look at Røyrvik, the bottom-most curve. Here it seems like the majority of runs in Røyrvik has an average reproduction number of either 1.0, 1.25 or 1.4. Between these, the whole spectrum is included. It can be concluded that the average reproduction number in a small municipality like Røyrvik is hard to pin down, even after 100 replicated simulations with similar starting conditions.

Finally, the average number of infected for each municipality is shown in Figure 4.2.6. The numbers look very similar to the results shown previously (Figure 4.2.3), with increasing variance for decreasing population size. It may be difficult to interpret from this plot, but the number of hospitalised and dead is small, so is the variance in these states. For this reason, these numbers are often used as an indicator for curve and model fitting since they display less stochasticity than infection numbers.

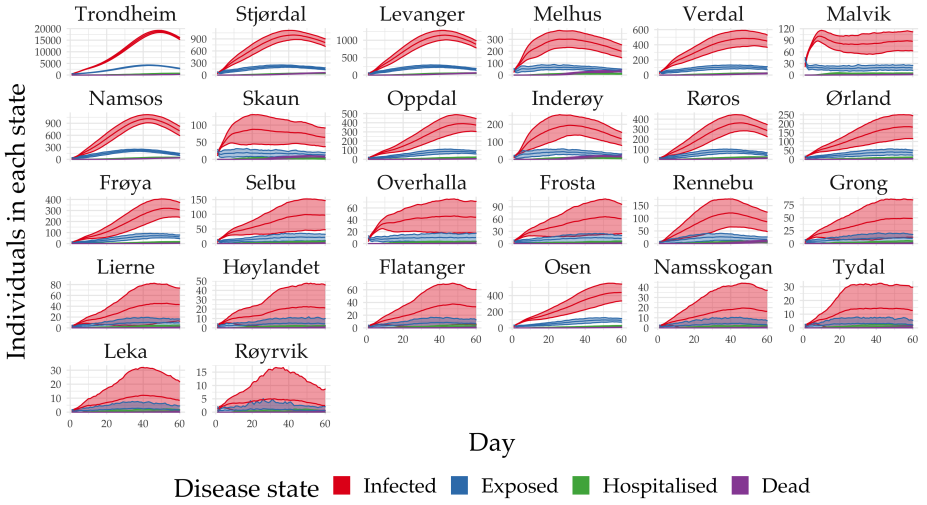


Figure 4.2.6: Mean daily state counts in Trøndelag. Values calculated after 100 runs. Standard deviation represented by the coloured ribbons.

4.2.1 Effect of Population Size

The previous results indicate that population size might be an essential factor in determining each municipality's reproduction number. To look closer at this, Figure 4.2.7 shows reproduction number as a function of population size, with a smoothed average displayed on top. The six same municipalities as before are highlighted.

Trondheim is by far the largest municipality and has the highest R-number, of around 1.4. Frøya and Oppdal, two similarly sized municipalities, have a slightly lower average, and Røyrvik has one of the lowest of close to 1.0. Nearly all municipalities have an average R-number of above 1, the median value is 1.25, and the average is 1.22.

The smoothed average might indicate an increasing linear trend, meaning a higher mean reproduction number as a function of population size. However, note the large grey area representing the variation, as well as the logarithmic horizontal axis. To investigate further, linear regression and ANOVA were performed on population size against the reproduction number. The resulting R^2

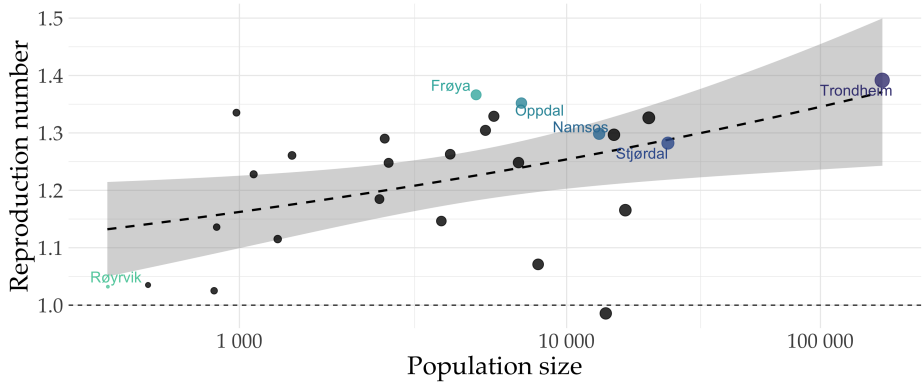


Figure 4.2.7: Mean reproduction number against population size in Trøndelag. The mean is calculated for each municipality after 100 runs. The smoothed mean is a linear regression of the values, but seems curved due to the logarithmic horizontal axis.

was 0.109, with a p-value of 0.1. This value means that the effect of population size on the average reproduction number for the municipalities in Trøndelag is not statistically significant, even though it may seem like it at first glance. However, this regression was done with only 27 municipalities, so it is difficult to get a significant result with so few values. This regression test will be repeated later for the larger national data set.

4.2.2 Effect of Commuting

One of the objects of this project has been implementing commuters between municipalities. The effect of these commuters has to be investigated. How much does the degree of commuting influence the infections in the region? Will a significant change in commuters affect some municipalities, or only the largest or smallest?

The degree of commuters present in the model was changed by varying the number of commuters for each municipality. The parameter is controlled by varying the commuter parameter between 0 and 1, where 1 is 100% of commuters present, while 0 is absolutely no commuters present in any municipalities.

Figure 4.2.8 shows the average reproduction number for each municipality after 100 simulations with five different commuter degree values.

First, looking only at the overall trends, there are few differences between different degrees, especially for the largest municipalities. One would first assume that commuters would impact the results by a more significant factor, but remember that even the largest municipalities only have a few thousand commuters, only a fraction of their total population size.

The highlighted municipalities are shown more clearly in Figure 4.2.9.

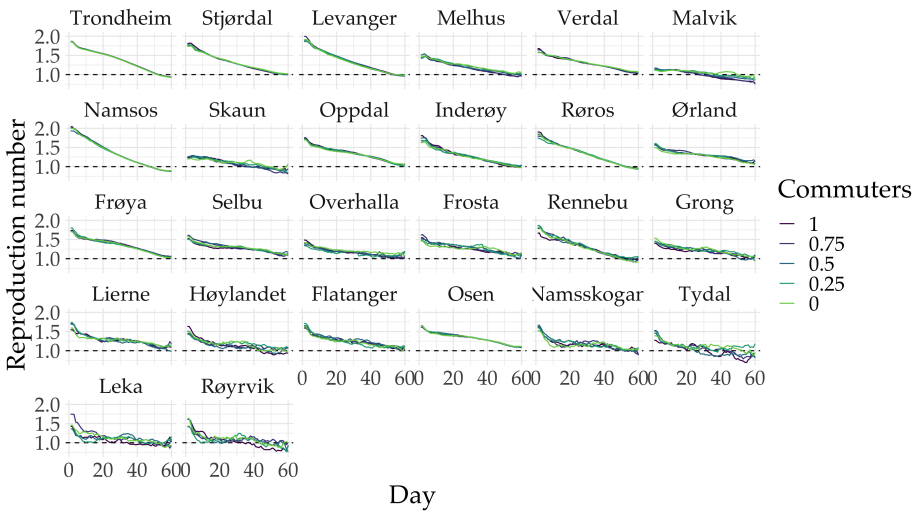


Figure 4.2.8: Average daily reproduction number as a function of commuter fraction. Daily mean for each municipality after 100 simulations of each infectivity value.

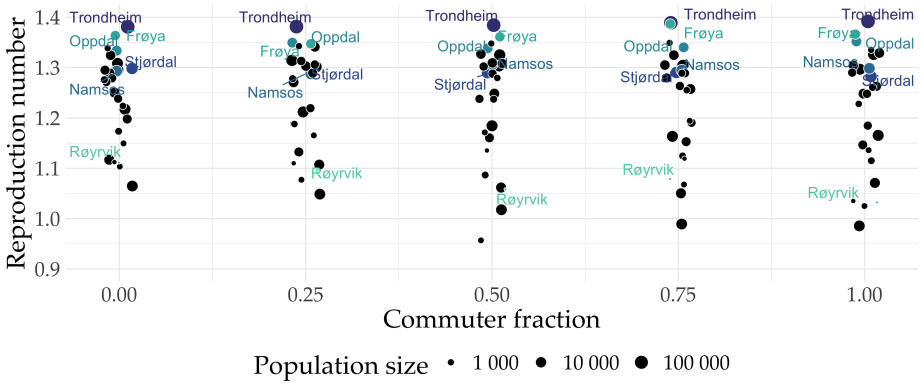


Figure 4.2.9: R-number as a function of commuter fraction. Mean for each municipality after 100 runs. The size of points is proportional to population size, and the six highlighted municipalities are coloured to be comparable across runs. There is added a slight amount of jitter in the horizontal direction, to better separate points.

Here the different commuter degrees are easier to separate, and it is even more apparent that there are at best minor differences between the simulations. Following the highlighted municipalities throughout the different parameter values, there are not any noteworthy differences. It might seem that the variance increases slightly as a function of more commuters, where mainly Røyrvik displays some decrease in the R-number.

The actual commuters removed from each run is drawn at random. However,

as there have been 100 repeated simulations for each parameter, this should not influence the outcome by a significant factor. All in all, it seems clear that the degree of commuters has no significant impact on the reproduction number in the Trøndelag region.

For completeness, the daily number of infected for each commuter degree can be found in Appendix C.2.1.

4.2.3 Effect of Mutations

During the last few months of the pandemic, a topic that has received increasing attention is the effect of new mutations. These new variants of the virus might have different abilities and infection probabilities. While this has not been a large part of this work, a simple mutation parameter was included in the model.

Mutations are modelled by increasing the base infection parameter of the virus by a factor of between 1 and 2. Mutation infectivity of 1 means the default probability, while 2 is a doubling of the base infection risk for every single encounter between infected and susceptible individuals.

First of all, a similar plot as the commuter degree was created, shown in Figure 4.2.10.

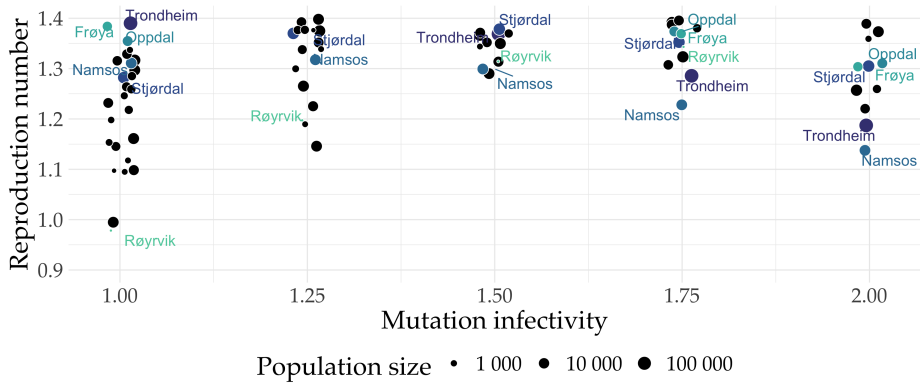


Figure 4.2.10: R-number as a function of mutation infectivity. Mean for each municipality after 100 runs. The size of points is proportional to population size, and the six highlighted municipalities are coloured to be comparable across runs. There is added a slight amount of jitter in the horizontal direction, to better separate points.

