Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

# Assessment Submission Form

| | |
|---|---|
| **Student Name** | **POOJA GANESH TEJE** |
| **Student ID Number** | [redacted] |
| **Course Title** | **Master of Science in Computer Science (Data Science)** |
| **Module Title** | **DISSERTATION** |
| **Lecturer(s)** | **Dr. Bahman Honari** |
| **Assessment Title** | **Analysis of Socio-Economic Impact due to COVID-19 on US using Geographically Weighted Regression** |
| **Date Submitted** | **07/09/2020** |
| **Word Count** | **14034** |

**I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at:** http://www.tcd.ie/calendar

**I have also completed the Online Tutorial on avoiding plagiarism 'Ready, Steady, Write', located at** http://tcd-ie.libguides.com/plagiarism/ready-steady-write

**I declare that the assignment being submitted represents my own work and has not been taken from the work of others save where appropriately referenced in the body of the assignment.**

Signed ……POOJA GANESH TEJE…… Date ……September 07, 2020……

# Analysis of Socio-Economic Impact due to COVID-19 on US using Geographically Weighted Regression

Pooja Ganesh Teje, MSc. Data Science

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Dr. Bahman Honari

September 7, 2020

# Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work. A section of this study has been researched upon as a part of Security and Privacy (CS7NS5) module.

Pooja Ganesh Teje

September 7, 2020

# Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Pooja Ganesh Teje

September 7, 2020

# Acknowledgments

To start with, I would like to express my gratitude to my instructor, Dr. Bahman Honari, for his valuable counselling and practical suggestions from the very start of my research process. I am also thankful to the Computer Science and Statistics Department of Trinity College Dublin for providing me with remote access to all the research resources to work on my thesis. I would also like to acknowledge my second reader Dr. Arthur White for his time and suggestions which additionally helped me in my research. I must express my profound gratefulness to my parents Ganesh Teje and Gayatri Teje for their unconditional support and motivation throughout my year of study. In the end, I would like to express my gratitude to all my friends and classmates who played a crucial role in making my time here at Trinity College Dublin a memorial experience.

Pooja Ganesh Teje

University of Dublin, Trinity College
September 7, 2020

# Analysis of Socio-Economic Impact due to COVID-19 on US using Geographically Weighted Regression

Pooja Ganesh Teje, Master of Science in Computer Science
University of Dublin, Trinity College, 2020

Supervisor: Dr. Bahman Honari

After the 2nd world war, The COVID-19 pandemic is the biggest and most critical global health calamity faced by the world. With an exponential rate of spread, COVID-19 outbreak is severely impacting the global economy as well as society as a whole. Analysis of socio-economic impact will help us to understand how each factor is getting affected differently by COVID-19 pandemic. However, as COVID-19 impacts vary from one location to another, it's difficult to analyse this impact using traditional machine learning and data analytics methods, as there is a large spatial variation in the data. The aim of this research is to apply the geographically weighted regression (GWR) models for analysis. GWR differs from traditional approaches of data analysis by considering the spatial variation of COVID-19 cases over a geographical area and by plotting the regression coefficients to and underlying patterns in data that are otherwise hidden when implementing a global regression model. GWR coefficient allows finding local estimates at each point which better explains the spatial variations and fits the data.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Pandemics are more of a catastrophic socio-economic and political crisis than being some public health concern. COVID-19, apart from being one of the greatest threats of the century to public health globally, is indicating the existence of deficiency and inequity of social advancement. As the name implies COVID-19, ‘CO’ is for ‘corona,’ ‘VI’ for ‘virus,’ and ‘D' for disease, and 19 shows the year of its occurrence. With 80 to 120 nm of diameter range, coronavirus is a single stranded RNA virus. COVID-19 pandemic was first discovered in Wuhan, Hubei province, China, by the end of 2019 and its source infection is from a wholesale market of seafood [1]. Since then, the disease rapidly circled the globe and has eventually affected almost every country. It has been categorized as a pandemic by the World Health Organization [2]. International Committee on Taxonomy of Viruses (ICTV) named the virus as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [3].

Shortly after, Iran and a few European countries, most notably Italy, experienced a significant increase in the number of cases and deaths. In the United States, the first COVID-19 case was discovered and confirmed on January 19, 2020, in Washington State [4]. Soon after, multiple states started to experience a sudden increase in the number of COVID-19 cases; New York State became one of the most affected areas of the disease spread [5]. On March 17, 2020, all 3000+

counties of the United States had confirmed cases of COVID-19 [6]. On March 26, 2020, the United States became the leading country in the number of COVID-19 cases worldwide, even taking over Italy which was in the lead of COVID-19 cases till time [5]. As of August, 2020, around 180,000 deaths and 6M cases have been confirmed in the United States [7].

To analyse the spatial distribution of such infectious diseases, geographic information is very essential [8] [9], such analysis can also help in facing the pandemic and curing the infection [10]. geographic information is an important feature in analysing and visualizing the spread of COVID-19. This study includes analysis using regressive spatial models to determine how well these models can explain the variations of COVID-19 in the continental United States based on several socioeconomic factors as explanatory variables. Using geographic information of COVID-19 this research aims to analyse socio-economic impact of COVID-19 using various Geographically Weighted Regression (GWR) models. Two different datasets have been used to implement 5 models all together for this analysis. The coefficients of regression from GWR models are plotted on a map to visually analyse underlying spatial variations among the indicators of the data that directly or indirectly impacted due to COVID-19. The research also aims to discuss the difference in results of geographically weighted models as opposed to traditional models of linear regression.

## 1.1 Motivation

Data science and modelling technology make it possible to analyse large amounts of data available across the world and find smart and informative patterns out of it. Using the knowledge of such tools and models can bring benefit to business as well as society. One area of application would be COVID-19 analysis.

From the beginning of coronavirus spread we have daily data of COVID-19 cases, this data includes a variety of features such as total number of cases, deaths, new cases each day, etc. Since the virus spread all over the globe and the unique properties of this virus of evolving itself based on the location made analysis of impact different from one area to another [36]. Including socio-economic features into the analysis helps us understand how covid-19 impacted on these features.

This research aims to analyse COVID-19 data to understand how GWR can be used to analyse impact of COVID-19 using socio-economic features. Since unemployment is one of the major impacts of COVID-19 on the economy, this study also includes analysis of how unemployment is impacted due to the spread of COVID-19. This research can also be useful to better understand how COVID-19 data and socio-economic features can be recorded more comprehensively for further use and analysis.

## 1.2 Research question

To analyse impact of COVID-19 on Socio-Economic features of the United States using Geographically Weighted Regression models

### 1.2.1 Research Aims

The aims of this research are as follows:

1. To analyse and map the COVID-19 data from counties of the United States

2. To implement Geographically Weighted Regression models that provides local relationship between COVID-19 data and socio-economic factors
3. To analyse local relationship between COVID-19 data and unemployment by implementing the Geographically Weighted Regression model

4. To analyse and map the coefficients of regression from the result of GWR models

5. To analyse how traditional regression analysis method differs from GWR analysis

### 1.2.2 Research Challenges

Implementing this research does include some challenges along the way. Challenges faced while implementing all of the GWR models are listed below:

1. Data Selection and collection was a major roadblock for the implementation of this research, as GWR requires unique nature of data i.e. all values of data should be in numerical form and recorded for each location or point in space (latitude and longitude)

2. Feature selection and data cleaning was one most time consuming process during the research, as there are more than 150 socioeconomic features present in the dataset but it's crucial to select only the most important features.

3. Data from various sources like COVID-19, socioeconomic and unemployment data were collected and collated for each location of the United States.

4. Understanding the results of a non-traditional regression model such as GWR and interpreting those results for analysis.

5. Tuning the GWR bandwidth to improve model's performance and mapping the results in a visually understandable format.

## 1.3 Thesis structure

The thesis is organized into the following sections. Section 2 represents the related work and literature review. Section 3 talks about basics of the GWR models used in this research. Section 4 talks about technical approach and implementation of research using GWR. Section 5 discusses the results of the experimental analysis of all models. Section 6 talks about the security and privacy considerations of this research. Section 7 talks about the limitations of this work and lastly Section 8 discusses the conclusion and future scope of the research.

# Chapter 2

# Literature Review

There are various studies available on the study of COVID-19 and study of the GWR models. Some of those studies are discussed here. In the research by Abolfazl Mollalo [11], various spatial models are studied, GWR being one of them focusing on COVID-19 data. In recent works across the globe has shown environmental conditions [12], air pollution [13] and smoking [14] may play a role in the spread of COVID-19. For example, Wu et al. [13] states that exposure to air pollution for long-term might affect COVID-19 treatment of such patients. Study from Taghizadeh-Hesary and Akbari [14] suggests that smoking may have a negative effect on COVID-19 treatment due to lack of a healthy immune system. Study from China by Wang et al. [12] says that some environmental conditions like humidity and temperature might influence the spread of COVID-19. Abolfazl's work show analysis of COVID-19 and its impact using global models, i.e. Ordinary least squares (OLS), Spatial lag model (SLM) and Spatial error model (SEM) as well as local models, i.e. Geographically weighted regression (GWR), Multiscale GWR (MGWR). On the assumption that multiscale approach will help in explaining the spatial variability of COVID-19 spread in a better way, MGWR model was compared with all other global and local models. Results of this research confirms the assumption with MGWR with the proof of its goodness-of-fit to the COVID-19 data. Since there is still significant lack in overall study of geographic modelling of COVID-19 data, this research acts as a basis for every future geographic modelling of COVID-19.

Another study of GWR models by Udita Retharekar [15], This work is a great resource for GWR introduction and basic understanding of GWR. This study analyses and predicts crime patterns in the United States with the help of various Geographically Weighted Models. This study also compares various global and local regression models with GWR being one of them. The work at NUI Maynooth by Chris Brunsdon et.al [16] also help in depth understanding of GWR models. Research extending the GWR model from a basic spatial regression model to a classification model by Chris Brunsdon et al. [17] is extremely helpful in understanding potential analysis using GWR models. Another helpful study is by Nezami and Khoramshahi [18] uses GWR for the analysis of the smuggling pattern in Iran.

After studies of GWR, some studies of COVID-19 are discussed ahead. Study from Indranil Chakraborty and Prasenjit Maity [19] discuss COVID-19 outbreak and its effect on society and global environment. Study also discusses important prevention from these effects as well. The outbreak of the new infection, COVID-19 was started from the Hunan seafood market of China in December 2019, and within a short period of time it has turned into a global health emergency [20]. The study focuses on some of the important aspects of the COVID-19, first, transmission dynamics of SARS-CoV-2 virus in people. Another is the impact of COVID-19 on the economy due to its prevention measures like social distancing and self-isolation. The study mainly focuses on the impact of COVID-19 on the global environment and its effect on greenhouse gases (CO2, CH4, N2O etc). Restricting mass gathering, Medicine, Forestation, controlling human population growth and Global ban on wildlife trade are the global strategy suggested in this study for prevention and control of COVID-19.

Lastly study by Nicola M et al. [21] is discussed for this literature review. This study also reviews the socio-economic impact of COVID-19 pandemic but on a global scale. Study summarises COVID-19 impact on aspects such as the world economy, industries working on raw materials, production and manufacturing companies and lastly, service-oriented organisations. This study is a great resource for the understanding the impact of COVID-19 on every possible aspect of life.

According to the study, the global drop in demand of materials from hotels and restaurants has resulted in 20% drop in prices of agricultural commodities [22]. On March 23rd, prices of Crude oil decreased by 24% from $34/barrel [23] to $25.70 [24]. Even though deceleration in total number of COVID-19 deaths has managed to keep some stabilisation of oil prices, there is much uncertainty still present in oil prices. British Plastics Federation (BPF) conducted a survey to explore the impact of COVID-19 on manufacturing businesses in the United Kingdom (UK). Over 80% of businesses expect a drop in total turnover for the next 2 quarters and 98% businesses show concern regarding the negative impact caused by the pandemic on business operations [25]. Reduction in the number of staff is one of the main concerns for businesses due to prevention policies like self-isolation and social distancing. In education, COVID-19 has affected completely from pre-school to tertiary education. COVID-19 has globally impacted communities, businesses and organisations, indirectly affecting the financial markets and the global economy.

Healthcare is most challenged due to COVID-19 pandemic as the risk to healthcare workers is that the worker is most vulnerable and exposed directly to the virus. As most healthcare workers are not able to work remotely, strategies to defend the frontline healthcare staff is important [26]. The tourism sector was hit hardest due to the outbreak of COVID-19, affecting both demand and supply of travel. The World Travel and Tourism Council has predicted 50 million jobs at risk in the global travel & tourism sector due to COVID-19 [27]. This study also discussed impact of COVID-19 on Sports, Information Technology (IT) industry and food industry.

This research takes all the above discussed studies into consideration while designing the GWR model of spatial analysis.

# Chapter 3

# GWR Model

This research study and implement spatial analysis of COVID-19 using geographically weighted regression models. In total 5 GWR models are studied and implemented throughout this research. The basics of a GWR model and its statistical parameters are discussed in this section. The book by Chris et al. [28] has provided a very simple and clear understanding of GWR the basics.

## 3.1 Introduction

This research study and analyses local relationship between COVID-19 and socio-economic factors thus analysing the impact of COVID-19 on these socio-economic factors of the United states. For this research, all together five geographically weighted regression models were implemented. Comparing linear regression model and GWR model will help in understanding the model more easily. This section discusses the basics of a GWR model along with its statistical considerations.

## 3.2 Regression analysis

A traditional linear regression model with n number of explanatory variables and a dependent variable can be represented as

$$y = \beta_0 + \beta_1 + \epsilon$$

In above expression $\beta_0$ and $\beta_1$ are the parameters of the model and represents error. In linear regression models the parameters are considered as non-varying or constant over the space, i.e. parameters are assumed to be spread over space in an independent manner. And for each location in space, it is assumed that the relationship between the dependent variable y and the independent variable x is constant and $\beta_0$ & $\beta_1$ are the same. The linear model also considers that the variables are distributed normally i.e. variables are centred around zero as a mean and data is stationary in space. Nonetheless, when there is a large spatial variation in data or if data varies significantly over space, using a linear regression model will provide results as a global estimator and this result is not capable of explaining the variation in space. The global estimate from linear regression models may not uncover the underlying relationship between explanatory and dependent variables due to geographical variation of data.

## 3.3 Local Estimates and Global Estimates

Unlike the traditional regression analysis model, GWR heavily focuses on local spatial statistics than global statistics. For example, average rainfall across a country is global statistics whereas local statistics means actual rainfall recorded at each location coordinates of the county. Here we can see that, average of data generally hides the geographical patterns of data. There might be other hidden environmental factors at each location which might affect the actual amount of rainfall. For instance, Nevada has a very dry climate and has almost no greenery in the area, hence receiving very less rainfall. Similar to climate, urbanisation and population can lead too little to no rainfall as well. This behaviour is considered as locally varying property. Regression analysis model with a global estimator will not be useful for such scenarios of data. Hence, GWR will be used for this analysis.

Values such as average, total, mean etc. are for global statistics and they are not mappable, on the other hand local statistics are mappable and have values like food environment index which vary from one location point to another. A regression analysis model for local statistics uses datasets which have geographical information of each location such as XY coordinates or latitude and longitude etc.

