Tampere University

Otto Heimonen

GEOSPATIAL ANALYSIS OF THE SPREADING OF COVID-19 IN THE UNITED STATES

Faculty of Information Technology and Communication Sciences
Master’s Thesis
September 2021

ABSTRACT

Otto Heimonen: Geospatial Analysis of The Spreading of COVID-19 In The United States
Master's Thesis
Tampere University
Master's Degree Programme in Computational Big Data Analytics
September 2021

The COVID-19 pandemic has been a big threat to public health and there is an increasing need for efficient modelling of pathogens, predicting the daily infection rates to reduce the spread of COVID-19.

The Moran's and Geary's statistics showed significant spatial autocorrelation in the infection counts for the US COVID-19 data. Spatial regression using the simultaneous autoregression (SAR) and conditional autoregression (CAR) models indicate clear association between the confirmed cases and the number of population and the population density in both national county and state specific analyses. The SAR model provided a better model fit with the low AIC value, leaving no significant autocorrelation for the residuals.

The approximate Bayesian computation (ABC) methods were used to provide a flexible posterior distribution of the infection rate for COVID-19 based on the first 100 days of the pandemic. Three different simulation methods such as ABC-Rejection, ABC-Markov Chain Monte Carlo (MCMC) and ABC-Sequential Monte Carlo (SMC) were employed and compared. These algorithms seem to give reasonable posterior estimates for the average daily infections when the likelihood calculations for the spread of a harmful pathogen become complex, or intractable entirely.

The posterior distributions of ABC-MCMC and ABC-SMC provided plausible estimations covering all of the observed infection rates at different time points.

Keywords: Approximate Bayesian computation, ABC, spatial regression, simultaneous autoregression, SAR, conditional autoregression, CAR, COVID-19, infection rate

CONTENTS

List of Figures

List of Tables

1 INTRODUCTION

Background

In the wake of the recent worldwide pandemic, caused by the SARS-CoV-2 virus, there is reason to consider more efficient epidemiological modelling methods. The pandemic has caused grand distress, causing millions to be infected and decease. The ramifications of this sudden outbreak has also brought big downturns in economic development and increased poverty rates in many regions.

An important aspect of the COVID-19 pandemic has been the level of action countries have taken to properly handle the spreading of the disease. This has brought a wide range of commentary from different civil sectors on their respective governmental elements, along with praise and criticism. Perceived satisfaction on the actions taken by these elements are an important feature to the overall atmosphere and may shed light on how well ordinary people are going to follow future guidelines. The objective statistical information on the other hand, gives a narrative into the numeric efficiency in handling the SARS-CoV-2 virus.

All of the aforementioned features are, in a larger picture, information on how to improve future efforts to limit the spread of a pathogen. While there has been global precaution and acknowledgement of virulent diseases, these including the H1N1 swine flu pandemic of 2009, COVID-19 is the first truly widespread pandemic of the 21st century. As long as there is life on Earth, there will also be different virulent pathogens that can and will have an effect on their surroundings. It is therefore important to learn from the occurrences of these cases to properly avoid the following instance from spreading as rapidly. COVID-19 is a phenomenon in time and space and can be studied as such.

In their article *GIS-based spatial modelling of COVID-19 incidence rate in the continental United States* by Mollalo et al. (2020) study the county level differences of COVID-19 incidence during the first 90 days of the pandemic, final date of the data being April 9th, 2020. In this article, the aim was to model incidence rates for the virus with geographically weighed regression (GWR) and multiscale GWR (MGWR) with 35 different possible independent variables, covering areas of topography, sosioeconomy, graphic and demography. In the study results, Mollalo et al. describe how the introduction of spatial autocorrelation could improve the performance of global OLS model significantly, the models gave poor results, when compared to results from local models. The highest explained variability by these models was with MGWR, achieving $R^2 = 68.1\%$. In May 2021, the COVID-19 pandemic has continued for almost one and a half years, with population being vaccinated against the virus. Therefore, there is much interest in extending the analysis further than the period of the first 90 days, which could result in additional inference of the pandemic.

It is highly valuable to try and prevent an epidemic or pandemic from spreading or the very least slow the infection rates down. To do this, governing members need estimations of how large infection numbers can get. It is very difficult or even impossible to give a confident estimate for the pathogenicity in an early stage, where there may be only few cases and they are spread in multiple different regions. It is still highly important to try to give an initial estimate of the possible effect size.

Approximate Bayesian computation (ABC) methods are a set of methods that approximate posterior distributions of variables in situations, where likelihood calculations are very complicated, or intractable entirely. In their article, *Bayesian epidemiological modelling over high-resolution network data*, Engblom et al. (2020) simulated the spread of Escherichia coli O157 bacteria in Swedish cattle with the use of approximate Bayesian computation method framework. Their aim was to provide a feasibility study for the potential Bayesian public health framework. The results of the article gave promising results, with the approach performing convincingly in every synthetic tests. The results from the study suggest that ABC methods can be applied on a larger scale in an attempt to assess disease spread and thus further proves applicability for the ABC methods in the field of epidemiology. These results also then raise to question the amount of data needed to effectively model future scenarios and how early can the methods be applied. ABC methods generally do not need a large library of data points in order to function and can give results from relatively tiny data. Therefore, the ABC methods seem promising in the context of epidemiological modelling for the spread of pathogens.

The focal goal of this thesis is to study the incidence of the COVID-19 virus and compare them in different regions of the United States. One of the main focuses is to describe the pandemic through statistical means and present ways of modelling its spreading, offering tools that can be used in future similar situations to estimate effective precautions.

Another aim is to try to predict the spreading of the disease based on the first few months of the pandemic, using Approximate Bayesian Computation (ABC) methods, focusing on a general scope of the incidence and giving a flexible distribution estimate for infection rates, rather than just a point estimate. Based on the aforementioned information, the study question can be worded as *"How well can the incidence of COVID-19 be modelled using ABC methods with a small amount of data from the first 100 days of the pandemic?"*. We also describe and model the spread of the disease using geospatial statistics.

The following sections will focus on the methods applied in the thesis, followed by description of the COVID-19 data. We present two types of modelling methods, which include spatial models for areal data and three different ABC methods. The empirical part of the work presents study results separately for spatial and ABC modelling. We also present practical issues related to these methods.

2 METHODS

Spatial data analysis offers methods for tying information into specific geographical locations. This information is often valuable in depicting these phenomena as closely as possible to our understanding of the real world. When data includes spatial information of multiple cities or counties, it is possible to portray that information with maps and include different spatial measures, such as a distance between two cities, whilst keeping the information in an easily presentable form, such as visually presenting the data in maps. Such methods as k-means and k-nearest neighbour can be applied to these data and then the analysis can be portrayed on a map image, solidifying the information to these locations.

Spatial statistics separates from traditional data description in its usage of variables that can be tied to locations. Such information can be, for example, GPS locations, social media, mobile phone tracking, satellite imagery, postal codes, street names, country and municipality names and other spatially recognized data that can be utilized in building a visualization of an area. This information may have potential to uncover more complex relationships in multidimensional data of phenomena and places. This information can also then be used in conjunction with time analysis, adding another dimension to the data, which can help produce a more complete scenario of an event. This can help avoiding some underlying patterns in the data represented only as numbers in a database being missed, because the human brain may intuitively notice patterns better, when they are represented in a form that they are able to effectively handle.

2.1 Spatial Data Analysis

Spatial data consists of information that is tied to locations, where they were observed. In comparison with regular statistical modelling, spatial modelling can offer additional information to the phenomena that is being studied. This can be applied to other events, such as natural disasters, disease, social events such as concerts and tourism. This spatial information can be divided into different categories such as Geostatistical data, lattice data and spatial point patterns.

Geostatistical data is assumed to be spatially continuous, such as rainfall measurements, temperature, pH, air pollutant measures and other physical measurements in a location. This information is visualized on a map and the size of the studied phenomena can be represented with the colour and size of any interesting measurements.

Lattice data, often described as areal data, presents averages or counts of a phenomenon that make a larger region as a dataset. This information can be, for example, number of a species of animal in a region, number of births in a hospital, recorded cases of a disease in different regions in an area, satellite imaging, median household income in a neighbourhood and other similar values. This type of data also is referenced as polygons, where the centroid of the units is used as spatial reference with the area of the polygon. The data can be displayed as a map and using colours to discern between different areal units. Analysis for this kind of areal data involves measures such as representation of spatial proximity and testing for the existence of spatial patterns, using autocorrelative measures such as Moran's I and Geary's C. The data can also be modelled with autoregressive models, such as simultaneous autoregression (SAR) and conditional autoregression (CAR). These measures will be presented later in the thesis.

Spatial point patterns describe a finite number of events in a region, such as locations of bird nests or craters born from meteor impact or volcanic activity and locations of homeless. Different categories for this information may be represented as differently coloured points in visualization and conclusions could be done solely from visualized data. Some of the goals for spatial point patterns are to see, whether a regular pattern appears in the points, if there are clear clusters of points, is there a process that could have produced the pattern and if so, then how intense is this process and if there are underlying distributions that could have affected the results to appear in a region.

Each of these data sets can also be referenced in time as spatio-temporal data. Each of these observations all have then a location, time and a value. Deciding on artificial borders for different phenomena may be difficult to justify and there may not be clear discernible borders to a phenomenon, like a disease in a city. Other objectives lie with inference for non-spatial structures, predictions of unobserved variables and clarifying design issues, such as physical locations for taking observations or how the arrangement of treatment should be handled.

2.2 Moran's and Geary's Statistics

In order to utilise neighbouring relationships in spatial statistics, a proximity matrix W must be constructed. A W matrix is a matrix for data points Z_{ij} so that

$$W_{ij} = \begin{cases} 1 & \text{if two cases i and j are neighbors} \\ 0 & \text{otherwise.} \end{cases}$$

There are three fundamental ways to define, which cases are neighbours. Two of the more commonly used are Rook's and Queen's Case. In Rook's case, the neighbour is strictly adjacent to the current case, while in Queen's case, the corners are accepted as well. The third option is the Bishop's case, where only the corners are taken into account. Changes in the interpretation of neighbourhood affects the results. Therefore, it is important to consider, how the neighbourhoods function in context of real world applicability instead of solely focusing on the data. In this thesis the data consists of US counties, which are connected with wide networks of roads, not constricting the neighbours to a situation like Rook's Case. Other ways to approach the neighbouring system is to use methods such as k-nearest neighbour, which finds k nearest values for each point. The analysis in this thesis used k-nearest neighbours approach to produce neighbours. This is achieved by the R function *knearneigh*, which produces k-nearest neighbours for spatial weights.

Moran's I statistic is a measure of spatial autocorrelation. The statistic describes the amount of correlation in close spatial locations. The expected $E(I)$ for Moran's I is $-\frac{1}{n-1}$, where n is the number of cases in the data. In an instance, where $I > E(I)$, we can expect positive spatial autocorrelation to exist, the further I is from the expected value, the stronger autocorrelation is. In an instance, where $I < E(I)$, spatial autocorrelation is negative. Moran's I can be defined in the following way:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j W_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_i (Z_i - \bar{Z})^2},$$

where Z_i are observations, S_0 describes standard deviation and is calculated by $S_0 = 2(2rc - r - c)$, where r and c describe the dimensions of a $r \times c$ lattice. The statistic is a measure of the existence of global autocorrelation and does not give information on where any clustering of a phenomenon exists.

Geary's C is compared to Moran's I as they both are measurements of spatial autocorrelation. Geary's C measures the autocorrelation of adjacent observations in a specific phenomenon. The two measures are related, but Geary's C is more sensitive to local fluctuations of spatial autocorrelation, whilst Moran's I focuses more on global autocorrelation. It can also be stated that the two measures are inversely related due to Geary's C emphasizing the amount of dissimilarity of adjacent observations. The values for the measure fluctuate between 0 and 2, with 1 being the middle point. If $C < 1$, this implies positive autocorrelation, due to the amount of dissimilarity being low. If $C > 1$, it can be expected that data has negative autocorrelation. The statistic can be calculated

$$C = \frac{n-1}{S_0} \frac{\sum_i \sum_j W_{ij}(Z_i - Z_j)^2}{\sum_i (Z_i - \bar{Z})^2}$$

The largest difference between the formulas of Geary's and Moran's statistics is how they handle the difference of data. Geary's C is calculated as the squared difference of values, while Moran's I first takes the difference of mean from the observations.

Both Moran's and Geary's statistic are measures for testing, whether spatial patterns exist in a data. They give insight, whether it is reasonable to expect different types of spatial autocorrelation. As it is already known, these measures do not, however, answer to the nature of these connections. One cannot make inference of the nature of the phenomena from these measures. For this, spatial modelling must be used. In the following sections, spatial autoregressive models are discussed that can be utilised in testing for predictive relations.

2.3 Spatial Models For Areal Data

In exemplary ordinary least square modelling, there are expectations of models existing in a vacuum of independence. However, a dependent variable tied to an independent variable is in most cases, also dependent of other variables or the observations are not fully independent of each other. Observations are all tied to a place and are not fully separate from each other. An outbreak of a disease can have multiple straightforward independent variables explaining it, but the outbreak is located in an area. If an area in a city suffers from an outbreak, it is rarely the case that the outbreak remains to that single location and that other cases of said disease are purely independent from the earlier cases in that area. Other areas have outbreaks of the disease and due to temporary interchanging population of regions caused by commuting for work, for example. Therefore, it is simply not always sufficient to consider only the dependent variables as independent. These variables of locational information are also properties of spatial statistics and thus cannot be ignored either. The following chapters will give exposition on the inclusion of spatial information.

2.3.1 SAR and CAR Models

Simultaneous autoregressive (SAR) model is a statistical model for spatial data. It has its roots in the time series autoregressive (AR) model, applied to spatial data. A time series autoregressive model can be expressed as $Y_t = \rho Y_{t-1} + \epsilon_t$, where the current value in time Y_t is the sum of the former time Y_{t-1} multiplied by an autoregressive correlation coefficient ρ and the error term in time ϵ. If we add a trend, we get an AR(1) model, which can be expressed thusly: $Y_t - \mu_t = \rho(Y_{t-1} - \mu_{t-1} + \epsilon_t)$. Instead of having the autoregression focus on a time variable and a lag effect, SAR models have a similar spatial effect for its surrounding areas, which means that the data is in two dimensions for spatial information. For this, the outcome Y_t needs to be expressed in another form to take the second dimension into consideration. Therefore, for spatial data and for SAR, Y_t is replaced with

$$Z_{u,v} = \frac{\rho}{4}(Z_{u-1,v} + Z_{u,v-1} + Z_{u+1,v} + Z_{u,v+1}) + \epsilon_{u,v},$$

where u, v describe the rows and columns for the surrounding observations and $u \in [2, R-1], v \in [2, C-1]$ in an $R \times C$ lattice. In putting this model to a matrix form, we get

$$Z = \rho W Z + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I),$$

where W is an $n \times n$ matrix with $n = RC$. Now we can set the SAR model for a matrix form with a trend so that

$$Z - \mu = S(Z - \mu) + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I),$$

where $S \equiv S_{ij}$ so that $S_{ii} = 0$ and $I - S$ is nonsingular and for SAR, the data is seen to follow the distribution:

$$Z - \mu \sim N(0, \sigma^2 (I - S)^{-1} (I - S')^{-1})$$

The Conditional autoregressive (CAR) model is related to SAR in similar ways that Geary's and Moran's statistics are related. Both are a way to create spatial regression models and thus they are often compared with each other. A CAR model requires symmetricity of the weighting matrix, which is already prevalent in the binary W proximity matrix. A CAR model can be written so that

$$(Z_i | Z_j, j \neq i) \sim N(\mu_i + \sum_j C_{ij}(Z_j - \mu_j), \sigma^2),$$

where $C \equiv C_{ij}$ so that $C_{ii} = 0$ and $I - C$ is symmetric and the data follows the distribution:

$$Z - \mu \sim N(0, \sigma^2 (I - C)^{-1}).$$

To form a maximum likelihood estimate (MLE) for regression, a log-likelihood estimate is formed. An MLE estimate is used in regression to find an estimate for a distribution, which maximises the likelihoods of observing the values in said distribution. This is an important step in regression analysis in forming the most suitable model for a data. The general MLE for a traditional linear Gaussian distribution, the log-likelihood model is

$$-\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j-1}^{n} (Z_j - \mu)^2,$$

but for SAR or CAR models, the specific characteristics of the matrices have to be taken into account. This brings the log-likelihood into the following form:

$$-\frac{n}{2} ln(2\pi\sigma^2) + \frac{1}{2} ln|B| - \frac{1}{2\sigma^2}(Z - \mu)' B (Z - \mu),$$

where

$$B = \begin{cases} (I-S')(I-S) & \text{for a SAR model} \\ (I-C) & \text{for a CAR model} \end{cases}$$

The differences between SAR and CAR models may seem subtle, but they give inference on different situations. A CAR model is best utilised in a situation, where there is low order dependency between the observations. This means that there is no strong autocorrelation among observations that are not adjacent to a specific observation. SAR models, on the other hand, tend to be more suitable in situations, where the spatial autocorrelative pattern is more global or there exists second order dependency between the observations. CAR models also requires symmetricity for the weighting matrix, which is something SAR models do not need.

