

The Evolving Relationship between Covid-19 Cases and Socio-demographic Factors in the Three Lockdowns in London

Chenxi Zhao

CASA0012, MSc Spatial Data Science and Visualisation Dissertation

Supervisor: Dr Huanfa Chen

This dissertation is submitted in part requirement for the
MSc in the Centre for Advanced Spatial Analysis,
Bartlett Faculty of the Built Environment, UCL

Word count: 9,701
2021-07-17

Abstract

With the continuous development of covid-19, it has brought great challenges to people's lives, the stability of the country and the world's economy. In order to better face this long war, more and more dissertations are studying the relationship between covid-19 and different factors, especially at the world and national levels.

This article focuses on the MSOA London range to discuss the changing relationship between covid-19 cases and socio-demographic factors in the three lockdowns in London in 2020. Linear regression is used as the main method, and regression tree and random forest are used as verification methods. We compare the core factors in different models and their changes during different lockdown periods. At the MSOA level, the development of the covid-19 has led to a stronger correlation between new cases per 100,000 population and socio-demographic. And cases have a strong correlation with adult skills score and income score. Then, among the relevant demographic factors, especially the proportion of 65-year-olds and the proportion of bame are also important factors

This article emphasizes the importance of multiple variables, multiple time periods and multiple models for understanding covid-19 cases and has made new discoveries. For example, among all factors, adult skills score has the strongest correlation with cases. However, this factor is rarely discussed in existing papers. At the same time, it also provides information on which variables are related or not related to covid-19 cases under different lockdown periods, and provides a certain basis for covid-19 control and preventive measures.

Declaration

I, Chenxi Zhao, here by declare that this dissertation is all my own original work and that all sources have been acknowledged. It is 9.701 words in length

Contents

1 Introduction	
1.1 Background	7
1.2 Research Question and Objectives	7
2 Literature Review	
2.1 The development and characteristics of the global pandemic	8
2.2 Covid-19 situation in London	9
3.Data Source and Processing	
3.1 Data selection and processing.	11
3.2 Data explanation	12
3.3 Data description.. . . .	14
3.4 Data advantages and limitations	15
4 Methodology	
4.1 Multicollinearity	16
4.2 Linear regression.	16
4.3 Regression Tree and Random Forest.	17
5 Results	
5.1 Covid-19 case rate under different lockdown periods.. . . .	19
5.2 Covid-19 case rate under different regions.. . . .	19
5.2.1 The result of the first lockdown	19
5.2.1.1 Linear regression	19
5.2.1.2 Regression Tree and Random Forest	21
5.2.2 The result of the second lockdown	22
5.2.2.1 Linear regression	22
5.2.2.2 Regression Tree and Random Forest	24
5.2.3 The result of the third lockdown	24
5.2.3.1 Linear regression	24
5.2.3.2 Regression Tree and Random Forest	25
6 Discussion	
6.1 Discussion on the results of the three lockdowns.	27
6.1.1 Comparison of the results of different lockdown periods.	27
6.1.2 Comparison of results of different models.	28
6.1.3 Visualization of important feature	29
6.2 Advantages and limitations	31
7 Conclusion	32

Appendix & Bibliography

List of Figures

Figure 1: Data Framework.	12
Figure 2: covid-19 case rate distribution maps.	19
Figure 3: Residual vs Fitted Plot of case rate at lockdown1.	20
Figure 4: Distribution of core variables	30
Figure 5: Income score and case rate distribution in different periods	31
Figure 6: Distribution of Adult skills score and case rate in different periods	31

List of Tables

Table 1: Lockdown Timeline.	12
Table 2: Description of Variables.	14
Table 3: Description of Data	15
Table 4: VIF of the deleted variable in lockdown1.	20
Table 5: Linear regression comparison in lockdown1	20
Table 6: Linear regression result in lockdown1	21
Table 7: The result of RT and RF in lockdown1	22
Table 8: Linear regression comparison in lockdown2	23
Table 9: Linear regression result in lockdown2.	23
Table 10: The result of RT and RF in lockdown2	24
Table 11: Linear regression comparison in lockdown3.	25
Table 12: Linear regression result in lockdown3	25
Table 13: The result of RT and RF in lockdown3	26
Table 14: Comparison of linear regression results.	28
Table 15: Comparison of random forest results.	28
Table 16: Comparison of LR and RF result.	29

Abbreviations

Abbreviations	Description
LR	linear regression
RT	regression tree
RF	random forest
Case rate	the new cases per 100,000 population
Cases1	the new cases per 100,000 population during the first lockdown
Cases2	the new cases per 100,000 population during the second lockdown
Cases3	the new cases per 100,000 population during the third lockdown
NUK	country of birth is not UK
O65	proportion of people over 65
Bame	all bame (Black, Asian and minority ethnic) proportion
Income	income score (rate)
HP	median house price
OO	proportion of owned outright
OML	proportion of owned with a mortgage or loan
SR	proportion of social rented
PR	proportion of private rented
NR	proportion of house old spaces with no usual residents
Geo	geographical barriers sub-domain score
Density	population density
IMD	index of multiple deprivation (IMD)
Health	health deprivation and disability score
AS	adult skills sub-domain score
Employment	employment score (rate)
Education	education, skills and training score
Crime	crime score
Environment	living environment score

Chapter 1 Introduction

1.1 Background

Covid-19 as a global epidemic, governments in various countries and international organizations have made a lot of efforts. At the beginning, almost all the propaganda was that we would fight against covid-19 together. However, the fact is that certain people or areas are more susceptible to the virus. For example, 'Covid triangle' (three London boroughs) has received a lot of attention due to the high infection and death rate. In Barking and Dagenham, 1 in 16 people was reported to be infected (Andrew Gregory, 2021). 'The virus does discriminate' as well as the adverse impacts of trying to control the pandemic and the economic consequences. Recently published data suggests that the most deprived areas of England have Twice the rate of deaths involving covid-19 than the most affluent (Palmer, 2021).

In addition to inequalities in spatial areas, we found that death rates were influenced by different factors at different stages of development over time. For example, at first it was thought that the main influencing factor was population density, so many governments required large-scale quarantine and wearing of masks (Cheng et al., 2020). Later it was found that there was also a strong correlation with some socio-demographic and social factors. This inequality of covid-19 cases should be worthy of attention and in-depth research, so that the society and government will continue to work hard to reduce the inequality in population health. This dissertation conducts related research in this context.

1.2 Research Question and Objectives

There are many studies on covid-19, but most of them are based on countries and regions. There are very few analyses that have detailed studies on various areas and cities in the country. However, only by understanding the relevant factors in these small areas can we make a more correct judgment on covid-19 and better reflect this global epidemic.

In this context, this dissertation will focus on London Middle Layer Super Output Areas (hereinafter referred to as MSOA) level areas and try to discuss the inequality and causes of covid-19 cases. Specifically to discuss the following two issues:

What is the relationship between London's covid-19 case rate and socio-demographic factors?

How is this relationship different in the three lockdowns in 2020?

Chapter 2 Literature Review

2.1 The development and characteristics of the global covid-19

Since the outbreak of the covid-19 in Wuhan, China, the world's perception of covid-19 has changed a lot. Including people's awareness of covid-19 at the beginning, with the development of covid-19, some trend characteristics have been discovered.

At first it was thought that the virus only occurred in China, but within a few weeks it spread to 213 countries and regions, and more than 5.5 million people were infected (Worldometer, 2020). As of 2020, the number of global cases will exceed 79.2 million, and the number of deaths will exceed 1.7 million (WHO). People realize that this disease is not limited to some regions, but affects the entire world.