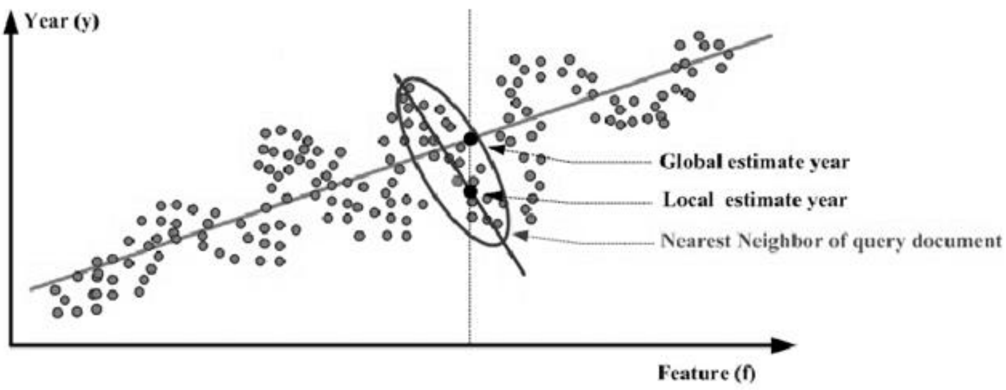


Fig 3.1 Local estimate and Global estimate [29]

## 3.4 Non-stationary nature of data

Generally, the nature of numerical data is stationary, which means irrespective of place of measurement all values hold true. For example, results from mathematical formulas or physics equations etc. Are not affected by external factors and they are the same even if they are measured at different locations. But, when we consider social data or socio-economic data, they are non-stationary almost every time. Population density, rainfall, unemployment, etc are some of the examples of such non-stationary data and this data varies from one place to another as they get measured. Fitting this type of data using a global model will be very unsatisfactory, as it will assume measurement of the data is the same across the locations and data is independent from any geographical factors.

However, non-stationarity may occur in data due to some miscalculations or mis-specification within and it doesn't mean that there is some definite spatial variation. Hence, pre-examination of such data is required in such scenarios, it can include visualisation of data by plotting it over graph or mathematically plotting the residual of linear regression will also help in spotting any spatial patterns.

## 3.5 GWR model

Extending a linear regression model mentioned in 3.2, GWR takes the form like below,

$$y = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)\, x_{ik} + \epsilon$$

Here, (ui,vi) represents a point in space, y is the estimated result and k is a continuous function which is summed over the space.

Hypothesis of this model is stated as, “the influence of points closer to each other in space on an explanatory variable will be similar. Therefore, the closer the points, the similar the weights given to them to indicate their influence on the target variable [30]”. However, the above equation of GWE has a set of unknown variables which need to be estimated. To estimate values of these variables’ weights, need to be assigned to each record and closer values can be assumed to be of almost equal magnitude and lesser as we move farther away. Relatively large datasets can help in avoiding any kind of biases while making such assumptions.

Hence, we can rewrite above equation with newly assigned weights as below,

$$\beta(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y$$

Here, W represents a diagonal matrix containing weights for each parameter measured at ( ui, vi ) location, i is a distance where the non-diagonal elements are zero.

$$\begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}$$

## 3.6 GWR Bandwidth

As stated in the above equation, k is a continuous function which is summed over the space, it needs to be converted into its discrete for further use. GWR uses a moving window technique which states that a region is described around every single regression point and based on this window region, weights for neighbouring points are calculated. GWR assign weights using kernel and weights are indirectly proportional to distance of points from the regression point.

Using the empirical data bandwidth assumptions can be made. For example, by considering the total area, one can set the bandwidth of analysis as 4km or so. But, GWR also provides a mathematically based method for bandwidth estimation from the data using various kernels. Gaussian kernel is the default kernel available for this calculation. In this kernel, as a continuous function with respect to the distance and tends to zero, the weight of points reduces. It forms a standard bell-shaped curve, represented in Figure 3.2 below.

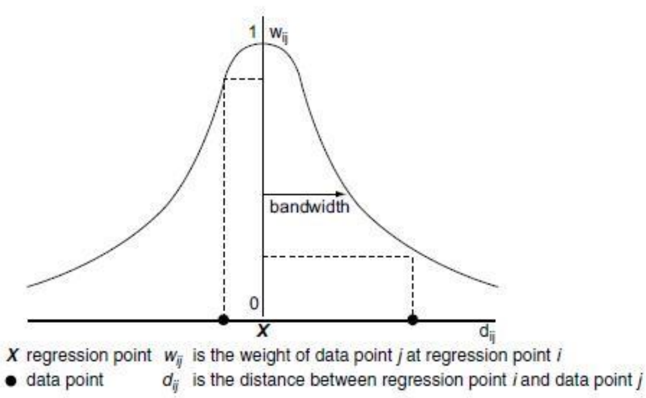


Fig 3.2: Bandwidth of the GWR model (Gaussian) [28]

Weights completely depend on distance in this spatial kernel. However, the curve of bandwidth will under-smooth the Gaussian distribution surface when density of data is sparse. Large bandwidth may lead to larger error in regression points, whereas, small bandwidth may lead to misclassification of the data. One can use an adaptive kernel to avoid such issues due to incorrect bandwidth.

In R, using the gwr.sel function from the spgwr package is very useful to select appropriate bandwidth for the regression model. Based on density of data, an adaptive kernel selects bandwidth for the model. Height of the kernel function may vary accordingly, resulting in the curve to be smooth in some parts and dense in other parts as shown in the graph below.

Choosing the type or shape of the kernel is another important task in GWR regression analysis. Gaussian kernel may not be suitable for all the datasets, as Gaussian kernel tends to take more computational time as the weights of each point gradually drops near zero. One can improve computational time using a bi-square kernel, after a fixed value this kernel clips the weights to zero. Both, Gaussian and bi-square kernel can be fixed or adaptive according to the requirements. Apart from these two kernels, there are other kernel types available as well, namely Uniform kernel and Exponential kernel, these kernels are not used as commonly as Gaussian and bi-square kernel. Below figure represents plots of these kernels.

$$Uniform\ kernel\ w(d_{ij}) = 1$$

$$Gaussian\ kernel\ w(d_{ij}) = exp(-1/2(d_{ij}/h)^2)$$

$$Exponential\ kernel\ w(d_{ij}) = exp\ \left(-1/2\ (|d_{ij}|\ /h)\right)$$

$$Bi - Square\ w(d_{ij}) = (1 - (d_{ij}\ /\ h)^2)^2 if \left|d_{ij}\right| < h, 0\ otherwise$$

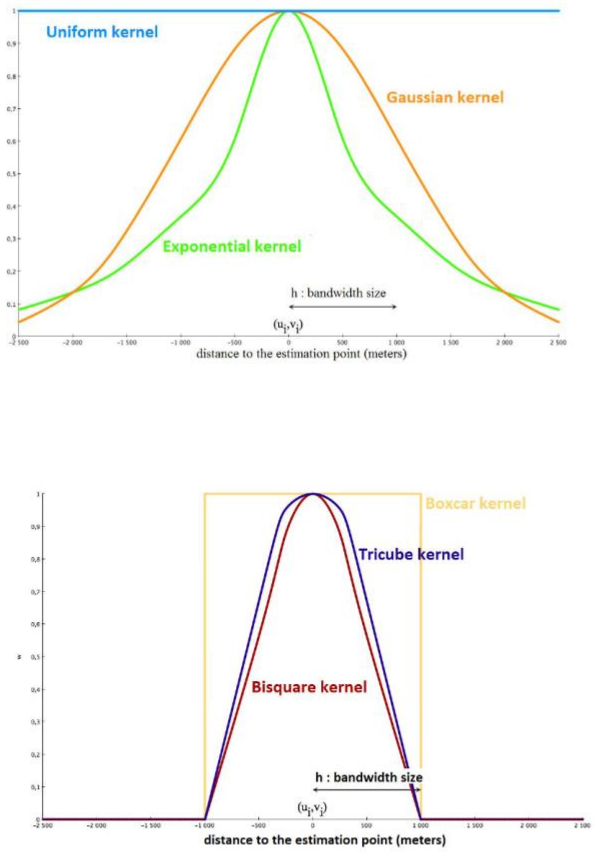


Fig 3.3: Kernel types [30]

Here, $|d_{ij}|$ is the distance.

A uniform kernel involves, calculating an ordinary least square regression at each point of data. One can also calculate the cross-validation score during bandwidth selection in R. Calculating the Cross-validation (CV) score also gives insight information regarding the predictive power of the regression model. Below equation can be used for calculation of the CV score

$$CV = \sum_{i=0}^{n} [y_i - \hat{y}_{i\neq 1}(h)]^2$$

Here $y_{i\neq 1}$ (h) is the resultant value of y at a point i. For these results, all points except $y_i$ are considered for the calculation. Value of $y_i$ - $\hat{y}_{i\neq 1}(h)$ will be 0 for h=0, hence all points except $y_i$ are taken into consideration here. Hence, value of h which reduces the cross-validation score if more effective bandwidth to predict the response [19].

## 3.7 GWR coefficients

In the GWR model one needs to provide input which includes Dependent variable one wants to model/explain/predict and all the Explanatory variables in the model. After implementing the model, use Spatial Points Data Frame (SDF) function to get details of the fitted model. Results from the SDF function includes: fit.points, GWR coefficient estimates, y value, predicted values, coefficient standard errors and t-values in its "data" slot [31]. From this result GWR coefficient estimates are very useful in understanding the spatial variation of data by simply mapping them.

Modelling the dependent variable using GWR models means mapping relations and understanding the pattern of values that play a role in the outcome of the dependent variable. One can also examine the stationary relationship between the dependent and explanatory variables and understand spatial variation in variables across the space [32].

Observing the coefficient distribution as a surface depicts where and how much variation is present within the data. Little regional variation means that variables are varying region wide. Strong regional variation means variables are varying from local location like area to area. When variables are not globally significant but have local significance, regionally these variables have positive relation in some regions and negative relation in another region [32].

Mapping GWR coefficients helps in understanding local relationships between predictor and dependent variables. For example, if the dependent variable is auto theft and the explanatory variable is parking lots in a GWR model. The resultant GWR coefficient contains positive coefficient values as well as negative coefficient values. The values of coefficient estimates are mainly the slopes of regression lines that fit the data at the location. So, if the coefficient value is positive then there are a higher number of auto thefts in places with a greater number of parking lots, and if coefficient value is negative then there are a smaller number of thefts where there are higher numbers of parking lots [33].

Since GWR works on a moving window called bandwidth and it also includes tuning of this window to get more accurate results.

# Chapter 4

# Approach and Implementation using GWR

## 4.1 Introduction

As stated in earlier sections, in this research all together five models are designed and implemented to analyse COVID-19 cases and its impact on socio-economic factors. This regression analysis task uses COVID-19 data of the United States along with socio-economic factors of the United States.

Implementation of this research involves following tasks/steps:

1. Data Selection: This step involves selecting relevant datasets for the research
2. Data Pre-processing: This step includes pre-processing steps such as data visualization, analysis of exploratory data, handling missing data etc.
3. Feature selection: This step includes feature selection methods such as correlation matrix and bandwidth calculation.
4. Model Implementation: This step deals with the implementation of all 5 GWR models
5. Result Evaluation: This Step involves recording of evaluation metrics such as AIC, $R^2$ scores etc.

## 4.2 GWR Model design

Below are all five GWR models designed in this research

1. COVID-19 cases V/S Covid-19 deaths
2. Unemployment V/S Socio-economic factors
3. Unemployment V/S COVID-19 cases
4. COVID-19 cases V/S Socio-economic factors
5. COVID-19 deaths V/S Socio-economic factors

Among all the models, model four and five are the most important GWR models.

## 4.3 Data Selection

This research uses two different datasets for the analysis of the socio-economic impact due to COVID-19. One dataset is from USAFACTS organisation and the other dataset is obtained from Kaggle. Each dataset focuses on a different category of information regarding the United States. Dataset from USAFACTS organisation talks about COVID-19 information [7]. This dataset is great for tracking daily cases of COVID-19 in the US. This data gets updated daily with the total number of cases and deaths of COVID-19. This county-level tracking data makes it very easy to follow COVID-19 cases on a granular level.

The next dataset for this research is from Kaggle. This data gives information of about 3,142 counties of the US, which includes social, economic, health and weather condition data [34]. Combining with COVID-19 data one can analyse if and how COVID-19 is affecting the health and social life of the US citizens. Transfer of COVID-19 might be affected by temperature and humidity of the United States. Also, in some of the warmer regions of the US, there are visibly different socio-economic and health impacts.

Hence, it's important to analyse factors like Labour force, Firearm Fatalities, Mental health providers, Social Association rate, Food environment index etc, since such factors likely play a role in COVID-19 fatalities and impacts. This dataset includes more than 200 columns of data representing different factors such as social, economic, health and weather conditions. For this research social and economic factor are taken into consideration. But even for social and economic factors there are more than 100 columns present in the dataset. Hence, correlation matrix has been used for feature selection and filtering of data. Some of the important socio-economic columns in the final dataset are mentioned in the table below.

| | |
|---|---|
| **Covid-19 features** | Total cases |
| | Total deaths |
| **Economic features** | Per capita income |
| | Income ratio |
| | Eightieth percentile income |
| | Twentieth percentile income |
| | Number of unemployed (CDC) |
| **Socio-Economic features** | Labour force |
| | Firearm Fatalities |
| | Mental health providers |
| | Social Association rate |
| | Total Population |
| | Food environment index |
| | Driving deaths |
| | Uninsured people |
| | Primary care physicians |
| | Dentists |
| | Household people |
| | Average of annual violent crimes |
| | Injury deaths |
| | Inadequate facilities |
| | Workers who drive alone |
| | Life expectancy |
| | Food insecure places |
| | People with limited access to healthy food |
| | Drug overdose deaths |
| | Motor vehicle deaths |
| | People enrolled in free/reduced lunch |
| | Segregation index |
| | Traffic volume per meter |
| | Homeowners |
| | Households with cost burden |

Table 4.1: The United States COVID-19 impact final dataset

Among all this data, COVID-19 features such as total cases and total deaths are used for initial analysis as well as for GWR models. Whereas, some economic features are also used for initial analysis to determine whether COVID-19 has any sort of impact on the economy of the United States.

## 4.4 Initial Analysis

As mentioned in the previous section, the research started with initial analysis of COVID-19 data for the United States. For the initial analysis COVID-19 data and some columns from economic data have been used. First task was to visualize the spread of COVID-19 over the United States on county level and the total population of each county.

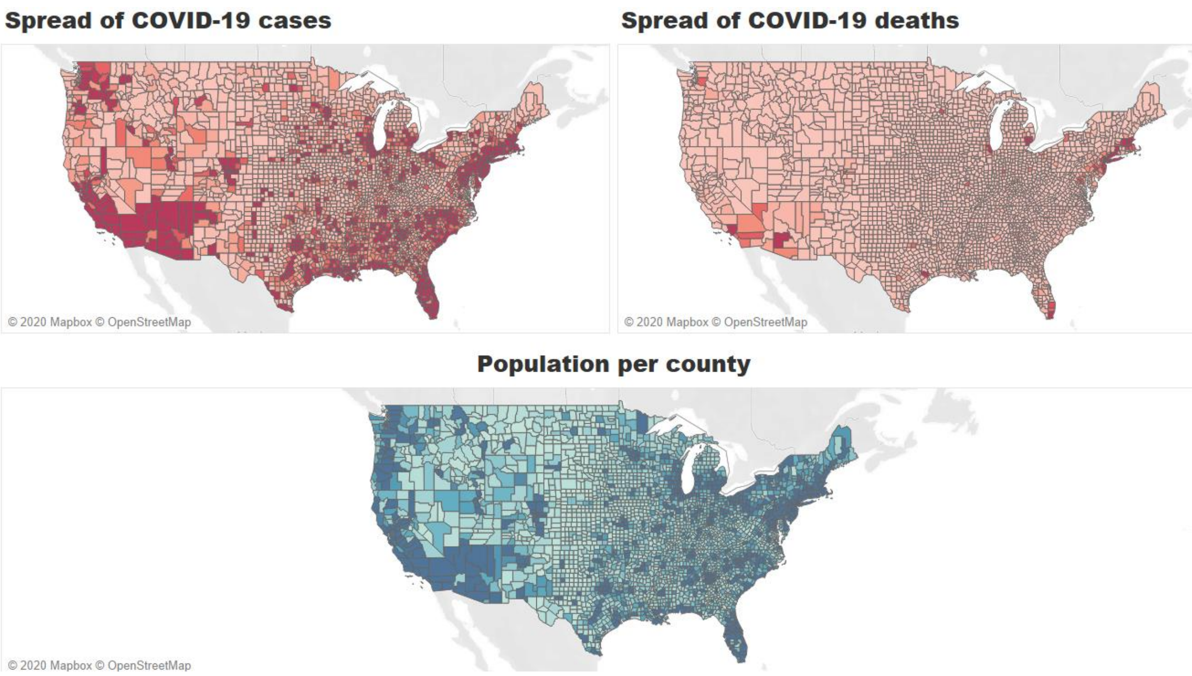


Fig 4.1 Initial analysis of COVID-19

Above figure represents the spread of COVID-19 cases and deaths on county level along with population of each county. It's very clear from above visualization that highly populated or dense counties have a greater number of COVID-19 cases that of less populated county. Also, deaths are more in highly affected counties here.