All the models presented here can be utilized with the *spatialreg* package in R. The package has a substantial coverage for spatial data modelling and can produce SAR and CAR models with the *spautolm* command. Modelling results will be presented in the results section for national, regional and county wide modelling for differing results and additional inference.

The methods used in this thesis do not give an R-squared measure, but a pseudo R-squared, which makes a direct comparison of the results to the ones from the study by Mollalo et al. more difficult. In regular linear regression models, the R^2 measure explains the amount of the variability of the data explained by the model. The pseudo R^2 measures are used as approximations for this measure. The Nagelkerke R^2 is a modified version of the Cox and Snell's R^2, which is based on the comparison of the log-likelihood for the used model and the baseline model. The measure is calculated by adjusting the scale of the statistics to cover the explainable variation of the model to cover the range from 0 to 1. Thus, the pseudo R^2 is not directly comparable to the regular R^2, which is received from $1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$.

2.4 Approximate Bayesian Computation (ABC) Methods

Approximate Bayesian computation (ABC) methods are a set of methods for approximating the posterior distributions for Bayesian inference. When likelihood calculations become so complex to calculate that they may even become intractable, ABC methods can be utilised. In this section, the ABC method family is introduced. We first introduce the ABC-Rejection algorithm and present other variation methods that have been developed in order to make the process more efficient or accurate.

In Bayesian inference, we obtain posterior distributions by combining prior information and likelihood of the data

$$p(\theta|y) \propto p(y|\theta)\pi(\theta),$$

where $\pi(\theta)$ represents the prior and $p(y|\theta)$ the likelihood function for a model. In ABC methods, the joint distribution is not calculated through traditional Bayesian methods, but with simulated data that follows the joint distribution. This simulated data is then compared to the observed data and is accepted, when the results are similar to the observed data. This results in the method approximating the likelihood results.

According to Yang et al., (2018) ABC methods can be divided into two main categories according to the way the data is simulated. These categories are sampling- and regression-based algorithms. Sampling-based algorithms directly approximate the likelihood function using simulated samples that are near the observations. The closeness of these simulated and observed values is measured with a similarity kernel. The regression-based ABC, on the other hand, establishes regression relationship of a model parameter and the conditional distribution $p(y|\theta)$. This thesis will focus mainly on the sampling-based ABC methods.

One of the questions on this algorithm is how to compare the simulated data and the observed one. One solution to this is to calculate a summary statistic from both data sets and compare them to each other to see, whether the summary statistic is accepted as a particle of the posterior distribution. Distances over ε are neglected and values under ε are accepted. This process is iterated, until a sufficient amount, denoted by n, of proposed posterior distribution particles has been reached. This practice shrinks down the information of the data sets, which oversimplifies the data in some aspect, Zheng et al. (2017). For example, the use of summary statistics lessens the amount of dimensionality of the data, which can reduce the number of discarded particles, but this does neglect the information given by the dimensions of the data. Sampling for individual proposal particles will in most cases lead to a more discarded particles, but can offer a wider proposal distribution, which in turn can offer more for inference. The ABC-Rejection algorithm, in its most simplistic form, can be written as follows:

1. Sample a candidate θ^* parameter vector from some proposed distribution $\pi(\theta)$.

2. Simulate dataset x^* from the model received from a conditional probability distribution $f(x|\theta^*)$.
3. Compare the simulated dataset x^* with experimental data x_0 using a distance function d and a tolerance level ε. If $d(x_0, x^*) \leq \varepsilon$, we accept θ^*. Otherwise reject θ^* .
4. If n particles are not accepted, return to step 1.

The choice of prior has influence on the efficiency of the algorithm. It is one of the disadvantages of the ABC-Rejection algorithm. If the prior sampling distribution is vastly different from the posterior, the acceptance rate of new samples will be low. This can cause accidental inaccurate results to occur by the user choosing non-informative priors. If the user is not informed or does not simply believe some outcomes to be possible for the variable estimated, it is easy to undermine the possible outcomes that can arise from the analysis. This results in the method taking excessive amounts of iteration and the possible range of the posterior distribution to be too small. These issues can be mended with different opinions from experts of different areas voicing opinions on the choice of priors. The ABC-Rejection algorithm is simplistic and easily computable in many softwares. This method has been improved on and there is extensive literature on the subject. In the following sections, two different improved variations of the ABC-Rejection algorithm are presented.

If we consider the distance measure d, it implies the possibility to compare the model output to the data directly. If there is an ongoing situation that is developing day by day, for example an epidemic, the method makes it possible to estimate the parameters for a Susceptible-Infected-Recovered (SIR) model. The sum of the squared differences between the actual and the predicted data can be calculated over time T in $d(x^*, x_0) = \sum_{t=1}^{T}(x_t - x_t^*)^2$. This can be helpful in a dynamical model following the developing situation.

One alternative possibility for reducing the number of iterations is to fix the amount of iteration, after which the algorithm stops and evaluates the accepted particles. This is much different from the original procedure of iterating, until a set number of proposal particles are accepted, fixing the time the algorithm uses instead. This process can introduce variability into the results. The algorithm may result in more proposed particles being chosen than one would have originally used in the iteration, but there is also a chance that the number of accepted particles will be low, which can result in lackluster information for inference. The posterior distributions may vary greatly because of this and can present a problem for inference. The fixed iterations approach can however, be used to give preliminary information on the choices for tolerance levels. Since the time used in the algorithm is fixed in these cases, there is more room for experimentation for different tolerance levels. These test runs can quickly give feedback on the choice of parameters.

2.4.1 Markov Chain Monte Carlo Algorithm

The main goal of the approximate Bayesian computation Markov chain Monte Carlo (ABC-MCMC) is to solve the problem of low acceptance rates often encountered in the ABC-Rejection algorithm. This is done through the use of Markov chain Monte Carlo methods. Markov chains are stochastic models that describe transitions from one state to another. A state z_i is dependent of its previous state z_{i-1} and the following state z_i is affected by the current state. Here we use θ_i instead of z_i.

The ABC-MCMC algorithm can be given as follows.

1. Initialize θ_i, i=0.
2. Sample a candidate θ^* parameter vector from some proposed distribution $\pi(\theta)$.
3. Simulate dataset x^* from the model received from a conditional probability distribution $f(x|\theta^*)$.
4. Compare the simulated dataset x^* with experimental data x_0 using a distance function d and a tolerance level ε. If $d(x_0, x^*) \leq \varepsilon$, we go to step 5. Otherwise $\theta_{i+1} = \theta_i$ and go to step 6 .
5. Set $\theta_{i+1} = \theta^*$ with probability $\alpha = min(1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)})$
   and $\theta_{i+1} = \theta_i$ with probability $1-\alpha$.
6. Set $i = i + 1$ and go to step 2.

The ABC-MCMC algorithm uses Metropolis-Hastings algorithm in step 5. α or the acceptance probability is calculated with the relation of the function of the priors times the likelihoods, given the θ values. Essentially this can be expressed in the form

$$\alpha = min(1, \frac{Prior(\theta_i)likelihood(\theta_{i-1}|\theta_i)}{Prior(\theta_{i-1})likelihood(\theta_i|\theta_{i-1})}),$$

where θ_i represents our proposed distribution. If the ratio of the distributions is larger than one, one will be chosen for α and the new value for θ is accepted. One way to select the proposed value for θ is to generate a value $u \sim U(0,1)$ and if $u < \alpha$, we set $\theta_{i+1} = \theta^*$. Otherwise, θ stays the same.

The Markov chain is formed by linking the candidate θ^* values. This ensures the process will always reach convergence to the target approximate posterior distribution. This is the shortest form of the Markov chain, because the future sample depends only on our current sample. If the ABC-MCMC is described more freely, it could be said that the process is very similar to the simple ABC, but the scope of the point, from which the proposed particles are taken from, is free to shift towards areas, where higher acceptance rates can be observed.

One of the potential disadvantages with ABC-MCMC is its dependence of the parameters chosen at the start. The priors given by the user at the beginning may influence greatly on how quickly convergence is reached. This is the same problem as with the ABC-Rejection algorithm, because the greatly differentiating priors may cause a lot of the early proposed values to be discarded. ABC-MCMC does however, get faster when the process comes closer to convergence.

The second disadvantage for the priors is the autocorrelation. The algorithm have more autocorrelation in the beginning of the iteration process and it can raise suspicion that the simulated values are too dependent on the previous iteration. One solution to this has been the introduction of burn-in period. A burn-in period is a term used to describe the practice of ignoring or throwing away a "beginning" part of the iteration process. The term *beginnig* in quotes is due to the loose definition for the beginning period. There has not yet been a strict rule on how to determine length of the burn-in period. It is therefore dependent on the user on how long the effects of the more heavily autocorrelated parts last. The burn-in period with a length of n tries to solve the autocorrelative notion from the beginning of a process defining the posterior distribution. The possible effects on the posterior distribution diminish with the amount of iterations however and it can be argued that the effects can be neglected and the burn-in is not needed.

Another potential disadvantage to the ABC-MCMC is that the correlative samples and low acceptance rate may cause very long chains that may remain stuck in regions of low probability for long periods of time. This can happen when the posterior leads to an area of lower likelihood and the next samples will not be accepted as likely and the ones accepted will not lead to a higher probability area.

2.4.2 Sequential Monte Carlo Algorithm

The approximate Bayesian computation Sequential Monte Carlo method (ABC-SMC) is an extension to the traditional ABC-Rejection algorithm. The aim of the ABC-SMC is to fasten the process of parameter sampling. The way ABC-SMC achieves this is through gradually decreasing the tolerance level ε for the proposal particles θ^* from the prior distribution $\pi(\theta)$. The tolerance level ε_i decreases so that $\varepsilon_1 > \varepsilon_2 > ... > \varepsilon_T > 0$. This ensures the distributions evolve towards a targeted posterior distribution. The ABC-SMC algorithm proceeds as follows:

1. Initialize $\varepsilon_1, ..., \varepsilon_T$ and set $t = 0$.
2. Set particle indicator $i = 0$.
3. If t = 0, sample θ^{**} independently from $\pi(\theta)$.
   Otherwise, sample θ^* from previous population $\{\theta_{t-1}^{(i)}\}$ with weights w_{t-1} and perturb θ^* to obtain $\theta^{**} \sim K_t(\theta|\theta^*)$, where K_t is a perturbation kernel.
   If $\pi(\theta^{**}) = 0$ return to beginning of step 3.
   Simulate a candidate dataset $x^* \sim f(x|\theta^{**})$.
   If $d(x^*, x_0) \geq \varepsilon_t$, return to beginning of step 3.
4. Set $\theta_t^{(i)} = \theta^{**}$ and calculate the weight for particle $\theta_t^{(i)}$,
   $$w_t^{(i)} = \begin{cases} 1, & if\ t = 0, \\ \frac{\pi(\theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta_{t-1}^{j}, \theta_t^{(i)})}, & if\ t > 0. \end{cases}$$
   If $i < N$, set $i = i + 1$, go to step 3.
5. Normalize the weights so that $\sum_{i=1}^{N} w_t^i = 1$. If $t < T$, set $t = t + 1$, go to step 2.

Single asterisk denotes the particles sampled from the previous distribution and double asterisk denotes particles after perturbation. The ABC-SMC first sets a tolerance level sequence, after which the particle indicator calculator is initialized. On the very first iteration of the tolerance level sequences, the proposed θ is sampled from some proposed distribution. Further iterations are received with weights and are perturbed by a perturbation kernel. If the proposed distribution is larger than 0, the algorithm will simulate a candidate data set and compare it to the data, accepting or discarding the particle received from the comparison. When a particle is accepted, it is given a weight, which is 1 on the first iteration of the tolerance level sequence. On the following iterations, the weights are calculated by dividing the proposal distribution with the sum of the former weights multiplied by the perturbation kernels used in step 3. This process is iterated until the sufficient amount of particles is reached and once the number of particles equals to N, a softmax normalization is done for the weights and the tolerance level is decreased. After this the process of creating a posterior distribution begins anew with the decreased tolerance level.

The behaviour of the ABC-SMC algorithm is linked to the sequence of tolerance levels and perturbation kernels. Small decreases between the tolerance levels $\varepsilon_{t-1} > \varepsilon_t$ will guarantee more particles to be accepted between iterations, but the process overall will need then more iteration to reach convergence. A longer sequence of intermediate posterior distributions will also increase the time needed to complete the algorithm, since with every new tolerance level, a new intermediate distribution is created. On the other hand, larger gaps between the tolerance levels result in lesser amounts of overall simulation, but with more discarded particles.

Considering the selection of the tolerance levels, Simola et al. (2020) explain that there are three primary ways of determining the tolerance sequence for ABC-SMC, which are to fix the values in advance, adaptively select a tolerance level ε_t based on a quantile of $\{d_{t-1}^{(J)}\}_{J=1}^{N}$, where d is the distances of the accepted particles from iteration $t-1$ and to adaptively select ε_t based on some quantile of the effective sample size values.

Selecting the tolerance sequence from a predetermined quantile can lead to the proposal particle sampling getting stuck in local modes. Thus, the tolerance levels have not only effect on the speed at which the algorithm functions, but to the convergence of the posterior. Silk et al. (2013) suggest using the adaptive selection of the tolerance sequence at every iteration. This is done through estimation of an optimal value on the threshold-acceptance rate curve (TAR curve). The method then balances the shrinkage of the tolerance level in relation to the acceptance rate. The idea is to select a value at the elbow of the TAR curve so that a vast majority of the proposed values will be rejected. This approach guarantees the method to converge to the real posterior. The calculation then requires the estimation of the TAR curve in each iteration of the algorithm.

The perturbation kernel for ABC-SMC is an important part of the process. In each intermediate distribution, a weighed sample of parameter vectors are chosen. In the first iteration, uniform weights are accepted, since there are no previous distributions to calculate the weights from. Successive distributions are then constructed through the sampling of parameters from the previous population and perturbing them through some kernel function, regarded here as perturbation kernel K Filippi et al (2012). In Machine learning, kernel trick is used to set a linear classifier for non-linear problems. In doing so, linearly inseparable data is transformed into linearly separable versions.

Often the kernel K_t can be chosen from a random walk process, either Gaussian or Uniform. In a simple random walk process, we could have an observation x to which we add a random effect moving the observation towards some direction, afterwards we can introduce another random effect to this point and so on, until we decide to stop the process. The resulting trajectory seems to wander randomly, but is limited to which distribution the random walk process follows. We can utilise such a process as our perturbation kernel. This adds some random mixturing components to our samples perturbing them slightly. A perturbation kernel with a large variance can prevent the algorithm from becoming stuck in a low probability area, but it can also lead to the algorithm rejecting a lot of the proposed particles, which in turn makes the process inefficient.

An advantage of the ABC-SMC in comparison to ABC-MCMC is that the particles are uncorrelated. This is due to the additional mixture component added by the perturbation kernel. The accepted particles on the first error threshold level are sampled and carried to the next intermediate level and perturbed, thus breaking the immediate dependancy of the former level and thusly the question for a burn-in period need not be considered.

One of the necessary steps for any ABC algorithm is to determine, when the algorithm should be stopped. In the earlier methods of ABC-Rejection and ABC-MCMC the solution was to stop the algorithm when the predetermined number of particles was accepted. In the sequential version, there is also the option of stopping the algorithm whenever the tolerance level decreases below a desired level, leaving the number of accepted particles more open. If ABC-SMC algorithm follows the same option as the previously introduced algorithms, there may be the risk of unnecessary iteration and thus increasing time consumption. Once the posterior stabilises and convergence is reached, further reduction to the tolerance level does not substantially improve the approximation of the posterior. The optimal tolerance level is chosen by the user, Simola et al. (2020).