It is the biggest disaster that has attacked mankind on a global scale since modern times (WHO, 2020), which has brought many difficult decisions to countries and regional governments as well as schools, businesses and families. How should resource-scarce countries allocate resources? How does the government guarantee the people's health and basic life? How do companies face continuous losses? As you can see, this impact is not only reflected in the infection rate and death rate, but also has a great impact on the social economy and even the social structure. From a localized outbreak to the blockade of the entire country, the world is facing the worst economic recession since the Great Depression. Globally, 155 million full-time jobs were lost in the first quarter of 2020, rising to 400 million in the second quarter, with lower- and middle-income nations bearing the brunt of the loss. A further 71 to 100 million people are being pushed into severe poverty as a result of the epidemic (IMF, 2020). Even according to simulations, the Human Development Index (HDI) is on the verge of a dramatic and historic fall, undermining six years of development (How COVID-19 is changing the world: a statistical perspective (Volume II) | Population Division, 2021).

In addition to the above-mentioned impacts on various aspects of society, there are some other findings from the perspective of cases. First, there are obvious regional differences. As mentioned above, the case rate in some specific countries or specific regions is surprisingly high. Even covid-19 is described as a heat-seeking missile speeding toward the most vulnerable in society. This means that rich countries or regions and poor countries or regions are experiencing two different epidemics (Schellekens and Sourrouille, 2021).

Secondly, as a long-term epidemic, covid-19 has changed over time and cannot be ignored. In the beginning, according to the data as of May 23, 2020, the paper showed that the cases and death rate were mainly concentrated in high-income countries. Although developing countries accounted for 85% of the global population, the death toll from covid-19 accounted for only 21%. But the paper argues that this phenomenon of excessive bias towards rich countries is inconsistent with demographics. Models

based on infectiousness and mortality indicate that the share of developing countries in global deaths will triple (from 21% to 69%) (Schellekens and Sourrouille, 2021). Facts show that this conjecture is correct. A report published on May 27, 2021 shows that the cumulative mortality rate in developing countries exceeds 50%. This means As time goes by, covid-19 has become an epidemic in developing countries (Schellekens, 2021).

Finally, there are findings related to some external factors, such as: the spread of covid-19 in the global age distribution is relatively slow, but it has a great relationship with demographics. The impact of the environment on the cases is not great Pan et al., 2021).

2.2 Covid-19 situation in London

As of 2020, Europe's cumulative cases and deaths rank second in the world, accounting for 31% of the world's total (Eurostat). The infection rate in the United Kingdom has always been of concern, because the cumulative number of cases in the United Kingdom is higher than in other European countries for most of the time. This article selects the capital London as the research sample to discuss some issues of case rate and social demographics.

London is not only a very representative area, but also a region with its own characteristics. First of all, the epidemic in London is the most serious, with the highest number of confirmed cases in a single day exceeding 40,000 in 2020. And in the first three months of the covid-19, London had the highest proportion of deaths, accounting for 1/5 of the total deaths in England (ONS). Studying an area with high cases and mortality rate is of great reference value to a certain extent.

Second, London is similar to the global case rate mentioned above. For example, the phenomenon of regional inequality is very serious. Some papers believe that this inequality is mainly due to wealth inequality, because three were in the most deprived quintile of the IMD, two were in the least deprived quintile (ONS). At the same time, A report published by WPI Economics considered for the first time the relationship between the mortality rate in London and the characteristics of the community (poverty, race, and age). According to this paper, we can see that areas with the following characteristics have higher infection rates and mortality rates than other areas: poor, elderly, Asians, and nursing home residents.

Third, there are some characteristics of London that are worthy of our investigation. According to ONS data, the majority of cases from covid-19 are people over 65 years of age. However, the average age of cases in London is younger than that in other parts of England (Weir and Oakley, 2020). This is why in 2020, the UK has almost the highest excess case rate under 65 years (ONS) in Europe. Finally, some papers believe that with the gradual development of the epidemic, poverty and population density are not as strong as they were at the beginning during this long period, and the correlation between obesity and air pollution must also be taken into account. Although their correlation coefficients are not so strong, considering these variables alone seems to be able to

predict the case rate that affects covid-19 (Bray, Gibson and White, 2020).

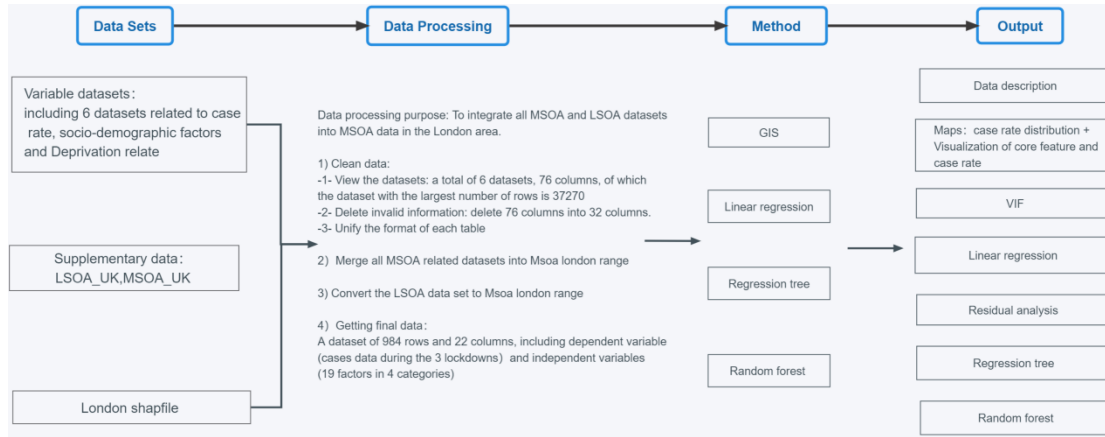
Chapter 3 Data Source and Processing

3.1 Data selection and processing

The data samples are obtained from three public non-profit government official websites, namely London Datastore, ONS and GOV.UK. We think it can be regarded as a reliable source of data. For variable data, 6 datasets of case rate, social demographic factors and privacy related are selected respectively. The dataset of sociodemographic factors includes demographic, economic and spatial aspects. Deprivation related datasets include Index of Multiple Deprivation (IMD) and deprivation index scores in different domains. In terms of time, the data of covid-19 uses the data that was blocked three times in 2020. And the other data uses the latest version of the public data, mainly the data of 2019 and 2011. The data sample is 983 areas of London MSOA level. The dimensions of the data are more complex, including MSOA and LSOA data from London, England and the United Kingdom. Need to select the London data in datasets and merge it into MSOA data.

In terms of data processing, the following four stages are mainly carried out. In the first step, find the relevant datasets including the 6 datasets of the above-mentioned related variables. The MSOA and LSOA correspondence table, and the London shapfile. In the second step, the final dataset is obtained through data processing. And the 19 independent variables are divided into 4 categories according to their characteristics. The specific data processing steps are as follows: firstly, each dataset is cleaned separately. Only useful variables are retained and the format is unified. The original 6 datasets have a total of 76 columns, and the dataset with the largest number of rows is 37270. After processing, 32 columns of useful data are left. Secondly, only the London part of all MSOA data is extracted. Then use the affiliation relationship between LSOA and MSOA to convert LSOA related data into MSOA.. Finally, merge the above data together and remove duplicate columns. A dataset of 984 rows and 22 columns is obtained, which includes 19 independent variables and 3 dependent variables during the 3 lockdown periods. The third and fourth steps of data processing are to process the verification data through different methods, and finally get the data results and visualization charts.

Figure 1: Data Framework



3.2 Data explanation

The dependent variable is the new cases by specimen date rolling rate of the weekly dimension in 2020. In order to compare areas or population groupings of different sizes, rates are determined. All current rates on GOV.UK are crude rates expressed per 100,000 population, which means that the count (for example, cases or deaths) is divided by the denominator population and then multiplied by 100,000, with no adjustments for additional factors (GOV.UK).

In the MSOA data of covid-19, the smallest unit of time is weekly. Meanwhile, the data selected in this article start with Sunday as the new week. In the case of a non-full week, more than half of the days will be calculated as a week. In other words, in the case of not a full week, if the total number of days between the start date and the start week and the number of days between the end date and the end week is greater than or equal to 4, the start week starts with the week of the start date, and the end week is the complete week. If the total is greater than or equal to 8 days, the start week starts with the week of the start date, and the end week ends with the week of the end date. The time frame studied in this article is the situation of three lockdowns in 2020. The start and end time of each lockdown comes from government and media releases. The specific dates are as follows:

Table1: Lockdown Timeline

Description	Start Date (2020)	End Date (2020)
First national lockdown	26th March	23rd June
Second national lockdown	5th November	2nd December
Third national lockdown	21st December	31st December

In terms of independent variables, 19 factors in 4 categories are selected.