After visualizing spread of COVID-19 over counties of the US, next task is to check the relation between income of COVID-19. For this visualization some economic features like per capita income and income ratio have been used. In figure 4.2 per capita income and COVID-19 data are simply plotted using ggplot in R with cases per hundred thousand residents in the US. It clearly shows that, per capita income and COVID-19 data are related to each other and people in the average income base line are exposed to more risk of covid-19 as they need to commute on a daily bases for work.

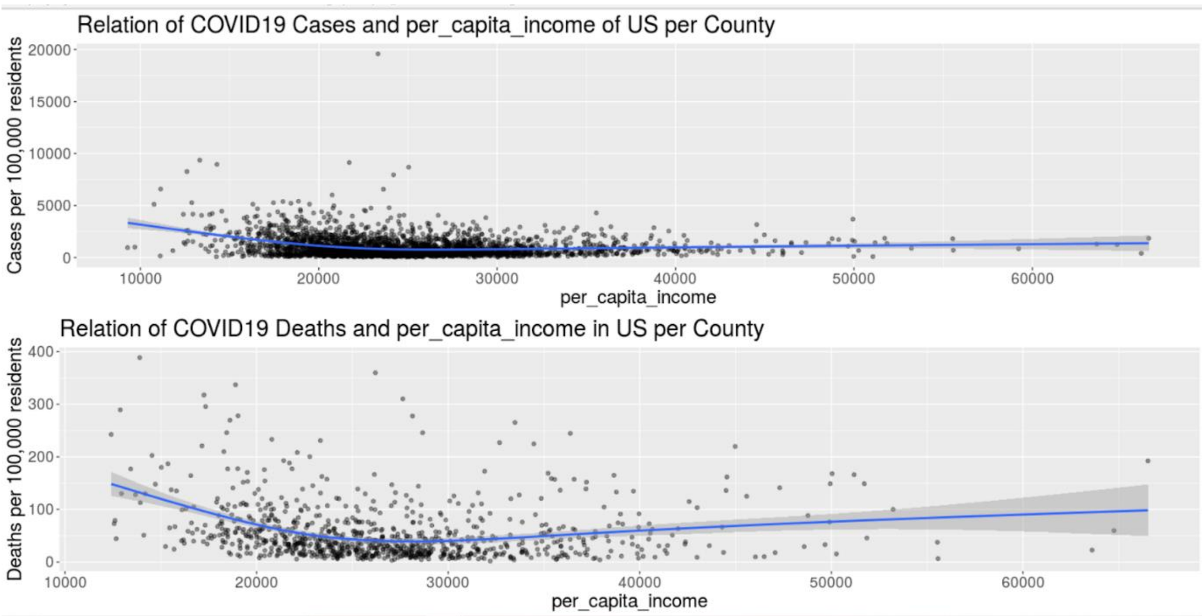


Fig 4.2 Relation of COVID-19 and Per capita income

Similar to per capita income, income ratio of the US and COVID-19 data is also plotted using ggplot in R with cases per hundred thousand residents in the US. It further explains the relationship of income and COVID-19 in the US. From fig 4.3 its clear that, income ratio and COVID-19 data are related to each other. It also shows that there are some records with high income ratio and large number of COVID-19 deaths.

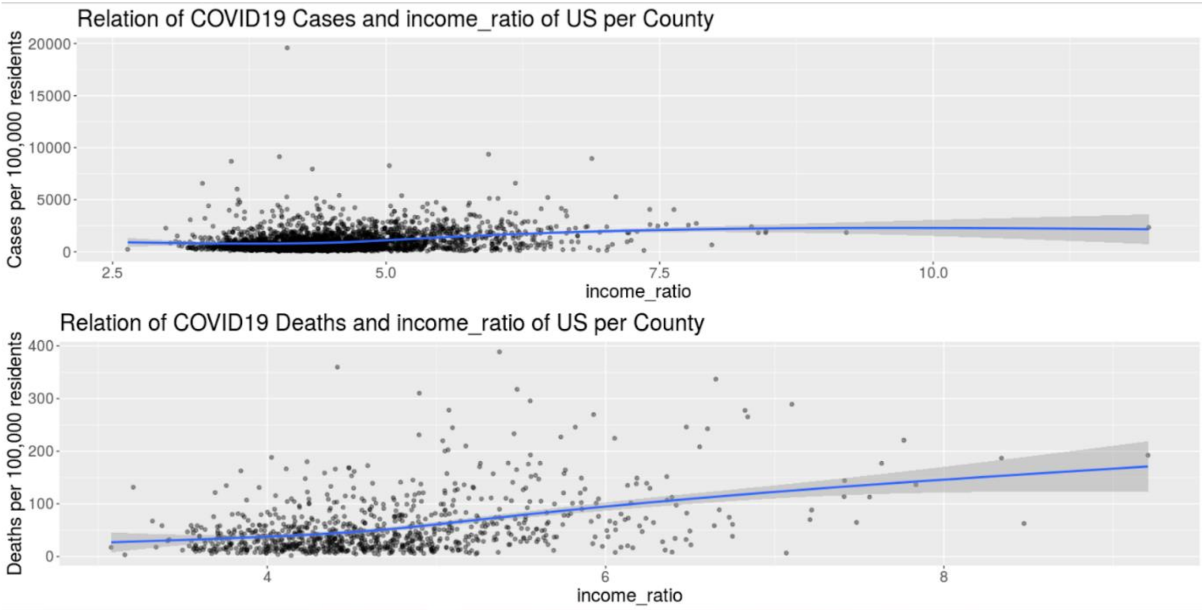


Fig 4.3 Relation of COVID-19 and Income ratio

Along with per capita income and income ratio, for initial analysis Eightieth percentile income and Twentieth percentile income is also plotted using ggplot in R and it further helps in understanding the relation between COVID-19 and income of the US.

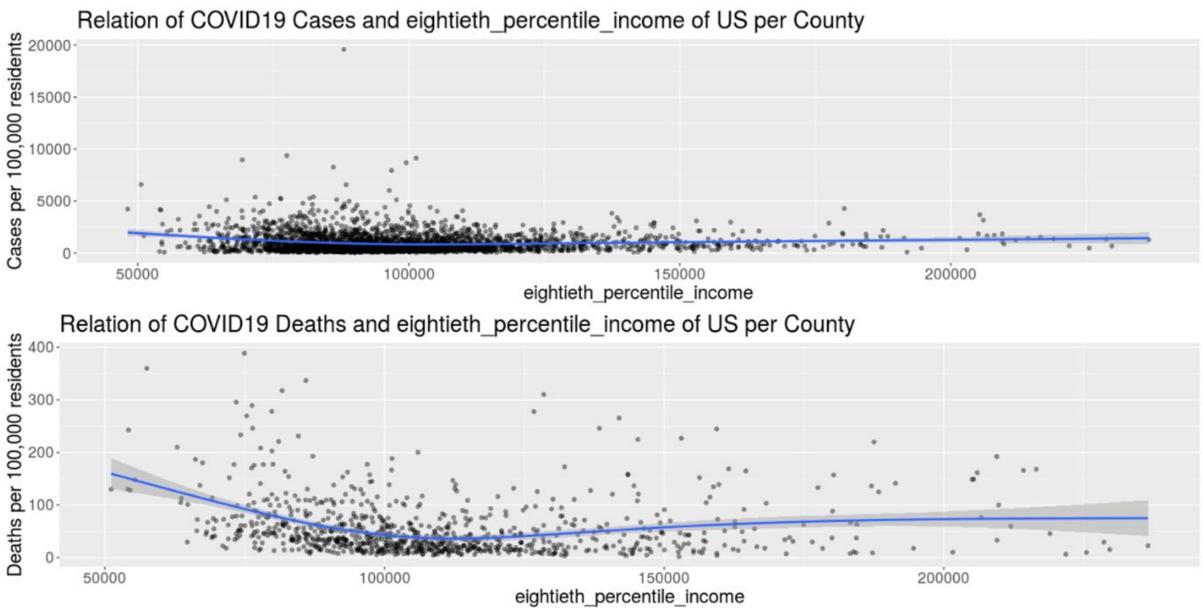


Fig 4.4 Relation of COVID-19 and Eightieth percentile income

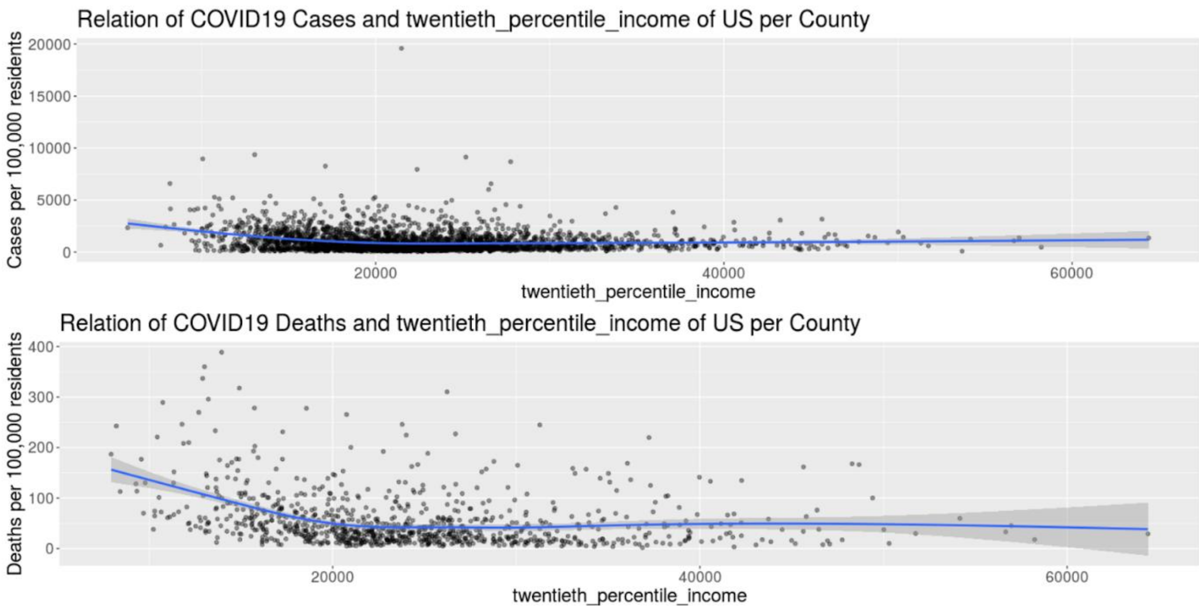


Fig 4.5 Relation of COVID-19 and Twentieth percentile income

## 4.5 GWR models' implementation

This research implements models for spatial analysis using geographically weighted regression (GWR). As mentioned in section 4.2 altogether 5 GWR models are implemented throughout this research.

### 4.5.1 Model 1: COVID-19 cases V/S Covid-19 deaths

The dataset obtained from USAFACTS has been used for this GWR model. As mentioned in table 4.1, COVID-19 data includes total cases and total deaths, hence the data was first filtered for relevant columns using R code. Data for total cases and total deaths is then joined using FIPS code of each county. Columns like County name and State name are eliminated due to the 'numeric value only' requirement of GWR function. Since this GWR model only takes about COVID-19 cases and deaths, all other economic and socio-economic features are not considered for this model.

```
total_deaths ~ total_cases
```

Using the above model design a GWR model and as a baseline to this GWR model, a simple linear regression model is implemented in this research. This GWR model is implemented to analyse COVID-19 cases and the impacted deaths due to COVID-19.

### 4.5.2 Model 2: Unemployment V/S Socio-economic factors

COVID-19 impacted many major areas of life and job being one of them. But they also present some impact on unemployment due to socio-economic factors. Unemployment has costs to a society that are more than just financial. Hence, this model is designed to analyse and understand the same.

```
num_unemployed_CDC ~ total_population +
food_environment_index + num_driving_deaths + num_uninsured +
num_primary_care_physicians + num_dentists +
num_mental_health_providers + labor_force +
num_households_CHR + social_association_rate +
annual_average_violent_crimes + num_injury_deaths +
inadequate_facilities + num_workers_who_drive_alone +
life_expectancy + num_food_insecure +
percent_limited_access_to_healthy_foods +
num_drug_overdose_deaths + num_motor_vehicle_deaths +
percent_enrolled_in_free_or_reduced_lunch + segregation_index
+ num_firearm_fatalities +
average_traffic_volume_per_meter_of_major_roadways +
num_homeowners + num_households_with_severe_cost_burden
```

Using the above model design a GWR model and a simple linear regression model is implemented in this research. This GWR model is implemented to analyse and understand the impact of socio-economic factors on unemployment in the US.

### 4.5.3 Model 3: Unemployment V/S COVID-19 cases

The COVID-19 pandemic has resulted in near 6 million cases and over 180,000 deaths in the US. It has also sparked fears of an unemployment crisis and recession. Prevention steps like social distancing, self-isolation and travel restrictions have caused a reduced workforce across every economic sector and caused many jobs to be lost.

```
num_unemployed_CDC ~ total_cases
```

Using the above model design a GWR model and a simple linear regression model is implemented in this research. This GWR model is implemented for the analysis of unemployment in the US caused due to COVID-19 pandemic.

### 4.5.4 Model 4: COVID-19 cases V/S Socio-economic factors

COVID-19 affected our health as well as society as a whole. Schools have been closed and need for manufactured products has lessened suddenly. On the other hand, demand for medical supplies and personal protective equipment (PPE) has increased significantly. There is also high demand in the food sector due to panic-buying and stockpiling of food products.

```
total_cases ~ total_population + food_environment_index +
num_driving_deaths + num_uninsured +
num_primary_care_physicians + num_dentists +
num_mental_health_providers + labor_force +
num_households_CHR + social_association_rate +
annual_average_violent_crimes + num_injury_deaths +
inadequate_facilities + num_workers_who_drive_alone +
life_expectancy + num_food_insecure +
percent_limited_access_to_healthy_foods +
```

```
num_drug_overdose_deaths + num_motor_vehicle_deaths +
percent_enrolled_in_free_or_reduced_lunch + segregation_index
+ num_firearm_fatalities +
average_traffic_volume_per_meter_of_major_roadways +
num_homeowners + num_households_with_severe_cost_burden
```

Using the above model design a GWR model and a simple linear regression model is implemented in this research. This GWR model is implemented for the analysis of the global outbreak due to COVID-19 and its deep effect on every individual.