3 EMPIRICAL DATA ANALYSIS

3.1 Data Description

The data used in this thesis consists of the daily updates of COVID-19 statistics upheld by the New York Times. The data describes daily cumulative infections and deaths related to the pathogen by each county in the United States. Along this, there is information on population size and population density, both received from US census bureau (https://covid19.census.gov/), and poverty percent, received from Economic Research Service (https://www.ers.usda.gov/).

The first incidence of COVID-19 in the data is dated to 21st of January 2020, in Snohomish County, Washington. The last date for all the cumulative cases in the data is for 23rd of May 2021, which is the date the information was exported from New York Times. This means that the duration of the pandemic, from the first incident of the virus, is 489 days. The first death related to COVID-19 was recorded on 29th of February 2020, in King County, Washington. The highest accumulation of incidents for COVID-19 cases was in Los Angeles, California, where there accumulated roughly 1.24 million cases. The highest cumulative figure for deaths related to COVID-19 was in New York City, with the number reaching over 33,000.

For the accumulation of cases of COVID-19 and the deaths related to it, the resulting time series can be seen as well as histograms with density plots in figure 3.1. In the figure, we can see that the cumulative graphs have a sharp increase after October of 2020 for infections and in November 2020 for deaths. The cumulative incidence for COVID-19 cases reaches to over 30 million, whereas the cumulative deaths seem to come close to 600,000 as of May 2021. The ending points in the data, and as seen in the graph, are $32,849,985$ infections and $586,031$ deceased.

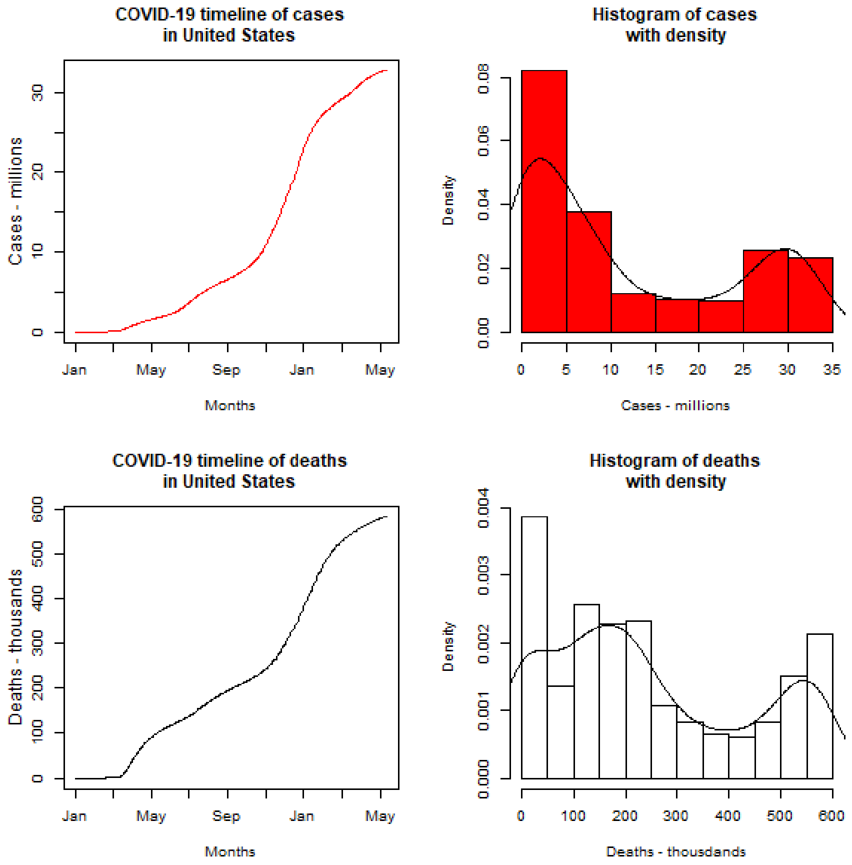


Figure 3.1 The cumulative cases for COVID-19 infections and deaths related to the virus from January 2020 to May 2021, with their respective histograms with density lines.

The probability density histogram shows that the data does not seem to be normally distributed, rather having tops at both ends of the distributions and a concave region around the middle. One possible explanation for this is the phenomenon that during the initial wave of COVID-19, the infection and death rates were low, due to the virus not being as widespread and thusly resulting in lower numbers, and after a certain period of reaching over the entire country, the virus spread much more rapidly, thus giving a sharp rise to the infection numbers and causing many days of high incidence rates.

Daily infections for the spread of COVID-19 can be seen in figure 3.2. We can see the corresponding phenomenon in the daily cases as with the cumulative cases, seeing a sharp rise to both after October. The maximum values seem to be reached around December 2020 and January 2021 for daily infections and around February

2021 for deaths.

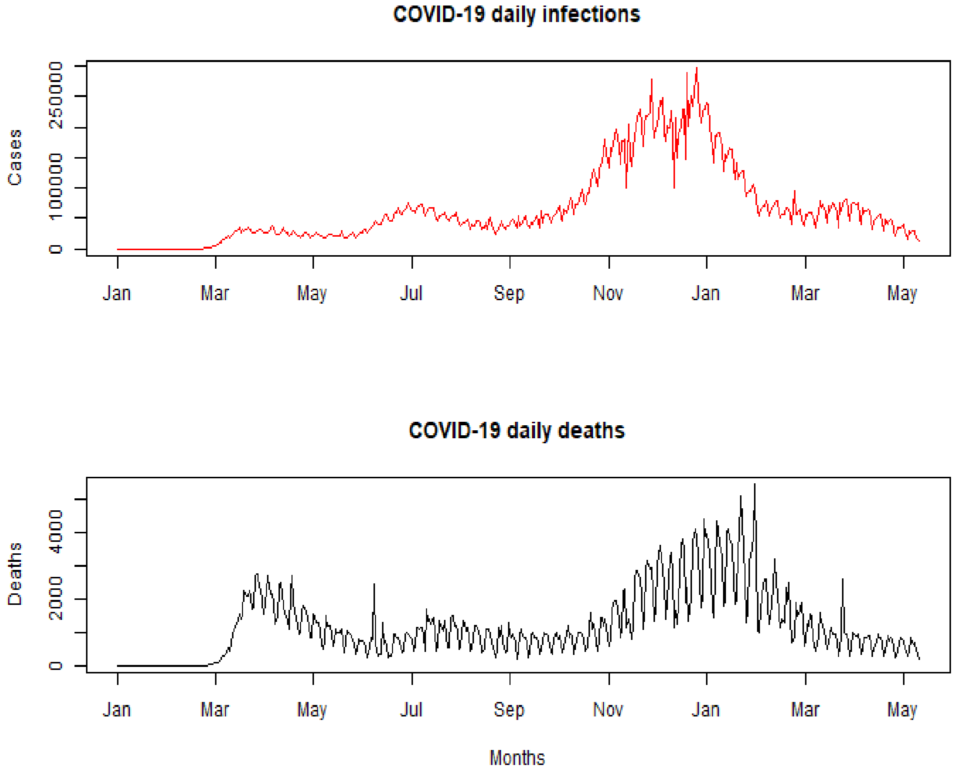


Figure 3.2 Daily infections and deaths related to COVID-19 from January 2020 to May 2021.

Maximum for daily infections was $297,799$ cases and $5,455$ for deaths. The averages for the same daily values are $67,177.88$ ($sd = 64,293.87$) for infections and 1198.43 ($sd = 1011.05$) for deaths. From the figure, it is possible to see that both began their acceleration roughly in March. After 60 days from the first recorded instance of COVID-19 infection, the overall infections reached over $20,000$ cases per week, which is 5 cases per $100,000$ people. This reaches over the official baseline limit for overall infections in Finland, and turns into acceleration phase. Before this, on the first 46 dates, the daily incidences do not surpass the limit of 100 cases and only on the 5 final days, the incidence passes 1000 cases per day.

The population density was recorded as a measure of people per square kilometres. The highest population density was in New York City, having $27,819.805$ people per square kilometres and the lowest was for Loving County, Texas, having 0.059

people/km^2. The visualization of COVID-19 cases can be seen in figure 3.3. In the figure, total amount of COVID-19 cases is represented as a grayscale map of every state. The state of California seems to be a relative hotspot for the infections. Surprisingly, the infections are much lower in the surrounding states. Other states, with relatively large amount of cases are Texas, Florida and New York. The states at the western side of midwestern region, North Dakota and South Dakota and Nebraska, and the states on the northeastern side of western area, which include Montana, Wyoming and Idaho, seem to have quite few cases. Another area, where the infections are relatively low, includes the states at the end of the northeastern region, which are Maine, New Hampshire and Vermont.

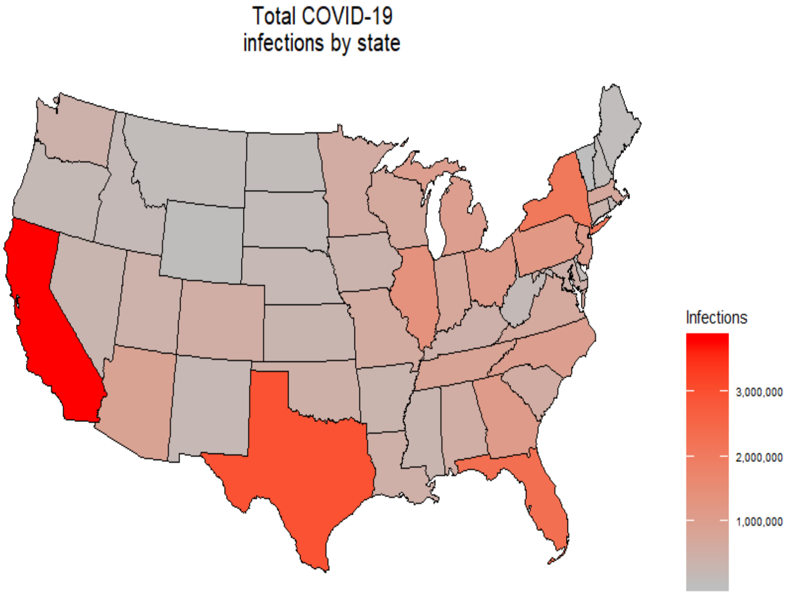


Figure 3.3 Overall infections of COVID-19 in every continental state by May 2021.

The visualization of the amount of infected individuals relative to the population of each state can be seen in figure 3.4. In comparison to figure 3.3, the relatively low infection numbers of the western side of the midwestern region, seems to have quite high infection percentages. North Dakota and South Dakota seem to have the highest infection percentage, whereas the northwestern and northeastern end of United States seem to have the lowest infection percentages. It seems from the figure, that the state of Vermont has the lowest infection percentage of all of the continental states.

The infection counts of each individual continental county in the US can be seen

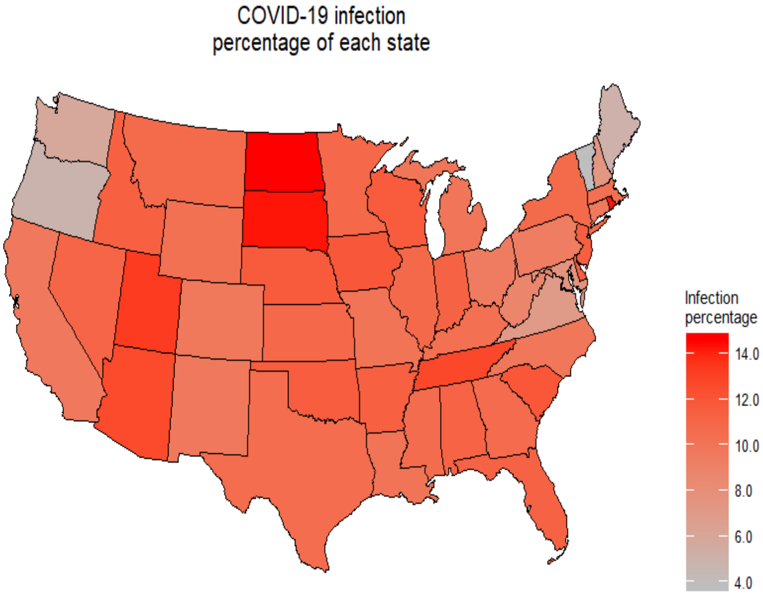


Figure 3.4 Infection percentage of COVID-19 in every continental state by May 2021.

in figure 3.5. In the grayscale map of the counties, it is clear that the most amount of infections are located in the counties of California and Arizona. There are relatively high infection counts in some counties of Texas, the southern counties of Florida, the state of Illinois at the shore of lake Michigan and the New York City.

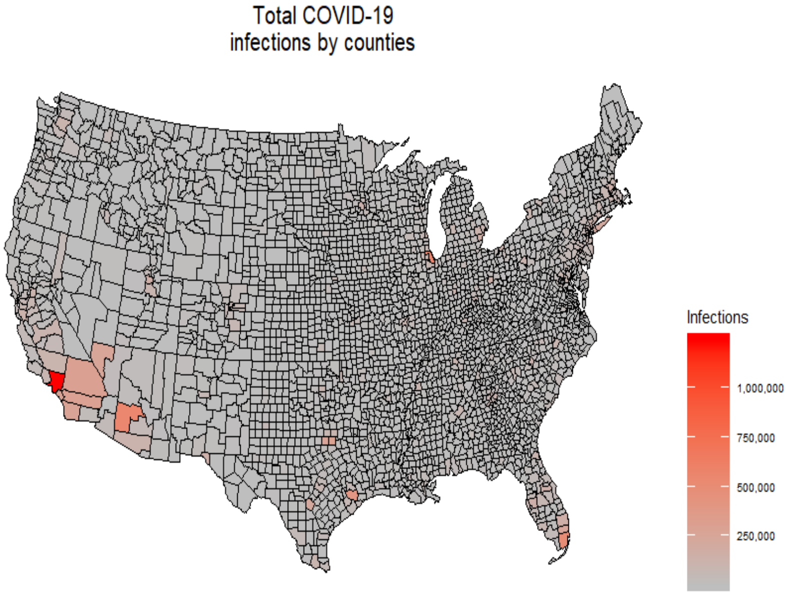


Figure 3.5 Overall infections of COVID-19 in every county by May 2021.

For the global feature inspection, the data was also divided into five different regions by geography, dividing the country to *Northeast, Southeast, Midwest, Southwest* and *West* regions. The northeastern region covers the least of the country, but includes multiple smaller states inside it. On the other hand, the western region covers the largest amount of land and all of the states inside seem relatively large in comparison to the eastern areas of the US. The southwest region has the least states, totalling at four, but covers a relatively large area at the southern border of the US.

Northeast	Southeast	Midwest	Southwest	West
Connecticut	Alabama	Illinois	Arizona	Caliornia
Delaware	Arkansas	Indiana	New Mexico	Colorado
Maine	District of Columbia	Iowa	Oklahoma	Idaho
Maryland	Florida	Kansas	Texas	Montana
Massachusetts	Georgia	Michigan		Nevada
New Hampshire	Kentucky	Minnesota		Oregon
New Jersey	Louisiana	Missouri		Utah
New York	Mississippi	Nebraska		Washington
Pennsylvania	North Carolina	North Dakota		Wyoming
Rhode Island	South Carolina	Ohio		
Vermont	Tennessee	South Dakota		
	Virginia	Wisconsin		
	West Virginia			

Table 3.1 Continental US regions and the names of states included in each region.

The division of each state into a specific region can be seen in table 3.1. In the table, we can see that the southeast region has the most states inside it, whilst the southwest has the least. Overall, the states in eastern US have smaller counties inside, while the west has fewer, but larger counties. This naturally results in larger population density in the eastern regions and lower in the west. The coastal regions tend to be more populated than inland states and the western US has much more arid areas, which usually have lower populations than areas with easier access to water. This is also evident in the population densities of these regions, where the mean population densities are shown in table 3.2. Interestingly, there is also the largest amount of different population densities in these regions, with the northeastern region having a standard deviation of 2301.14 and the western region having a standard deviation of 408.60. These large standard deviations may be explained with grand cities. Both, the northeastern and the western region have grand hot-spot cities, New York City and Los Angeles, which adds skewness and variation to the population density.