At a glance, this looks surprisingly similar to the corresponding plot of commuter degree. The rightmost plot in the previous figure corresponds to the leftmost plot in this figure. However, there are some key differences. There is a decreased variance and slightly lower reproduction number as the mutation infectivity increases. Comparing, for instance, Namsos and Røyrvik, they seem to follow opposite paths through the different parameters. Namsos has a slight decrease in reproduction number as mutation infectivity increases, while Røyrvik displays a slight increase.

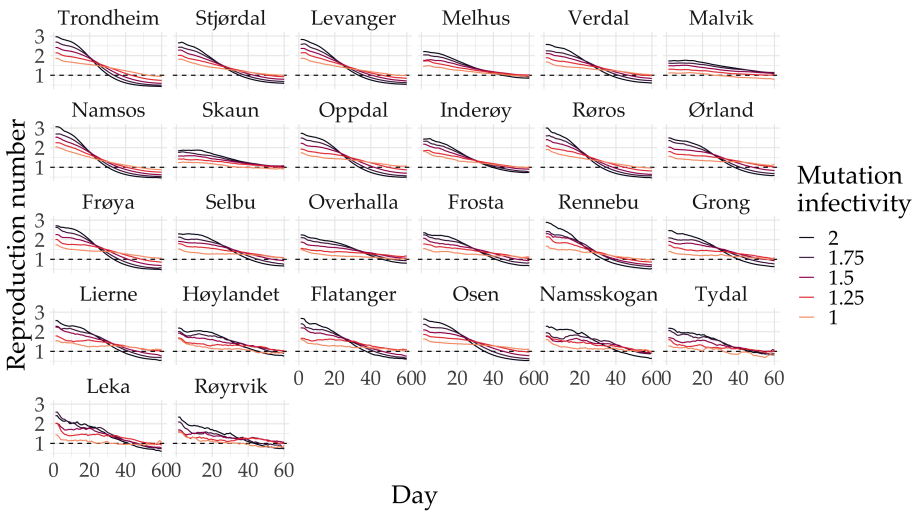


Figure 4.2.11: Average daily reproduction number as a function of mutation infectivity. Daily mean for each municipality after 100 simulations of each infectivity value.

It can be challenging to understand the course of the disease in the different municipalities by this aggregated figure of the mean values. Going back to looking at the evolution of daily R-numbers, this can be remedied, and the reason is more evident. Figure 4.2.11 shows the average daily reproduction number for each municipality and shows a fascinating pattern.

Here, while the total average reproduction number is not that different, the daily average starts high for the highest mutation infectivity but quickly falls after about 20 days. This fall is probably because the initial infected spread the virus to a very high number of people in their first few days, which in turn will lead to several cliques becoming saturated with infected individuals, leading to fewer new cases after a few weeks. If everyone is already sick or immune from a previous infection, the available amount of new susceptible falls quickly, and herd immunity is achieved fast.

The overall number of infected can be seen in Figure 4.2.12, where a doubling in mutation infectivity leads to somewhere between 3 and 4 times as many infected for most municipalities.

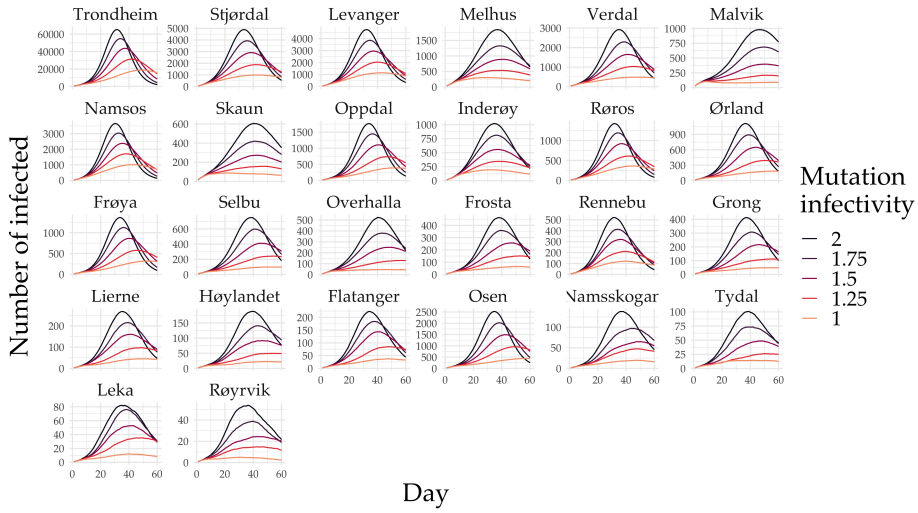


Figure 4.2.12: Average daily number of infected as a function of mutation infectivity. Daily mean for each municipality after 100 simulations of each infectivity value.

4.2.4 Effect of Seed Municipality

Few real-world pandemics start with a large number of infected spread evenly throughout a given geographical region. Instead, a much more realistic course of events is that a handful of infected individuals in a small area catch the disease from somewhere else, and it gradually spreads through the social network in a region. This timeline was accurate for several European countries during the winter of 2020, where an outbreak of SARS-CoV-2 started in Italy and was spread unknowingly by tourists when they travelled home to their own countries.

Different seeding conditions were simulated by initialising a single municipality with 100 infected on day one to see how the virus diffused throughout the region. The previously highlighted municipalities were chosen as the seeding municipality, and the mean reproduction number was calculated for the entire region afterwards. The results are shown in Figure 4.2.13.

The mean reproduction number is around 1.4 for both Trondheim and the default parameter values. A group of infected in the largest city in the region will have a higher chance of infecting the rest of the region than a much smaller city. Interestingly, Stjørdal gives a slightly higher average reproduction number than the default value. The reason for this might be found when looking back at Figure 4.1.2, where Stjørdal has the highest fraction of outgoing commuters in Trøndelag, in addition to a relatively large population size and density.

The differences are significant between the three seeds mentioned in the previous paragraph and the remaining four municipalities. Worth noting is that Røyrvik is not present because of lack of data, which is the cause of two different things. First, the mean reproduction number is only calculated for

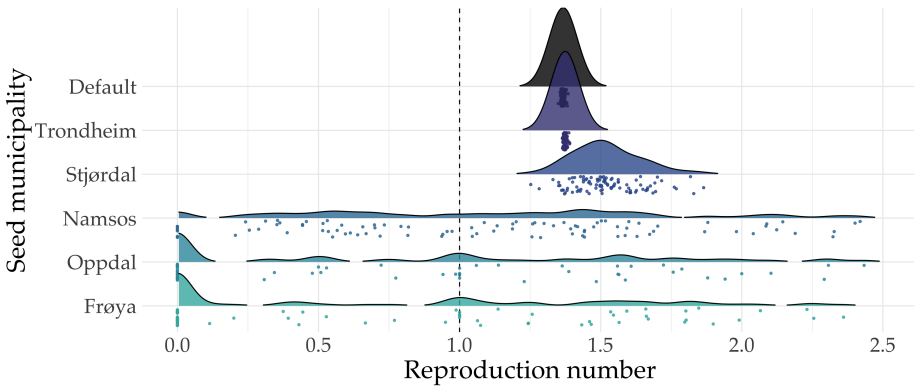


Figure 4.2.13: Density ridge plot of mean reproduction number for Trøndelag for different seed municipalities. The data points show only the average for the entire region, not individual municipalities.

simulations where the infected individuals spread the disease to others. Second, days without new infections are defined as not a number instead of zero. These two factors explain why, on average, infected in Røyrvik do not spread the disease further after the first 15 days. Based on the SSB data, Røyrvik has no incoming commuters. This lack of infected in the plot demonstrates that the commuters in the model work as intended, at least for the cases where a municipality have no incoming commuters.

Namsos, Frøya and Oppdal are pretty similar, but the mean reproduction number is slightly lower for the last two. The variation is significant when the infection starts in one of these municipalities. Figure 4.2.14 shows the daily R-numbers for the different seeds.

The variance is significant, and it can be challenging to interpret the plot. The main takeaway of this is that an infection in the smaller municipalities, namely Namsos, Oppdal and Frøya, on average do not spread the virus to all other municipalities. The reproduction number is often low when they do, indicating that few infected are moving between the different areas. An infection in Trondheim and Stjørdal usually spread to the other municipalities, but there are substantial variances here.

The default model works by initialising each municipality with a percentage of infected individuals, which means that Trondheim sees many more infected individuals than Leka and Røyrvik. For these simulations, a set number of 100 infected was infected for all seeds, to be comparable between runs. However, this implies that these results are not directly comparable to the other parameters inspected in this chapter.

Appendix C.2.2 includes the daily number of infected for each seed municipality.

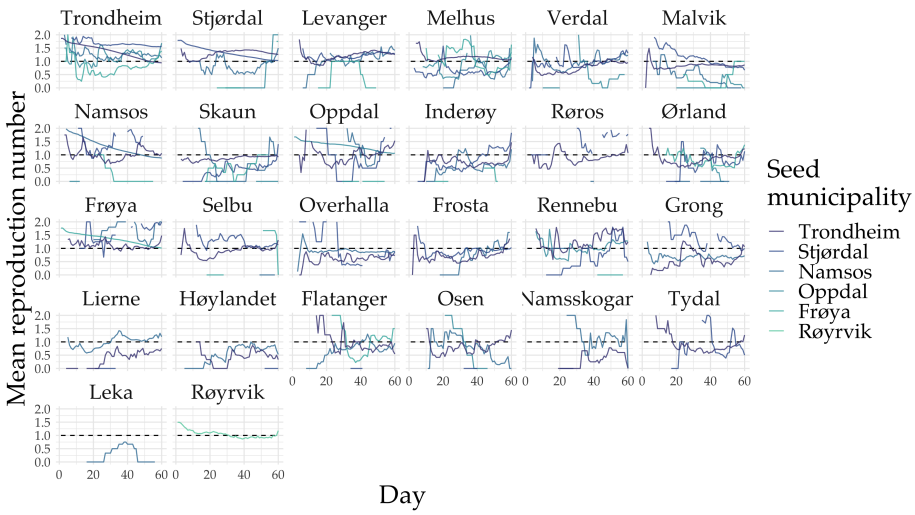


Figure 4.2.14: Average daily reproduction number for different seeding municipalities. Calculated for each municipality after 100 simulations of each different seed.