### 4.5.5 Model 5: COVID-19 deaths V/S Socio-economic factors

It's not only the confirmed cases but COVID-19 deaths have affected the society as well. Coronavirus have potential to evolve fast and in a nastier direction [36]. With unknown cure and precautionary practices such as social distancing and self-isolation, coronavirus have become more dangerous than cancer [37].

```
total_deaths ~ total_population + food_environment_index +
num_driving_deaths + num_uninsured +
num_primary_care_physicians + num_dentists +
num_mental_health_providers + labor_force +
num_households_CHR + social_association_rate +
annual_average_violent_crimes + num_injury_deaths +
inadequate_facilities + num_workers_who_drive_alone +
life_expectancy + num_food_insecure +
percent_limited_access_to_healthy_foods +
num_drug_overdose_deaths + num_motor_vehicle_deaths +
percent_enrolled_in_free_or_reduced_lunch + segregation_index
+ num_firearm_fatalities +
average_traffic_volume_per_meter_of_major_roadways +
num_homeowners + num_households_with_severe_cost_burden
```

Using the above model design a GWR model and a simple linear regression model is implemented in this research. This GWR model is implemented for the analysis of deaths due to COVID-19 and its impact on society.

## 4.6 Linear Regression

To act as a baseline for implementing and evaluating a GWR model, a simple linear regression model is applied to each model to check heteroskedasticity of data as well as to have a metric for comparison between traditional regression analysis models and GWR by comparing scores of the $R^2$ of each model.

A linear regression model was applied on exact same variable as that of GWR. The summary stats for each model were recorded and the graph of residuals V/S fitted values were plotted to check the heteroskedasticity of data for spatial patterns. GWR working includes the assumption that the variability of data for the second set of predictor variables is not equal and it's scattered or dispersed differently. This is called heteroscedasticity property of data [35].

If the standardized residuals plot is looking dispersed randomly then the data is homoscedastic. Whereas, along the regression line if the residuals are scattered in a conical shape, then the data is heteroscedastic. Heteroscedastic data means that the variability of the data is not even across the set of predictors, second set to be precise. Here heteroscedasticity is quite evident in each model and will be discussed ahead in results section.

## 4.7 GWR

After analysis of a linear regression model, a GWR regression model was implemented on the same set of features to evaluate the impact of COVID-19 in the US. Same set of features were used for each model as the ones used for simple linear regression. Values such as AIC, adjusted $R^2$ were recorded on every stage of implementation. To run the GWR function in R, value for bandwidth was needed to select. One can select the value of bandwidth by simple assumption based on location and spread of data or by mathematical calculation.For this research, bandwidth was mathematically calculated with the help of the gwr.sel function in R from the spgwr package in. Cross-validation was carried out while selecting the best bandwidth for all models. Default kernel i.e. Gaussian kernel was used for these GWR models in R.

## 4.8 Evaluation of GWR model

To evaluate the performance of the simple linear model, values of the adjusted $R^2$, AIC were checked along with p-value and the F-statistic value of each model. The value of $R^2$ is used while estimating that how well the model fits the data, which further helps in explaining the variance in the data. Closer the $R^2$ value to 1, better the model to explains the variance of the data. But it can be said that the $R^2$ score is still very low than that of the GWR model. Along with the $R^2$ value, the p-value for each model is less with p < 0:05 for each model of simple linear regression.

The Akaike Information Criterion (AIC) value shows how well the model fits the data on the bases of complexity of the data. Lower the AIC value, better is the model. But AIC value is not completely an absolute measure for fitness of data. The model with less AIC value is a better fit for the data. Finally, the F-statistic gives the information regarding how strong the relationship is among the independent and dependent variables for the model. It’s another test for statistical significance of data where, higher the value of F-statistic stronger the statistical significance for data. For some models in this research F-statistic is very large than 1 but for some models is much smaller than 1, which might be due to size of data for analysis.

GWR models are much interested in finding patterns and relationships with data. To evaluate the performance of the GWR model, values like AICc, AIC, residual sum of squares and Quasi global $R^2$ are checked. Higher the value of the global $R^2$, model better explains the variance in the data. Goodness of the fit of the data is measured by AICs, lower AIC values is a good fit to the data. Lastly, as the name suggests residual sum of squares is a sum of squares of each residual in the model. A model with a lower value of sum represents a good fit model.

# Chapter 5

# Results and Analysis

## 5.1 Introduction

As stated in earlier sections, in this research all together five models are designed and implemented to analyse relationship between COVID-19 cases, unemployment and socio-economic factors. This regression analysis task uses COVID-19 data of the United States along with data of unemployment and socio-economic factors of the United States. This research analyses impact on COVID-19 using GWR models and results from each GWR model is explained below.

## 5.2 Model 1: COVID-19 cases V/S Covid-19 deaths

As mentioned in section 4.5.1 this model completely talks about COVID-19 cases and deaths. First a simple linear regression is applied which acts like a baseline for further analysis. Fig 5.1 represents the summary statistics for this model whereas fig 5.2 shows the graph of residuals V/S fitted values for model 1 and it can be used to check for spatial patterns in the data.

```
> #LR
> model1 <- lm(total_deaths ~ total_cases)
> summary(model1)

Call:
lm(formula = total_deaths ~ total_cases)

Residuals:
    Min      1Q  Median      3Q     Max
-2398.5    -6.1    -0.5     1.0  5038.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0966923  3.4069846  -0.322    0.748
total_cases  0.0354434  0.0005441  65.142   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 186.5 on 3140 degrees of freedom
Multiple R-squared:  0.5747,	Adjusted R-squared:  0.5746
F-statistic:  4243 on 1 and 3140 DF,  p-value: < 2.2e-16
```

Fig 5.1 LR summary – Model 1

Linear regression model 1 has $R^2$ score and Adjusted $R^2$ score of 57% which is low. Also, p-value is very small too.

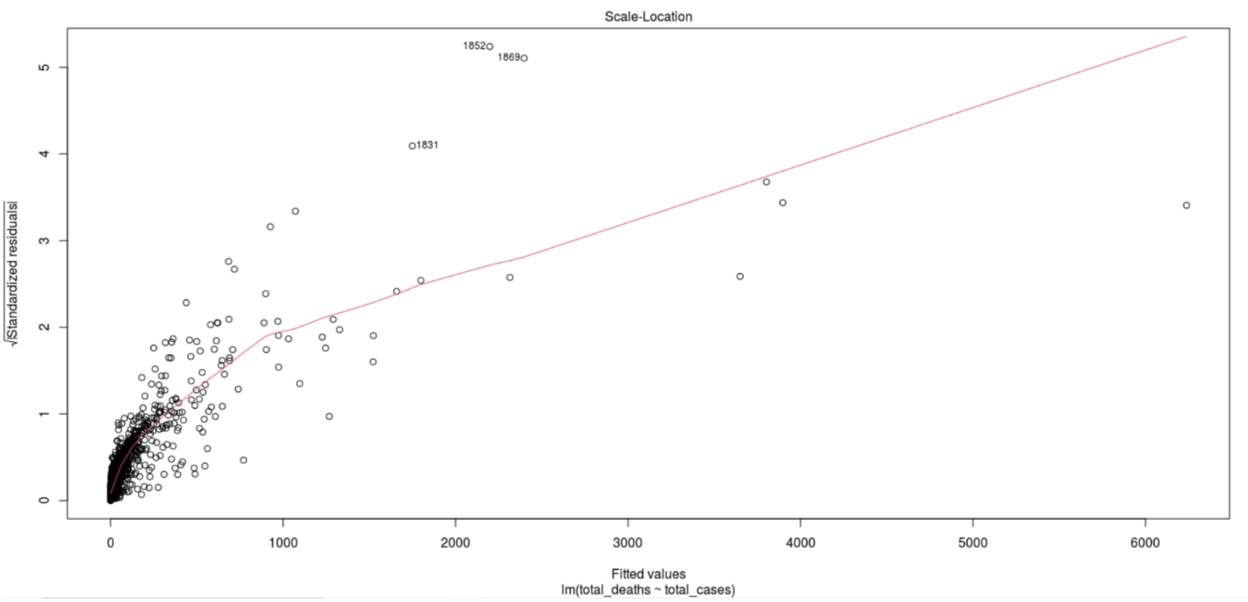


Fig 5.2 Standardised Residuals Plot LR – Model 1

In fig 5.2 the residuals are clearly scattered in a conical shape along the regression line which shows variability of the data.

Now, let's take a look at the summary of the GWR model using fig 5.3, it shows that the value of AICc is 40629.66 and AIC is 40624.28. GWR regression model 1 has a Quasi-global $R^2$ score of 70% which is far better than the $R^2$ score of the linear model.

```
Call:
gwr(formula = total_deaths ~ total_cases, data = gwr_1, coords = cbind(lat,
    lon), adapt = GWRbandwidth, hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.618034 (about 1941 of 3142 data points)
Summary of GWR coefficient estimates at data points:
                   Min.    1st Qu.     Median    3rd Qu.       Max.  Global
X.Intercept. -11.783170  -9.105888  -5.923376  -1.959076  -0.248799 -1.0967
total_cases    0.028175   0.030889   0.038508   0.046239   0.051135  0.0354
Number of data points: 3142
Effective number of parameters (residual: 2traceS - traceS'S): 4.235408
Effective degrees of freedom (residual: 2traceS - traceS'S): 3137.765
Sigma (residual: 2traceS - traceS'S): 155.3867
Effective number of parameters (model: traceS): 3.363649
Effective degrees of freedom (model: traceS): 3138.636
Sigma (model: traceS): 155.3651
Sigma (ML): 155.2819
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 40629.66
AIC (GWR p. 96, eq. 4.22): 40624.28
Residual sum of squares: 75761405
Quasi-global R2: 0.7050245
```

Fig 5.3 GWR summary – Model 1

The output from the GWR model reveals how the coefficients vary across all the 3000+ counties in the United States. Using GWR coefficients are checked using Spatial Points Data Frame (SDF) function of GWR. The output from the SDF function provides GWR coefficients which can be attached to the original data frame, this is quite easy to achieve as the coefficients for each county appear in the same order in SDF output as they do in the original data frame.

In this particular model, taking coefficient of total_cases causing deaths into the consideration, the coefficients range from a minimum value of 0.028175 to 0.051135. Hence using the inter-quartile range, it can be said that for half counties the dataset, number of total cases rises by 1 point, percentage of cases causing death are increasing from 3% to around 4.5%.

Fig 5.4 represents a visual pattern extracted from the GWR coefficients that shows COVID-19 cases causing deaths per county of the US.

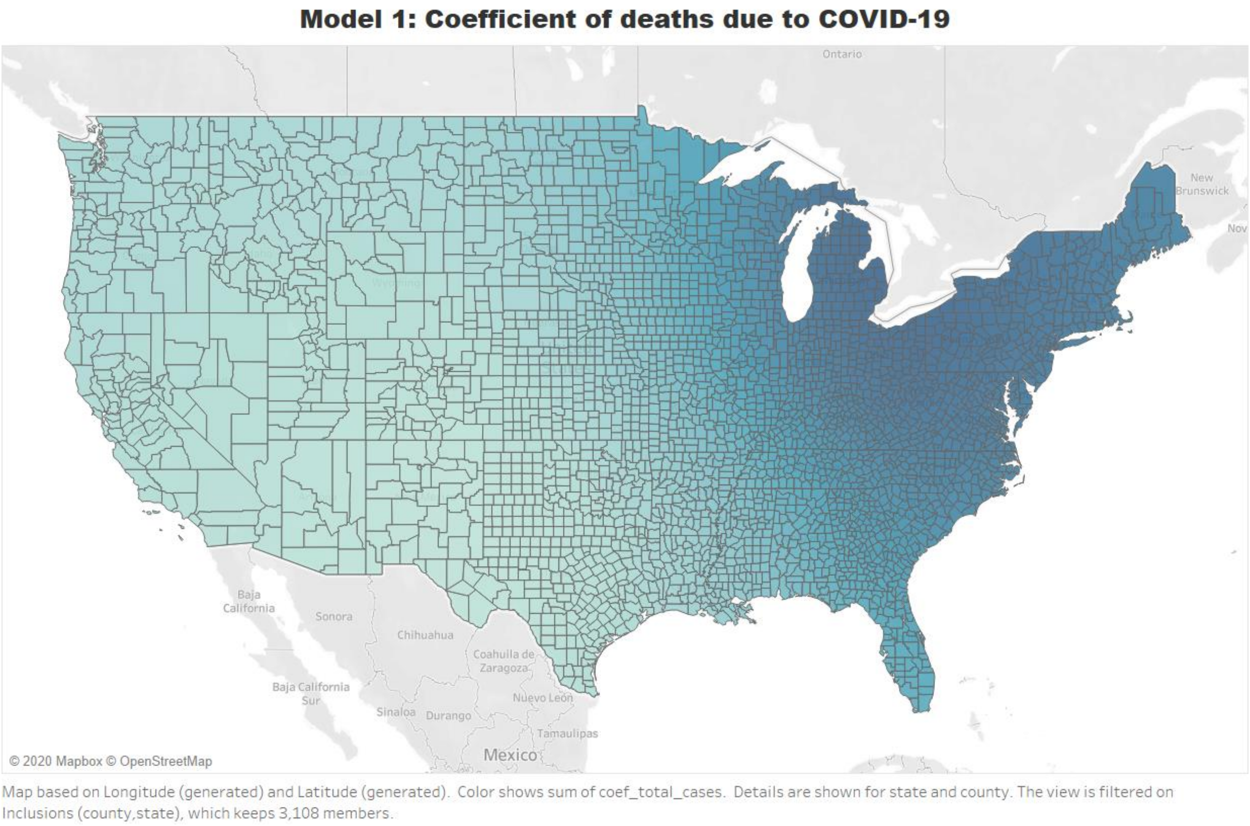


Fig 5.4 GWR coefficients – Model 1

## 5.3 Model 2: Unemployment V/S Socio-economic factors

As explained in section 4.5.2, this model analyses unemployment caused due to socio-economic factors in the United States. Simple linear regression is also applied to act as a baseline for further analysis. Fig 5.5 represents the linear regression summary statistics for this model whereas fig 5.2 shows the graph of linear regression residuals V/S fitted values for this model and it's been used to check for spatial patterns in the data.

```
Call:lm(formula = num_unemployed_CDC ~ total_population + food_environment_index +
    num_driving_deaths + num_uninsured + num_primary_care_physicians +
    num_dentists + num_mental_health_providers + labor_force +
    num_households_CHR + social_association_rate + annual_average_violent_crimes +
    num_injury_deaths + inadequate_facilities + num_workers_who_drive_alone +
    life_expectancy + num_food_insecure + percent_limited_access_to_healthy_foods +
    num_drug_overdose_deaths + num_motor_vehicle_deaths + percent_enrolled_in_free_or_reduced_lunch +
    segregation_index + num_firearm_fatalities + average_traffic_volume_per_meter_of_major_roadways +
    num_homeowners + num_households_with_severe_cost_burden)

Residuals:
     Min       1Q   Median       3Q      Max
-11863.7   -209.5     29.5    248.3  19240.1

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       -1.831e+02  2.859e+02  -0.640  0.52193
total_population                                   4.928e-02  3.392e-03  14.531  < 2e-16 ***
food_environment_index                            -2.440e+01  2.626e+01  -0.929  0.35291
num_driving_deaths                                 1.388e+01  1.584e+00   8.761  < 2e-16 ***
num_uninsured                                     -3.035e-02  2.873e-03 -10.565  < 2e-16 ***
num_primary_care_physicians                       -4.126e+00  6.016e-01  -6.859 8.35e-12 ***
num_dentists                                      -4.885e-01  6.307e-01  -0.775  0.43865
num_mental_health_providers                        1.195e+00  8.775e-02  13.616  < 2e-16 ***
labor_force                                        1.256e-02  4.487e-03   2.798  0.00517 **
num_households_CHR                                 3.746e-02  5.539e-03   6.764 1.60e-11 ***
social_association_rate                            7.703e+00  3.955e+00   1.948  0.05155 .
annual_average_violent_crimes                      6.169e-01  6.166e-02  10.003  < 2e-16 ***
num_injury_deaths                                 -8.282e+00  3.112e-01 -26.611  < 2e-16 ***
inadequate_facilities                              3.337e+01  1.413e+01   2.362  0.01822 *
num_workers_who_drive_alone                       -9.111e-02  5.919e-03 -15.393  < 2e-16 ***
life_expectancy                                   -2.910e-01  2.056e+00  -0.142  0.88744
num_food_insecure                                 -3.476e-02  3.840e-03  -9.052  < 2e-16 ***
percent_limited_access_to_healthy_foods           -1.023e+00  3.821e+00  -0.268  0.78897
num_drug_overdose_deaths                           1.356e+01  6.030e-01  22.494  < 2e-16 ***
num_motor_vehicle_deaths                          -3.642e+00  1.435e+00  -2.538  0.01119 *
percent_enrolled_in_free_or_reduced_lunch          4.550e+00  1.213e+00   3.751  0.00018 ***
segregation_index                                 -3.228e+00  1.003e+00  -3.217  0.00131 **
num_firearm_fatalities                             2.089e+01  6.221e-01  33.581  < 2e-16 ***
average_traffic_volume_per_meter_of_major_roadways -1.295e-01  1.062e-01  -1.219  0.22305
num_homeowners                                     4.159e-02  4.387e-03   9.480  < 2e-16 ***
num_households_with_severe_cost_burden             3.035e-01  1.413e-02  21.477  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1253 on 3118 degrees of freedom
Multiple R-squared:  0.9934,  Adjusted R-squared:  0.9934
F-statistic: 1.882e+04 on 25 and 3118 DF,  p-value: < 2.2e-16
```

Fig 5.5 LR Summary – Model 2

Linear regression model 2 has $R^2$ score and Adjusted $R^2$ score of 99.3% which is very good. But, the p-value is very small too and the value of F-statistic is quite small too.