Region	Population Density (sd)	Poverty Percent (sd)
Northeast	528.76 (2301.14)	11.38 (3.75)
Southeast	74.80 (226.14)	17.60 (6.32)
Midwest	49.02 (150.92)	12.17 (4.48)
Southwest	37.75 (114.28)	16.13 (5.11)
West	71.85 (408.60)	12.61 (4.36)

Table 3.2 Mean population densities and poverty percent for each US region with standard deviations (sd).

The poverty percent was assessed as population in a county, whose income levels are below the poverty line. The lowest poverty percent was for the northeastern region with 11.38% and the highest for southeast region, with the average being 17.60%. The highest standard deviation (sd) was also in the southeast region, with sd being 6.32%. This would imply that there is the most heterogeneity in that region in regards of financial income levels. The coastal and northern regions overall seem to have the least amount of poverty and the southern regions the highest poverty rates. Poverty percent was not used in the modelling, since it never had significance in the preliminary testing for the models.

3.2 Spatial Autocorrelation

To assess the general autocorrelation in the data, Moran's and Geary's statistics were taken over three different levels of the United States. The first two sectors were state specific, in which the spatial autocorrelation was measured, when the areal units were states, and county specific, in which the units were counties. The final level was county level, which took the measurements from each continental state separately on county level. This approach aimed to give more freedom to local differences in the spread of the COVID-19 virus and to help spot these differences, instead of remaining on a national county wide perspective.

On national county specific level, Moran's I was 0.2946 ($E(I) = -0.0003$), with p-value less than $2.2e-16$ and thus significant. Geary's C was 0.7704 ($E(C) = 1.0$), p-value 0.0029 and significant ($p < 0.01$). The inference from both of these measures would suggest the existence of spatial autocorrelation. In the state specific analysis, Moran's I was 0.0354 ($E(I) = -0.0208$) and $p = 0.3193$, significantly lowering the amount of autocorrelation detected and making it not significant. On the other hand, Geary's C was 0.7189 ($E(C) = 1.0$) and $p = 0.0328$, thus significant.

For the separate inspection of each state, some states had to be dropped out from the material. For example, the District of Columbia cannot be considered independently, since it not an official state, but the capital. DC does not have other counties inside its border and results for Moran's or Geary's statistics cannot be achieved separately from other states. The state of Delaware suffers from the same issue, having three counties, only one of which has two adjacent counties. Since the analysis for autocorrelation is made with k-nearest neighbour method, where $k = 2$, it is impossible to determine two nearest neighbours for the two end counties of the state of Delaware, which in turn makes it impossible to calculate Moran's or Geary's statistics. Other states, such as Alaska, Hawaii, Virgin Islands and Puerto Rico were also dropped for this same reason, limiting the material to the continental states.

For the separate state analysis, the complete autocorrelation structure for each state is reported in the appendix section. The states, which both, Moran's and Geary's statistics, gave a significant value under $p < 0.01$, were *Alabama, Arkansas, Colorado, Georgia, Indiana, Maine, Massachusetts Michigan, Minnesota, Missouri, New Jersey, Ohio, Oregon, Pennsylvania, Texas, Virginia* and *Washington*. More importantly, the individual inspection shows that when both measures are taken into account, the pattern of significant autocorrelated states is not indicating a distinct cluster.

3.3 Spatial Regression Modelling

The spatial regression was conducted for the three different areal units, two of which were on the national level. These two national levels were county and state specific. For the final analysis, each state was modelled separately. The regions described in table 3.1 were used as a trend for the models.

3.3.1 County Specific Analysis

On a national level the spatial regression model used population size and population density as continuous and the regions as categorical coefficients. The overall SAR model was

$$\begin{aligned} Z_i = \rho \times \frac{\sum_{j \in N_i} Z_j}{|N_i|} + \beta_1 \times \text{population} + \beta_2 \times \text{population density} \\ + \beta_3 \times \text{region} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2 I), \end{aligned}$$

where Z_i represents an observation in a geographical location and Z_j represents the surrounding locations. $|N_i|$ represents the amount of neighbour the location i has and N_i is the set of neighbouring locations. ρ in this model is the autoregressive correlation coefficient between the neighbouring units. The β_i describe regression coefficients for population, population density and the effect of different regions. The results from this model found the coefficient for population to be 0.1027 and 8.7434 for population density. The regional coefficients were −5692.30 for northeast, −411.30 for southeast, −245.72 for midwest, 403.61 for southwest and −2480.2 for west. Most of the coefficients were significant on $p < 0.001$ level. The significance level for southeast, midwest and southwest regions was $p > 0.05$, thus non-significant. The lowest p-value for these regions was for the southeast region for $p = 0.4029$. The coastal regions for northeast and west were both significant and surprisingly, their effect was a negative. This means that the only significant trend was found in the west and northeast regions and the inner areas of the country did not have significance for the effect of region. The CAR model is similar, as presented in chapter 2.3.1 so that we get Z_i, given their surrounding values Z_j. The resulting coefficients for continuous variables were 0.1030 for population and 8.3303 for population density. For the regional coefficient, the values were −5072.10 for northeast, −474.08 for southeast, −302.29 for midwest, 295.81 for southwest and −2414.2 for west. The same coefficients were significant in the CAR model as in the SAR model, with only the effect of southeast, midwest and southwest regions not being significant. The Nagelkerke pseudo R^2 reached over 0.80 for both SAR and CAR models, SAR being 0.8310 and CAR being 0.8312. This means that the CAR model explained the variance in the data slightly better than the SAR model. The AIC was 68,507 for the SAR model and 68,504 for CAR. The fitted values for each county from both, SAR and CAR models, can be seen in figure 3.6. The resulting patterns for infections seem to be similar to the observed values. The middle and northern sections of the country have more counties with fewer infections in them. The areas of western coast and southern border, where the states of California and Arizona are located, have counties with more infections in the results from both models, which also corresponds with the observed infection counts. The autocorrelation of the residuals from the comparison of fitted and observed values was also inspected. This gives more insight on the goodness of fit for the models. For the national level, considering the counties as units, Moran's I was 0.0012, $(E(I) = -0.0003)$ and Geary's C was (1.1949, $(E(C) = 1)$ for the SAR model. The p-values were $p = 0.4617$ for Moran's

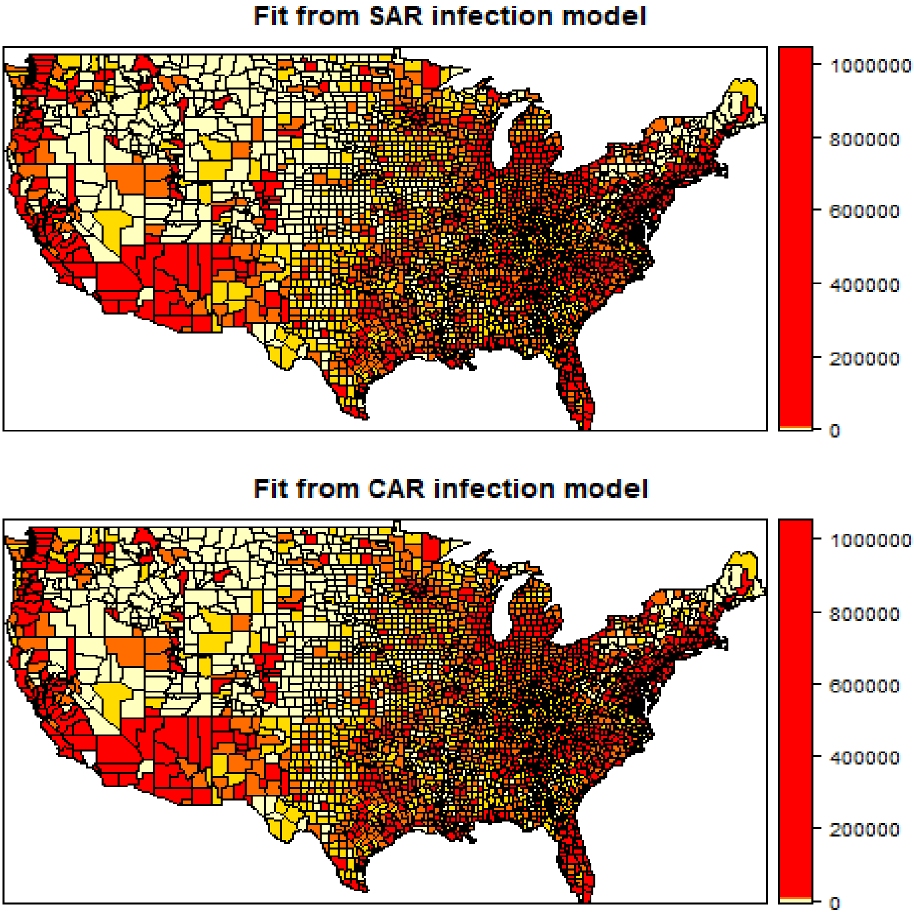
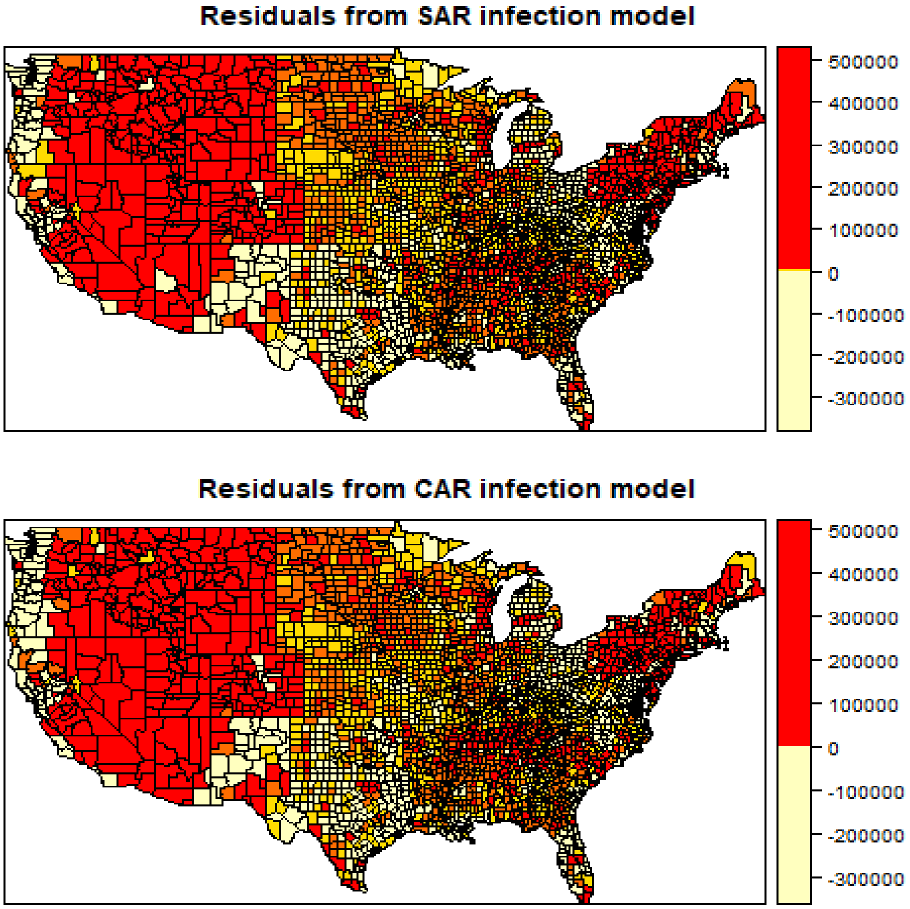


Figure 3.6 Grayscale map of fitted infection values for each county in the US from SAR and CAR models.



Residuals from SAR infection model

Residuals from CAR infection model

Figure 3.7 Map of residuals for each county from SAR and CAR models.

I and $p = 0.9826$ for Geary's C, respectively. Since both of the p-values were not significant, it can be said that the residuals did not have sufficient autocorrelation and thus the model formed a relatively good fit. For the CAR model, Moran's I was 0.0996, $(E(I) = -0.0003)$ and Geary's C was (1.0964, $(E(C) = 1)$. The p-values for these measures were $p < 0.001$ for Moran's I and $p = 0.8581$ for Geary's C. Since Moran's I was significant, it signals that there was autocorrelation left in the residuals. This then means that the SAR model performed relatively better than the CAR model in modelling the infections on the national level, when the units are counties. Interestingly, the results from Moran's I and Geary's C differ greatly for the CAR model.

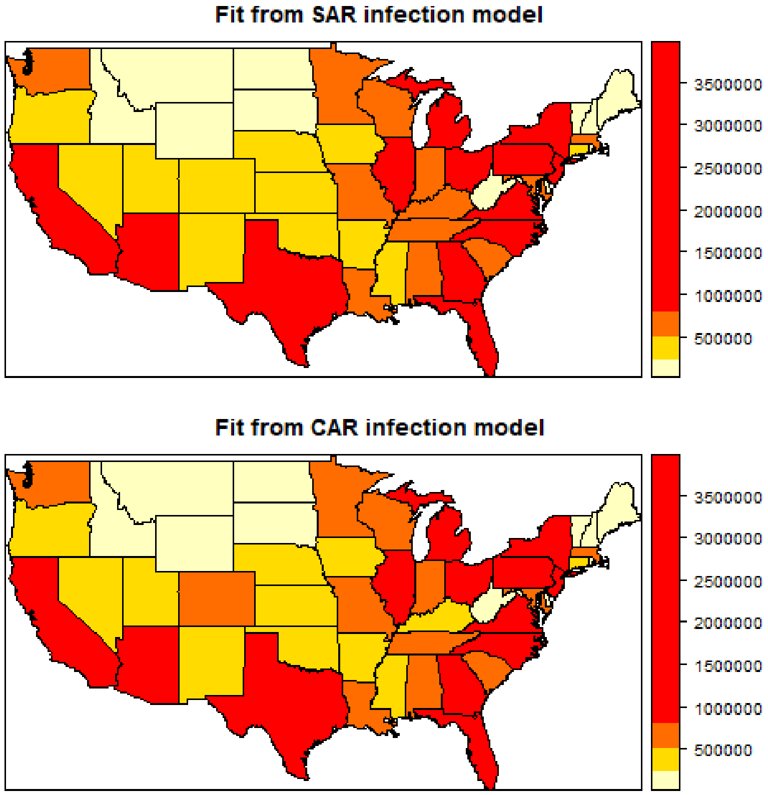
The residuals from the fitting process were plotted to inspect the pattern of the residuals. The results from this plot are in figure 3.7. There seems to be some areas, where the residuals are distinctly lower than elsewhere. These areas are in the middle of the country, spanning from north to south. The western areas of the US seem to have relatively many areas, where the residuals were high. Some western coastal areas seem to have negative residual values as well. The residuals in the northeastern region seem to be relatively high as well.

3.3.2 State Specific Analysis

For the national state specific SAR and CAR models, the coefficients for SAR were 0.1008 for population and 5.5439 for population density. The coefficients for regions were -7493.10 for northeast, $35,382$ for southeast, $21,218$ for midwest, $61,029$ for southwest and $-63,645$ for west. The coefficient for population was significant for $p < 0.001$, while none of the other coefficients were significant. This means that on the state level, the only significant factor for the infections was the population in the states. The coefficients for CAR model were 0.1002 for population and -1.4428 for population density. The p-value for population was significant ($p < 0.001$) and population density was not significant. None of the coefficients for regions were significant for the CAR model, the lowest being $p = 0.0684$ for southwest region. Nagelkerke R^2 was 0.9856 for SAR and 0.9838 for the CAR model, respectively. Interestingly, the SAR model explained the variance in the data better than the CAR model did. The AIC was 1273.40 for the SAR model and 1279.20 for CAR.

The fitted estimates are shown in figure 3.8 respectively for SAR and CAR models. Both SAR and CAR models produced similar results, with some differences from each other. The highest infection numbers seem to appear in California, Texas, Florida and New York, while the lowest were in the northern regions, including the states of North Dakota and South Dakota, Nebraska, Idaho, Wyoming from inner parts of the country and Maine, New Hampshire and Vermont in the northeast coastal area.