4.2.5 Effect of Initial Outbreak Size

A starting prevalence of 0.5% (the default value) means around 900 infected in Trondheim, but only 2-3 in Røyrvik. Similar patterns have been observed during the pandemic in Norway. Small municipalities have little disease spread until suddenly 1 or 2 inhabitants get infected from another area and spread the disease to their home place, starting local outbursts. It is vital to see how the initial outbreak size effects the reproduction number and variance between municipalities. Four different prevalence values were simulated, and the results are shown in Figure 4.2.15.

The results show significant variance in the average R-number for some municipalities, but not for all. In Trondheim, for example, the difference between a starting prevalence of 0.05 and 0.005 is slight, but the R-number quickly fall for values below this. The stochasticity increases significantly with the lowest two prevalence values , which makes sense as a random draw of 0.0005 would mean only a handful infected in a small municipality.

When comparing the number of infected individuals as a function of starting prevalence, on the other hand, the story is different. Figure 4.2.16 shows the average number of infected, and the results vary by a considerable factor. A starting prevalence of 0.05 gives a massive number of infections early, while a prevalence of 0.005 gives a much smoother and slower rising curve. For a simulation of 60 days, this seems to work out well, as there are also enough infected in the smaller municipalities. The overall trends for each municipality are surprisingly similar for the different prevalence values, indicating the impact of initial values for the simulations once more.

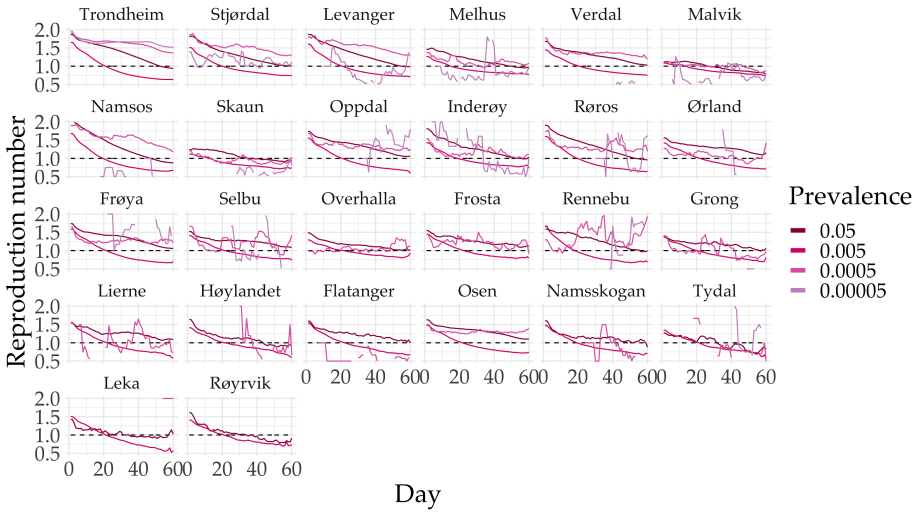


Figure 4.2.15: Average daily reproduction number for different starting prevalences. Calculated for each municipality after 100 simulations of each different prevalence value.

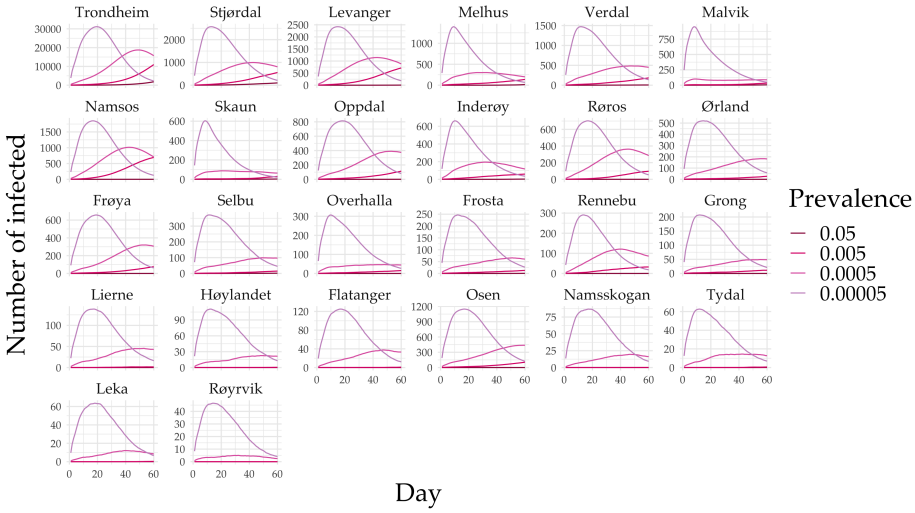


Figure 4.2.16: Average daily number of infected as a function of starting prevalence. Daily mean for each municipality after 100 simulations of each prevalence value.

4.2.6 Effect of Containment Measures

The initial model was created to emulate Norway in the spring of 2020, with a rising number of Covid-19 infections and strict containment measures to combat the disease. The base infection parameter was set based on values from the literature and adjusted according to observed numbers of infected in the country in the first few months of the pandemic. The containment measures were implemented by a strategy factor between 0 and 1 to simulate the degree of general containment measures like increased hand-washing, social distancing, fewer daily contacts and working remotely. In the model, this means a reduction in the basic infection number for the generic contact layer.

For the original model, this value was usually in the range of 0.1 to 0.2, which gave sufficient curve fitting to actual hospitalisation data. The default strategy factor has been set to 0.25 in this project to have a sufficient number of infected in municipalities of all sizes. Figure 4.2.17 shows the results for the five different strategy levels.

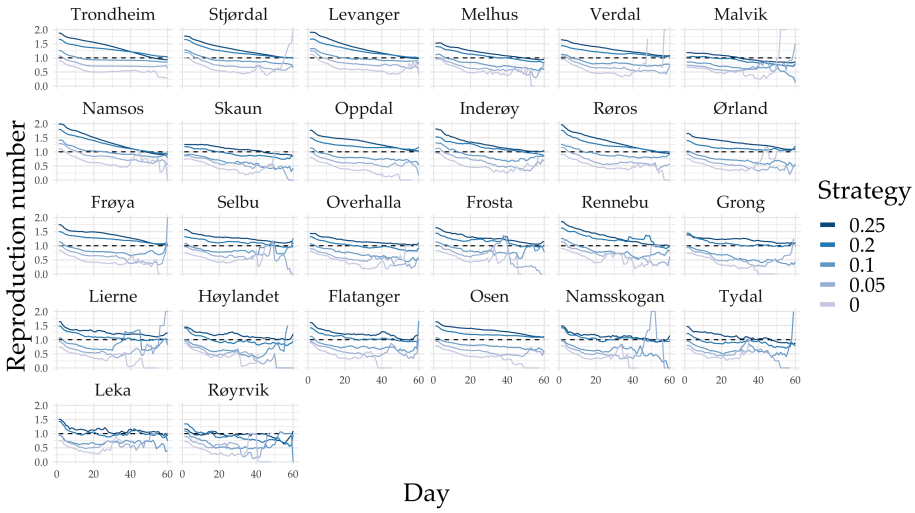


Figure 4.2.17: Average daily reproduction number for different containment strategies. Calculated for each municipality after 100 simulations of each different prevalence value.

As can be seen, a lower number gives a slightly lower reproduction number, and the decrease is relatively steady for a slight decrease in the input values. A strategy value of 0.25 versus 0.2 seems to give a slight difference in reproduction number the first few days, but not in the long run. This difference might be because the number of infections happening in the generic contact layer is large the first few days, but later it is more impacted by the spread in workplaces and households, for instance.

However, when looking at the number of infected for the different values, the story is different. Figure 4.2.18 shows a massive increase in the number of

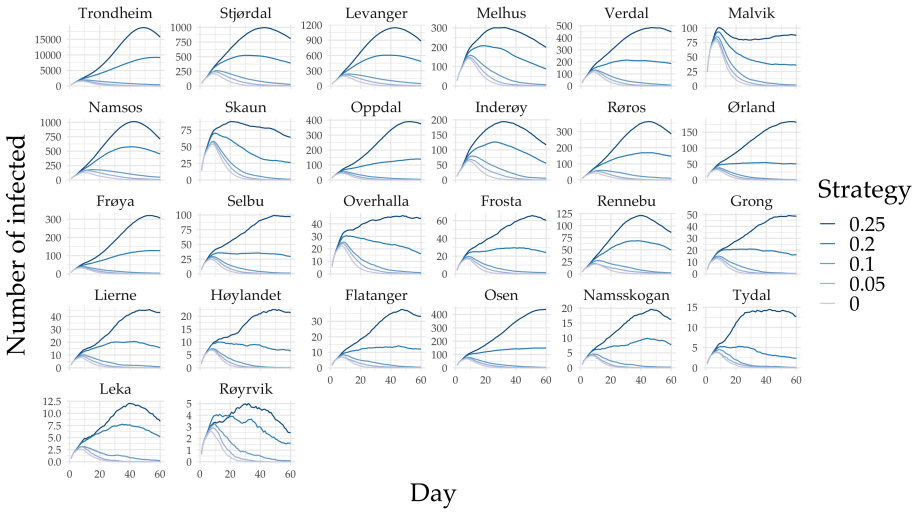


Figure 4.2.18: Average daily number of infected as a function of containment measures. Daily mean for each municipality after 100 simulations of each prevalence value.

infected for even slight changes in the strategy parameter. A decrease from 0.25 to 0.2 gives a 50% reduction in the number of infected for many municipalities. A small change in the effective reproduction number from 1.3 to 1.2, for instance, will give an exponential decrease in the new number of infected. Compare the two upper curves of Stjørdal, where the decrease in the reproduction number is nearly negligible, but the number of infected on day 40 decreases from around 1000 to under 500. A further halving in the strategy value gives close to no infected after day 30. The same trend can be seen for most of the municipalities.

In other words, the infection chance in the generic contact layer influences the number of infected in the model significantly.

4.2.7 Fractional Experiment

The final simulation was a simple fractional experimental setup, where the variables were set to either high or low, and 100 repetitions were done to explore the differences between variables. Table 4.2.1 shows the values used for the different variables.

Table 4.2.1: Values used in the fractional simulation setup.

Variable	High	Low
Strategy	0.10	0.25
Commuters	0	1
Mutation infectivity	1	2
Prevalence	0.0005	0.005

A linear regression model was created, and the results are shown in Table 4.2.2. It can be seen that all the coefficients are significant except the commuter degree. This trend is in line with the previous results. The differences between the estimates show that the strategy and mutation parameter has a greater influence on the reproduction number than the prevalence factor. An analysis of variance was also done, with results indicating the same trend, shown in Table 4.2.3.