In fig 5.6, the residuals are somewhat scattered in a conical shape along the regression line which shows little variability of the data.

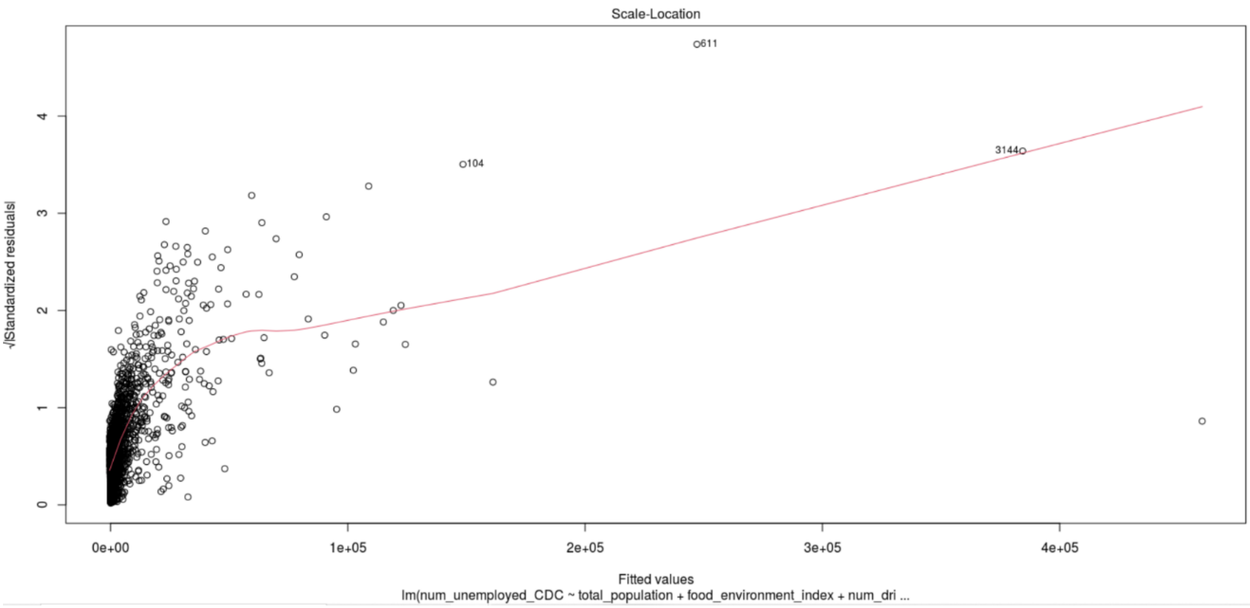


Fig 5.6 Standardised Residuals Plot LR – Model 2

Now, looking at summary of the GWR model using fig 5.7, it shows that the value of AICc is 53270.69 and AIC is 53228.35. GWR regression model 2 has Quasi-global $R^2$ score of 99.4% which is little better than $R^2$ score of the linear model. Summary of GWR coefficients is omitted here

```
> gwr.model
Call: gwr(formula = num_unemployed_CDC ~ total_population + food_environment_index +
    num_driving_deaths + num_uninsured + num_primary_care_physicians +
    num_dentists + num_mental_health_providers + labor_force +
    num_households_CHR + social_association_rate + annual_average_violent_crimes +
    num_injury_deaths + inadequate_facilities + num_workers_who_drive_alone +
    life_expectancy + num_food_insecure + percent_limited_access_to_healthy_foods +
    num_drug_overdose_deaths + num_motor_vehicle_deaths + percent_enrolled_in_free_or_reduced_lunch +
    segregation_index + num_firearm_fatalities + average_traffic_volume_per_meter_of_major_roadways +
    num_homeowners + num_households_with_severe_cost_burden,
    data = gwr_2, coords = cbind(lat, lon), adapt = GWRbandwidth,
    hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.618034 (about 1943 of 3144 data points)
Number of data points: 3144
Effective number of parameters (residual: 2traceS - traceS'S): 47.30729
Effective degrees of freedom (residual: 2traceS - traceS'S): 3096.693
Sigma (residual: 2traceS - traceS'S): 1149.942
Effective number of parameters (model: traceS): 39.26512
Effective degrees of freedom (model: traceS): 3104.735
Sigma (model: traceS): 1148.452
Sigma (ML): 1141.258
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 53270.69
AIC (GWR p. 96, eq. 4.22): 53228.35
Residual sum of squares: 4094962694
Quasi-global R2: 0.994494
```

Fig 5.7 GWR Summary – Model 2

Following figures represents visual pattern extracted from the GWR coefficients, how each socio-economic factor caused unemployment in each county of the US.

Positive coefficients are represented by blue colour and yellow colour represents negative GWR coefficients. Colour density has been used to represent how local coefficients vary across all the counties in the United States. Global coefficients of simple linear regression differ from the local coefficients of GWR.

Fig 5.8 shows visual pattern extracted from the GWR coefficients of some socio-economic factors causing unemployment such as Labour force, Firearm fatalities, Segregation index and Social association rate.

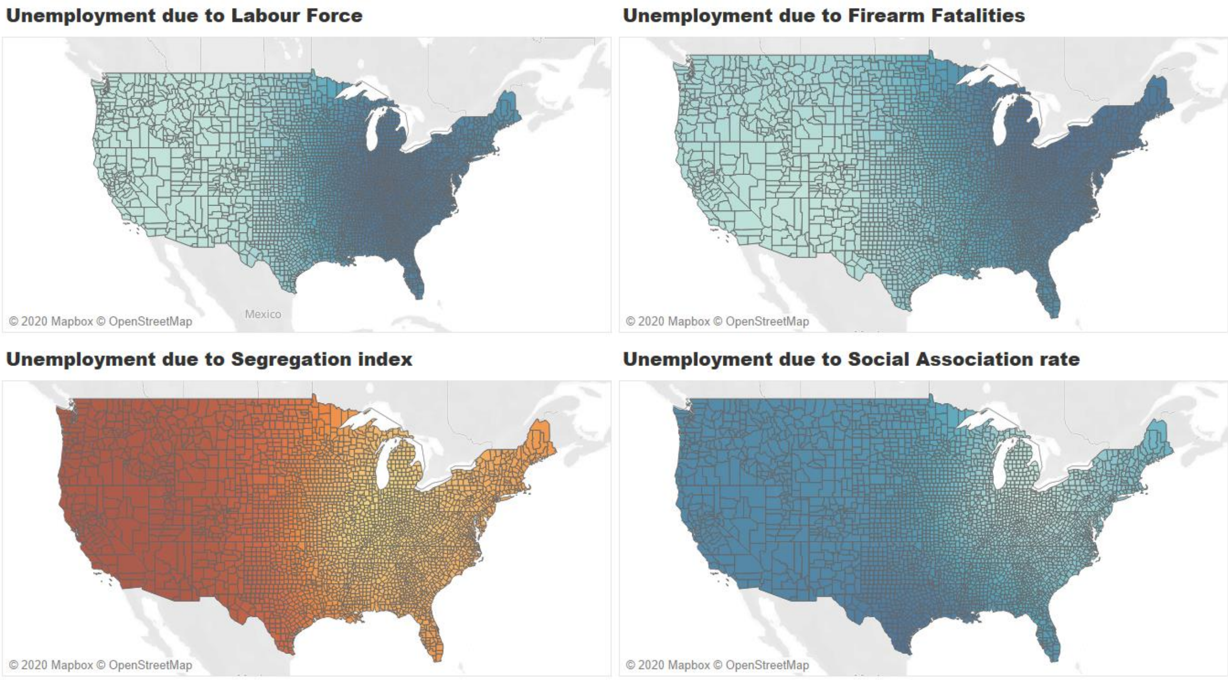


Fig 5.8 GWR Coefficients – Model 2

For Labour force, the coefficients range from a minimum value of 0.000514 to 0.03805. For Firearm fatalities, the coefficients range from a minimum value of 17.954 to 24.132. For Segregation index, the coefficients range from a minimum value of -3.746 to -1.008. Lastly for Social association rate, the coefficients range from a minimum value of 3.865 to 8.418.

## 5.4 Model 3: Unemployment V/S COVID-19 cases

As mentioned in section 4.5.3 this model completely talks about COVID-19 impact on economic sector, unemployment to be exact. Fig 5.9 represents the summary statistics for this model and fig 5.10 shows the graph of residuals V/S fitted values for model 1 and it can be used to check for spatial patterns in the data.

```
Call:
lm(formula = num_unemployed_CDC ~ total_cases)

Residuals:
    Min      1Q  Median      3Q     Max 
-110767    -896    -612     121   90397 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 934.20850   96.47605   9.683   <2e-16 ***
total_cases   2.09759    0.01541 136.143   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5281 on 3140 degrees of freedom
Multiple R-squared:  0.8551,	Adjusted R-squared:  0.8551 
F-statistic: 1.853e+04 on 1 and 3140 DF,  p-value: < 2.2e-16
```

Fig 5.9 LR Summary – Model 3

Linear regression of this model has $R^2$ score and Adjusted $R^2$ score of 85% which is good. But, p-value is very small. In fig 5.10 the residuals are scattered in a conical shape along the regression line which shows variability of the data.

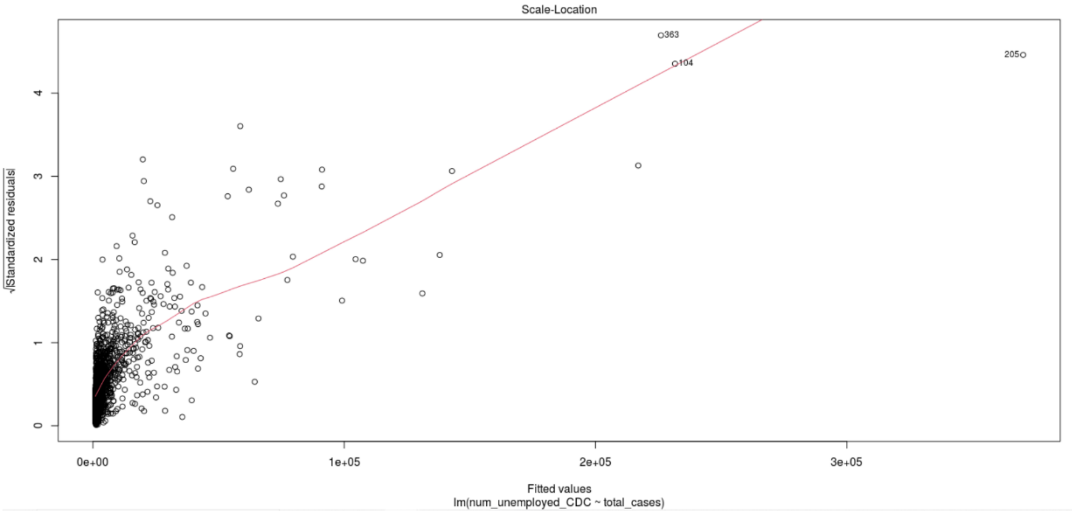


Fig 5.10 Standardised Residuals Plot LR – Model 3

Fig 5.11 shows the summary of the GWR model, it shows that the value of AICc is 62396.17 and AIC is 62390.79. GWR regression model 1 has a Quasi-global $R^2$ score of 87% which is little better than the $R^2$ score of the linear model.

```
Call:
gwr(formula = num_unemployed_CDC ~ total_cases, data = gwr_3,
    coords = cbind(lat, lon), adapt = GWRbandwidth, hatmatrix = TRUE,
    se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.618034 (about 1941 of 3142 data points)
Summary of GWR coefficient estimates at data points:
                  Min.   1st Qu.    Median   3rd Qu.      Max.   Global
X.Intercept.  631.0728  708.1679  822.8216 1011.9857 1104.4512 934.2085
total_cases     1.8653    1.9198    2.0627    2.1534    2.2269   2.0976
Number of data points: 3142
Effective number of parameters (residual: 2traceS - traceS'S): 4.235408
Effective degrees of freedom (residual: 2traceS - traceS'S): 3137.765
Sigma (residual: 2traceS - traceS'S): 4962.755
Effective number of parameters (model: traceS): 3.363649
Effective degrees of freedom (model: traceS): 3138.636
Sigma (model: traceS): 4962.065
Sigma (ML): 4959.409
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 62396.17
AIC (GWR p. 96, eq. 4.22): 62390.79
Residual sum of squares: 77279792501
Quasi-global R2: 0.872177
```

Fig 5.11 GWR Summary – Model 3

The output from this GWR model is extracted using Spatial Points Data Frame (SDF) function resultant coefficients' data reveals how the coefficients vary across all the 3000+ counties in the United States.

In this particular model, taking coefficient of total_cases causing unemployment into the consideration, the coefficients range from a minimum value of 1.8653 to 2.2269. Hence using the inter-quartile range, it can be said that for counties of the United States, number of total cases rises by 1 point, percentage of cases causing unemployment are elevating from 1.9% to around 2.1%.

Fig 5.12 represents a visual pattern extracted from the GWR coefficients that shows pattern of COVID-19 cases causing unemployment per county of the US.