1) Demography related factors (X1: NUK, X2: O65, X3: Bame)

X1: NUK is country of birth is not UK. This variable is added because it determines whether people can get the same social rights and benefits. Meanwhile, in covid-19, we have seen that the infection is related to certain groups of people, especially bame and the elderly (Kakkar, Dunphy and Raza, 2021). Approximately 90% of coronavirus deaths occur in those over the age of 65 (Van Rens and Oswald, 2021).

Therefore, proportion of people over 65 (X2: O65) and all bame (Black, Asian and minority ethnic) proportion (X3: Bame) are also considered as variables

- 2) Economy& house related factors (X4: Income, X5: HP, X6: OO, X7: OML, X8: SR, X9: PR, X10: NR) are mainly divided into three parts: income, house price and tenure.

X4: Income is a proxy for the percentage of the population that suffers from deprivation as a result of low income. The term "low income" is used to refer to both those who are unemployed and those who are employed but earn a low wage. X5: HP uses the median, because the price difference in London is larger. The median is more representative of the overall level than the average (Laerd Statistics, 2021). Housing related expenditure is an important part of people's life pressure, so here is a breakdown of relevant data for housing scenarios such as renting a house and buying a house.

- 3) Spatial related factors(X11: Geo, X12: Density).

X11: Geo is the Geographical Barriers sub-domain, which measures the comprehensive index of Road distance to a post office, a primary school, a general store or supermarket and a GP surge y. X12: Density refers to population density, divided by the total population of the same area by the total area.

- 4) Deprivation related factors (X13: IMD, X14: Health, X15: AS, X16: Employment, X17: Education, X18: Crime and X19: Environment).

The Multiple Deprivation Index (IMD) is an overall relative measure of deprivation, which is composed of seven areas of deprivation (such as income, education, etc.) with different weights. It can comprehensively represent the deprivation situation in a region.

The Health Deprivation and Disability Domain (X14: Health) assesses the likelihood of premature death and the impact of poor physical or mental health on one's quality of life. The domain collects data on sickness, disability, and premature mortality, but not on behavioral or environmental factors that may be predictive of future health deprivation.

The Adult Skills sub-domain (X15: AS) is a non-overlapping count of adult skills and English language proficiency. The targets are women aged 25-59 and men aged 25-64.

The Employment Deprivation Domain (X16: Employment) quantifies the percentage of working-age residents who are involuntarily excluded from the labor market in a certain location. This category comprises individuals who wish to work but are prevented from doing so owing to unemployment, illness or disability, or caring duties.

The Education, Skills and Training Domain (X17: Education) comprehensively represents the comprehensive educational skills from children to adults in a certain area. Including reference to the absenteeism rate of middle school students, the situation of receiving higher education and so on.

X18: Crime assesses the risk of victimisation, including assault, burglary, theft, and criminal damage, at the local level.

The Deprivation of a Living Environment Domain (X19: Environment) quantifies the quality of the local environment. Housing quality, air quality, and road traffic accidents are all factors to consider.

Table2: Description of Variables

Dimensions	Name	Variable Description	Unit	Data Source
Demography	Y1: Cases1	the new cases per 100,000 population during the first lockdown	Percentage	GOV.UK, 2020
	Y2: Cases2	the new cases per 100,000 population during the second lockdown	Percentage	GOV.UK,2020
	Y3: Cases3	the new cases per 100,000 population during the third lockdown	Percentage	GOV.UK,2020
	X1: NUK	country of birth is not UK	Percentage	London Datastore,2011
	X2: O65	proportion of people over 65	Percentage	London Datastore,2019
	X3:Bame	all bame (Black, Asian and minority ethnic) proportion	Percentage	London Datastore,2011
	X4: Income	income score (rate)	Percentage	GOV.UK, 2019
	X5: HP	median house price	Pounds	ONS, 2019
	X6: OO	owned outright	Percentage	London Datastore,2011
	X7: OML	owned with a mortgage or loan	Percentage	London Datastore,2011
Economy& house	X8: SR	social rented	Percentage	London Datastore,2011
	X9: PR	private rented	Percentage	London Datastore,2011
	X10: NR	house old spaces with no usual residents	Percentage	London Datastore,2011
	X11: Geo	geographical barriers sub-domain	Score	GOV.UK, 2019
	X12: Density	population density	Value	London Datastore,2019
	X13: IMD	index of multiple deprivation (IMD)	Score	GOV.UK, 2019
	X14: Health	health deprivation and disability	Score	GOV.UK, 2019
	X15: AS	adult skills sub-domain score	Score	GOV.UK, 2019
	X16: Employment	employment score (rate)	Percentage	GOV.UK, 2019
	X17: Education	education, skills and training score	Score	GOV.UK, 2019
Deprivation related	X18: Crime	crime score	Score	GOV.UK, 2019
	X19: Environment	living environment score	Score	GOV.UK, 2019

3.3 Data Description

The overall data distribution is reasonable and no additional processing is required. However, there is a large gap between the maximum and minimum values of some variables, such as the case rate of three lockdowns and X12: Density. It shows that the data distribution of these variables is relatively scattered, and the degree of dispersion is relatively large. This means that there is a big gap between the case rate and population density of different regions. On the other hand, the case rate of the three lockdowns, the median of X5: HP and X12: Density is significantly smaller than the average. This shows that the difference between the top 50% and the bottom 50% of the data is relatively large, and more than half of the data are below the average. While, these data are true and reliable, they just show the characteristics of different data, so no additional processing is done.

Table3: Description of Data

Dimensions	Name	Count	Std	Mean	Median	Min	Max
Demography	Y1: Cases1	972	9020.14	578.86	265.30	33.70	281613.70
	Y2: Cases2	983	32480.71	2072.90	1001.30	388.20	1019866.40
	Y3: Cases3	983	49572.99	3163.75	1535.10	503.40	1556563.60
	X1: NUK	983	13.83	36.21	37.17	4.39	69.23
	X2: O65	983	0.05	0.12	0.12	0.02	0.28
	X3: Bame	983	0.19	0.39	0.37	0.04	0.94
Economy& house	X4: Income	983	0.06	0.14	0.13	0.01	0.34
	X5: HP	983	243318.41	526463.37	465000	214000	2850000.00
	X6: OO	983	10.91	21.68	20.60	2.20	53.70
	X7: OML	983	10.22	27.66	28.30	5.90	53.70
	X8: SR	983	16.55	23.68	19.90	0.50	74.10
	X9: PR	983	10.62	24.44	24.30	3.40	57.70
Spatial	X10: NR	983	2.77	3.32	2.50	0.50	28.60
	X11: Geo	983	0.49	-0.60	-0.61	-2.04	0.69
	X12: Density	983	54.14	91.55	79.85	2.89	286.43
Deprivation related	X13: IMD	983	9.45	21.56	21.20	4.06	50.42
	X14: Health	983	0.60	-0.38	-0.29	-3.04	1.04
	X15: AS	983	0.10	0.26	0.26	0.05	0.49
	X16: Employment	983	0.04	0.09	0.08	0.01	0.24
	X17: Education	983	8.72	13.10	11.72	0.21	44.75
	X18: Crime	983	0.44	0.26	0.30	-1.66	1.55
	X19: Environment	983	9.02	29.18	28.95	7.66	61.35

3.4 Data advantages and limitations

The data set used in this article has the following advantages and disadvantages. First of all, the data comes from the government's public data official website, and the security, authenticity and source of the data are guaranteed to a certain extent. Secondly, the data dimensions of 100 make subsequent results more in line with actual conditions. Because MSOA data has more samples than other dimensions such as ward and borough. The outcomes of aggregate data analysis are well known to be reliant on the size and shape of the zones utilised to convey the data (Lloyd, 2015). Therefore, MSOA level data can reflect the real situation in a small area. But it is also because the data is more detailed. Some relevant data will not be disclosed, such as the number of mortality in the weekly, population movements, etc. Because this kind of data involves privacy, after communicating with the official email, it is not recommended as a graduate student for academic research. Therefore, the privacy risk data is not involved in this article. At the same time, because of the lack of the above data, the research direction of this paper will not get the maximum fitting result.