Table 4.2.2: Coefficient estimates for the linear regression model. The model was created from the average value in Trøndelag after 100 runs. All values are scaled and centred. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Coefficients	Estimate	Std. Error	t value	*p*-value	Signif.
(Intercept)	1.793	0.005	356.757	0.000	***
Commuters	0.009	0.005	1.871	0.061	.
Mutations	0.381	0.005	75.819	0.000	***
Strategy	0.490	0.005	97.396	0.000	***
Prevalence	-0.035	0.005	-6.971	0.000	***

Table 4.2.3: ANOVA results of Trøndelag. The model was created from the average value in Trøndelag after 100 runs. All values are scaled and centred.
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Variable	Df	Sum Sq	Mean Sq	F value	*p*-value	Signif.
Commuters	1	0.142	0.142	3.502	0.061	.
Mutations	1	232.433	232.433	5748.504	0.000	***
Strategy	1	383.551	383.551	9485.947	0.000	***
Prevalence	1	1.965	1.965	48.596	0.000	***
Residuals	1595	64.492	0.040	-	-	

4.3 National Model

The previous section looked at the different parameters in the model and their effect on simulations on the Trøndelag region. However, the model was designed to be scalable based on a list of input municipalities, and one goal has been to simulate the entire country of Norway. The following section will go through some of the main results from these simulations.

All the parameters from the previous section are assumed to have similar impacts on the national model simulations. The default setup will be used for the national results unless otherwise noted.

The first result is shown in Figure 4.3.1, showing the daily reproduction number for each county, with each line representing a municipality. This figure gives an impression of the different number of municipalities in each county and the overall trends. They are very similar to what has been presented previously. Like the one large bump in Nordland, some outliers can be noted, but all in all, nothing much surprising.

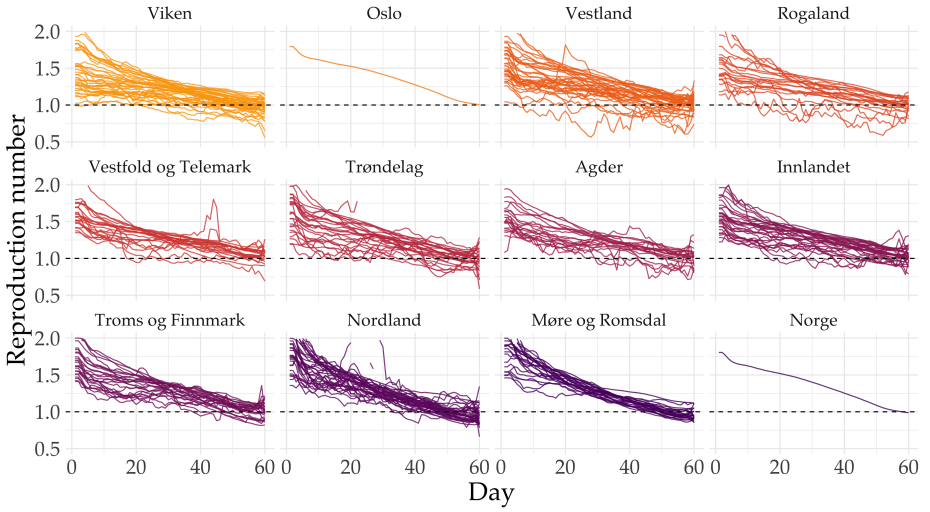


Figure 4.3.1: Average daily reproduction number for each county in Norway. Calculated for each municipality after 40 runs. The lower right plot shows the overall trend for the entirety of Norway.

The number of infected for the same simulations are shown in Figure 4.3.2. Here it shows significant variations between and within counties, but this comes as no big surprise, as the population size varies tremendously between municipalities.

Taking a closer look at the distributions of the eleven counties, Figure 4.3.3 shows the average reproduction number for all the municipalities in Norway. The counties are again sorted after decreasing population size, and the average reproduction number for the default parameter values seems to be between

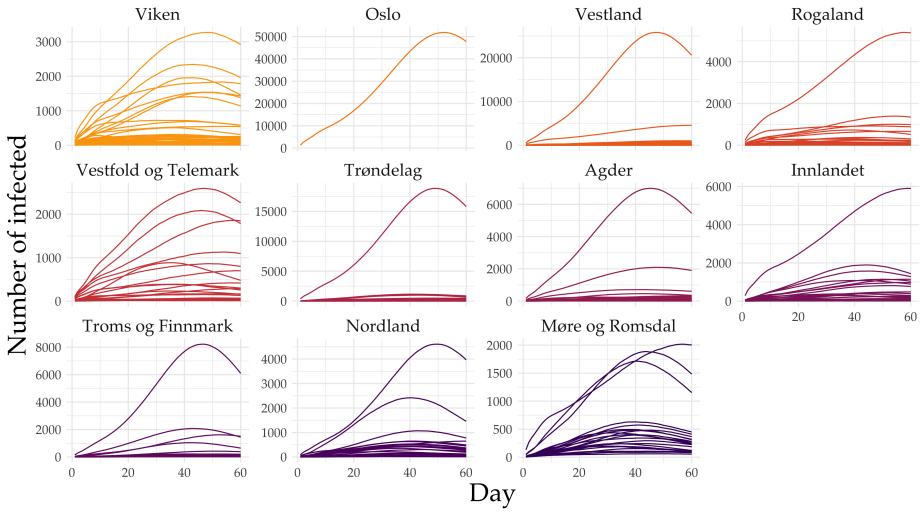


Figure 4.3.2: Average daily number of infected for each county in Norway. Calculated for each municipality after 40 runs.

around 1.3, quite similar to the Trøndelag simulations. While it may seem to be a trend that variation increases as a function of county size, this may be a side-effect of larger counties having more municipalities and of more varying sizes. There are quite a few municipalities with an average R-number of below 1.0 and a few with very high numbers of over 1.5.

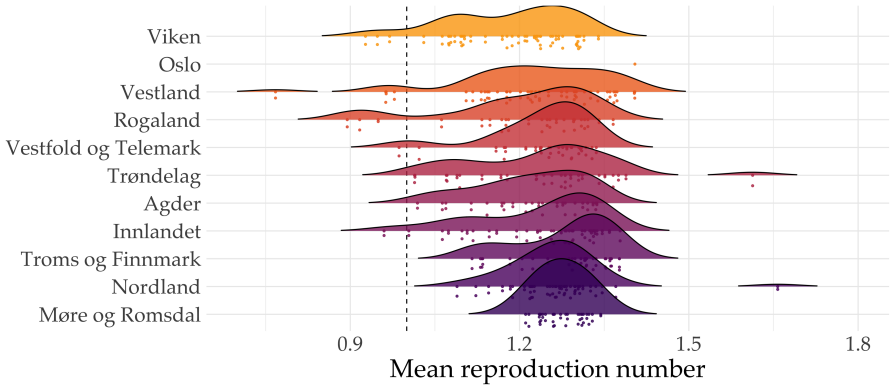


Figure 4.3.3: Distribution of mean reproduction number for all counties in Norway. Calculated after 40 simulations.

4.3.1 Effect of Population Demographics

The variance in the reproduction number for different counties and municipalities is large, but how significant is it? Which demographic parameters influence the reproduction number the most? While most of the analysis done on the Trøndelag data could be repeated for the Norway simulations, the following section will take a closer look at the overall trends for different population demographics instead. Figure 4.3.4 shows the average reproduction number as a function of population size, population density and mean age.

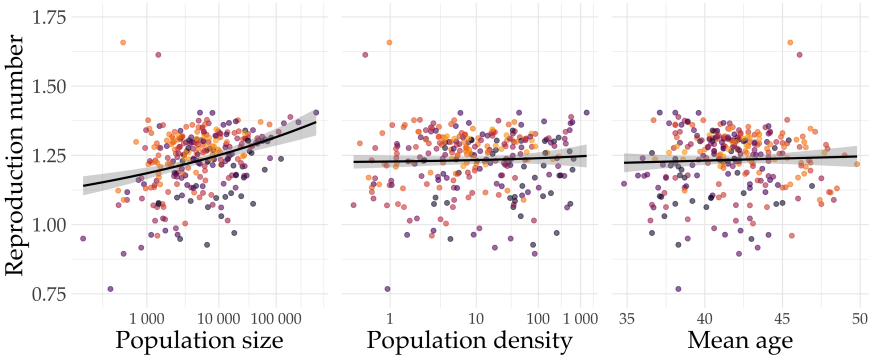


Figure 4.3.4: Reproduction number as a function of different population demographics. All municipalities in Norway, average after 40 runs.

The leftmost plot shows the effect of population size on the average reproduction number. There seems to be a correlation between population size and reproduction number, with an increasing size means a slightly higher R-number. However, there is some variance in the trend, shown by the grey area. In addition, while the trend line is increasing steadily, there is a large spread in the municipalities in the middle of the plot, where most of the municipalities are placed by population size. To conclude, there seems to be some correlation for the smallest and largest municipalities in Norway, but the effect on medium-sized municipalities is not apparent.

The middle plot shows the population density, which is another factor that might influence the average reproduction number. A dense population would mean more people in a smaller area, which could, in theory, increase the disease spread. However, there is no significant trend showing on the plot, and it seems population density does not explain the variance in R-numbers between municipalities.

The rightmost plot shows the distribution of mean ages for all municipalities. As with the population density plot, there is no significant correlation between the two, and the effect seems to be negligible.

Note that these plots only display linear trends of the data. There could be higher-order effects or interactions between several variables, but this has not been investigated further.

4.3.2 Effect of Commuters

Finally, the effect of commuter degree was explored in Figure 4.3.5. This plot shows the fraction of incoming commuters on the left-hand side and outgoing commuters on the right side.

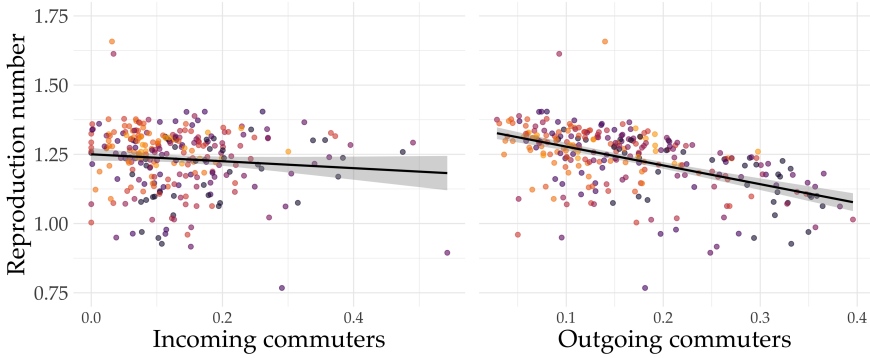


Figure 4.3.5: Reproduction number as a function of commuter fraction. All municipalities in Norway, average after 40 runs.