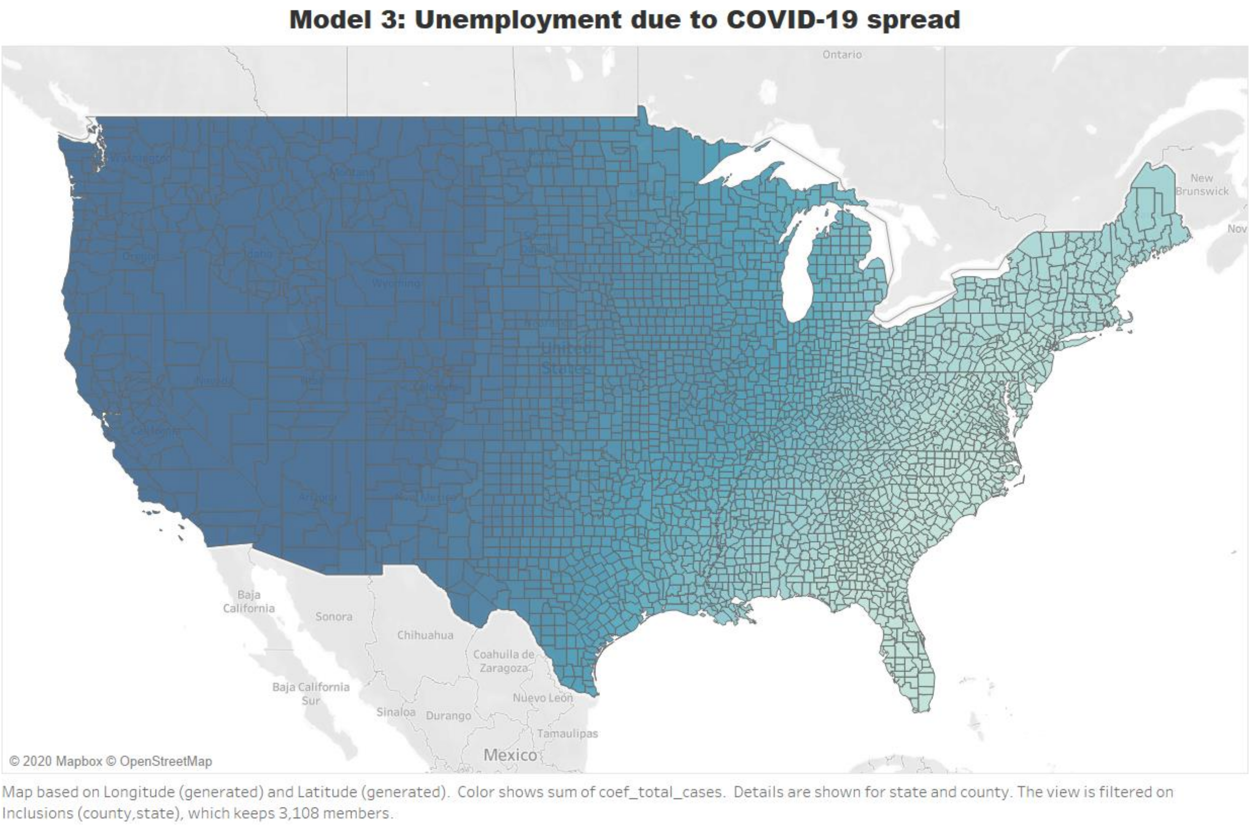


Fig 5.12 GWR Coefficients – Model 3

## 5.5 Model 4: COVID-19 cases V/S Socio-economic factors

As explained in section 4.5.4, this model analyses impact of COVID-19 cases on socio-economic factors in the United States. Due less accuracy results, simple linear regression won’t be discussed here, only GWR regression model will be discussed. As this model is one of the most important GWR models in research, effect of COVID-19 cases on each socio-economic factor is visualized individually here. Fig 5.13 shows the summary of the GWR model, it shows that the value of AICc is 53392.14 and AIC is 53349.71. GWR regression model has a Quasi-global $R^2$ score of 96%.

```
Call:
gwr(formula = total_cases ~ total_population + food_environment_index +
    num_driving_deaths + num_uninsured + num_primary_care_physicians +
    num_dentists + num_mental_health_providers + labor_force +
    num_households_CHR + social_association_rate + annual_average_violent_crimes +
    num_injury_deaths + inadequate_facilities + num_workers_who_drive_alone +
    life_expectancy + num_food_insecure + percent_limited_access_to_healthy_foods +
    num_drug_overdose_deaths + num_motor_vehicle_deaths + percent_enrolled_in_free_or_reduced_lunch +
    segregation_index + num_firearm_fatalities + average_traffic_volume_per_meter_of_major_roadways +
    num_homeowners + num_households_with_severe_cost_burden,
    data = gwr_3, coords = cbind(lat, lon), adapt = GWRbandwidth,
    hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.618034 (about 1941 of 3142 data points)
Number of data points: 3142
Effective number of parameters (residual: 2traceS - traceS'S): 47.45632
Effective degrees of freedom (residual: 2traceS - traceS'S): 3094.544
Sigma (residual: 2traceS - traceS'S): 1178.712
Effective number of parameters (model: traceS): 39.35177
Effective degrees of freedom (model: traceS): 3102.648
Sigma (model: traceS): 1177.172
Sigma (ML): 1169.777
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 53392.14
AIC (GWR p. 96, eq. 4.22): 53349.71
Residual sum of squares: 4299443272
Quasi-global R2: 0.96341
```

Fig 5.13 GWR Summary – Model 4

The output from the SDF function provides GWR coefficients which can reveal information about how the coefficients of the socio-economic factors vary across all the 3000+ counties in the United States by plotting the coefficients over the map. Following figures represents visual pattern extracted from the GWR coefficients to how each socio-economic factor has affected due to COVID-19 cases in the US.

Positive coefficients are represented by blue colour and yellow colour represents negative GWR coefficients. Colour density has been used to represent how local coefficients vary across all the counties in the United States. Global coefficients of simple linear regression differ from the local coefficients of GWR.

Fig 5.14 shows visual pattern extracted from the GWR coefficients of some socio-economic factors affected due to COVID-19 cases such as Food environment index, Inadequate facilities, Insecure food and People enrolled in free/reduced lunch. For Food environment index, the coefficients range from a minimum value of -105.05 to -62.02. For Inadequate facilities, the coefficients range from a minimum value of -4.73 to 17.52. For Insecure food, the coefficients range from a minimum value of -0.14175 to -0.11073. Lastly for People enrolled in free/reduced lunch, the coefficients range from a minimum value of 1.639 to 3.354.

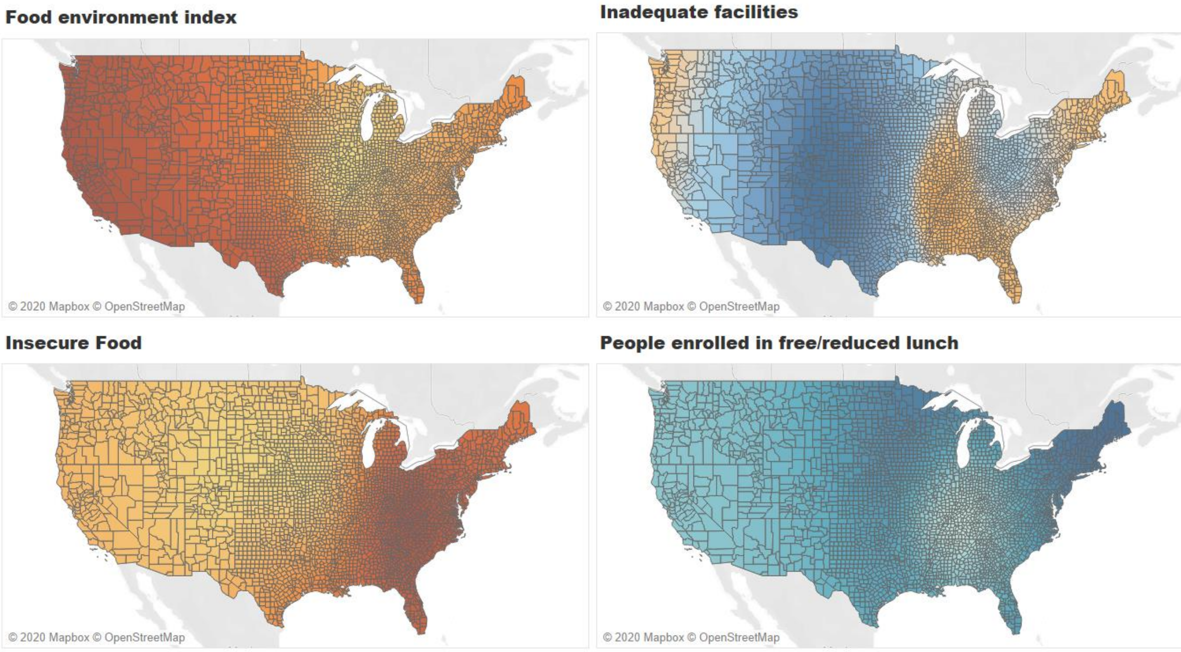


Fig 5.14 GWR Coefficients (set 1) – Model 4

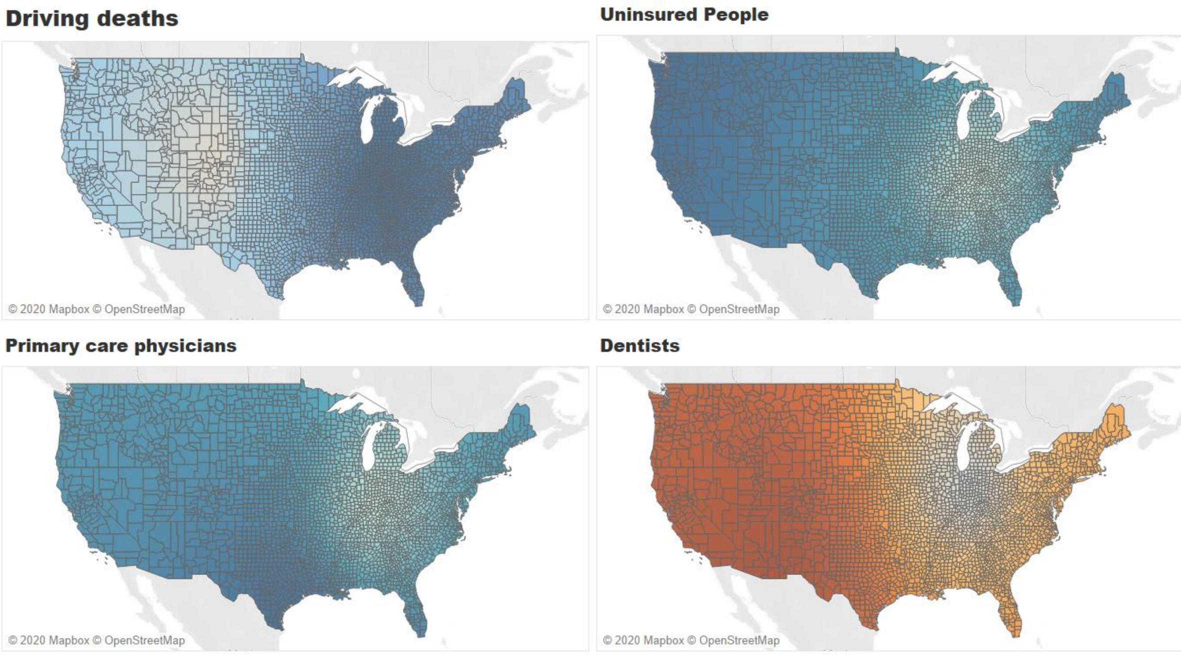


Fig 5.15 GWR Coefficients (set 2) – Model 4

Fig 5.15 shows visual pattern extracted from the GWR coefficients of another set of socio-economic factors affected due to COVID-19 cases such as Driving deaths, Uninsured people, Primary care physicians and dentists. For Driving deaths, the coefficients range from a minimum value of -0.041 to 5.912. For Uninsured people, the coefficients range from a minimum value of 0.01786 to 0.05490. For Primary care physicians, the coefficients range from a minimum value of 3.933 to 8.971. Lastly for Dentists, the coefficients range from a minimum value of -4.850 to 0.059.

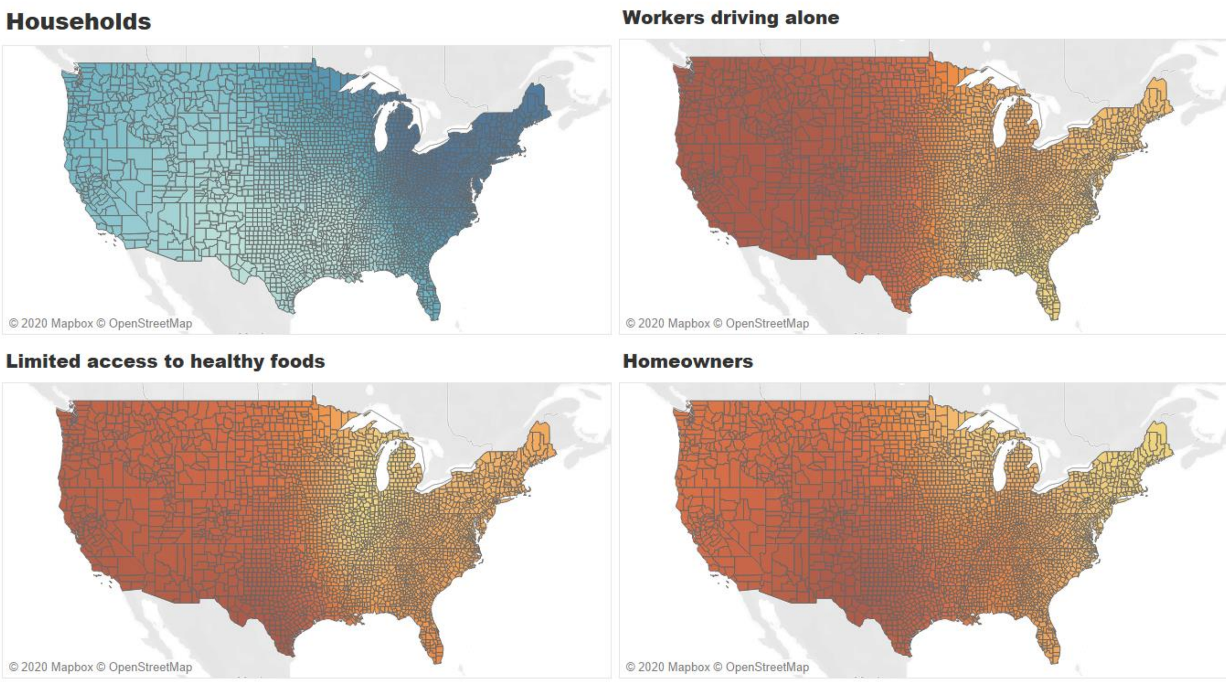


Fig 5.16 GWR Coefficients (set 3) – Model 4

Fig 5.16 shows visual pattern extracted from the GWR coefficients of some more socio-economic factors affected due to COVID-19 cases such as Households, Workers driving alone, Limited access to healthy foods and Homeowners. For Households the coefficients range from a minimum value of 0.03117 to 0.05764. For Workers driving alone, the coefficients range from a minimum value of -0.10380 to -0.04423. For Limited access to healthy foods, the coefficients range from a minimum value of -4.128 to -1.178. Lastly for Homeowners, the coefficients range from a minimum value of -0.05789 to -0.00304.

The Last set of socio-economic factors contains Life expectancy, Injury deaths, Drug overdose deaths and Motor/Vehicle deaths, Fig 5.17. For Life expectancy, the coefficients range from a minimum value of -1.803 to -0.277. For Injury deaths the coefficients range from a minimum value of -1.392 to 2.887. For Drug overdose, the coefficients range from a minimum value of 1.290 to 4.523. Lastly for Motor/Vehicle deaths, the coefficients range from a minimum value of -3.99 to 6.05.