The plot for the residuals of the state specific model can be seen in figure 3.9. Based on the figure, there seems to be a similar area of lower residuals in the middle of the US as was with the county specific analysis. The western coastal states seem to have strongly negative residuals and the inland states have positive residuals. The northeastern region has some negative values around Virginia and North Carolina.



Fit from SAR infection model

Fit from CAR infection model

Figure 3.8 Grayscale map of fitted infection values for each state from SAR and CAR models.

The inspection for autocorrelation in the residuals for the state specific models produced somewhat similar results to those from the county specific analysis. Moran's I was 0.0654, $(E(I) = -0.0208)$ and Geary's C was 0.8272, $(E(C) = 1)$ for the SAR model. The p-values for these measures were $p = 0.2467$ and $p = 0.1121$. The residuals did not have sufficient autocorrelation and thus the model gave a relatively good fit. For the CAR model, Moran's I was 0.2551, $(E(I) = -0.0208)$ and Geary's C was 0.6296, $(E(C) = 1)$. The respective p-values for these measures were $p = 0.0146$ and $p = 0.0042$. Both, Moran's I and Geary's C are significant, which means that there was autocorrelation in the residuals for the CAR model. It can be concluded that the SAR model formed a better fit for the data in the state specific analysis as well. There does not seem to be difference for the interpretation between the county specific and state specific models for the national level. In both approaches, the residuals from the CAR model seemed to have some autocorrelation, while no significant autocorrelation for either of the SAR models was detected.

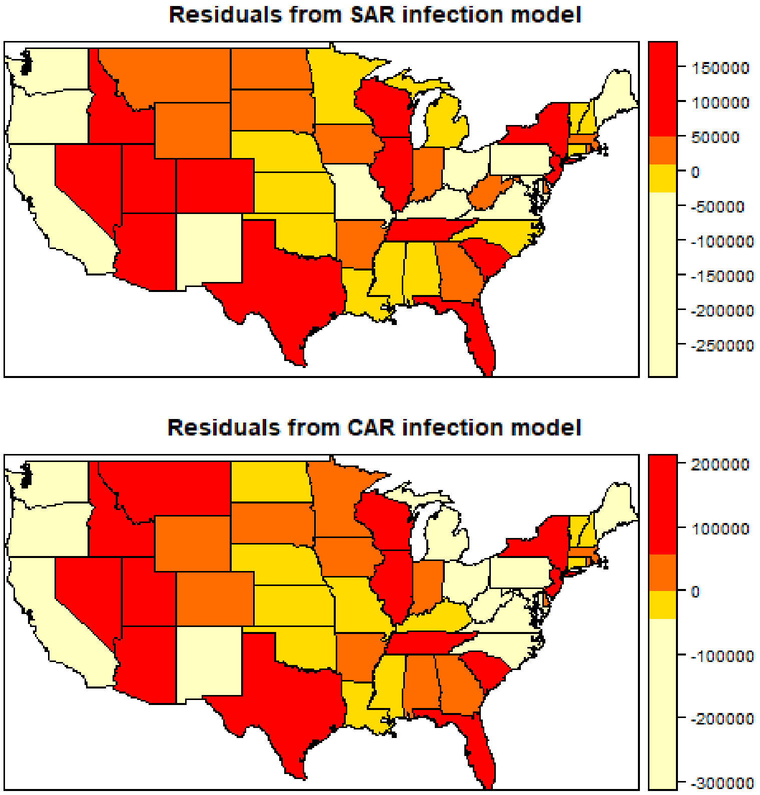


Figure 3.9 Map of residual for each state from SAR and CAR models.

As stated in the introduction, Mollalo et al. (2020) aimed to model the spread of COVID-19 with GWR, which yielded a result of $R^2 = 68.1$ for the US. On a national level, the SAR and CAR models, when considering the counties as units, covered over 80% of the variability, which seems somewhat satisfactory, considering the model held only population size and population density as continuous independent variables, and regional categories as a categorical variable. Moran's and Geary's statistics were not significant for the residuals for the SAR model and only Geary's C was significant for the CAR model. The states as units approach produced similar results. The overall Nagelkerke R^2 was 0.99 for SAR and 0.98 for CAR models, which suggests a very good fit. However, in the residual inspection of the models it was noticed that Geary's C was significant with $p < 0.05$ for the CAR model. The SAR model formed a better fit for the data in both national levels.

3.3.3 Analysis For Each State

The full table for the regression coefficients for counties of each state received from SAR and CAR models, is presented in the appendix section. The maximum positive coefficient for population in the SAR model was for Rhode Island (population = 0.1626 and the minimum was for Vermont (population = 0.0258). The maximum for population density was for Arizona (population density = 760.4765) and minimum for New Mexico (population density = −121.3584). The maximum coefficient for population size in the CAR model was for Florida, the value being 0.1613 and the minimum for Vermont 0.0246. The maximum-minimum values for population density in the CAR models was 99.9058 for Arizona and −67.4166 for New Mexico. There seems to be some differences in the effect sizes for the coefficients from the two models.

There seems to be a greater amount of variation for the effect of population density, when compared to population size. The effect seems to behave very differently depending on the region, while the coefficient for population stays more or less the same throughout the data. The coefficients reported here are only for the ones, where $p < 0.05$. Non-significant coefficients were found for example for Colorado and Kansas, where the coefficients for the effect of population density were both above 0.1. Altogether, the coefficient for population was found to be significant in 46 states in both SAR and CAR models, while population density was significant in 29 SAR and 27 CAR models.

The Nagelkerke pseudo R^2 was varied between the models for all states. The median for SAR models' pseudo R^2 was 0.9717, while the minimum and maximum values were 0.6002 and 0.9995 respectively. For CAR models, the median for pseudo R^2 was 0.9809 and the minimum-maximum values were 0.6144 and 0.9995. There is some variety in how much variance is explained by the spatial regression models, but the lower-end seems to include mainly outliers, which the models could not cover properly. The median is very high for both models, which suggests that the models generally cover the variations in the states rather accurately.

3.4 ABC Modelling

ABC methods are frequently used in studies that involve creating data by pseudorandom sampling, also known as simulation studies. Often in these studies, the aim is to evaluate the behaviour of a statistical model, when mathematical proof is difficult or impossible to find. ABC methods are used similarly, but the aim of the simulation is to study the behaviour of some variable, the behaviour of which is difficult or impossible to calculate. An example of this is shown in a blog post by Rasmus Bååth, where the ABC-Rejection algorithm was used to give an estimation of the total amount of socks in the laundry, given a sample of 11 discreet socks. The results from this analysis gave accurate results, the estimated total amount of socks being 44, when the total amount of socks in the data was 45. The full information for this study is given in the appendix A.

In order to model the incidence and to compare the results, it is necessary to model the current circumstances based on a time window from the start of the pandemic. As stated in chapter 3.1, the incidence rate for the infection numbers did not begin climbing until after 60 days from the first registered case of COVID-19 had passed. These first 60 days of low activity can skew the interpretation for the analysis. Thus, it is reasonable to not include the first 60 days of the data, when the pandemic was not yet in its acceleration phase. This assumption is strengthened by inspecting the infection rate from the first 100 days. The average of infections for the full data from the first 100 days is $10,417.84$ ($sd = 13,362.16$) and $25,596.42$ ($sd = 8146.873$) for the data from the last 40 days. As expected, the mean for daily infections increased greatly after the exclusion, with the size being nearly 2.5 times larger. The standard deviation also decreased by over 5000, but the change is not as drastic as with the mean. The length of the original data spans to 489 days and after the exclusion the length is 429 days. The new starting point for the analysis will then be the starting point of the acceleration phase, which is the 61st day from the first case of COVID-19 in January 2020.

When the infection rate is divided by the population of the US, we get 0.000032 with the full data from the first 100 days, and 0.000078 from the 61st to 100 days. This means that based on the first 100 days, from 0.0032% to 0.0078% of the entire population of the US gets infected daily. If the infection rate is then divided by the infection percent, we get the original US population again, meaning $\frac{10{,}417.84}{0.000032} =$

$327,409,635$. When the estimated amount of infections is divided by the computed percentage of the total population, we get a figure representing the US population. Formally, we set

$$\frac{\text{infection rate}}{\text{infection percentage}} = \text{population}.$$

This value can be used to measure, how close is the computed population to the true population of US. The reason for this is that at 100 days, there is no further comparison point for the daily infections and thus an estimated average infection rate is not a viable estimate to compare to. The total population of US is known however, which is directly linked to the infection rate and this makes it suitable for comparing the estimates from the ABC analysis. When the distance between the true US population and the estimated population is acceptably small, the estimate for infection rate can be accepted to the posterior distribution. Thus, the more likely an estimate is, the more likely it will repeat in the simulations, thus accumulating a higher point in the posterior distribution.

When the daily infection rate is multiplied by days, the resulting values should have similarity with the number of infected on that specific day. Therefore, when the posterior distribution for estimated daily infections is then multiplied by days, a distribution of possible infections at that time is calculated. This can be then compared to the observed values at that date. Together, the posterior distribution of infection rates and the predictions based on these values produce a window of possible outcomes for the spread of COVID-19, which in turn functions as a possible threat assessment. The next sections will focus on these estimations and the results produced by the different ABC methods.

3.4.1 Infection Rate Approximation

In this thesis, the ABC methods were used as a way to predict the future spread of SARS-CoV-2 with limited amount of data from the early days of the pandemic. The full aim was not to accurately state, which prediction was going to happen, but to offer an array of possible infection rates based on the early data in order to provide a threat analysis fo the spread of the disease. Different ABC methods were then used for creating these estimates.

To acquire candidate samples for average daily infections, a $Gamma$ distribution was used. The variance for the analysis was not limited to that of the data from the 61st to 100 days, since it would introduce too little variability for the possible daily infections. Therefore, the variance was multiplied by 2.5. The proposal estimates followed a $Gamma(\alpha, \beta)$ distribution, where α and β were calculated with the mean from the data from the last 40 days and the multiplied variance. This proposal distribution was chosen, because it gave only positive values as estimates for the infection rate. Since there is no knowledge of the observed infection percentage after the first 100 days, the infection percentage has to be estimated. Each infection percentage was sampled from $Beta(4, 22000)$ distribution. This distribution provides samples, where the observed infection percentage of 0.000032 or 0.000078 are below the 1st quarter of samples, signifying that the daily percentage from the first 100 days assumed to be relatively small in the overall distribution of infection percentage. The upper limits of the percentage sampling distribution produce suggested infection percentages of roughly 10 times the observed percentage from the first 100 days, which offer more variability to the estimation of the possible outcomes for infection rates.

The sequential Monte Carlo version of ABC focused on calculating the posterior distribution of infection rates by forming intermediate posterior distributions, the limits of which are shrunk throughout each iteration loop to get a more accurate representation of the posterior distribution. The user chooses each error tolerance threshold and the sequence chosen for this thesis was $\varepsilon \in [1000000, 500000, 100000]$. These error thresholds describe the acceptable limit for the difference between the estimated population and the population of United States. The perturbation kernel K followed the $Uniform$ distribution with $K \sim U(0.85, 1.25)$. The last threshold of 100,000 was the limit used for each of the ABC algorithm variants. This means that if the estimated population differs from the real US population by 100,000 or less, we can accept the proposed infection average to the posterior distribution.

The ABC-Rejection algorithm needed 59,514,392 iterations to reach 10,000 particles. The overall mean for the ABC-Rejection algorithm was 34,833 cases per day, which corresponds to an estimation of 14,943,363 cases at 429 days. The maximum value was 95,407 which, when multiplied by 429 days, produced the closest estimate to the observed infection size of 32,832,058 at 429 days from the ABC-Rejection approach. The posterior distribution of infection rates can be seen in figure 3.10. In the plot, the vertical red lines depict the observed average of daily infections at different days, which are indicated with text. The black curve depicts the posterior distributions of the estimates produced by the ABC-Rejection approach. In the case of ABC-Rejection algorithm, the peak of the distribution is quite far from the observed infection rate at 429 days, which would suggest that the ABC-Rejection algorithm does not give accurate predictions that far into the future, even though the observed infection rate is below the maximum value. The kurtosis of the posterior distribution seems quite high in comparison to the distributions of the other approaches and with the least variability. The infection rate at 429 days situatea almost to the tail end of the distribution on the right-hand side. On the other hand, the ABC-Rejection approach could predict the infection rate at 100, 200 and 250 days relatively well, with their respective vertical lines situating well inside the posterior distribution of infection rates. The overall time spent on the ABC-Rejection analysis took roughly 7 minutes to produce the posterior distribution.

The estimates given by ABC-MCMC are more accurate in comparison with the results from ABC-Rejection, the mean being 66,121, which is closer to the observed infection rate than in the ABC-Rejection algorithm. The observed amount of infected at 429 days is closer at 3rd quarter of the posterior distribution at an estimated 34,460,941 infected in 429 days, with the maximum reaching over 73 million. The estimate at 34 million is roughly only 2 million off from the observed amount of infected at 429 days. The posterior distribution for the estimates of infection rates, given by the ABC-MCMC approach can be seen in figure 3.11. In the figure, it is possible to see that unlike in the rejection approach, the observed infection rate at 429 days is much further inside the distribution of daily infection rate estimates. However, the tail of the ABC-MCMC on the right-hand side seems to continue much further after the elbow of the curve. This does not pose great problems, because in most cases, the wider tails have relatively low densities. It is also evident in the figure that the estimates of the ABC-MCMC approach gave the most accurate representation of the three approaches.

For the other infection rates, ABC-MCMC seems to provide different results. The infection rate at 250 days seems to be little behind the peak of the posterior distribution. The lower infection rate estimates do not seem all that likely anymore and the general interpretation from the ABC-MCMC posterior distribution is that the original infection rate at roughly 25,000 could be only the smallest infection rates, where the average infection rates can reach. The ABC-MCMC approach was more

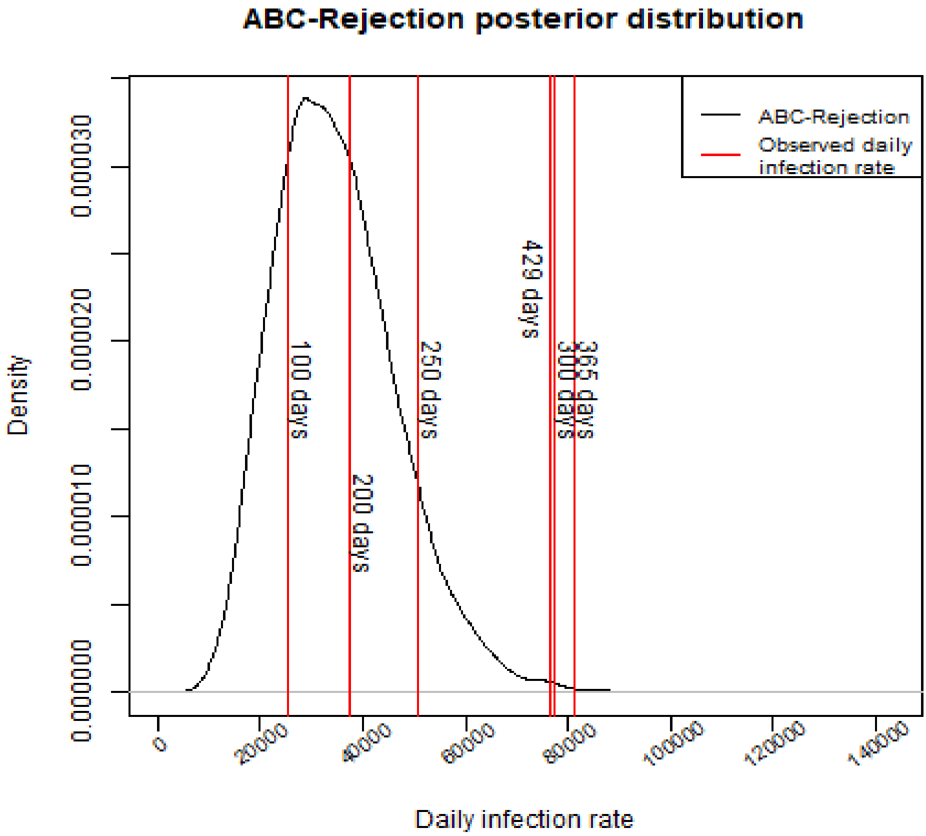


Figure 3.10 The distribution of estimates from ABC-Rejection algorithm for infections rates for COVID-19. The red vertical lines depict the average of daily infections at different points from the start of the acceleration phase.

time-effective, when compared to the other approaches and produced the posterior distribution in roughly 4 minutes. The iterations needed to reach 10,000 accepted estimates for infection rates was the second lowest, requiring 30,305,687 iterations.