Chapter 4 Methodology

This article mainly uses linear regression to study the relationship between covid-19 and social demographic factors, and uses regression tree and random forest for testing and prediction. By comparing the advantages and disadvantages of different methods, we critically look at the results of different methods and models.

In linear regression, VIF is used to deal with multicollinearity, and the confidence of P-value is discussed. Finally residual analysis is performed. In order to achieve a better fitting effect, continuously adjust the range of different indicators and remove outliers in this process.

In the two models of regression tree and random forest, we try to continuously adjust the hyperparameters by using cross-validation and GridSearchCV methods to optimize the model. In this process, the depth and number of the tree are adjusted through max_depth and n_estimators. At the same time, because the dataset has a small sample size, internal nodes such as min_samples_split and min_samples_leaf are not divided.

4.1 Multicollinearity

Multicollinearity refers to the existence of linear correlation between independent variables, that means the relationship between independent variables is strong. Multicollinearity will lead to instability in the estimation of regression coefficients and intercept coefficients, which leads to instability of the model. Therefore, when the multicollinearity is serious, appropriate methods should be adopted to adjust.

In an ordinary least squares regression analysis, **Variance Inflation Factor (VIF)** quantifies the severity of multicollinearity. It represents the ratio of the variance of the regression coefficient estimator to the variance when the independent variables are assumed to be non-linearly correlated. The specific formula is as follows:

$$VIF_k = \frac{1}{1 - R_k^2} \quad (1)$$

The value of VIF is greater than 1. The closer the VIF value is to 1, the lighter the multicollinearity, and vice versa. If the VIF of the independent variable is displayed in an appropriate range, the problem of multicollinearity in the model can be ignored (Voltes-Dorta and Schez-Medina, 2020). Variables with a vif score greater than 5 are considered to have a strong correlation, there may be serious collinearity (Shrestha, 2020). Therefore, this article deletes these variables with vif greater than 5.

4.2 Linear regression

Linear regression is a statistical method for modeling the connection between a scalar response and one or more explanatory variables (also known as dependent and independent variables). Multiple linear regression is a specific example of generic linear models with only one dependent variable, and is a generalization of simple linear regression with more than one independent variable. Meanwhile, multiple linear regression is a model that uses the best combination of two or more independent

variables to predict or estimate dependent variables (MAXWELL, 1975). The equation is as follows: (when n is the number of observation, k is the number of independent variables):

$$y = X\beta + \varepsilon \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

OLS is used to estimate model parameters since it minimizes the sum of squared errors. As a result, the coefficient matrix can be solved as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y = Hy \quad (3)$$

After getting the regression equation, we have to test the significance of the regression equation. The significance test here mainly includes four parts. The first is the **F-test**, which is to test whether all independent variables have a significant effect on dependent variable as a whole. And it is a variance test of the regression model as a whole. The second is the significance test of the coefficient of a single variable by the **T-test**. At the same time, the **P-value** is a measure used for T-test and F-test. P-value less than 0.05 means that the variable rejects the null hypothesis and has a significant effect on dependent variable. The third is to judge the goodness of fit through **The coefficient of determination** (R^2) :

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \quad (4)$$

The value is between 0 and 1. The closer to 1, the better the effect of regression fitting, and the closer to 0, the worse the effect. However, the R^2 will increase due to the increase of variables, so the adjusted R^2 is introduced. The adjusted R^2 is useful for comparing a model with and without a particular variable in order to determine whether or not the variable improves the model. The formula is as follows:

$$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \quad (5)$$

Finally, use **residual analysis** to check the linear model. The residual is the difference between the observed value and the fitted value, that is, the difference between the actual observed value and the regression estimate. Normality test, linearity test, independence test, homoscedasticity test and other methods to residual analysis. And remove outliers whose absolute value of standardized residual is greater than 3. Because these values are considered "outliers", they will affect the final results and accuracy of the model (Gray and Woodall, 1994). If there are no obvious outliers, all data will be kept, because they will not affect the r-square too much.

4.3 Regression Tree and Random Forest

The **regression tree(RT)** is traversed by traversing the branches of the tree and selecting the next branch according to the decision of the node. Regression Tree Induction takes a set of training examples as input, determines which attributes are most suitable for segmentation, and divides the data set, and loops on the divided data set until all training examples are classified, and the task ends.

But the regression tree has a tendency to over-fit, which means that new data cannot be

used. Therefore, **random forest (RF)** will be used for optimization. It is a collection of simple regression tree, and the input vector runs on multiple regression tree. For the regression problem involved in this article, the output values of all regression trees are average.

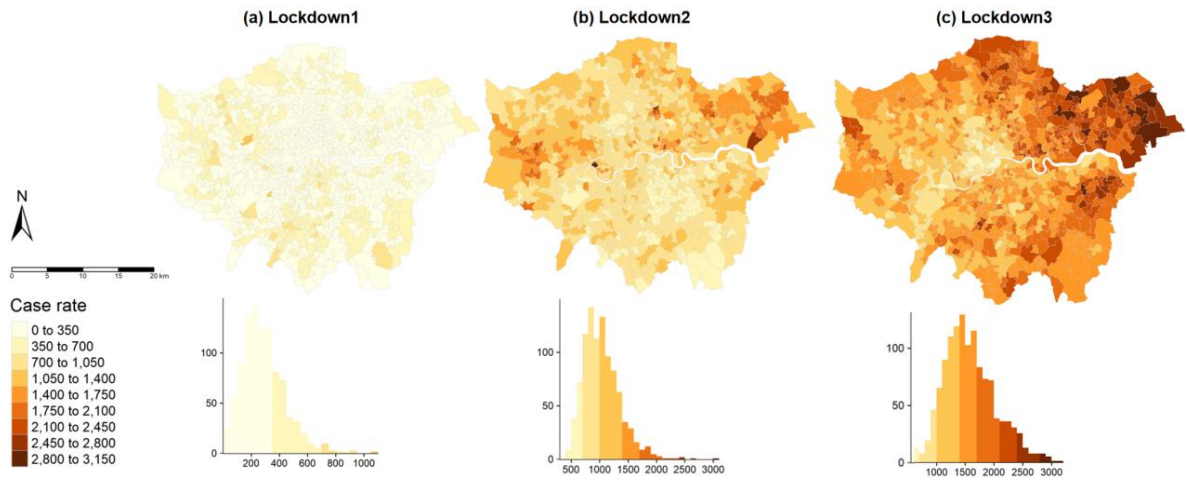
In general, the regression tree constructs a tree structure by subdividing predictors. While, random forest generates a large number of trees at random and then aggregates their forecasts.

Chapter 5 Results

5.1 Covid-19 case rate under different lockdown periods

The following three pictures respectively show the distribution of infection rates during the first, second and third London lockdown. The regional division standard is the MSOA mentioned above. It can be seen that over time, the overall case rate in London is gradually increasing. At the same time, the infection rate in London's North East is getting worse. Specifically, the case rate is mainly concentrated at around 200 in lockdown1, around 1000 in lockdown2, and around 1500 in lockdown3. Therefore, although the case rate has become more and more serious over time, the growth rate has slowed down from lockdown2.

Figure 2: covid-19 case rate distribution maps



5.2 Covid-19 case rate under different regions

The following will show the results of linear regression, regression tree and random forest for each lockdown. As the main method of the article, linear regression will be explained in detail. The model will be fitted in various stages of removing multicollinearity, outliers, residual analysis, etc., to obtain the final model. Regression tree and random forest use hyperparameters and cross-validation methods to continuously optimize tree branches and parameters to get the final model. Compared with regression tree, random forest can be less affected by outliers and can reduce the possibility of overfitting. Therefore, we focus on the results of random forest.