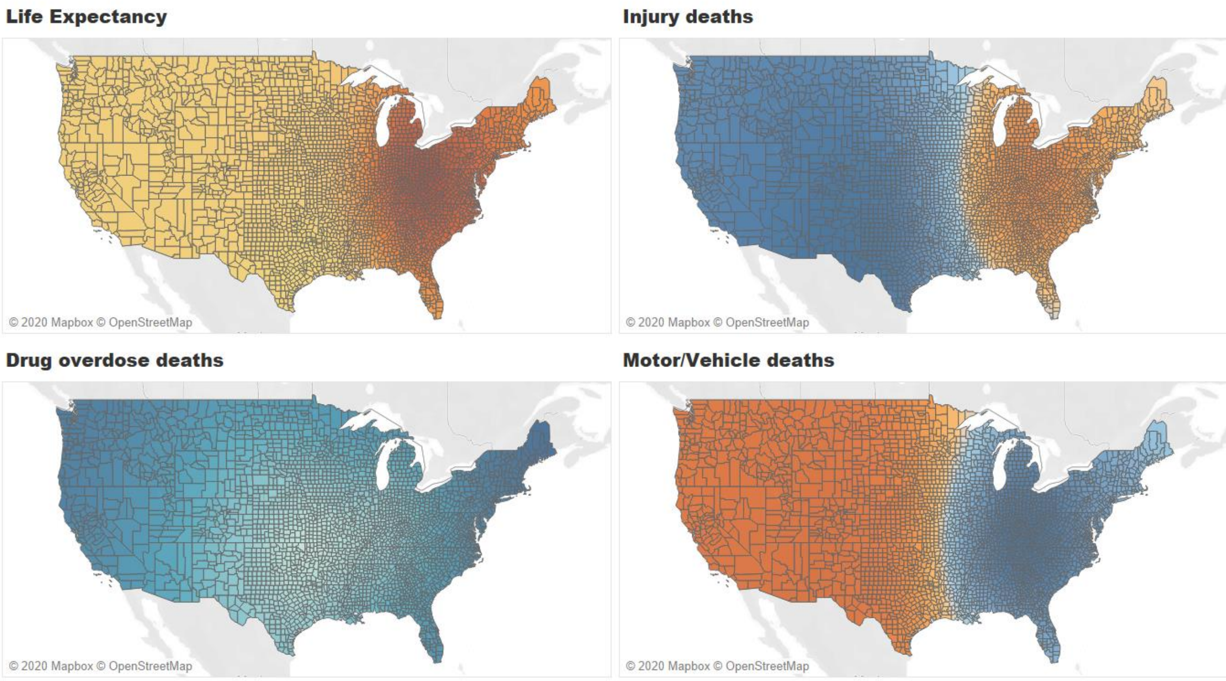


Fig 5. 17 GWR Coefficients (set 4) – Model 4

## 5.6 Model 5: COVID-19 deaths V/S Socio-economic factors

As explained in section 4.5.5, this model analyses impact COVID-19 deaths on socio-economic factors in the United States. This model is another most important GWR model in research, effect of COVID-19 deaths on each socio-economic factor is visualized individually here. Fig 5.18 shows the summary of the GWR model, it shows that the value of AICc is 37280.32 and AIC is 37237.89. GWR regression model has a Quasi-global $R^2$ score of 90%.

```
gwr(formula = total_deaths ~ total_population + food_environment_index +
    num_driving_deaths + num_uninsured + num_primary_care_physicians +
    num_dentists + num_mental_health_providers + labor_force +
    num_households_CHR + social_association_rate + annual_average_violent_crimes +
    num_injury_deaths + inadequate_facilities + num_workers_who_drive_alone +
    life_expectancy + num_food_insecure + percent_limited_access_to_healthy_foods +
    num_drug_overdose_deaths + num_motor_vehicle_deaths + percent_enrolled_in_free_or_reduced_lunch +
    segregation_index + num_firearm_fatalities + average_traffic_volume_per_meter_of_major_roadways +
    num_homeowners + num_households_with_severe_cost_burden,
    data = gwr_3, coords = cbind(lat, lon), adapt = GWRbandwidth,
    hatmatrix = TRUE, se.fit = TRUE)
Kernel function: gwr.Gauss
Adaptive quantile: 0.618034 (about 1941 of 3142 data points)
Number of data points: 3142
Effective number of parameters (residual: 2traceS - traceS'S): 47.45632
Effective degrees of freedom (residual: 2traceS - traceS'S): 3094.544
Sigma (residual: 2traceS - traceS'S): 90.76152
Effective number of parameters (model: traceS): 39.35177
Effective degrees of freedom (model: traceS): 3102.648
Sigma (model: traceS): 90.64291
Sigma (ML): 90.07349
AICc (GWR p. 61, eq 2.33; p. 96, eq. 4.21): 37280.32
AIC (GWR p. 96, eq. 4.22): 37237.89
Residual sum of squares: 25491781
Quasi-global R2: 0.9007483
```

Fig 5.18 GWR Summary – Model 5

The output from the SDF function gives GWR coefficients which can reveal information about how the coefficients of the socio-economic factors vary across all the 3000+ counties in the United States by plotting the coefficients over the map. Following figures represents visual pattern extracted from the GWR coefficients to how each socio-economic factor has affected due to COVID-19 deaths in the US.

Positive coefficients are represented by blue colour and yellow colour represents negative GWR coefficients. Colour density has been used to represent how local coefficients vary across all the counties in the United States. Global coefficients of linear regression differ from the local coefficients of GWR.

Fig 5.19 shows visual pattern extracted from the GWR coefficients of some socio-economic factors causing affected due to COVID-19 such as Population, Primary care physicians, Life expectancy and Household People. For Population, the coefficients range from a minimum value of -0.000375 to 0.004371. For Primary care physicians, the coefficients range from a minimum value of -0.2396 to 0.1652. For Life expectancy, the coefficients range from a minimum value of -0.2168 to 0.0415. Lastly for Household People, the coefficients range from a minimum value of 0.000313 to 0.001973.

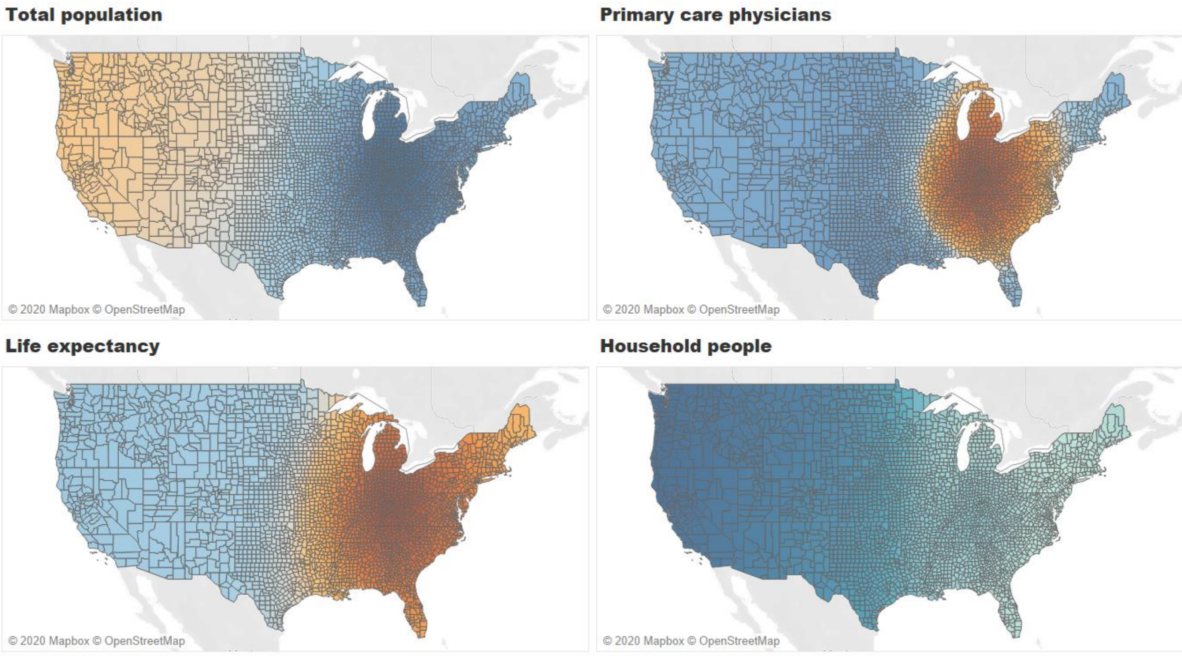


Fig 5.19 GWR Coefficients (set 1) – Model 5

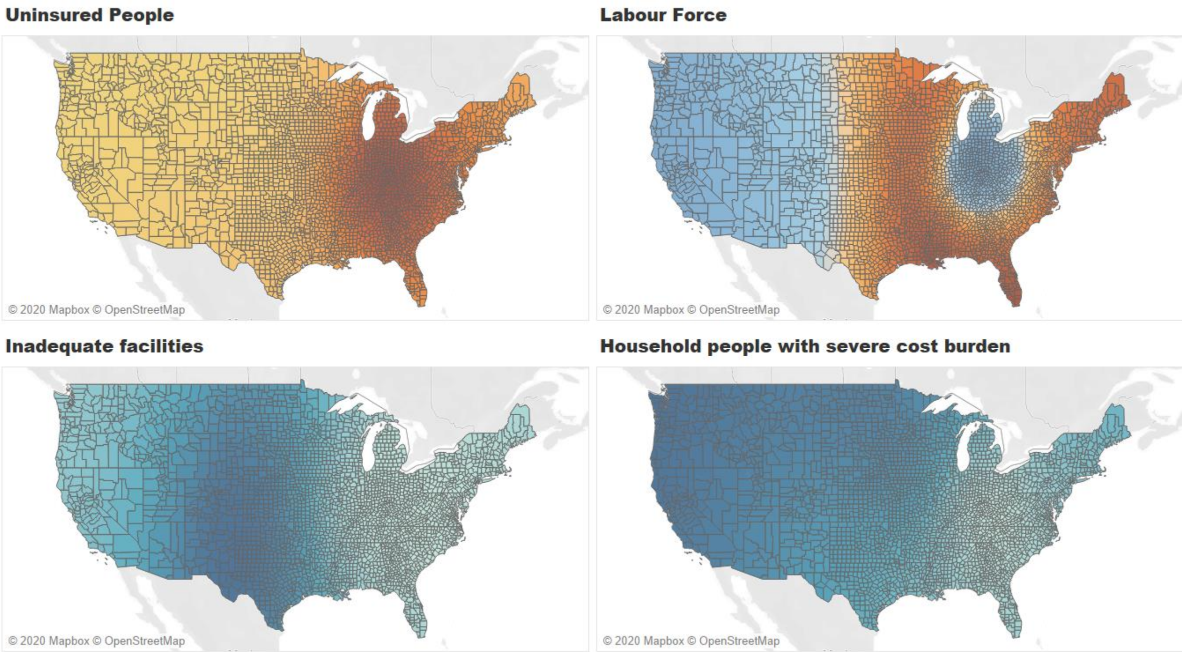


Fig 5.20 GWR Coefficients (set 2) – Model 5

Fig 5.20 shows visual pattern extracted from the GWR coefficients of some more socio-economic factors affected due to COVID-19 deaths such as Uninsured people, Labour force, Inadequate facilities and Household people with severe cost burden. For Uninsured people the coefficients range from a minimum value of -0.005723 to -0.000594. For Labour force, the coefficients range from a minimum value of -0.002076 to 0.001171. For Inadequate facilities, the coefficients range from a minimum value of 0.346 to 4.554. Lastly for Household people with severe cost burden, the coefficients range from a minimum value of 10.731 to 13.187.

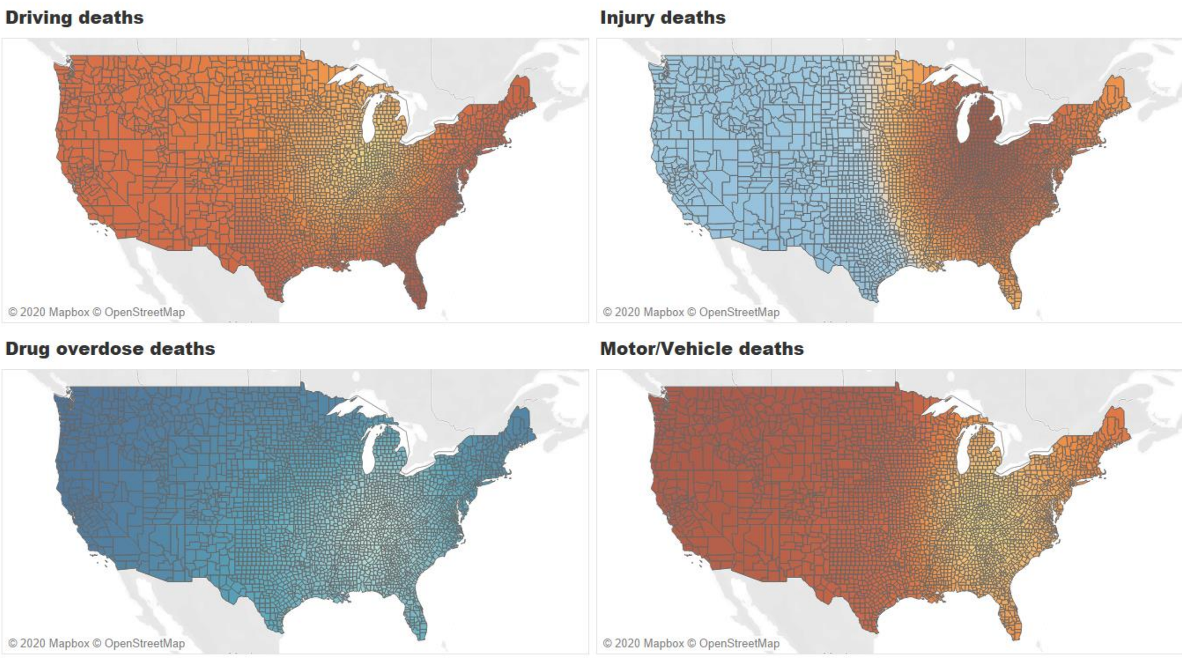


Fig 5.21 GWR Coefficients (set 3) – Model 5

Fig 5.21 shows visual pattern extracted from the GWR coefficients of some more socio-economic factors affected due to COVID-19 deaths such as Uninsured people, Labour force, Inadequate facilities and Household people with severe cost burden. For Uninsured people the coefficients range from a minimum value of -0.005723 to -0.000594. For Labour force, the coefficients range from a minimum value of -0.002076 to 0.001171. For Inadequate facilities, the coefficients range from a minimum value of 0.346 to 4.554. Lastly for Household people with severe cost burden, the coefficients range from a minimum value of 10.731 to 13.187.

Fig 5.22 shows visual pattern extracted from the GWR coefficients of last set of socio-economic factors affected due to COVID-19 deaths such as Social association rate, Violent crimes on an average, Segregation index and Homeowners. For Social association rate, the coefficients range from a minimum value of -1.4396 to -0.5369. For Violent crimes on an average, the coefficients range from a minimum value of 0.03659 to 0.09614. For Segregation index, the coefficients range from a minimum value of -0.0956 to 0.3376. Lastly for Homeowners, the coefficients range from a minimum value of -0.002228 to 0.003361.