The differences between the estimates of ABC-Rejection and ABC-MCMC reflect the moving sampling point of the MCMC approach. This caused the mean of the infection rates to be much further than in the results of the ABC-Rejection algorithm. Both of the algorithms have the same starting point, but since the MCMC has the freedom to shift the sampling point for infection rate, it will move towards an area of higher acceptance rate. The results of each approach can also be seen in table 3.3, in which the posterior distribution characteristics are described for the results

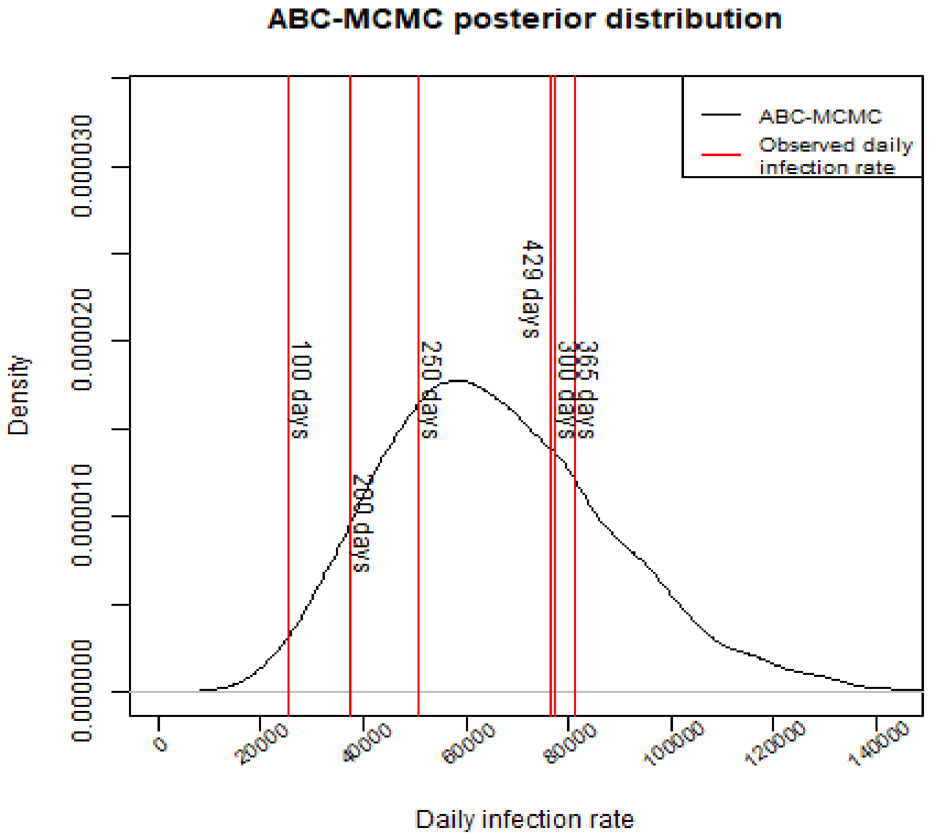


Figure 3.11 The distribution of estimates from ABC-MCMC algorithm for infections rates for COVID-19. The red vertical lines depict the average of daily infections at different points from the start of the acceleration phase.

from each ABC approach along with the iterations the approach needed to achieve 10,000 accepted estimates. Under the summaries of the infection rates, there is also the prediction for infected people at 429 days, which was calculated by multiplying the infection rate by 429 days.

The estimates from ABC-SMC approach were closer to the observed infection rate at 429 days than those from ABC-Rejection algorithm and relatively similar to those from ABC-MCMC. The mean for ABC-SMC was 54,037, which corresponded to 23,181,899 cases at 429 days. The 3rd quarter estimate of 64,054 was much closer to the true size of 76,531.60, but was still off by 12,477.60. This produced an estimated value of 27,479,166 infections at 429 days, which is off from the real value by

Measurement	ABC-Rejection	ABC-MCMC	ABC-SMC
Iterations	$59,514,392$	$30,305,687$	$28,012,107$
Min (at 429 days)	6230 $2,672,581$	6540 $2,805,456$	5875 $2,520,315$
Max (at 429 days)	$95,407$ $40,929,405$	$171,049$ $73,380,102$	$134,199$ $57,571,405$
Mean (at 429 days)	$34,833$ $14,943,363$	$66,121$ $28,365,925$	$54,037$ $23,181,899$
1st Quarter (at 429 days)	$26,205$ $11,241,868$	$49,635$ $21,293,488$	$42,600$ $18,275,569$
Median (at 429 days)	$33,572$ $14,402,180$	$63,905$ $27,415,084$	$52,805$ $22,653,139$
3rd Quarter (at 429 days)	$42,107$ $18,064,048$	$80,329$ $34,460,941$	$64,054$ $27,479,166$

Table 3.3 Summaries for posterior distributions of infection rates from ABC-Rejection, ABC-MCMC and ABC-SMC methods. The value below each estimate depicts the estimated amount of infected at 429 days, which were obtained by multiplying the estimate for infection rates by 429 days.

roughly 5 million. The ABC-SMC needed the least amount of iteration to get this estimate as well, requiring 28,012,107 iterations for 10,000 particles. This result is however, only for the final iteration level, where the tolerance threshold was 100,000 and with earlier intermediate levels having already been estimated, which means that the total amount of iteration is much higher, which is also reflected on the time spent on producing the posterior distribution. The ABC-SMC algorithm needed the most amount of time for the iterations, spending roughly 12 hours on the process. This is vastly more time-consuming in comparison to the earlier methods. The increase in time deficiency may be explained by the combination of multiple factors. Firstly, the ABC-SMC developed two other intermediate posterior distributions before the final results. Secondly, the approach used weighed sampling from these previous distributions, which most likely skewed the amount of represented proposal particles. Thirdly, the process of calculating new proposal infection rates required to calculate new values for α and β for the $Gamma$ sampling distribution. All of these steps produce a more complex process than in ABC-Rejection and ABC-MCMC approaches, which in turn makes the process less time efficient.

The observed infection rate at 429 days seems to be much closer to the centre of the distribution in the ABC-SMC approach, when compared to the ABC-Rejection approach. The estimates resulting in the same outcome seem to be much more

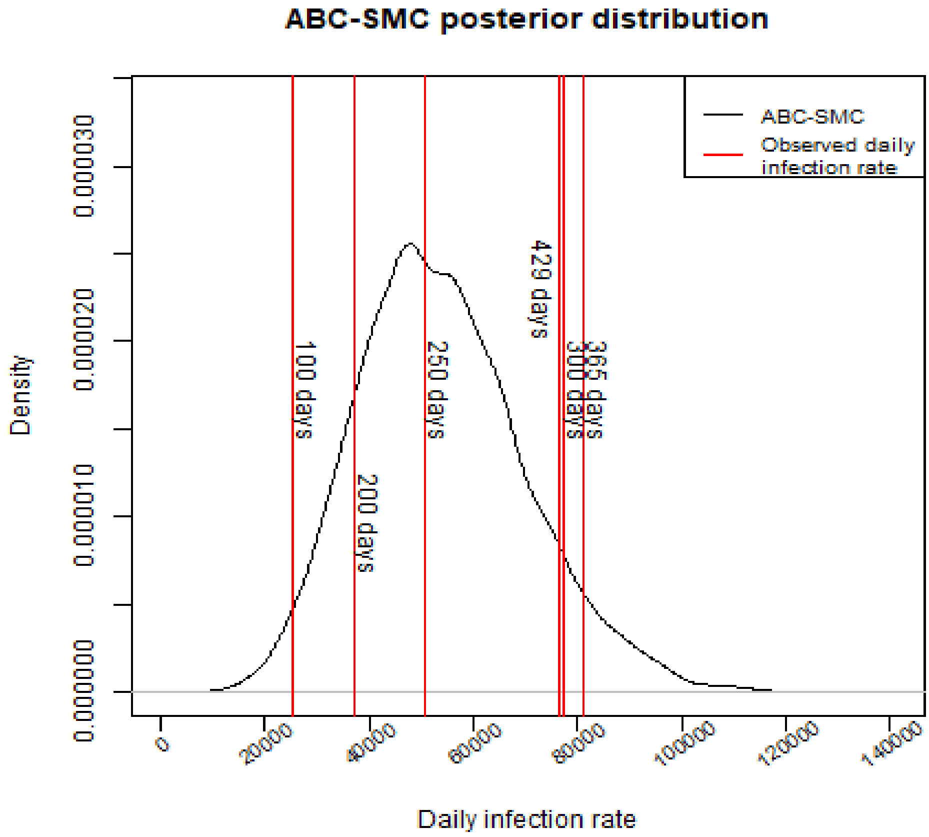


Figure 3.12 The distribution of estimates from ABC-SMC algorithm for infections rates for COVID-19. The red vertical lines depict the average of daily infections at different points from the start of the acceleration phase.

common in the ABC-SMC approach and the observed infection rate seems much more plausible. One notable difference between the results from ABC-Rejection and ABC-MCMC approaches, is that while the ABC-SMC approach provided a wider window for possible infection rate outcomes, the posterior distribution stayed relatively condensed, having much shorter tails than in the posterior distribution of the ABC-MCMC approach. The posterior distribution given by ABC-SMC can be seen in figure 3.12.

From figures 3.10, 3.11 and 3.12, it is possible to deduce that at the minimum, in the context of the information given at the first 100 days, using the models with the prior expectations of the spread of the COVID-19, it would seem that the posterior distributions given by ABC-SMC and ABC-MCMC captured the observed daily infection rate at 429 days as a plausible outcome. The observed rate at 429 days was very near the 3rd quarter estimate of the ABC-MCMC and while the the observed infection rate was outside the 3rd quarter of the ABC-SMC approach, it is still relatively well inside the posterior distribution.

The results may be a merit of the ABC method family to expect more unlikely developments for infection rates to occur. In figure 3.2, it was seen that there was a sequence, where there were much more daily infections at and after October of 2020 and lasting until January 2021. This sequence most likely introduced temporal skew into the average infections, which is also evident in the red vertical lines depicting these average infections at different points of time in the figures representing the posterior distributions from the different ABC approaches. Before day 250, the infection rates seem to be increasing at steady intervals, but there is a large gap after the infection rate at 250 days. Incidentally, the period of higher daily infections begun after 250 days. This period affected the average of the infections, which is why the red vertical lines for 300 and 365 days are further in the distribution than for 429 days, when the period had passed. These higher averages are also present inside the posterior distributions, which is why it can be argued that the ABC methods could predict the possible period of higher infection rates, but found it less likely.

The basic method framework of ABC is flexible and offers a lot of room for various implementations and finetuning. One of the main merits of the ABC methods is that they are easy to apply and modify for different applications. This is evident from the different variants, which all have unique approaches to the approximation process. This manageability and cost-effectiveness of the method makes its use appealing.

The simplest ABC method, ABC-Rejection, produced the most inaccurate results. This is most likely due to the most simplistic nature of the method, with the mean of the distribution, from which the proposal particles are sampled, remaining stationary. This may have contributed to the estimates of the distribution being relatively small in comparison to the ones produced by ABC-MCMC and ABC-SMC. The suggested maximum for the infection rate was 95,407, with the mean being 34,833 cases a day, which is not even half of the observed infection rate of 76,531.60 at 429 days. The density plot for the results of ABC-Rejection algorithm, as seen in figure 3.10, show that the kurtosis of the posterior distribution is the steepest of any of the results from the other ABC algorithm results. This result is also evident in the relatively high number of iteration needed in order to reach 10,000 estimates. The curve is very steep, with the lower-ends being areas of very low probability. The highest estimate of the ABC-Rejection algorithm captures the observed infection rate at 429 days under the distribution of estimates, but the estimate is clearly at the tail-end of the distribution, visually showing itself as almost flat on the y-axis. It would seem from these results that while the ABC-Rejection algorithm did succeed technically in having the observed infection rate at 429 days as a possible outcome for the progress of the epidemic, the method provided a more limited estimate window of the possible outcomes for the infection rates, when compared to the results from ABC-MCMC and ABC-SMC. The ABC-Rejection algorithm did however, successfully capture the predictions for the average infections at 100, 200 and 250 days, which in turn could suggest that the ABC-Rejection algorithm may be more suitable in predicting outcomes for data sets with less variability.

ABC-MCMC provided much more accurate results than ABC-Rejection, having the mean be 66,121 cases a day. This estimate is still 10,000 off from the observed infection rate, but the estimate at 3rd quarter of the distribution seems to be quite accurate. The observed daily infection rate at 429 days is well inside the posterior distribution. This was also evident in figure 3.11. ABC-MCMC provided a much wider range for the estimates, clearly showing the approach moving the proposal particle sampling point towards an area of higher acceptance rate and thus approaching the observed rate. Interestingly, the minimum value for ABC-MCMC is the highest of all of the ABC variants. This also points to one of the drawbacks of the ABC-MCMC. Since the starting points are similar with ABC-Rejection, depending on the run of the ABC-MCMC, the estimate range will always vary and if the algorithm happens to receive a couple of successful estimates from a lower probability area, the algorithm moves its sampling point towards it, which temporarily traps it to an area of lower acceptability. This is also evident in the relatively long tail of the ABC-MCMC posterior distribution. This issue highlights the possibility for the need of the burn-in period mentioned earlier in chapter 2.4.1 for the introduction for ABC-MCMC. Since the method takes some time to gravitate towards the convergence point, fluctuation in the estimates at the start of the algorithm cause some increase in the range of the estimates. It is, however, a matter of debate, how many of the estimates should be neglected, should the burn-in period be introduced. The estimates may very well be valid and excluding them could omit valuable estimates. The interpretation of the density plot for the infection rate estimates given by ABC-MCMC leave much more room for different outcomes for the spread of COVID-19. The biggest difference between ABC-Rejection and ABC-MCMC is the kurtosis between the two densities. The top of the distribution, given by ABC-MCMC, is much lower and the range of the distribution is larger than the one given by ABC-Rejection.

The ABC-SMC approach gave more similar results to ABC-MCMC than ABC-Rejection, but with few notable differences. The 3rd quarter of estimates is as high as 64,054, which in turn provides a prediction of 27,479,166 infections at 429 days, which is much closer to the observed amount of infections than the estimates produced by ABC-Rejection approach were. From the density plots, it is discernible that the curve of ABC-SMC posterior distribution is much more condensed, when compared to the plot of the ABC-MCMC posterior distribution. This is most likely due to the prior intermediate distributions of the ABC-SMC approach. Since the approach re-runs the particle assigning loops with diminishing acceptance thresholds, it is possible to shake off parts of the ends of the distributions, which in turn provides a sharper posterior distribution. In the context of the analysis of this thesis, the ends of the distribution received much lower weights in sample selection in the intermediate distributions. Thus, the lower estimates were selected much more rarely than their higher counterparts, resulting in the lower-ends of the posterior distribution having only minimal representation, if not completely excluded. This is a property missing entirely in the two other algorithms. One possibility to improve the accuracy of the ABC-SMC would be to add additional tolerance threshold levels for error. This would result in additional intermediate posterior distributions, which in turn would also quickly increase the time required to complete the run of the ABC-SMC algorithm. However, the analysis used only 3 iterations for the algorithm, which is a relatively small amount of intermediate posterior distributions.

There are reasons, why the results from the ABC approaches can be criticized, one of them being a possibility for the models used to get the estimates not depicting the spread of COVID-19 accurately. With limited data only from the first 100 days, the expected curve of the spread of the epidemic may be very difficult to estimate. Here, a $Gamma$ distribution was used for sampling proposed infection rates and a $Beta$ distribution was applied for sampling total infection percentage, which could have provided incorrect results. However, at 100 days, there is no further information that will indicate accurately, which distribution would most accurately depict the spread of the virus and therefore it is left for the user to decide the generative model for the ABC method applied.

In addition to the models themselves, a point of criticism is towards the parameters inputted by the user. Parameters, such as the particle sampling point, the acceptance threshold for the error and the amount of accepted posterior estimates required, until the iteration is stopped, are set by the user and each can have effect on the accuracy of the estimates.