It is not simple to draw conclusions from the left plot, as most municipalities have a fraction of between 0.0 and 0.2 incoming commuters, with a few having much larger fractions. There seems to be a slight linear trend. However, the deviation is larger than the decrease in the trend line.

For the outgoing commuters, on the other hand, the linear trend is more apparent. There is a significant decrease in reproduction numbers as the fraction of outgoing commuters increase. At first, this trend might seem strange since more commuters would logically imply more infections, even though the effect of commuters has not been significant previously. However, the fraction of commuters is only one part of the picture.

Figure 4.3.6 shows the actual number of commuters instead of the fraction. Here it seems like a slight increase for the incoming commuters on the left plot and a similar trend on the right plot. Therefore, a possible explanation for the last plot is that municipalities with a high fraction of outgoing commuters might also often be small municipalities, which typically exhibit slightly lower reproduction numbers.

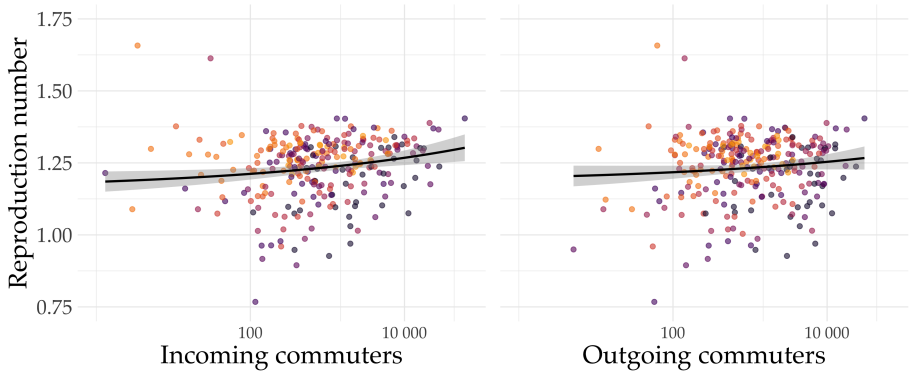


Figure 4.3.6: Reproduction number as a function of number of commuters. All municipalities in Norway, average after 40 runs.

4.3.3 Statistical Analysis

A linear regression model was created to test the effect of all demographic variables simultaneously. The resulting R^2 was 0.026, with a p-value of zero. The regression coefficients are shown in Table 4.3.1, and the results after ANOVA is shown in Table 4.3.2.

Table 4.3.1: Coefficient estimates in the linear regression model. The model was created from the average value for each municipality after 40 runs on the national model. All values are scaled and centred.
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Coefficient	Estimate	Std. Error	t value	p-value	Signif.
(Intercept)	1.227	0.004	338.109	0.000	***
Population	0.018	0.015	1.216	0.224	
Population density	-0.013	0.008	-1.627	0.104	
Age	-0.019	0.004	-4.218	0.000	
Commuters in	-0.044	0.016	-2.700	0.007	**
Commuters out	0.059	0.010	5.758	0.000	***
Commuter fraction in	0.006	0.005	1.285	0.199	
Commuter fraction out	-0.076	0.005	-16.448	0.000	***

The regression model shows a significant p-value for both the number and fraction of outgoing commuters. There is a relatively significant p-value of 0.007 for the number of incoming commuters (assuming a significance level of 0.95).

An ANOVA was performed to test the variance between the groups further. It turns out the variables explaining the variance in the data set best was the population size and density, with the outgoing commuter fraction being significant as well. Mean age in a municipality is not significant, and the number of commuters is not significant.

Comparing the two tables, several points are worth noting. While population size and density are significant, they do not contribute to the estimate as much as the commuter fraction. It is also interesting that the outgoing commuter fraction is significant, but not the number of outgoing commuters. This fact might point to a non-linear relationship between the two.

Table 4.3.2: ANOVA results of Norway demographic data. The model was created from the average value for each municipality after 40 runs on the national model. All values are scaled and centred.
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Variable	Df	Sum Sq.	Mean Sq	F value	*p*-value	Signif.
Population	1	6.5	6.5	43.57	0.000	***
Population density	1	2.4	2.4	16.0	0.000	***
Age	1	0.0	0.0	0.23	0.629	
Commuters in	1	0.7	0.7	4.98	0.026	*
Commuters out	1	0.0	0.0	0.03	0.867	
Commuter fraction in	1	0.4	0.4	2.75	0.097	.
Commuter fraction out	1	40.1	40.1	270.531	0.000	***
Residuals	11252	1667.6	0.148	-	-	

ANOVA rests on the assumption of a linear relationship in the data, which might not be valid for all variables presented here, like the number of commuters and the population size. Still, the statistical analysis done hopefully helps explain some of the variance in the reproduction numbers of the simulations, even though some of the assumptions in the underlying analysis might be uncertain.

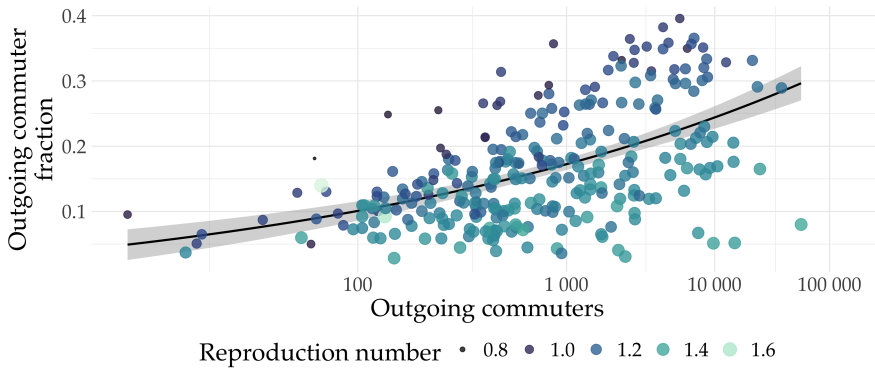


Figure 4.3.7: The relationship between outgoing commuter fraction and numbers. The line indicates the smoothed linear trend. Logarithmic horisontal axis.

To further investigate the relationship between outgoing commuter fraction and number, Figure 4.3.7 was created. The plot shows a fascinating pattern where the linear smoothed average seems to divide the results into two. The

municipalities with many outgoing commuters but a low commuter fraction have R-numbers much higher than for the opposite case. With one notable exception, all the highest R-numbers are on the lower side of the curve. The same visualisation was done for the incoming commuters, but a similar trend was not visible. The plot can be found in Appendix C.3.1.

Similar higher-order interactions may exist in the results, but this has not been investigated further due to a lack of time.

CHAPTER 5
DISCUSSION

"All models are wrong, but some are useful" - George Box[51]

As George Box now famously said in 1976, models have always been approximations of reality. They try to describe a system as best as possible but have shortcomings, either in their assumptions, complexity or understanding of the underlying system.

So is the case with the modelling framework presented throughout this thesis. The model tries to simulate the spread of SARS-CoV-2 in a social network of Norway. It includes several assumptions and simplifications to make simulations feasible and efficient while having the ability to implement containment measures and different strategies. This chapter will discuss some of the most critical assumptions, as well as list possible further work.

First of all, the disease parameters. While the SARS-CoV-2 pandemic has lasted for close to 1.5 years now, extensive research is still ongoing. Several key aspects of the virus and disease are still under debate, like the effect of airborne transmission[52], the degree of asymptomatic patients[53, 54], and why some people experience severe complications or so-called "Long-COVID"[55, 56].

Virus transmission parameters used in the model has not been a focus during this project. Except for the simplified mutation infectivity, which aims to simulate a linear increase in transmission chance, the viral parameters have been kept from the original model values. Time could be spent comparing emerging research throughout the world to get the most up-to-date characteristics of the virus, but this has not been prioritised.

Second, the population and social network in the model has been based on the already existing data set, which includes several simplifying assumptions. The databases from Statistics Norway are assumed to be regularly updated and sufficiently accurate, but the social network is only a representation of the

population in Norway. In addition to the issues discussed in Section 5.2.2, the activity of each individual is a parameter influencing model output significantly. This is drawn from two different distributions, one for children and the elderly, and one for the rest of the population, a divide that might be too simplified. An example can be seen in Norway, where young people in their twenties have been overrepresented in the disease statistics for several months. This parameter has been held constant in the model and therefore does not consider a more heterogeneous age distribution, day of the week, public holidays, or geographic differences.

Another simplification worth noting is that the model, first and foremost, has been used to predict differences in the reproduction number for different containment strategies. While statistics like the daily number of contacts, the number of individuals in each disease state and in which layers infections happen can be extracted from the model, this has not been a focus. This simplification might overlook the complicated dynamics happening on a micro-scale in different municipalities, layers and cliques, and only focus on the macro-scale results. The overall results should hopefully still be valid, but insight into important epidemic characteristics might be missed.

In summary, these three simplifications and limitations might influence some conclusions drawn from this model. However, all models are only representations of a system. It is assumed that the modelling framework can still give valuable insight into the dynamics of a regional or national epidemic of SARS-CoV-2.

5.1 Key Assumptions

In addition to the overall simplifications described above, there are some key assumptions included in the model. This section will discuss the most important ones.

Simplified contact patterns: The generic contact layer is intended to simulate what might be called "random contacts", or infections from an unknown source in a disease setting. This layer simplifies many types of possible disease transmission, like going to the store, bumping into someone on public transport, or meeting an old acquaintance on your way home. These examples will have very different possibilities of disease transmission, which is not explicitly modelled. As we have learned during the pandemic, there is a significant difference in infection based on where people interact and for how long. Factors like ventilation and indoor area, for example, might influence the risk of disease spread tremendously[57].
In addition, the heterogeneity of social interactions might be too simplified. There is no explicit simulation of different types of social contact outside cliques, for example. Let us say a group of friends from different cliques meet one afternoon weekly and interact for a couple of hours. This form of contact, which might be described as somewhere in-between daily clique contacts and the generic contact layer, is not possible.

No travel except commuting: The model includes commuters travelling between two municipalities, but no other form of travel within the country. At the start of the pandemic, this assumption was relatively realistic. Later, as Norway opened up more, in combination with vacations and holidays, not so much. However, as commuters in the model have not significantly impacted the average reproduction number, it is assumed this is negligible for most simulation cases tested in this project.
In addition, it is assumed that there are no imported cases of Covid-19 from abroad.

No geographical dimension in the model: The lack of a geographical component in the model means that in a given municipality, every single inhabitant has the potential to meet anyone else each day. This assumption might be realistic for smaller municipalities but not for large cities like Oslo, with several hundred thousand inhabitants. In addition, no within-municipality geographical information is explicitly included, like population density or area size. A consequence of this is that contact between two individuals is the only way the virus can spread. There have been reports of Covid-19 transmission without direct human-to-human interaction, and the impact of disease spread on surfaces should not be neglected[58].