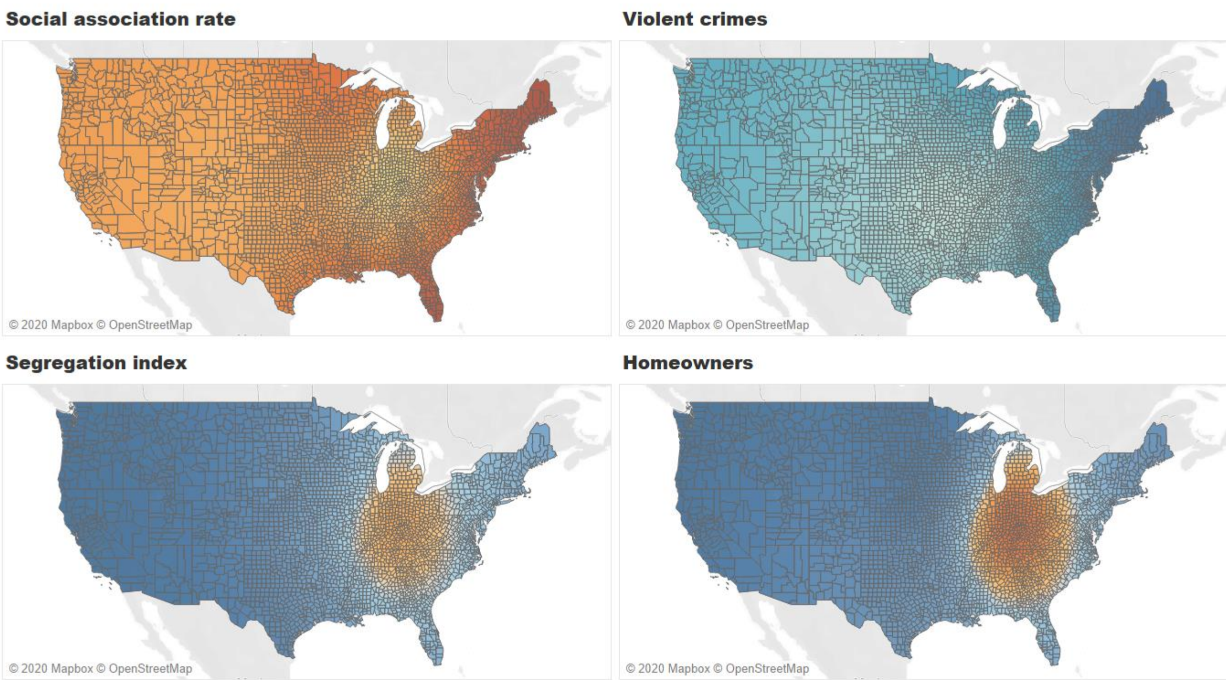


Fig 5.22 GWR Coefficients (set 4) – Model 5

# Chapter 6

# Security and Privacy Considerations

## 6.1 Data Security

Datasets used for this research contain county-level information of COVID-19, income, unemployment and statistics of socio-economic factors. These datasets do not have explicit personal details of any COVID-19 patient or unemployed individual; hence data theft is not a primary interest here. COVID-19 datasets are available on open public platforms and are updated daily by the USAFACTS. Whereas, the income and socio-economic dataset are getting updated weekly by the author of the Kaggle competition. However, as the COVID-19 data and unemployment data are correlated, it is possible to guess the information regarding COVID-19 patients based on different factors from the dataset like county name, unemployment status etc. along with location information. Hence, encrypting the data before analysis is always a good practice. Next scenario to consider is, if an attacker manages to access nelly created dataset within the analysis, this new data can get corrupted and may cause error in final output of function. This situation might lead to wrong analysis of COVID-19 data and may show some non-existing impact on socio-economic data. Hence, encryption of data is very essential here. Data encryption can be achieved by using the encryption algorithm such as RSA, SHA-1 or MD5. Encryption of data helps in maintaining the integrity and security of data by allowing only authorized people to have access to the data using a decryption mechanism.

As this research aims to analyse the relation between income and COVID-19 cases along with socio-economic factors, this data needs to be monitored or it may harm society by spreading rumours of some false pattern across the population and may strike fear among them. To prevent such scenarios from occurring, it's very important to protect the data using encryption techniques and only allowing access purely for study purposes.

## 6.2 Privacy Considerations

Even though the research doesn’t include any direct personal information of any COVID-19 patient of the US, such analysis needs some kind of regulations to be followed. In a guiding article from Episerver [38] talks about the impact of COVID-19 on data privacy, protection and security. It says, sharing personally identifiable information (PII) of a person on the name of COVID-19 should be monitored closely. With the reason of COVID-19, lots of agencies are interested in collecting data of individuals. It is suggested that hospitals and health care should review existing privacy policies of organisations before sharing any information to ensure that the privacy policies cover the disclosure of PII of patients to only governmental agencies for requested emergency purposes, including public health.

With almost every employee working from home due to COVID-19, there are definitely increases in cybersecurity issues and risks. Working remotely due to COVID-19 employees are experiencing some challenges such as poor internet bandwidth issues, increase in critical organization data to personal devices, attackers taking advantage of the COVID-19 situation and increase in security issues due to new or inexperienced remote-working employees.

Since this is academic research, the above-mentioned situation should not arise. Therefore, this research can be used to analyse COVID-19 impact on the US rather than monitoring and controlling people. However, as to the future scope of this study, use of this research in the real world needs to be monitored.

# Chapter 7

# Limitations

Even if it's very clear that GWR regression models give much better performance than global models like simple linear regression, there are certain assumptions along the way in implementation of GWR models which can act as a limitation for analysis. In GWR, Gaussian kernel and Gaussian distribution is selected by default for bandwidth selection for the variables. However, Gaussian distribution doesn't hold true in the real world, meaning that not all the data follow Gaussian distribution each time. Also, there are requirements such as, the data should include the location coordinates for each record in space for GWR study. Another requirement is numeric data only rule, considering categorical values for GWR is an issue and categorical value needs encoding. Most of the time data for analysis does include more than 50% categorical values, in such scenarios encoding this data and then applying GWR is a hassle. Hence, GWR regression models are currently not enough to handle all types of data. Since, GWR function performs multiple local linear regression for each of regression points over the data, the model takes a long amount of time for execution to complete. Hence, GWR is quite expensive in terms of execution. For large datasets parallel processing needs to be implemented.

Now regarding COVID-19 data, there are various limitations to the data. It includes lack of data for factors such as age, sex and daily habitat of individual. COVID-19 data is getting updated daily and it's not always correct, which may affect the analysis.

# Chapter 8

# Conclusion & Future work

## 8.1 Conclusion

Aim of this research included analysis of the socio-economic impact due to COVID-19. Also, this analysis focuses on the United States COVID-19 data for the same. From this study, it was very clear that economic factors, majorly unemployment, were affected deeply due to COVID-19. Along with unemployment, other socio-economic factors such as food environment index, life expectancy, total population etc. are also affected by COVID-19. Aim also include use of spatial regression technique i.e. Geographically weighted regression (GWR) analysis. Using GWR for the analysis of COVID-19 helped in understanding how the impact of COVID-19 is varying from one location to another. From the analysis of these experiments, It can be concluded that geographically weighted models are better suited for spatially varying data of COVID-19. This analysis will help in predicting social patterns or behaviour of society over COVID-19 pandemic. For example, in GWR model 3, unemployment is increasing along with the number of COVID-19 cases. This information will clearly help in predicting exactly which county is suffering more from unemployment currently and which county will be next in near future. GWR model 1 will be helpful in similar types of prediction i.e. which county in the US needs more medical care based on the coefficient of death due to COVID-19. Now focusing on the most important models from this research, the first model is COVID-19 case and socio-economic impact. This model includes around 25 socio-economic factors into analysis, which will be helpful in prediction of various social aspects. For example, factors like food environment index, insecure food, number of people enrolled in free/reduced value food, etc. will be helpful in prediction of demand and supply chain in the food industry and how the food industry is affected due to COVID-19. Similarly, Life expectancy and deaths due to drug overdose will help in predicting some of the negative effects of home isolation. Moving to the second most important model of this research, COVID-19 deaths and socio-economic impact. In this model, there is mixed variation in GWR coefficients. Since socio-economic data is from Kaggle competition, it’s not always completely accurate data. For example, in the total population, primary care physicians, life expectancy, labour force, injury deaths, etc. have negative as well as positive coefficients in the result and such variation in coefficients is not easy to use directly for future prediction.

## 8.2 Future Work

One can find various patterns using the GWR analysis model across a variety of data like human development, environment effects and so on. However, GWR also has some existing issues which need to be worked on. GWR needs more robustness and it also needs time efficiency to be implemented. GWR needs an alternative way of using categorical data for analysis, because more than half of the data in real word is nothing but categorical data.

In terms of this research, there are a lot of socio-economic factors in real life. This study only touches the tip of the iceberg here. For future research, one can include factors like education, manufacturing industry, environment and global warming into the analysis. This analysis can be expanded to include not just the United States but other countries as well.

Last but not the least, COVID-19 data is getting updated daily and it's evolving as the time goes, just like the virus itself [36]. COVID-19 impact on society might differ with time. Hence it is very much important to keep analysing the impact of COVID-19 time to time, also spatial data just expands all the possibilities and Borden’s your horizon.

# Bibliography

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al. “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China Lancet”, 395 (2020), pp. 497-506, 10.1016/S0140-6736(20)30183-5

[2] Wang, D., Hu, B., Hu, C., Zhu, F., Liu, X., Zhang, J., Wang, B., Xiang, H., Cheng, Z., Xiong, Y. and Zhao, Y., 2020. “Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. Jama”, 323(11), pp.1061-1069.

[3] Gorbalenya, A.E., Baker, S.C., Baric, R., Groot, R.J.D., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., Neuman, B.W. and Penzar, D., 2020. “Severe acute respiratory syndrome-related coronavirus: The species and its viruses–a statement of the Coronavirus Study Group”.

[4] Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., ... Diaz, G., 2020. “First case of 2019 novel coronavirus in the United States. New England Journal of Medicine” 382, 929–936. ttps://doi.org/10.1056/NEJMoa2001191.

[5] @Center for Infectious Disease Research and Policy, “US COVID-19 cases surge past 82,000, highest total in world”, 2020.

[6] Abir, M., Nelson, C., Chan, E.W., Al-Ibrahim, H., Cutter, C., Patel, K., Bogart, A., 2020. “Critical Care Surge Response Strategies for the 2020 COVID-19 Outbreak in the United States”. Retrieved from RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA100/RRA164-1/RAND_RRA164-1.pdf.

[7] @USAFACTS, “ US Coronavirus Cases and Deaths - Track COVID-19 data daily by state and county”, 2020

[8] Mollalo, A., Sadeghian, A., Israel, G.D., Rashidi, P., Sofizadeh, A., Glass, G.E., 2018. “Machine learning approaches in GIS-based ecological modeling of the sand fly Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran”. Acta Trop. 188, 187–194.

[9] Mollalo, A., Mao, L., Rashidi, P., Glass, G.E., 2019. “A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States”. Int. J. Environ. Res. Public Health 16 (1), 157.

[10] Lovett, D.A., Poots, A.J., Clements, J.T., Green, S.A., Samarasundera, E., Bell, D., 2014. “Using geographical information systems and cartograms as a health service quality improvement tool”. Spatial and Spatio-temporal Epidemiology 10, 67–74.

[11] Mollalo, Abolfazl & Vahedi, Behzad & Rivera, Kiara. (2020). “GIS-based spatial modeling of COVID-19 incidence rate in the continental United States”. Science of The Total Environment. 728. 138884. 10.1016/j.scitotenv.2020.138884.

[12] Wang, Jingyuan, Tang, Ke, Feng, Kai, Lv, Weifeng, 2020. “High Temperature and High Humidity Reduce the Transmission of COVID-19”.

[13] Wu, X., Nethery, R.C., Sabath, B.M., Braun, D., Dominici, F., 2020. “Exposure to Air Pollution and COVID-19 Mortality in the United States”. https://doi.org/10.1101/ 2020.04.05.20054502.

[14] Taghizadeh-Hesary, F., Akbari, H., 2020. “The powerful immune system against powerful COVID-19: a hypothesis”. Preprints 2020, 2020040101. https://doi.org/10.20944/preprints202004.0101.v1.

[15] Udita Retharekar “Crime Pattern Analysis and Prediction using various Geographically Weighted Models” Trinity, 2019

[16] B. Lu, P. Harris, M. Charlton, and C. Brunsdon, “The gwmodel r package: Further topics for exploring spatial heterogeneity using geographically weighted models,” 2013.

[17] B. Chris, F. Stewart, and C. Martin, “Geographically weighted discriminant analysis - brunsdon - 2007 - geographical analysis - wiley online library,” 2019.

[18] S. Nezami and E. Khoramshahi, 2-4 June 2016 Date Added to IEEE Xplore: 23 August 2016 ISBN Information: INSPEC Accession Number: 16251112 DOI: 10.1109/BGC.Geomatics.2016.47Publisher: IEEE Conference Location: Gdansk, Poland.

[19] Indranil Chakraborty, Prasenjit Maity, “COVID-19 outbreak: Migration, effects on society, global environment and prevention”, Science of The Total Environment, Volume 728, 2020, 138882, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138882.

[20] World Health Organization, 2020. “Clinical management of severe acute respiratory infection when novel coronavirus ( nCoV) infection is suspected: interim guidance”, 25 January 2020 (No. WHO/nCoV/Clinical/2020.2). World Health Organization.

[21] Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., & Agha, R. (2020). “The socio-economic implications of the coronavirus pandemic (COVID-19): A review. International journal of surgery (London, England)”, 78, 185–193. https://doi.org/10.1016/j.ijsu.2020.04.018

[22] @Rediff Realtime News, “Prices of agricultural commodities drop 20% post COVID-19 outbreak”, 2020

[23] @National Public Radio, “Oil prices, stocks plunge after Saudi Arabia stuns world with massive discounts”, 2020

[24] @ Markets Insider, “GmbH finanzen net ‘There is no escape’: stocks, oil, and bitcoin plunge as US lawmakers fight over coronavirus rescue package”, 2020

[25] @British Plastics Federation, “Plastics trade body publishes first study of coronavirus impact on UK manufacturing”,

[26] Tanne J.H., Hayasaki E., Zastrow M., Pulla P., Smith P., Rada A.G. BMJ; 2020 Mar 18. “Covid-19: How Doctors and Healthcare Systems Are Tackling Coronavirus Worldwide” .https://www.bmj.com/content/368/bmj.m1090

[27] @World Economic Forum, “This is how coronavirus could affect the travel and tourism industry”, 2020.

[28] B. Chris, F. Stewart, and C. Martin, “Geographically Weighted Regression: The Analysis of Spatially Varying Relationships”. 2019.

[29] Sheng, He & Samara, Petros & Burgers, Jan & Schomaker, Lambert. (2014). Towards Style-Based Dating of Historical Documents. 10.1109/ICFHR.2014.52.

[30] V. Loonis and M.-P. de BELLEFON, “INSEE Editorial - reader ’ s guide,” no. Krugman 1991, 2015

[31] @ R Documentation, “gwr.basic”

[32] @ArcGIS Desktop, “Interpreting GWR results”

[33] @GIS StackExchange, “Interpreting result of Geographically Weighted Regression (GWR)?”, 2016

[34] @Kaggle’ “US Counties: COVID19 + Weather + Socio/Health data”, by John Davis, 2020

[35] Meijer and M. Wessels, “Predictive policing: Review of benefits and drawbacks,” https://doi.org/10.1080/01900692.2019.1575664, 2019.

[36] @ Science AAAS , “The pandemic virus is slowly mutating. But is it getting more dangerous?”, Kai Kupferschmidt, 2020

[37] @MEDPAGE TODAY, “COVID-19 More Deadly Than Cancer Itself?”, Ian Ingram, 2020

[38] @Episerver, “COVID-19’s impact on data privacy, protection and security”

# Appendix

## Abbreviations

| | |
|---|---|
| GWR | Geographically Weighted Regression |
| MGWR | Multiscale GWR |
| LR | Linear Regression |
| BW | BandWidth |
| CV | Cross Validation |
| GIS | Geographic Information System |
| WHO | World Health Organisation |
| ICTV | International Committee on Taxonomy of Viruses |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome Coronavirus 2 |
| COVID-19 | Coronavirus Disease 2019 |
| OLS | Ordinary least squares |
| SLM | Spatial lag model |
| SEM | Spatial error model |
| NUI | The National University of Ireland |
| BPF | British Plastics Federation |
| IT | Information Technology |