Another point of criticism is towards the interpretation of the nature of the spread of COVID-19. It is possible that the nature of the pandemic is erratic and therefore unforeseeable, resulting in the more highball estimates of infection rates also being rather low. Some of the open discussion online and in different news outlets concerning the pandemic in the US has created a belief that the COVID-19 has spread with surprising vigour. There may be inaccuracy in the estimates, because the scale of the spread of SARS-CoV-2 was unpredictable based on the 100 days. On the other hand, it can be argued that the ABC-MCMC and ABC-SMC approaches gave somewhat accurate results, since all of the infection rates at different time points were inside the posterior distributions. The later time points after 300 days after the period of high infection counts, the infection rates are not situated even at the very tail-end of the distributions, making them seem plausible outcomes. It is not yet known, where the final average of the daily infections will be at the end of the COVID-19 pandemic, so the question for that average infection rate has to be left somewhat open-ended.

3.4.2 Approximation With Uniform Distribution

The ABC analysis was also run with a noninformative proposal distribution for comparison with the results from $Gamma$ sampling distribution. In the analysis the $Uniform$ distribution used the same mean for infections from the 61st to 100 days. The minimum and maximum values for the sampling range of the $Uniform$ distribution were calculated by adding and reducing the doubled standard deviation to the average infections. This limit was not enough for the ABC-Rejection algorithm to cover the extent of the infections. The mean from the 61st to 100 days was 25,596.42 and the standard deviation, calculated from the variance multiplied by 2.5, was 12,048.20. This limits the possible upper limit for proposed infection rates to $25,596.42 + 2 \times 12,048.20 = 49,692.83$, which is well below the observed value of 76,531.60. This limit would have severely hindered the accurate approximation results. In order for the ABC-Rejection algorithm to be able to give reasonable estimates for infection rates, the standard deviation was multiplied by 3 for the ABC-Rejection algorithm. The standard deviation used for ABC-MCMC was calculated from the original variance multiplied by 2.5. The comparative analysis was not done with the ABC-SMC algorithm, since it took very long to form the posterior distribution for infection rates. The results for the ABC-Rejection algorithm with a non-

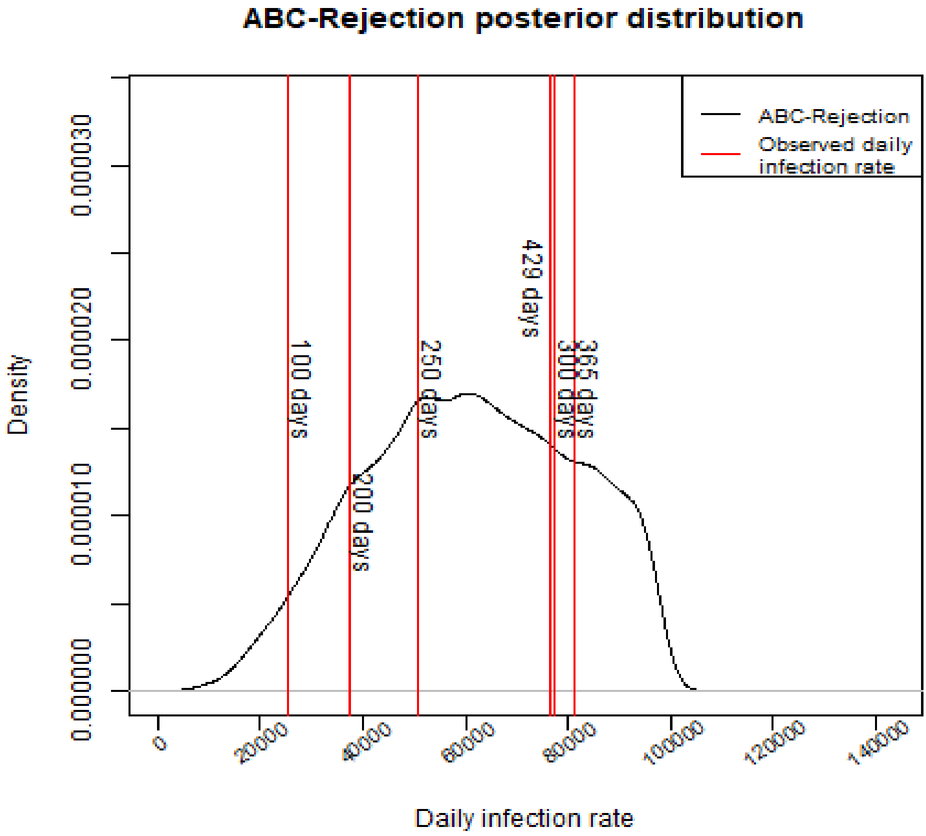


Figure 3.13 The distribution of estimates from ABC-Rejection algorithm for infections rates for COVID-19 using noninformative prior distribution. The red vertical lines depict the average of daily infections at different points from the start of the acceleration phase.

informative sampling distribution can be seen in the figure 3.13. With the enlarged standard deviation, the ABC-Rejection algorithm produced better estimates for the infection rate. The mean for the comparative analysis was 60,956, which is almost 2 times the estimate from the earlier 34,833. The 3rd quarter estimate for the infection rate was 77,068, which very near the observed rate of 76,531.60. It is also noticeable that the tail of the distribution on the right-hand side declines very quickly. This is because of the maximum value achievable by the ABC-Rejection algorithm with the updated standard deviation was 97885.65, which is very close to the maximum accepted estimate of 97,884. The comparative analysis needed less iteration to get 10,000 accepted estimates, ending after 50,123,728 iterations. In the figure 3.14, the

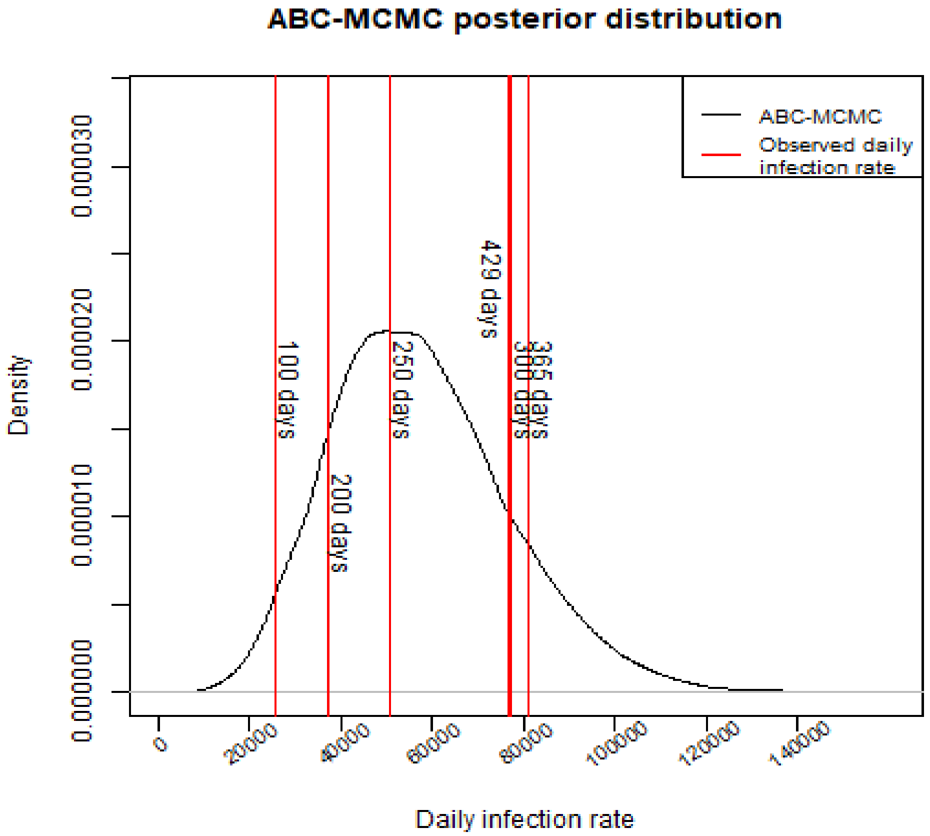


Figure 3.14 The distribution of estimates from ABC-MCMC algorithm for infections rates for COVID-19 using noninformative prior distribution. The red vertical lines depict the average of daily infections at different points from the start of the acceleration phase.

peak of the distribution seems to be at around 60,000, after which the estimates begin to lower. This is very different from the results of the previous ABC-Rejection algorithm. The shape of the posterior distribution is distinctly different from the previous posterior distribution for infection rates. It would seem that the noninformative proposal distribution, with a larger range for proposal values, produced more accurate estimates than the original.

The results of the ABC-MCMC algorithm, with a noninformative proposal distribution, were closer to the values of the posterior distribution with a $Beta$ proposal distribution. The mean for the ABC-MCMC was 57,462, when the original was 66,121. The 3rd quarter estimate was 69,363, when the original was 80,329. The comparative ABC-MCMC algorithm needed 30,559,687 iterations for 10,000 estimates for infection rates, which is only 254,000 iterations more than the original. The distribution of these estimates can be seen in the figure 3.14. The differences between the figures 3.11 and 3.14 are not as distinct as with the rejection algorithm. In the noninformative sampling distribution, the peak of the posterior distribution seems to be earlier than in the previous one. The peak seems to be quite near the infection rate at the 250 days, while the estimates for the later dates seem to be situated lower in the distribution.

4 DISCUSSION

In this thesis, the US COVID-19 data was explored, spatially modelled, and estimated by using both frequentist spatial methods and Bayesian computational methods. Moran's and Geary's statistics showed significant spatial autocorrelation in the areal data for COVID-19 infections. SAR and CAR models were used to model the spatially autocorrelated data. The SAR model formed a more accurate representation for the infections, showing an association between population and population density in the county specific and state specific analysis. The SAR model left no spatial autocorrelation for the residuals, suggesting a good fit for the data. The variability covered by the models on the national level was 0.83 for the county specific and 0.99 for the state specific SAR models according to Nagelkerke R^2. These results are most likely due to the more complex nature of the county specific analysis, in which there were more data units than in the state specific analysis. For future analyses, methods that provide mean squared error measures for models could be utilised. These methods could give different useful insight on the modelling of the incidence of COVID-19.

Approximate Bayesian computation methods were used to form a prediction for the posterior distribution of the average infections for the spread of COVID-19 in the United States. The aim was to use a small amount of data from the first 100 days of the pandemic to form these posterior distributions for the infection rate. The three ABC methods used, were ABC-Rejection, ABC-MCMC and ABC-SMC algorithms. The results show that the ABC-Rejection algorithm gave the most inaccurate results and the ABC-MCMC the most accurate. The results for the ABC-Rejection algorithm did however improve, when an noninformative proposal distribution was chosen for the infection rate estimates.

The downside to the ABC-MCMC was that the range of the posterior distribution was relatively large, compared to the other methods. The ABC-SMC worked relatively well, but took a very long time to run, which made its use much less appealing. The ABC-MCMC and the ABC-SMC formed viable posterior distributions, where all the infection rates from different time points of the 429 days period were inside the distributions and could have be plausible outcomes for the spread of COVID-19, which means that the method can accurately give a threat assessment for the spread of a pathogen. This information is valuable in the early stages of an outbreak of a potentially harmful disease to help decide required actions to enact in order to limit any harm caused by the pathogen as early as possible.

The time of COVID-19 pandemic has given unprecedented events, which were mostly unforeseeable. It has been debated whether the precautions and reaction to the onset of COVID-19 were adequate in the US, with many claiming that more could have been done to discourage the spread of the disease. The United States itself is also a vast country with distinctly different geographical regions, with various cultural circles and financial cells, which all produce diversity. Therefore, producing a precise estimate for the incidence in such an environment could be seen to be disingenuous. There may be too much variability even in a certain region of the country to produce a point estimate with a confidence interval for the spread of the disease. It may also be that when a similar situation arises in the future, there is not enough information to produce such estimates with enough reliable data. Data points may be scattered to a few cases in few hotspots or be located in a single location entirely, which creates difficulties in giving an estimate suitable for completely different areas. ABC can be utilised in these situations to give a more flexible estimate, which could share vital information for preliminary precautions.

Possible future study aspects would be to further develop the ABC methods and study, whether some other features, such as regression-based algorithms would produce more viable estimates for infection rates. These models could potentially account for more elements, such as the decline of susceptible people for the disease. It would also be valuable for these future studies to take into account the variants of COVID-19 and how differently these variants behave in the models.

5 REFERENCES

Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L. (2017) *Disease Prediction by Machine Learning Over Big Data From Healthcare Communities*, IEEE.

Cooper, I., Mondal, A. and Antonopoulos, C. (2020) *A SIR model assumption for the spread of COVID-19 in different communities*, Elsevier.

Engblom, S., Eriksson, R. and Wifgren, S. (2020), *Bayesian epidemiological modeling over high-resolution network data*, Elsevier.

Filippi, S., Barnes, C., Cornebisse, J. and Stumpf, M. (2012) *On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo*, Centre for Integrative Systems Biology and Bioinformatics, Imperial College London, London SW7.

Lawson, A. and Lee, D. (2017) *Bayesian Disease Mapping for Public Health*, ResearchGate.

Minter, A. and Retkute, R. (2019) *Approximate Bayesian Computation for infectious disease modelling*, Elsevier.

Mollalo, A., Vahedi, B. and Rivera, K. (2020) *GIS-based spatial modeling of COVID-19 incidence rate in the continental United States*, Elsevier.

Silk, D., Filippi, S. and Stumpf, M. (2013) *Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems*, Imperial College London.

Simola, U., Cisewski-Kehe, J., Gutmann, M. and Corander, J. (2021) *Adaptive Approximate Bayesian Computation Tolerance Selection.* Bayesian Anal. 16(2): 397-423

Sunnåker, M., Busetto, AG., Numminen, E., Corander, J. and Foll, M. (2013) *Ap-*

proximate Bayesian Computation. PLoS Comput Biol 9(1): e1002803.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M. (2009) *Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems*, Royal Society.

Vasishth, S. (2020) *Using approximate Bayesian computation for estimating parameters in the cue-based retrieval model of sentence processing*, Elsevier.

Yang, Y., Dai, B., Kiyvash, N. and He, N. (2018) *Predictive Approximate Bayesian Computation via Saddle Points*, Advances in Neural Information Processing Systems 31 (NeurIPS 2018).

Zheng, Y. and Aris-Brosou, S. (2017) *Approximate Bayesian Computation Algorithms for Estimating Network Model Parameters*, bioRxiv.