Simplified containment measures: The containment measures implemented in the model mainly consist of two types. The first is reducing the basic reproduction number in the generic contact layer, meant to simulate increased handwashing, social distancing and a reduced number of contacts. The second is the partial or complete close-down of different layers. This can be a fraction of workplaces closing down or schools closed for given age groups, for example. However, real-life containment measures have been complex, differentiated and varying throughout the entire pandemic. How do we turn the different rules into quantitative and tunable parameters in the agent-based model? This is no easy task, and there is no correct answer.

Compliance to containment measures: The model assumes that all individuals follow the rules and measures in place. If a strategy implements reductions of commuting closes down workplaces, all agents in the model comply. Looking at the situation the last year in Norway, one area of interest is the differences in rules and containment measures in different municipalities, combined with the compliance. The variances have been considerable and could be interesting to simulate.

No testing: The original model also included different testing measures. Testing is implemented in the national model but has not been turned on for the simulations in this project. In a real-life scenario, extensive testing would be a pivotal factor to combat the pandemic. It would presumably reduce the average reproduction number significantly, as a vast number of infected are tested and put into quarantine before spreading the virus further.

5.2 Challenges

There have been several small and large challenges throughout the project. Two of the more impactful ones have been the run time of the simulations and issues with the population data sets.

5.2.1 Run Time

The goal of the model is to simulate entire regions or even the country across dozens of parallel simulations with different parameter values. This makes hundreds of different simulations on a population of several hundred thousands or even millions of individuals. It comes as no surprise then that the model's run time is a significant factor influencing the ability to do simulations.

During the first part of the project, shortly after the model was made object-oriented, the model was profiled extensively to investigate possible bottlenecks in the code or underlying logic. Initially, the model looped through every single individual for every single simulation day, which quickly turns into million of calculations even for small regions or municipalities. Two particular points of interest were discovered.

Clique Infections

The first is how infections in each clique were simulated. Previously, this was done by counting the number of infected in a clique and then calculating the infection risk for the remaining nodes. However, most cliques do not have any infected on a typical day during a simulation with default parameters. The solution was to give each clique an attribute that kept count of the number of infected. If this was zero, the entire clique calculations were skipped for that day. This simple check reduced the number of daily clique simulations by several orders of magnitude and sped up the overall simulation times tremendously.

Generic Contact Layer

The second point of interest was the contacts in the generic contact layer. Instead of calculating the number of contacts for every infected individual present in the layer, an average prevalence was calculated for the entire layer, and newly infected nodes determined by a random draw against this value as a function of the node's activity. This simplification reduces the number of calculations made for every simulation day, which can be especially useful for more extensive simulations.

5.2.2 Population Data Issues

As mentioned briefly in Section 4.1, there were some challenges with the underlying datasets the social network is created from. Several municipalities are non-functional or have some unusual features. It is assumed most of these issues

arise from unforeseen problems with the network generating script created a year ago, in combination with the municipality amalgamation. The non-functional municipalities were simply removed from the model, but another issue with the data becomes clear when plotting actual population sizes against the population in the model. This issue is illustrated in Figure 5.2.1, where the worst offenders are labelled.

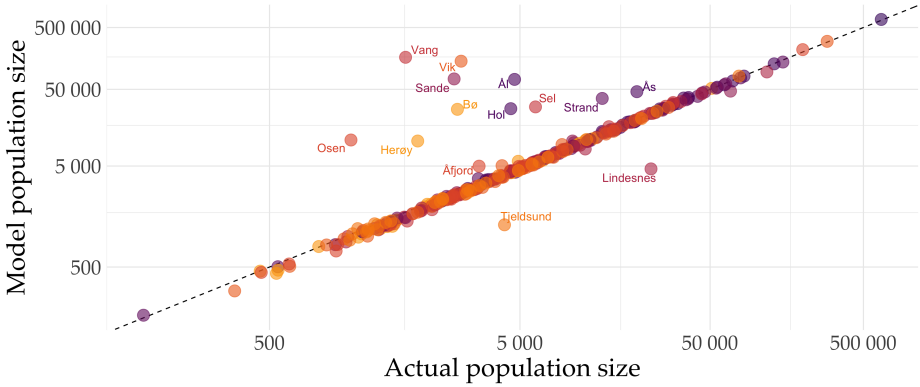


Figure 5.2.1: Actual population size against model population size. Municipalities that are over 50% larger or smaller than they should be are labelled by name.

As can be seen, there are a dozen municipalities with significantly larger population size in the model compared to actual population sizes. These were removed from the network, as several large artificial municipalities could have unforeseen consequences. The most probable explanation is naming conflicts, as several of the troublesome municipalities have short names that are subsets of some of the largest municipalities in Norway, for instance, Ål (Ålesund), Sande (Sandefjord) and Vang (Stavanger). However, this naming conflict does not explain the smaller municipalities, but the possible reasons could be several. Merged or split municipalities, non-functional data queries from Statistics Norway, or

Table 5.2.1: Actual population size against model population size. The "Actual" row is data from January 2020. "Unaltered" are the numbers with all functional municipalities, and "Corrected" are the numbers with the removed municipalities mentioned above.

	Model population	Actual population	Municipalities
Actual	-	5 367 580	356
Unaltered	4 972 444	4 517 136	308
Corrected	4 344 374	4 427 237	294

Table 5.2.1 summarises the population size and number of municipalities in the model before and after correction. After corrections, the model includes 86%

of the population of Norway and 82% of the municipalities. Sixty-two municipalities are removed, either due to data issues or too large or small populations. For this project's scope, this was decided to be acceptable but is a clear area of improvement for further work.

5.3 Further Work

There are several possibilities for further extending the model. Many of the points mentioned in Section 5.1 are obvious candidates for improvements, like simulating testing and intermittent travelling within the country, and differences in compliance to containment measures.

One interesting simulation protocol to look closer at would be a more realistic representation of the pandemic in Norway this last year. Elements to investigate could be an initial number of infected in Oslo and Viken, and gradual closure of different municipalities throughout the pandemic. Factors like different containment strategies in different counties and municipalities as a function of infected in a given geographical region could be implemented, in addition to prioritised test capacity and vaccinations.

In addition to the aspects as mentioned above, some other areas of improvement include, but are not limited to:

Geographical dimension: The model includes a temporal dimension, but an explicitly modelled spatial dimension would increase the realism. A geographical dimension would also add the possibility of different types of disease transmission, not only from human-to-human interactions, but include surface transmission, for instance.

Network analysis: Further analysis of the network dynamics of the national model has not been a priority, but the possibilities are many. How do infected individuals spread the disease throughout a clique, a layer and a region? How many commuters are infected, and how much do they influence spread in their home municipalities. How does the virus diffuse through a region after a single area is infected?
A simple edge network was created from the agents in the model, but this has not been analysed thoroughly and could yield valuable insight into the model and disease dynamics.

CHAPTER 6
CONCLUSION AND OUTLOOK

SARS-CoV-2 has ravaged the world for close to one and a half years. The pandemic has turned life upside down for a majority of the population. While vaccinations have come far in many countries, several years might remain until everything is back to normal. Creating models to get a deeper insight into infectious diseases like Covid-19 is critical, both to understand the current situation and future pandemics from undiscovered pathogens.

In this project, the agent-based Covid-19 model created at NTNU in the spring of 2020 has been extended to simulate the entire country of Norway based on actual population demographics and properties of SARS-CoV-2. Commuter travelling and a municipality network has been implemented in the Python code. The main parameters can be tuned by user input, and the model framework has been tested with different parameter values. The Trøndelag region has been used as a case study to investigate different model aspects like commuters, mutations, initial conditions, containment measures, and demographic data.

It was found that the average reproduction number varied significantly based on model input and the demographic properties of the simulated municipalities. The most important factors determining the average reproduction number in a municipality were found to be population size, population density, and the fraction of outgoing commuters. The number of commuters in a municipality was not a statistically significant factor, and changes in the amount of commuters present did not affect the simulation results in Trøndelag considerably.

The agent-based model is a robust and theoretical simulation framework for investigating SARS-CoV-2 spread in Norway, but it has not been tested against current developments in the pandemic. Disease dynamics are complex and influenced by stochastic processes, and the results presented here are not generalisable to other diseases or countries with different population demographics.

Further work should explore different simulation situations in Norway, with more realistic parameter values and initial conditions, in addition to heterogeneity

in containment strategies between municipalities and counties. Implementing an explicit geographical dimension in the model is a possible long-term goal for future research.

The results presented in this project lay the groundwork for a flexible and large-scale agent-based model that can be utilised to simulate different diseases in diverse regions. Furthermore, they can also give insight into the ongoing pandemic and the spread of future human pathogens.

BIBLIOGRAPHY

[1] ECDC. *COVID-19 situation update worldwide, as of week 21, updated 10 June 2021*. 2021. URL: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases.

[2] Samuel Asumadu Sarkodie and Phebe Asantewaa Owusu. "Global assessment of environment, health and economic impact of the novel coronavirus (COVID-19)". In: *Environment, Development and Sustainability* 23.4 (2021), pp. 5005–5015. URL: https://doi.org/10.1007/s10668-020-00801-2.

[3] Robert Rowthorn and Jan Maciejowski. "A cost–benefit analysis of the COVID-19 disease". In: *Oxford Review of Economic Policy* 36.Supplement_1 (2020), S38–S55. URL: https://doi.org/10.1093/oxrep/graa030.

[4] Mohammad Yamin. "Counting the cost of COVID-19". In: *International Journal of Information Technology (Singapore)* 12.2 (2020), pp. 311–317. URL: https://doi.org/10.1007/s41870-020-00466-0.

[5] J N Hays. *Epidemics and Pandemics: Their Impacts on Human History*. ABC-CLIO, 2005. URL: https://books.google.no/books?id=GyE8Qt-kS1kC.

[6] Noelle-Angelique M Molinari et al. "The annual impact of seasonal influenza in the US: Measuring disease burden and costs". In: *Vaccine* 25.27 (2007), pp. 5086–5096. URL: https://www.sciencedirect.com/science/article/pii/S0264410X07003854.

[7] Velislava N. Petrova and Colin A. Russell. "The evolution of seasonal influenza viruses". In: *Nature Reviews Microbiology* 16.1 (2018), pp. 47–60. URL: http://dx.doi.org/10.1038/nrmicro.2017.118.

[8] Wuqi Qiu et al. "The Pandemic and its Impacts". In: *Health, Culture and Society* 9 (Dec. 2017), pp. 1–11.

[9] J. M. Gran et al. “Estimating influenza-related excess mortality and reproduction numbers for seasonal influenza in Norway, 1975–2004”. In: *Epidemiology and Infection* 138.11 (2010), pp. 1559–1568.