Because each lockdown data will be processed with the same steps as above, there is a certain degree of repetition. Thus, some details of the display will be omitted in the lockdown2 and lockdown3 parts.

5.2.1 The result of the first lockdown

5.2.1.1 Linear regression

In linear regression, the VIF method is first used to sequentially delete independent variables with high multicollinearity. It is found that the VIF of 7 independent variables such as proportion of social rented, IMD is higher than 5, indicating that the

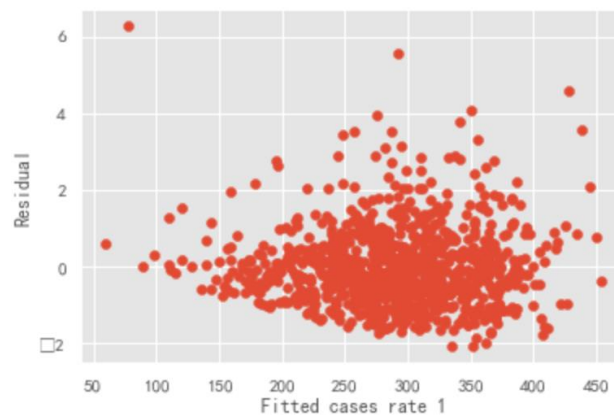
multicollinearity is very strong and will affect the results of the model. The 7 independent variables whose VIF is higher than 5 are deleted, and 12 independent variables remain. The following table shows the VIF value of the deleted variables:

Table 4: VIF of the deleted variable in lockdown1

Variable	X7: SR	X12: IMD	X16: Employment	X15:AS	X1: NUK	X5: OO	X6: OML
VIF	266.163	61.070	24.708	16.912	10.843	7.618	5.195

Secondly, the fitting of linear regression was performed twice. The first time is the fitting result after removing multicollinearity, and the second time is after removing outliers. Among them, after the first fitting, a residual analysis was carried out. It is found that the overall residual distribution is relatively uniform, and there is no obvious non-linear relationship. However, there are some separate group values, and the final fitting is performed after removing the outliers that will affect the results and accuracy of the model. Normally, the standardized residuals with absolute values greater than 3 in the observations are considered "outliers" (Gray and Woodall, 1994). Through figure3, we can see that there are many outliers that exceed the standard residual by 3 times, so we delete it. The following is the figure of the residual analysis

Figure 3: Residual vs Fitted Plot of case rate at lockdown1



After deleting the outliers, refit the linear regression. On the whole, we can see from the table that the sample size after the processing of collinearity and residuals is 959 (13 outliers are deleted), and the independent variable is 12 (7 independent variables with high multicollinearity are deleted). And the final regression, whether it is the result of R-squared or Adj.R-squared is better than the previous fitting situation, which are 0.186 and 0.176 respectively. F-statistic is within a reasonable range, Prob (F-statistic) less than 0.05 indicates that the null hypothesis is rejected and the model is significant.

Table 5: Linear regression comparison in lockdown1

Regression	Before removing outliers	After removing outliers
Number of samples	972	959
Number of independent variables	12	12
R-squared	0.154	0.186
Adj.R-squared	0.144	0.176
F-statistic	14.59	18.04
Prob(F-statistic)	2.42e-28	2.27e-35

From the perspective of P-value, the values of income score and population density are greater than 0.1, indicating that these two independent variables are not very explanatory for the infection rate during lockdown1. The P-values of proportion of house old spaces with no usual residents, health score and education score are between 0.05 and 0.1, indicating at least a 90% confidence level, which is highly significant. The P-values of other independent variables are all less than 0.05, indicating at least 95% confidence level, which means a strong correlation.

From the perspective of coefficients, when all the independent variables take 0, the predicted value of the dependent variable is 297.174. Demographic related factors can affect the case rate more than other factors. Specifically, the coefficients worth paying attention to are proportion of people over 65 (256.1346), proportion of bame (254.0734), income score (-68.8467), Geo score (48.621) and crime score (-34.0174). This shows unitproportion of people over 65, proportion of bame and Geo score by one unit may increase the case rate by 256.134, 254.0734 and 48.621 units. Income score and crime score are negatively correlated. When other independent variables remain the same, with an 1% increase in the proportion of residents aged over 65 and proportion of bame, the the new cases per 100,000 population would increase by 256.13% and 254.07%. The coefficient of house price is extremely low, indicating that the impact on the dependent variable is small. The specific regression results are as follows:

Table 6: Linear regression result in lockdown1

Dimensions	Name	Coefficient	Std error	T-value	P-value
Demography	(Intercept)	297.1742	35.843	8.291	0
	X2: O65	256.1346	113.505	2.257	0.024
	X3:Bame	254.0734	29.727	8.547	0
Economy& house	X4: Income	-68.8467	165.173	-0.417	0.677
	X5: HP	-9.87e-05	2.79e-05	-3.541	0
	X9: PR	-3.3659	0.633	-5.317	0
	X10: NR	3.7471	2.18	1.719	0.086
Spatial	X11: Geo	48.621	11.726	4.147	0
	X12: Density	-0.0106	0.095	-0.112	0.911
Deprivation related	X14: Health	25.8371	14.374	1.798	0.073
	X17: Education	-1.5223	0.807	-1.888	0.059
	X18: Crime	-34.0174	14.601	-2.33	0.02
	X19: Environment	1.7739	0.792	2.239	0.025

5.2.1.2 Regression Tree and Random Forest

Use other methods of regression tree and random forest to verify and predict the above results. First, the data is divided into training, verification, and test sets, where the proportion of data is 7: 1.5: 1.5 respectively. Then through cross-validation and GridSearchCV methods to continuously adjust the hyperparameters. In order to adjust the number of iterations and depth of trees to a better value, two parameters are mainly adjusted namely depth and n_estimators. Among them, because the data has more features, max_depth is set to 8. At the same time, in order to ensure the unity of the

dataset and the reliability of the results, the datasets of these two methods are the same as those in the final regression. That means using the dataset after removing the outliers. Follow-up lockdown 2 and 3 also follow the same idea and operation method.

In general, these two results are slightly better than linear regression. From the optimization results of the regression tree model, we can see that when best_depth is 4, the best prediction result is 0.2043. At the same time, when n_estimators is 150, the random forest has the best prediction result with a score of 0.28. The score here refers to the prediction score, which can be understood as the prediction accuracy that can be obtained through the existing dataset.

From the point of view of feature importance, the importance of 9 features such as NUK and income score in the regression tree is 0, while each feature in the random forest has a certain importance. At the same time, in the regression tree, the most important feature is the adult skills score of 0.352 related to the economy & house. The importance of adult skills score in the random forest is also the highest at 0.1127.

Table 7: The result of RT and RF in lockdown1

Dimensions		Regression Tree		Random Forest	
Demography	Core index	best_depth: 4	Core index	n_estimators: 150	
	Score	0.2043	Score	0.28	
	X1: NUK	0	X1: NUK	0.04202	
	X2: O65	0.04703	X2: O65	0.06535	
Economy& house	X3: Bame	0.12319	X3: Bame	0.0853	
	X4: Income	0	X4: Income	0.02125	
	X5: HP	0.05972	X5: HP	0.07466	
	X6: OO	0.04864	X6: OO	0.04179	
Spatial	X7: OML	0	X7: OML	0.04308	
	X8: SR	0.07865	X8: SR	0.03518	
	X9: PR	0	X9: PR	0.05251	
	X10: NR	0.02313	X10: NR	0.05049	
Deprivation related	X11: Geo	0.15365	X11: Geo	0.06849	
	X12: Density	0	X12: Density	0.05506	
	X13: IMD	0	X13: IMD	0.02563	
	X14: Health	0	X14: Health	0.0422	
	X15: AS	0.35213	X15: AS	0.11277	
	X16: Employment	0	X16: Employment	0.0304	
	X17: Education	0	X17: Education	0.05164	
	X18: Crime	0.08772	X18: Crime	0.05102	
	X19: Environment	0.02614	X19: Environment	0.05115	