A A SIMULATION STUDY FOR THE ABC-REJECTION ALGORITHM

In a blog post *Tiny Data, Approximate Bayesian Computation and the Socks of Karl Broman* by Rasmus Bååth, a researcher from Lund University Cognitive Science, Sweden, the effectiveness of ABC methods are studied on a set of tiny data (*link: http://www.sumsar.net/blog/2014/10/tiny-data-and-the-socks-of-karl-broman/*). The data consists of a tweet, made by Karl Broman, a statistician and a professor from the University of Wisconsin-Madison. In his tweet, Karl Broman states *That the 1st 11 socks in the laundry are each distinct suggests there are a lot more socks.* Based on this data, Bååth presents a study question *Given the Tiny dataset of eleven unique socks, how many socks does Karl Broman have in his laundry in total?*

Bååth solves the problem by using Approximate Bayesian Computation (ABC) methods. The process needs a few parameters, which were `n_socks` = total amount of socks, `n_picked` = the number of socks going to be picked, `n_pairs` = total amount of paired socks and `n_odd` = the amount of odd socks. The vector for socks can be represented then as

```
socks <-rep(seq_len(n_pairs+n_odd),rep(c(2,1),c(n_pairs,n_odd)))
socks
## [1] 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 9 10 11
```

The simulation for `n_picked` can be set by

```
picked_socks <- sample(socks, size = min(n_picked, n_socks))
sock_counts <- table(picked_socks)
sock_counts
## picked_socks
## 1 3 4 5 7 8 9 10 11
## 1 2 2 1 1 1 1 1 1
c(unique = sum(sock_counts == 1), pairs = sum(sock_counts == 2))
## unique pairs
## 7 2
```

The `n_socks` variable needs to be positive and discrete, since there is a finite amount of socks. The chosen distribution to sample socks is the Negative Binomial distribution, given by the `rnbinom` function in R. The parameters for the model are `mu` and `size`, where `size` is the relationship between `mu` and variance `s^2` described as

```
size = -mu^2 / mu - s^2.
```

For the analysis, it is presumed that in a family of 3-4 people and a change of socks 5 times a week, it can be estimated that there is on average 15 pairs of socks in the laundry. Therefore, the prior for `n_socks` is set to follow the Negative Binomial distribution with mean `prior_mu = 30` and standard deviation `prior_sd = 15`.

```
prior_mu <- 30
prior_sd <- 15
prior_size_param <- -prior_mu^2 / (prior_mu - prior_sd^2)
n_socks <- rnbinom(1, mu = prior_mu, size = prior_size_param)
```

Instead of direct prior distribution over `n_pairs` and `n_odd`, the proportion of paired socks was modelled with `prop_pairs`, which was sampled from a `Beta` prior distribution, where most of the range of paired socks is between 0.75 and 1.0. The results from the sampling were rounded to discreet `n_pairs` and `n_odd`.

```
prop_pairs <- rbeta(1, shape1 = 15, shape2 = 2)
n_pairs <- round(floor(n_socks / 2) * prop_pairs)
n_odd <- n_socks - n_pairs * 2
```

When the prior information was set, the ABC-Rejection algorithm was set with a fixed amount of iteration set to 100,000, which gives 100,000 samples. The following program was then used

```
n_picked <- 11 # The number of socks to pick out of the laundry
sock_sim <- replicate(100000,
# Generating a sample of the parameters from the priors
prior_mu <- 30
prior_sd <- 15
prior_size <- -prior_mu^2 / (prior_mu - prior_sd^2)
n_socks <- rnbinom(1, mu = prior_mu, size = prior_size)
prop_pairs <- rbeta(1, shape1 = 15, shape2 = 2)
n_pairs <- round(floor(n_socks / 2) * prop_pairs)
n_odd <- n_socks - n_pairs * 2

# Simulating picking out n_picked socks
socks<-rep(seq_len(n_pairs+n_odd),rep(c(2,1),c(n_pairs,n_odd)))
picked_socks <- sample(socks, size = min(n_picked, n_socks))
sock_counts <- table(picked_socks)

# Returning the parameters and counts of the number of matched
# and unique socks among those that were picked out.
c(unique = sum(sock_counts == 1), pairs = sum(sock_counts == 2),
n_socks = n_socks, n_pairs = n_pairs, n_odd = n_odd, prop_pairs =
prop_pairs))

# just translating sock_sim to get one variable per column
sock_sim <- t(sock_sim)
head(sock_sim)
```

After the ABC-Rejection algorithm was run, the samples not matching the original data of 11 unique socks were discarded. This was achieved through the following code:

```
post_samples <- sock_sim[sock_sim[, "unique"] == 11 &
sock_sim[, "pairs" ] == 0 , ]
```

Of the 100,000 samples, 11,506 gave suitable results in `post_samples`. The median value of the posterior distribution was 19 pairs of socks and 6 odd socks. This then gives a total estimate of $19 \times 2 + 6 = 44$ socks. The total amount of socks from the data given by Karl Broman, was 21 pairs and 3 singletons. This then results in $21 \times 2 + 3 = 45$. The difference between the estimated total amount of socks and the observed amount was a total of 1 sock. Therefore, the ABC-Rejection algorithm gave quite accurate results. The composition of the total amount of socks is criticised, as the odd amount of socks was higher than in the observed data, but it is commented as resulting from the difference in the organizational skills between Bååth and Broman.

B MORAN'S AND GEARY'S STATISTICS FOR EACH COUNTY

State	Moran's I	p-value	Geary's C	p-value
Alabama	0.2438	0.0033	0.5741	0.0034
Arizona	-0.0852	0.5699	1.218	0.8187
Arkansas	0.415	0	0.637	0.0063
California	0.2929	0	0.6986	0.0748
Colorado	0.4046	0	0.6743	0.0083
Connecticut	-0.1493	0.5094	0.9846	0.4755
Florida	0.4011	0	0.6036	0.0106
Georgia	0.6372	0	0.3886	0
Idaho	0.2783	0.0014	0.7118	0.0836
Illinois	0.2391	0	0.5356	0.0128
Indiana	0.3041	0.0001	0.6617	0.0099
Iowa	0.1807	0.0048	0.7725	0.034
Kansas	0.1716	0.0077	0.8255	0.0795
Kentucky	0.1132	0.0148	0.8891	0.2861
Louisiana	0.1113	0.1233	1.0025	0.5073
Maine	0.5902	0.0004	0.3546	0.0041
Maryland	0.3843	0.0087	0.7128	0.0578
Massachusetts	0.7074	0.0002	0.3545	0.0034
Michigan	0.5894	0	0.4231	0
Minnesota	0.3616	0	0.5173	0.0016
Mississippi	0.1132	0.0988	0.7866	0.0441

Missouri	0.5618	0	0.5621	0.0013
Montana	0.1901	0.0326	0.8573	0.1658
Nebraska	0.2489	0	0.7111	0.0436
Nevada	-0.048	0.41	0.7712	0.1704
New Hampshire	0.1764	0.0812	0.8146	0.2385
New Jersey	0.6621	0.0001	0.3248	0.0002
New Mexico	0.1283	0.0721	1.2098	0.8421
New York	-0.0352	0.6447	1.2197	0.8002
North Carolina	0.3313	0	0.7665	0.0514
North Dakota	-0.123	0.8413	1.2734	0.9555
Ohio	0.2428	0.0021	0.6652	0.0076
Oklahoma	0.2015	0.0049	0.6219	0.0144
Oregon	0.4166	0.0006	0.4989	0.0014
Pennsylvania	0.4976	0	0.5511	0.0023
Rhode Island	-0.1952	0.3524	0.9375	0.4092
South Carolina	0.2819	0.0088	0.687	0.0174
South Dakota	0.1682	0.0068	0.8437	0.1849
Tennessee	0.1416	0.0344	0.8414	0.1375
Texas	0.1796	0.0002	0.702	0.0048
Utah	0.2988	0.0038	0.8354	0.2066
Vermont	-0.2252	0.8822	1.1288	0.7405
Virginia	0.5351	0	0.4893	0.0002
Washington	0.4433	0	0.5369	0.0073
West Virginia	0.0531	0.2693	0.8706	0.1899
Wisconsin	0.465	0	0.7343	0.0592
Wyoming	-0.1631	0.7506	1.1678	0.8004

C SAR AND CAR REGRESSION RESULTS FOR US STATES

	SAR model			CAR model		
State	population (p-value)	population density (p-value)	Nagelkerke R^2	population (p-value)	population density (p-value)	Nagelkerke R^2
Alabama	0.1185 $(p < 0.001)$	−14.789 $(p = 0.2579)$	0.9873	0.1114 $(p < 0.001)$	3.4323 $(p = 0.7739)$	0.9864
Arizona	0.0980 $(p < 0.001)$	760.4765 $(p = 0.0029)$	0.9984	0.1210 $(p < 0.001)$	162.0454 $(p = 0.5634)$	0.9983
Arkansas	0.1326 $(p < 0.001)$	−43.5420 $(p = 0.0386)$	0.9798	0.1250 $(p < 0.001)$	−27.6071 $(p = 0.1866)$	0.9789
California	0.1216 $(p < 0.001)$	−10.1543 $(p = 0.0083)$	0.9707	0.1263 $(p < 0.001)$	−8.1755 $(p = 0.0241)$	0.9724
Colorado	0.0961 $(p < 0.001)$	2.3841 $(p = 0.1326)$	0.9817	0.0963 $(p < 0.001)$	2.8519 $(p = 0.0690)$	0.9809
Connecticut	0.08110 $(p < 0.001)$	45.1900 $(p = 0.0432)$	0.9914	0.1002 $(p < 0.001)$	1.6210 $(p = 0.9638)$	0.9858
Florida	0.1617 $(p < 0.001)$	−82.7240 $(p < 0.001)$	0.9328	0.1613 $(p < 0.001)$	−81.3736 $(p < 0.001)$	0.9328
Georgia	0.1087 $(p < 0.001)$	−4.7360 $(p = 0.0293)$	0.9840	0.1045 $(p < 0.001)$	−1.1501 $(p = 0.6158)$	0.9833
Idaho	0.1076 $(p < 0.001)$	30.7206 $(p < 0.001)$	0.9950	0.1034 $(p < 0.001)$	39.4932 $(p < 0.001)$	0.9943
Illinois	0.1094 $(p < 0.001)$	−9.7041 $(p < 0.001)$	0.9995	0.1087 $(p < 0.001)$	−7.5968 $(p < 0.001)$	0.9995
Indiana	0.1127 $(p < 0.001)$	−2.2099 $(p = 0.6332)$	0.9917	0.1055 $(p < 0.001)$	5.8726 $(p = 0.2196)$	0.9914
Iowa	0.1101 $(p < 0.001)$	14.2923 $(p = 0.2642)$	0.9871	0.1019 $(p < 0.001)$	26.3087 $(p = 0.0426)$	0.9869

	SAR model			CAR model		
State	population (p-value)	population density (p-value)	Nagelkerke R^2	population (p-value)	population density (p-value)	Nagelkerke R^2
Kansas	0.1061 ($p < 0.001$)	−0.5136 ($p = 0.8093$)	0.9911	0.1062 ($p < 0.001$)	−0.7000 ($p = 0.7415$)	0.9917
Kentucky	0.1101 ($p < 0.001$)	−1.811 ($p = 0.2877$)	0.9961	0.1077 ($p < 0.001$)	0.1323 ($p = 0.9371$)	0.9960
Louisiana	0.1093 ($p < 0.001$)	−10.7699 ($p < 0.001$)	0.9857	0.1106 ($p < 0.001$)	−11.9894 ($p < 0.001$)	0.9855
Maine	0.04664 ($p < 0.001$)	16.1334 ($p = 0.2021$)	0.9658	0.0401 ($p < 0.001$)	37.6702 ($p = 0.0041$)	0.9434
Maryland	0.0764 ($p < 0.001$)	1.6882 ($p = 0.3332$)	0.9716	0.0763 ($p < 0.001$)	1.7962 ($p = 0.3040$)	0.9716
Massachusetts	0.0901 ($p < 0.001$)	3.5942 ($p = 0.0969$)	0.9483	0.0909 ($p < 0.001$)	3.7241 ($p = 0.0756$)	0.9469
Michigan	0.0823 ($p = 0.0024$)	24.2249 ($p = 9892$)	0.9892	0.0805 ($p < 0.001$)	27.3112 ($p < 0.001$)	0.9892
Minnesota	0.1058 ($p < 0.001$)	−4.6266 ($p = 0.0011$)	0.9922	0.1055 ($p < 0.001$)	−3.1095 ($p = 0.032$)	0.9921
Mississippi	0.0621 ($p < 0.001$)	59.8182 ($p < 0.001$)	0.9720	0.0550 ($p < 0.001$)	71.4591 ($p < 0.001$)	0.9695
Missouri	0.0870 ($p < 0.001$)	−8.7678 ($p < 0.001$)	0.9509	0.0934 ($p < 0.001$)	−8.7094 ($p < 0.001$)	0.9361
Montana	0.1149 ($p < 0.001$)	−36.4934 ($p = 0.4172$)	0.9683	0.1127 ($p < 0.001$)	−21.1376 ($p = 0.6431$)	0.9682
Nebraska	0.0912 ($p < 0.001$)	29.8833 ($p < 0.001$)	0.9979	0.0919 ($p < 0.001$)	29.2089 ($p < 0.001$)	0.9979
Nevada	0.1160 ($p < 0.001$)	1.5309 ($p = 0.9109$)	0.9109	0.1161 ($p < 0.001$)	2.0730 ($p = 0.8802$)	0.9989
New Hampshire	0.08289 ($p < 0.001$)	−5.1982 ($p = 0.8110$)	0.9691	0.0826 ($p < 0.001$)	−4.4000 ($p = 0.8401$)	0.9691

State	SAR model			CAR model		
	population (p-value)	population density (p-value)	Nagelkerke R^2	population (p-value)	population density (p-value)	Nagelkerke R^2
New Jersey	0.1097 ($p < 0.001$)	2.8380 ($p = 0.0136$)	0.9718	0.1095 ($p < 0.001$)	2.9467 ($p < 0.001$)	0.9715
New Mexico	0.1246 ($p < 0.001$)	−121.3584 ($p < 0.001$)	0.9787	0.1095 ($p < 0.001$)	−67.4166 ($p < 0.001$)	0.9775
New York	−0.0080 ($p = 0.1923$)	19.5392 ($p < 0.001$)	0.6002	0.0022 ($p = 0.9074$)	18.2151 ($p < 0.001$)	0.6144
North Carolina	0.0870 ($p < 0.001$)	12.9279 ($p = 0.0014$)	0.9808	0.0812 ($p < 0.001$)	21.7766 ($p < 0.001$)	0.9795
North Dakota	0.1274 ($p < 0.001$)	91.4759 ($p = 0.0759$)	0.9927	0.1304 ($p < 0.001$)	76.6258 ($p = 0.1315$)	0.9927
Ohio	0.0933 ($p < 0.001$)	3.5997 ($p = 0.2628$)	0.9964	0.0883 ($p < 0.001$)	9.8432 ($p = 0.0030$)	0.9964
Oklahoma	0.1079 ($p < 0.001$)	9.6769 ($p = 0.0934$)	0.9981	0.1068 ($p < 0.001$)	11.7862 ($p = 0.0041$)	0.9981
Oregon	0.0458 ($p < 0.001$)	3.3143 ($p = 0.4427$)	0.9594	0.0464 ($p < 0.001$)	2.5270 ($p = 0.5605$)	0.9591
Pennsylvania	0.0870 ($p < 0.001$)	3.3063 ($p < 0.001$)	0.9890	0.0859 ($p < 0.001$)	4.5483 ($p < 0.001$)	0.9887
Rhode Island	0.1626 ($p < 0.001$)	9.6593 ($p < 0.001$)	0.9989	0.1553 ($p < 0.001$)	−6.7936 ($p = 0.2878$)	0.9795
South Carolina	0.1203 ($p < 0.001$)	8.1698 ($p = 0.7031$)	0.9750	0.0978 ($p < 0.001$)	52.5641 ($p = 0.0430$)	0.9674
South Dakota	0.1252 ($p < 0.001$)	82.0982 ($p < 0.001$)	0.9897	0.1305 ($p < 0.001$)	73.9278 ($p < 0.001$)	0.9900
Tennessee	0.0801 ($p < 0.001$)	62.3365 ($p < 0.001$)	0.9832	0.0711 ($p < 0.001$)	76.0967 ($p < 0.001$)	0.9829
Texas	0.0699 ($p < 0.001$)	97.9762 ($p < 0.001$)	0.9756	0.0696 ($p < 0.001$)	99.9058 ($p < 0.001$)	0.9757

	SAR model			CAR model		
State	population (p-value)	population density (p-value)	Nagelkerke R^2	population (p-value)	population density (p-value)	Nagelkerke R^2
Utah	0.1608 ($p < 0.001$)	−45.7393 ($p < 0.001$)	0.9955	0.1603 ($p < 0.001$)	−45.1433 ($p < 0.001$)	0.9955
Vermont	0.0258 ($p < 0.001$)	22.5539 ($p = 0.0085$)	0.9515	0.0246 ($p < 0.001$)	24.0358 ($p = 0.0034$)	0.9522
Virginia	0.0768 ($p < 0.001$)	−1.4238 ($p = 0.0108$)	0.9578	0.0749 ($p < 0.001$)	−0.9801 ($p = 0.0796$)	0.9573
Washington	0.0568 ($p < 0.001$)	−15.8311 ($p = 0.3074$)	0.9407	0.0571 ($p < 0.001$)	−17.1757 ($p = 0.2691$)	0.9408
West Virginia	0.0822 ($p < 0.001$)	6.1380 ($p = 0.0010$)	0.9783	0.0829 ($p < 0.001$)	5.5067 ($p = 0.0039$)	0.9783
Wisconsin	0.0999 ($p < 0.001$)	14.6756 ($p < 0.001$)	0.9879	0.0996 ($p < 0.001$)	14.4249 ($p < 0.001$)	0.9880
Wyoming	0.1139 ($p < 0.001$)	−114.9679 ($p = 0.0322$)	0.9708	0.1141 ($p < 0.001$)	−116.8884 ($p = 0.0617$)	0.9687