[10] WHO. *Pandemic influenza*. 2021. URL: https://www.euro.who.int/en/health-topics/communicable-diseases/influenza/pandemic-influenza.

[11] Peter Spreeuwenberg, Madelon Kroneman, and John Paget. “Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic”. eng. In: *American journal of epidemiology* 187.12 (Dec. 2018), pp. 2561–2567. URL: https://pubmed.ncbi.nlm.nih.gov/30202996%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7314216/.

[12] Frank R DeLeo and B Joseph Hinnebusch. “A plague upon the phagocytes”. In: *Nature Medicine* 11.9 (2005), pp. 927–928. URL: https://doi.org/10.1038/nm0905-927.

[13] Marc P Girard et al. “The 2009 A (H1N1) influenza virus pandemic: A review”. In: *Vaccine* 28.31 (2010), pp. 4895–4902. URL: https://www.sciencedirect.com/science/article/pii/S0264410X1000719X.

[14] Wei Duan et al. “Mathematical and computational approaches to epidemic modeling: a comprehensive review”. In: *Frontiers of Computer Science* 9.5 (2015), pp. 806–826.

[15] Zhen Z. Shi, Chih-Hang Wu, and David Ben-Arieh. “Agent-Based Model: A Surging Tool to Simulate Infectious Diseases in the Immune System”. In: *Open Journal of Modelling and Simulation* 02.01 (2014), pp. 12–22.

[16] Thomas E. Gorochowski. “Agent-based modelling in synthetic biology”. In: *Essays in Biochemistry* 60.4 (2016). Ed. by Vitor B Pinheiro, pp. 325–336. URL: https://doi.org/10.1042/EBC20160037.

[17] Sameera Abar et al. “Agent Based Modelling and Simulation tools: A review of the state-of-art software”. In: *Computer Science Review* (2017).

[18] Helge Bergo. *Agent-Based Modelling of SARS-CoV-2 Spread in a Public Transport System*. Tech. rep. Trondheim: Norwegian University of Science and Techonology, 2020.

[19] Nicholas C. Grassly and Christophe Fraser. “Mathematical models of infectious disease transmission”. In: *Nature Reviews Microbiology* (2008).

[20] Abby Helen. “An Examination of the Reed-Frost Theory of Epidemics”. In: *Human Biology* 24.3 (1952).

[21] W. O. Kermack and A. G. Mckendrick. “A contribution to the mathematical theory of epidemics”. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 115.772 (1927), pp. 700–721.

[22] Herbert W. Hethcote. “The Mathematics of Infectious Diseases”. In: *SIAM Review* 42.4 (2000), pp. 599–653.

[23] Albert László Barabási. "The origin of bursts and heavy tails in human dynamics". In: *Nature* 435.7039 (2005), pp. 207–211.

[24] NPHI. *Coronavirus modelling at the NIPH*. 2020. URL: https://www.fhi.no/en/id/infectious-diseases/coronavirus/coronavirus-modelling-at-the-niph-fhi/.

[25] Fiona M. Guerra et al. "The basic reproduction number (R0) of measles: a systematic review". In: *The Lancet Infectious Diseases* 17.12 (2017), e420–e428. URL: http://dx.doi.org/10.1016/S1473-3099(17)30307-9.

[26] Brian J. Coburn, Bradley G. Wagner, and Sally Blower. "Modeling influenza epidemics and pandemics: Insights into the future of swine flu (H1N1)". In: *BMC Medicine* 7 (2009), pp. 1–8.

[27] Steven Sanche et al. "RESEARCH High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2". In: *Emerging Infectious Diseases* 26.7 (2020), pp. 1470–1477.

[28] J. M. Heffernan, R. J. Smith, and L. M. Wahl. "Perspectives on the basic reproductive ratio". In: *Journal of the Royal Society Interface* 2.4 (2005), pp. 281–293.

[29] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[30] Stephen Eubank et al. "Modelling disease outbreaks in realistic urban social networks". In: *Nature* 429.6988 (2004), pp. 180–184.

[31] Guillaume Fournie et al. "Interventions for avian influenza A (H5N1) risk management in live bird market networks". In: *Proceedings of the National Academy of Sciences of the United States of America* (2013).

[32] Marcelo N. Kuperman. "Invited review: Epidemics on social networks". In: *Papers in Physics* 5.0 (2013).

[33] Muaz Niazi and Amir Hussain. "Agent-based computing from multi-agent systems to agent-based models: a visual survey". In: *Scientometrics* 89.2 (2011), p. 479. URL: https://doi.org/10.1007/s11192-011-0468-9.

[34] Emilio Serrano, Carlos A. Iglesias, and Mercedes Garijo. "A novel agent-based rumor spreading model in twitter". In: *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*. 2015.

[35] Patrick Manser et al. "Designing a large-scale public transport network using agent-based microsimulation". In: *Transportation Research Part A: Policy and Practice* 137.September 2019 (2020), pp. 1–15. URL: https://doi.org/10.1016/j.tra.2020.04.011.

[36] Nora Rosvoll Finstad. *Computational SIS Modeling of the Spread of Antibiotic Resistance within Bacterial Metapopulation Networks*. Tech. rep. May. NTNU, 2018.

[37] Theodore Gordon and David Greenspan. "The management of chaotic systems". In: *Technological Forecasting and Social Change* 47.1 (1994), pp. 49–62.

[38] Marco Hernandez, Simon Scarr, and Manas Sharma. *The Korean clusters*. 2020.

[39] Elizabeth Bruch and Jon Atwell. "Agent-Based Models in Empirical Social Research". In: *Sociological Methods and Research* 44.2 (2015), pp. 186–221.

[40] Eric Shook and Shaowen Wang. "Investigating the Influence of Spatial and Temporal Granularities on Agent-Based Modeling". In: *Geographical Analysis* 47.4 (2015), pp. 321–348.

[41] Steven Manson et al. "Methodological issues of spatial agent-based models". In: *Jasss* 23.1 (2020).

[42] Hazhir Rahmandad and John Sterman. "Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models". In: *Management Science* 54.5 (2008), pp. 998–1014.

[43] Ronald E. Walpole et al. *Probability & Statistics for Engineers and Scientists*. Ed. by Deirdre Lynch. 9th editio. Pearson, 2007.

[44] Python Software Foundation. *Python, version 3.8*. 2020.

[45] André Voigt et al. "Containing pandemics through targeted testing of households". In: *BMC Infectious Diseases* 21.1 (2021), p. 548. URL: https://doi.org/10.1186/s12879-021-06256-8.

[46] NTNU COVID-19 Taskforce. "Description of modelling framework per April 29 , 2020". In: (2020), pp. 1–21. URL: https://www.ntnu.edu/documents/36652962/0/NTNU+epidemic+spread+model+2020+v290420-2.pdf/18cfb5ac-488b-311d-046d-373b6d0e3a4d?t=1588092351148.

[47] Robert Verity et al. "Estimates of the severity of coronavirus disease 2019: a model-based analysis". In: *The Lancet Infectious Diseases* 20.6 (2020), pp. 669–677.

[48] Robert C. Martin. *Clean Code : A Handbook of Agile Software Craftsmanship*. Pearson, 2009, p. 464.

[49] Nikolay Martyushenko et al. "Containing pandemics through targeted testing of households". In: *medRxiv* (2020). URL: http://medrxiv.org/content/early/2020/11/04/2020.10.30.20219766.abstract.

[50] Regjeringen. *Navn på nye kommuner*. 2020. URL: https://www.regjeringen.no/no/tema/kommuner-og-regioner/kommunereform/nye-kommuner/id2470015/.

[51] George Box. "Science and statistics". In: *Journal of the American Statistical Association* (1976), p. 71.

[52] Michael Klompas, Meghan A Baker, and Chanu Rhee. "Airborne Transmission of SARS-CoV-2: Theoretical Considerations and Available Evidence". In: *JAMA* 324.5 (Aug. 2020), pp. 441–442. URL: https://doi.org/10.1001/jama.2020.12458.

[53] Aki Sakurai et al. “Natural History of Asymptomatic SARS-CoV-2 Infection”. In: *New England Journal of Medicine* 383.9 (June 2020), pp. 885–886. URL: https://doi.org/10.1056/NEJMc2013020.

[54] Victor M Corman et al. “SARS-CoV-2 asymptomatic and symptomatic patients and risk for transfusion transmission”. In: *Transfusion* 60.6 (June 2020), pp. 1119–1122. URL: https://doi.org/10.1111/trf.15841.

[55] Carole H Sudre et al. “Attributes and predictors of long COVID”. In: *Nature Medicine* 27.4 (2021), pp. 626–631. URL: https://doi.org/10.1038/s41591-021-01292-y.

[56] Jeffrey P Kanne et al. “COVID-19 Imaging: What We Know Now and What Remains Unknown”. In: *Radiology* 299.3 (Feb. 2021), E262–E279. URL: https://doi.org/10.1148/radiol.2021204522.

[57] Kenichi Azuma et al. “Environmental factors involved in SARS-CoV-2 transmission: effect and role of indoor environmental quality in the strategy for COVID-19 infection control”. In: *Environmental Health and Preventive Medicine* 25.1 (2020), p. 66. URL: https://doi.org/10.1186/s12199-020-00904-2.

[58] S Rawlinson, L Ciric, and E Cloutman-Green. “COVID-19 pandemic - let’s not forget surfaces”. In: *Journal of Hospital Infection* 105.4 (Aug. 2020), pp. 790–791. URL: https://doi.org/10.1016/j.jhin.2020.05.022.

Appendix

APPENDIX A
THEORY SUPPLEMENTARY

A.1 Theory Presented in the Project Report

Several of the sections presented in Chapter 2 (the theory) of this thesis is obtained from or based on material presented in the project report *"Agent-Based Modelling of SARS-CoV-2 Spread in a Public Transport System"*[18]. An overview is listed below:

Section 2.1: The material covering epidemic modelling is obtained from [18], with some textual alterations and updated figures.

Section 2.1.1: The theory about compartmental models is obtained from [18].

Section 2.1.2: The material related to network models is obtained from [18].

Section 2.1.3: The theory covering agent-based models is obtained from [18].

Section 2.2: The section on statistics is obtained from [18], with some alterations and sections removed, as well as updated figures.

APPENDIX B
METHOD SUPPLEMENTARY

B.1 Python Modules

The project was written in Python version 3.9.5, but should work fine for everything above Python 3.8. The modules used in the project are shown in Table B.1.1.

Table B.1.1: Python modules used in the project. The first modules are included in the Python Standard Library, which is why they have no version number.

Module	Description	Version
os	Operating system interfaces	-
pickle	Object serialisation	-
random	Pseudo-random numbers	-
time	Time-related functions	-
sys	System-specific parameters and functions	-
re	Regular expression operations	-
numpy	Scientific computing	1.20.3
pandas	Data analysis	1.2.1
scipy	Statistics	1.6.0
matplotlib	Visualisations	3.3.3

B.2 R Libraries

The project was written in R version 4.1.0. The libraries used in the project are shown in Table B.2.1.