5.2.2 The result of the second lockdown

5.2.2.1 Linear regression

Since the process of processing data will be consistent with that in 5.2.1, some details will be omitted to avoid repetition. In the linear regression within the time range of lockdown2, two linear regression fittings were also performed. The first time is the fitting after removing the multicollinearity variable. The independent variables greater

than 5 are deleted here, and it is found that the independent variables that need to be deleted are consistent with those in the lockdown1 model. After processing 19 variables, the remaining 12 independent variables. Then remove the outliers in the residual analysis, which refers to the observations with the absolute value of the standard residuals greater than 3. After the final regression fitting, the results are shown in the table:

Table 8: Linear regression comparison in lockdown2

Regression	Before removing outliers	After removing outliers
Number of samples	983	974
Number of independent variables	12	12
R-squared	0.163	0.164
Adj.R-squared	0.153	0.153
F-statistic	15.74	15.65
Prob(F-statistic)	9.56e-31	1.50e-30

The sample size of the final model was 974 (9 outliers were deleted), and the independent variable was 12 (7 independent variables with high multicollinearity were deleted). And the final regression effect is better than the model before residual analysis, the results of R-squared and Adj.R-squared are 0.164 and 0.153. F-statistic is within a reasonable range, Prob (F-statistic) less than 0.05 indicates that the null hypothesis is rejected and the model is significant.

In terms of P-value, the overall effect is not as good as the model in lockdown1. Among them, the P-value of 7 independent variables such as house price, percentage of private rented is greater than 0.1, indicating that these independent variables are not very explanatory for the case rate in lockdown2 period. The P-values of the other five independent variables are all less than 0.05, which has a strong correlation.

From the perspective of coefficients, when all the independent variables take 0, the predicted value of the dependent variable is 1050.436. The coefficient of income score is surprisingly high at -2677.65. This means that with other independent variables remaining constant, with an 1% increase in income score, the the new cases per 100,000 population would decrease by 2677.65% at lockdown2. At the same time, the variables percentage of bame and proportion of people over 65 related to people also affect the infection rate more than other characteristics, with coefficients of 517.35 and 166.54, respectively. And it is found that the coefficient of house price is also very small, which is the same as lockdown1, indicating that house prices in these two periods have a small impact on the infection rate. The specific regression results are as follows:

Table 9: Linear regression result in lockdown2

Dimensions	Name	Coefficient	Std error	T-value	P-value
Demography	(Intercept)	1050.436	77.385	13.574	0
	X2: O65	166.5387	246.626	0.675	0.5
	X3:Bame	517.3545	64.656	8.002	0
Economy& house	X4: Income	-2677.65	361.096	-7.415	0
	X5: HP	2.06E-05	6.00E-05	0.344	0.731
	X9: PR	-1.2581	1.372	-0.917	0.359

	X10: NR	-1.8452	4.703	-0.392	0.695
Spatial	X11: Geo	-56.9937	25.387	-2.245	0.025
	X12: Density	-0.4595	0.204	-2.249	0.025
Deprivation related	X14: Health	36.0244	30.815	1.169	0.243
	X17: Education	15.4932	1.812	8.549	0
	X18: Crime	13.5358	31.178	0.434	0.664
	X19: Environment	-1.4238	1.712	-0.831	0.406

5.2.2.2 Regression Tree and Random Forest

Use the same methods and operations as 5.2.1 to fit regression tree and random forest. The specific steps will not be repeated. Unlike lockdown1, the results of RT are far worse than the fit of LR, indicating that this method is not very suitable. The situation of RF is better than LR. When n_estimators is 150, the score is 0.231. As in lockdown1, the most important feature is also adult skills score, where the importance is 0.18. And it is found that the importance of economy& house related features is generally high.

Table 10: The result of RT and RF in lockdown2

Dimensions		Regression Tree		Random Forest	
	Core index	best_depth: 2	Core index	n_estimators: 150	
	Score	0.0688	Score	0.2312	
Demography	X1: NUK	0	X1: NUK	0.06176	
	X2: O65	0	X2: O65	0.04718	
	X3: Bame	0.09375	X3: Bame	0.09897	
Economy& house	X4: Income	0	X4: Income	0.02242	
	X5: HP	0	X5: HP	0.04671	
	X6: OO	0	X6: OO	0.04262	
	X7: OML	0	X7: OML	0.03331	
	X8: SR	0.32001	X8: SR	0.06883	
	X9: PR	0	X9: PR	0.04195	
	X10: NR	0	X10: NR	0.02857	
Spatial	X11: Geo	0	X11: Geo	0.04785	
	X12: Density	0	X12: Density	0.04962	
Deprivation related	X13: IMD	0	X13: IMD	0.0241	
	X14: Health	0	X14: Health	0.03317	
	X15: AS	0.58624	X15: AS	0.18047	
	X16: Employment	0	X16: Employment	0.04554	
	X17: Education	0	X17: Education	0.0489	
	X18: Crime	0	X18: Crime	0.04179	
	X19: Environment	0	X19: Environment	0.03624	

5.2.3 The result of the third lockdown

5.2.3.1 Linear regression

In the two linear regression fittings in lockdown3, the final fitting result is slightly better than the first one. In this process, the same as lockdown1 and 2 processing standards, still delete the data with vif greater than 5 and the absolute value of the standardized residual greater than 3. In the last fitting, the sample size was 979 (removed 4 outliers), and the independent variable was 12 (removed 7 independent variables with high

multicollinearity). And the goodness of fit is slightly better than that before processing the residuals, the r-square is 0.302 and the model is significant. The result is shown in the table:

Table 11: Linear regression comparison in lockdown3

Regression	Before removing outliers	After removing outliers
Number of samples	983	979
Number of independent variables	12	12
R-squared	0.301	0.302
Adj.R-squared	0.292	0.293
F-statistic	34.74	34.76
Prob(F-statistic)	2.87e-67	2.91e-67

Specifically, the P-value performance of each variable is better than lockdown1 and lockdown2. Among them, Geo score and health score are not very explanatory for the infection rate during lockdown2, because the p-value is greater than 0.1. While, the remaining 5 independent variables P-values are all less than 0.05, which has a strong correlation.

From the point of view of coefficients, the intercept of the model is 2460.2. It is worth noting that the coefficients of the two negatively correlated independent variables are particularly high, income score is -3573.85, and proportion of people over 65 is -934.49. This means that under the condition that other characteristics remain unchanged, when an 1% increase income score or the proportion of residents aged over 65, the the new cases per 100,000 population would increase by 3573.85% and 934.49%. At the same time, house price is still the variable with the smallest coefficient of -0.0003. The specific regression results are as follows:

Table 12: Linear regression result in lockdown3

Dimensions	Name	Coefficient	Std error	T-value	P-value
Demography	(Intercept)	2460.206	103.824	23.696	0
	X2: O65	-934.492	331.587	-2.818	0.005
	X3:Bame	189.2974	86.656	2.184	0.029
Economy& house	X4: Income	-3573.85	482.273	-7.41	0
	X5: HP	-0.0003	8.05E-05	-4.242	0
	X9: PR	-7.6214	1.838	-4.146	0
	X10: NR	-19.0532	6.32	-3.015	0.003
Spatial	X11: Geo	-47.694	34.082	-1.399	0.162
	X12: Density	-0.5753	0.275	-2.093	0.037
Deprivation related	X14: Health	57.8922	41.337	1.401	0.162
	X17: Education	17.2098	2.361	7.289	0
	X18: Crime	136.3015	41.858	3.256	0.001
	X19: Environment	-4.6989	2.302	-2.041	0.042

5.2.2.2 Regression Tree and Random Forest

In the process of using RT and RF to test the results, it was found that their scores were higher than those of LR, which were 0.275 (best_depth is 6) and 0.45 (n_estimators is 150). It shows that there is a certain non-linear relationship in this data. adult skills

score and independent variables related to economy& house are still the most important features. See the table below for specific data:

Table 13: The result of RT and RF in lockdown3

Dimensions	Regression Tree		Random Forest	
	Core index	best_depth: 6	Core index	n_estimators: 150
	Score	0.2749	Score	0.4332
Demography	X1: NUK	0.06898	X1: NUK	0.10436
	X2: O65	0.00194	X2: O65	0.03616
	X3: Bame	0.039	X3: Bame	0.04168
Economy& house	X4: Income	0.0005	X4: Income	0.01662
	X5: HP	0.07132	X5: HP	0.06396
	X6: OO	0.15668	X6: OO	0.07425
	X7: OML	0.00794	X7: OML	0.04228
	X8: SR	0.03667	X8: SR	0.03574
	X9: PR	0.01495	X9: PR	0.04435
	X10: NR	0.03231	X10: NR	0.03997
Spatial	X11: Geo	0.06909	X11: Geo	0.04451
	X12: Density	0.0005	X12: Density	0.04485
Deprivation related	X13: IMD	0.0037	X13: IMD	0.01981
	X14: Health	0.01718	X14: Health	0.02838
	X15: AS	0.3915	X15: AS	0.24653
	X16: Employment	0	X16: Employment	0.02435
	X17: Education	0.02417	X17: Education	0.022
	X18: Crime	0.0201	X18: Crime	0.04051
	X19: Environment	0.04346	X19: Environment	0.02969

Chapter 6 Discussion

6.1 Discussion on the results of the three lockdowns

The following will compare the results of different time, different models and different regions. The model will select linear regression and random forest for comparative analysis. Because the random forest has more weighted average or arithmetic average than the regression tree, it will be more refined and accurate fitting, so the comparative analysis of the regression tree is omitted.

6.1.1 Comparison of the results of different lockdown periods

As the lockdown time changes, no matter which model is used, the overall fit has improved. The specific trend is a slight decrease from lockdown1 to lockdown2, but a substantial increase in lockdown3. However, there are certain differences in the feature importance rankings of different models. The following will compare the results of linear regression and random forest.

In the linear regression model, overall r-square is increasing. It shows that with the development of covid-19, the relationship between the independent variable and the case rate is getting bigger and bigger, and the fitting effect is getting better and better. During the first lockdown, all variables could only account for 18.6% of the case rate. In the second lockdown, it dropped to 16.4% by a small margin. In the third lockdown, it increased sharply to 30.2%.

From the perspective of coefficient ranking, whenever proportion of people over 65, proportion of bame and income score are the three independent variables with the largest coefficients. In addition, the coefficients of health score and Geo score have also been ranked relatively high.

There are also some findings from the perspective of coefficient trends. First of all, with the change of time and the development of covid-19, the correlation of income score is getting stronger and stronger, much higher than other factors. In the first lockdown, the coefficient of income score ranked third, but it became the highest ranked coefficient in the second and third lockdowns. And in the second and third lockdown, the coefficient of income score was much higher than the other independent variables, which were -2677.65 and -35663.85, respectively. This means that when other independent variables remain the same, for income score decrease by 1%, the new cases per 100,000 population will increase by -2677.65% (lockdown2) and -35663.85% (lockdown3). Secondly, the coefficient ranking of some variables is decreasing, which means that their importance is gradually decreasing. Such as: proportion of people over 65, proportion of bame and Geo score. Finally, the correlation of health score has become stronger and stronger, and the coefficient ranking has also remained at 5 from 6 at the beginning. The detailed results are compared in the table below.

Table 14: Comparison of linear regression results

Model Name	Linear Regression					
	Lockdown1		Lockdown2		Lockdown3	
Number of samples	959		974		979	
Number of independent variables	12		12		12	
R ²	0.186		0.164		0.302	
Coefficient (top 6)	X2: O65	256.1346	X4: Income	-2677.65	X4: Income	-3573.85
	X3:Bame	254.0734	X3:Bame	517.3545	X2: O65	-934.492
	X4:Income	-68.8467	X2: O65	166.5387	X3:Bame	189.2974
	X11: Geo	48.621	X11: Geo	-56.9937	X18: Crime	136.3015
	X18: Crime	-34.0174	X14: Health	36.0244	X14: Health	57.8922
	X14: Health	25.8371	X17: Education	15.4932	X11: Geo	-47.694

In the random forest model, the overall score is also increasing. From 0.28 at the first blockade to 0.433 at the third time.

From the perspective of importance ranking, adult skills score is the most important feature at any time. Even when it is blocked for the third time, the importance is as high as 0.24, far exceeding other features. In addition, proportion of bame has always been ranked second in importance in lockdown1 and lockdown2, which shows that this feature also has a strong correlation with case rate. A similar variable is population density, which has always been ranked fifth in the ranking of the importance of 19 variables.

Table 15: Comparison of random forest results

Model Name	Random Forest					
	Lockdown1		Lockdown2		Lockdown3	
Number of samples	959		974		979	
Number of independent variables	19		19		19	
Score	0.28		0.2312		0.4332	
Importance score (top 6)	X15: AS	0.11277	X15: AS	0.18047	X15: AS	0.24653
	X3: Bame	0.0853	X3: Bame	0.09897	X1: NUK	0.10436
	X11: Geo	0.06849	X8: SR	0.06883	X6: OO	0.07425
	X2: O65	0.06535	X1: NUK	0.06176	X5: HP	0.06396
	X12: Density	0.05506	X12: Density	0.04962	X12: Density	0.04485
	X9: PR	0.05251	X17: Education	0.0489	X11: Geo	0.04451

6.1.2 Comparison of results of different models

We collate the results of two of the three models. It is found that there is a certain score difference, and the order of the coefficients of the features is also different. We can see that the fitting of random forest is better than linear regression during different blockade periods. This shows that the data has a certain non-linear relationship. But this cannot be used as the only basis, because the overall sample size is about 1,000, and the amount of data is not sufficient for the random forest method.

From the perspective of feature importance, the important features of linear regression are more concentrated than random forest. In the top3 ranking of importance,

regressions are all concentrated in income score, proportion of bame and proportion of people over 65, while random forest is mainly concentrated in adult skills score and proportion of bame. Although the importance of random forest ranking does not focus on certain three variables every time the top three, adult skills score ranks first in the three blockades. And proportion of bame has appeared in second place twice, respectively in lockdown1 and lockdown2.