Table B.2.1: R libraries used in the project. The first 8 are included in the tidyverse collection.

Library	Description	Version
ggplot2	A declarative plotting library based on *"The Grammar of Graphics"*.	3.3.3
dplyr	A tool for working with data frames.	1.0.6
tidyr	Tools for creating tidy data.	1.1.3
stringr	String manipulations.	1.4.0
readr	Read rectangular data like csv and txt files.	1.4.0
tibble	Provides the tibble data frame.	3.1.2
forcats	Helpers for reordering and modifying factor levels.	0.5.1
purrr	Functional programming toolkit for R.	0.3.4
scales	Graphical scales functions for ggplot2.	1.1.1
ggrepel	Text and label geoms for ggplot2.	0.9.1
ggridges	Ridgeline plots for ggplot2.	0.5.3
janitor	Cleaning and formatting data frames and tables.	2.1.0

B.3 Population Data

To generate the network in the model, high-resolution demographic data for each municipality in Norway was used. This comes from Statistics Norway (SSB), and all data tables can be found by accession number from www.ssb.no/en. In addition, school information was downloaded from the Norwegian National School Registry, where the API can be found on https://data-nsr.udir.no. Table B.3.1 shows the different data tables used.

Table B.3.1: Demographic data tables from Statistics Norway used to generate the social network.

Accession number	Data table text description
3321	Employed persons (aged 15-74) per 4th quarter, by municipality of work, municipality of residence, contents and year
4469	Residents in dwellings for nursing and care purposes, by age (M) 2002 - 2019
6070	Private households, by type of household (M) 2005 - 2019
6079	Private households and persons in private households, by size of household (per cent) (M) (UD) 2005 - 2019
6206	Children 0-17 years, by number of siblings and the child's age 2001 - 2019
6445	Employed persons, by place of residence, sex and age (per cent). 4th quarter (M) 2005 - 2019
8947	Pupils, apprentices, students and participants in upper secondary education, by sex, age and type of school/institution 2006 - 2019
9169	Children in kindergartens, by age, hours of attendance per week and ownership (M) 1999 - 2019
9220	Kindergartens, by ownership (M) 1987 - 2019
9929	Nursing and care institutions and beds, by ownership (C) 2009 - 2018
10308	Establishments, by the enterprises sector and number of employees (M) 2012 - 2020
11933	Care institutions - rooms, by region, contents and year
12562	Selected key figures kindergartens, by region, contents and year

APPENDIX C
RESULTS SUPPLEMENTARY

C.1 Data Exploration Supplementary

Supplementary figures and results from the Section 4.1 is shown on the following pages.

C.1.1 Highlighted Municipalities in Trøndelag

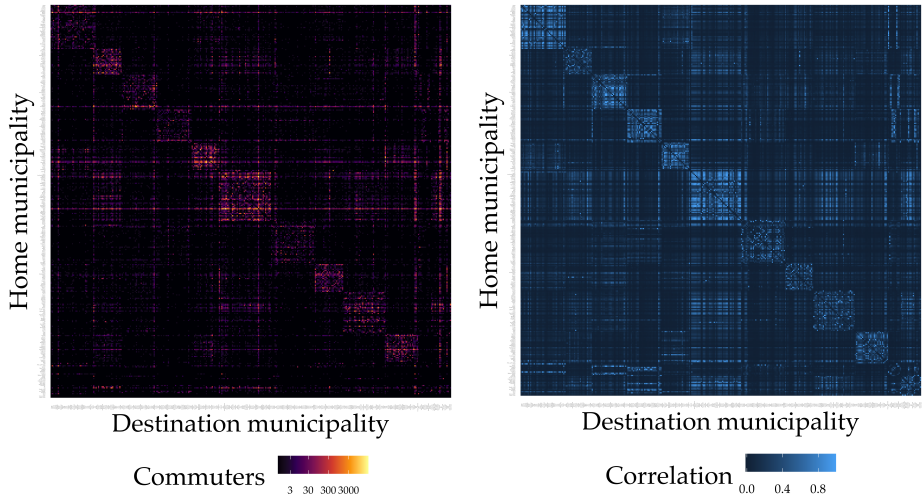
The demographic statistics for the highlighted municipalities described in Section 4.1.1 are shown in Table C.1.1.

Table C.1.1: Demographic information about the six highlighted municipalities in Trøndelag. Sorted by decreasing population size. The commuter columns are fractions of the population size.

Municipality	Population	Area	Population Density	Commuters in	Commuters Out	Mean age
Trondheim	190464	342	556.5	0.15	0.05	38
Stjørdal	23625	938	25.1	0.20	0.18	39
Namsos	13051	2132	6.1	0.12	0.07	41
Oppdal	6973	2274	3.0	0.07	0.08	42
Frøya	4937	241	20.5	0.09	0.03	39
Røyrvik	469	1584	0.3	0.00	0.12	42

C.1.2 Commuters in Norway

In Section 4.1.2, Figure 4.1.7 showed heatmaps of the commuters travelling between and within counties. Figure C.1.1 shows the same underlying data, but for every municipality in Norway, sorted by population size of the counties.



(a) Number of commuters.

(b) Commuter correlation.

Figure C.1.1: Heatmaps of commuters in all municipalities in Norway. The municipalities are sorted by largest population size, within each county.

C.2 Regional Model Supplementary

Supplementary figures and results from the Section 4.2 follows.

C.2.1 Effect of Commuters

Figure C.2.1 shows the average daily number of infected for different commuter degrees, supplementary to Section 4.2.2.

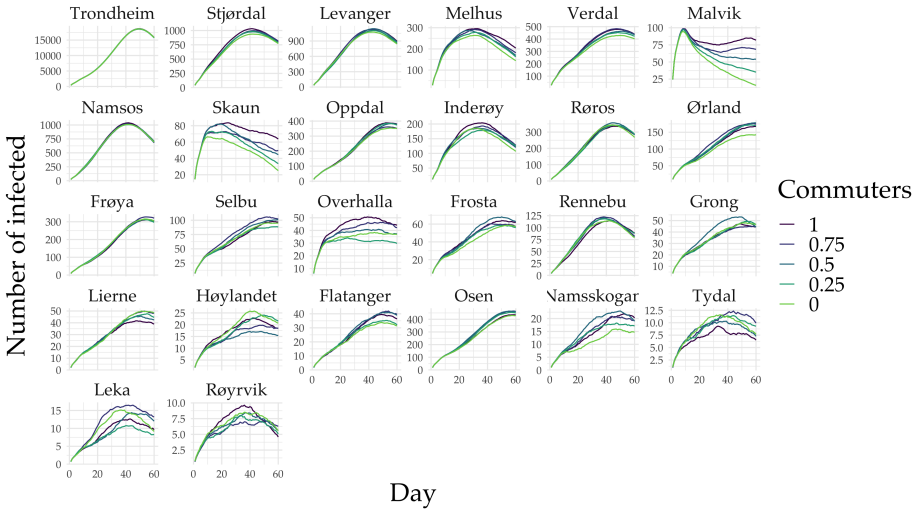


Figure C.2.1: Average daily number of infected as a function of commuter degree. Daily mean for each municipality after 100 simulations of each commuter fraction.

C.2.2 Effect of Seed Municipality

Figure C.2.2 shows the average daily number of infected for different seeds, supplementary to Section 4.2.4.

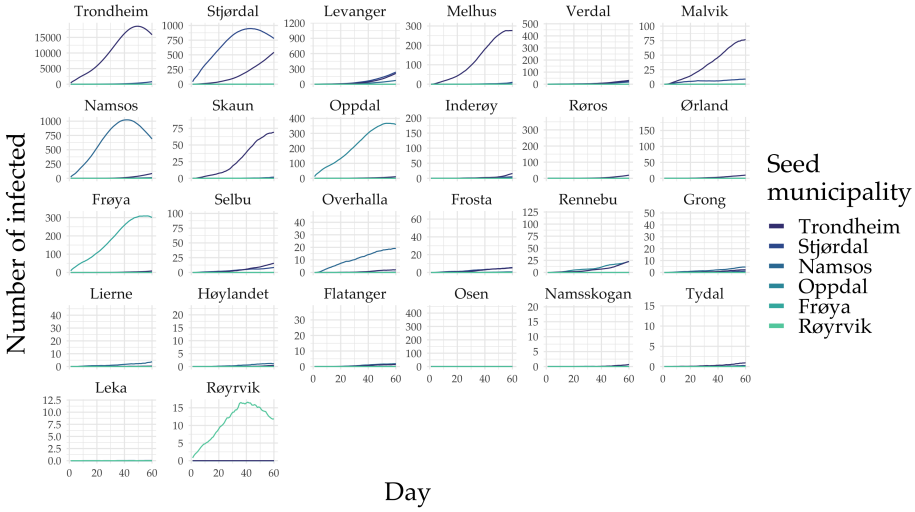


Figure C.2.2: Average daily number of infected as a function of seed municipality. Daily mean for each municipality after 100 simulations of each commuter fraction.

C.3 National Model Supplementary

Supplementary figures and results from the Section 4.3 follows.

C.3.1 Statistical Analysis

Figure 4.3.7 showed the relationship between outgoing commuters and commuter fraction. A similar plot is created in Figure C.3.1, with incoming commuters instead.

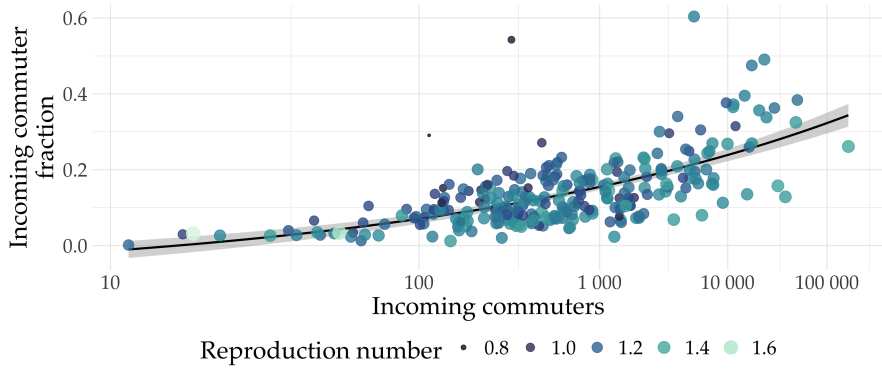


Figure C.3.1: The relationship between incoming commuter fraction and numbers. The line indicates the smoothed linear trend. Logarithmic horizontal axis.

Helge Bergo

Master's thesis

NTNU
Norwegian University of
Science and Technology