Table 16: Comparison of LR and RF results

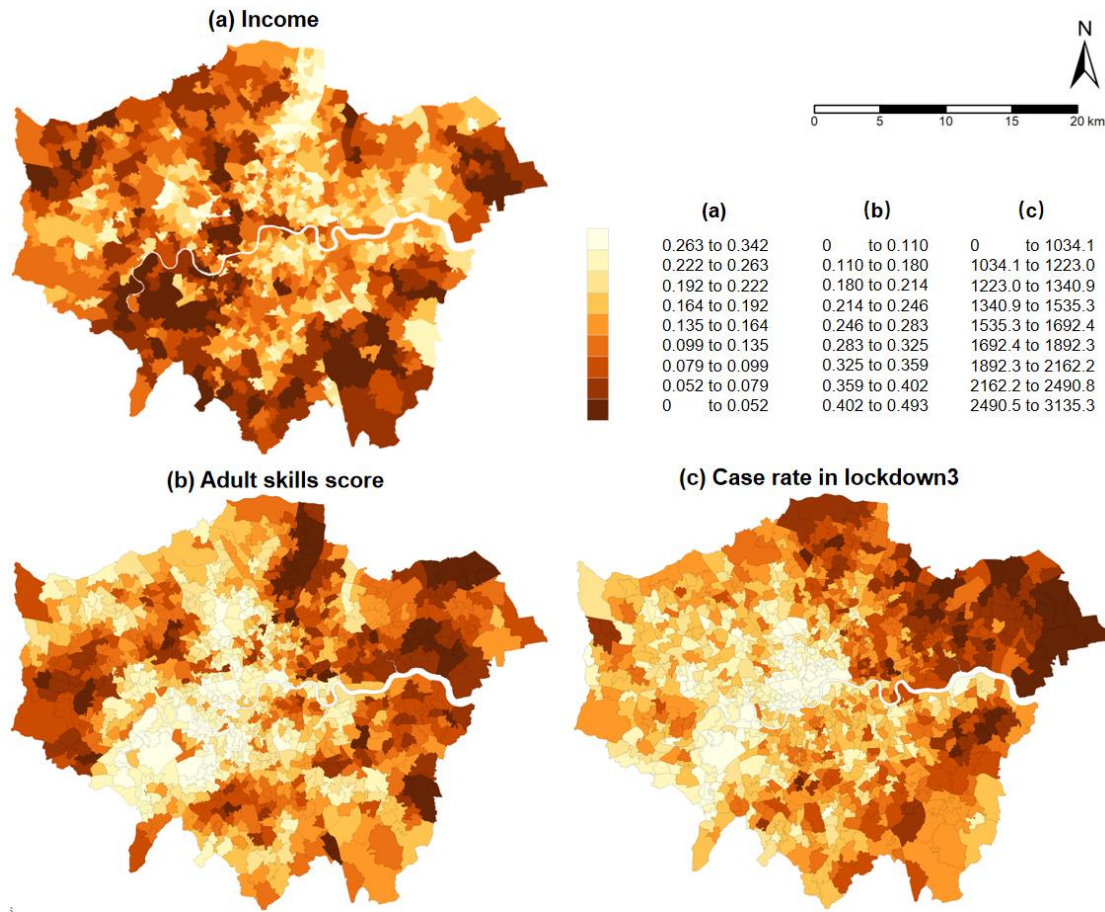
	Model	Score /R ²	Coefficient/Importance (top3)					
			Top 1		Top 2		Top 3	
Lockdown1	LR	0.186	X2: O65	256.1346	X3:Bame	254.0734	X4:Income	-68.8467
	RF	0.28	X15: AS	0.11277	X3:Bame	0.0853	X11: Geo	0.06849
Lockdown2	LR	0.164	X4: Income	-2677.65	X3:Bame	517.3545	X2: O65	166.5387
	RF	0.2312	X15: AS	0.18047	X3: Bame	0.09897	X8: SR	0.06883
Lockdown3	LR	0.302	X4: Income	-3573.85	X2: O65	-934.492	X3:Bame	189.2974
	RF	0.4332	X15: AS	0.24653	X1: NUK	0.10436	X6: OO	0.07425

6.1.3 Visualization of important features

Based on the above conclusions, we can know that income score is the most important feature in linear regression, and random forest is adult skills score. The following will be displayed visually the above two features and case rate to compare which factor has a stronger correlation with the independent variable, and compare the trends in different lockdown periods.

From figure 3, we can see that compared with income score, the correlation between adult skills score and case rate is stronger. The core variables and independent variable distributions of the two different methods are shown in figure3. Among them, according to the results, we know that income score and case rate are negatively correlated. In order to facilitate the comparison and visualization of the results, the lower the income score is set to the darker color. And choose the case rate of lockdown3 as the representative of the independent variable, because during this period the new cases per 100,000 population is the most, there will be a better visualization effect, which is convenient for comparison with other factors. Specifically, we can see that in figure3 (a), dark colors are mainly concentrated in the southern and northeast regions at the edge of the map, which means that the income score in these areas is relatively low. In figure 3 (b), the areas with high adult skills scores represented by dark colors are mainly concentrated in the northeast and west regions. The dark color in figure3 (c) means that there is a higher case rate, which is mainly concentrated in the northeast region. Based on the pictures in figure3, we can know that adult skills score and case rate in lockdown3 have a stronger correlation and show a positive correlation trend.

Figure 4: Distribution of core variables



The following figures 4 and 5 will specifically show the distribution relationship between different variables and case rates during different lockdown periods. Among them, the darker the background color, the lower the income score and the higher the adult skills score. The larger the dot, the more the new cases per 100,000 population. According to the following two figures, the following results can be obtained. First, each lockdown is more serious than the previous lockdown of covid-19. It can be seen that the blue dots in (a), (b) and (c) in Figure 4 and Figure 5 are getting bigger and bigger and more obvious. This shows that with the development of covid-19 and the passage of time, the new cases per 100,000 population are increasing. Second, compared to figure 4, the dark background in figure 5 has a higher degree of overlap with the large dots. This means that compared with income score, the correlation between adult skills score and case rate is higher, and it is mainly concentrated in the northeast. In figure 4, taking the most obvious (c) as an example, although the low income score and high case rate overlap to a certain extent, the overlap is relatively small, mainly in the northeast area at the edge of the map.

Figure 5: Income score and case rate distribution in different periods

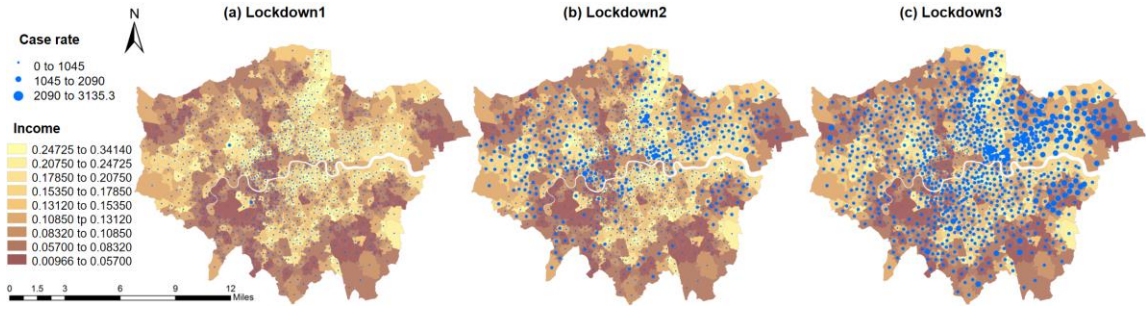
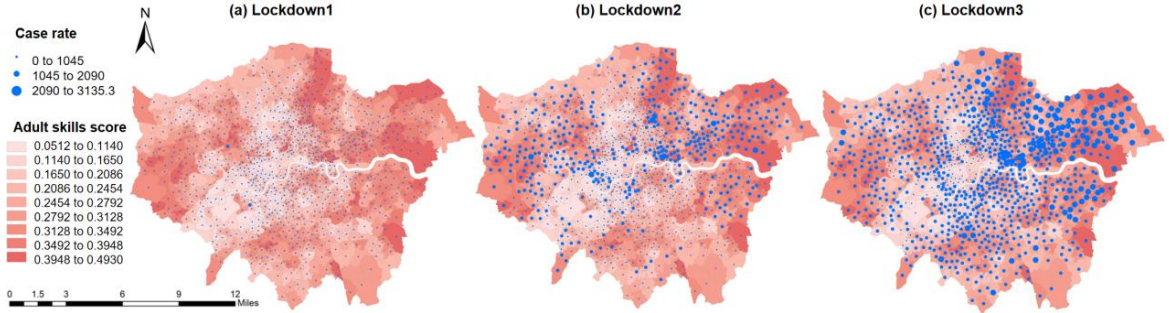


Figure 6: Distribution of Adult skills score and case rate in different periods



6.2 Advantages and limitations .

In general, the results have some advantages and disadvantages shown. The advantage is that the data set selects 983 samples of MSOA, which can discuss topics in more detail. Second, using three different methods to verify and compare analysis conclusions, we can look at the results and different models more dialectically. Third, compare and analyze the situation of the three lockdowns in 2020. Then we can know the changing trends of covid-19 at different times.

At the same time, if we make the following improvements, we may get better results. First, supplement the data of the independent variables. Because MSOA is a relatively small area unit, and the week is also a relatively small time division unit. After communicating with the official email, many covid-19-related data involve privacy and are not suitable for academic research for graduate students. In the absence of relevant datasets, the fitting effect cannot be optimized. If we can add other data such as the number of deaths and population movements in the weekly dimension, the results will be better. Secondly, for the latter two methods of machine learning, the amount of data is not very large, and it is impossible to confirm whether the fit is optimal. Third, although outliers have been deleted in this article, some data does have a large span in different regions and times.

Chapter 7 Conclusion

This article mainly studies the changing relationship between covid-19 cases and socio-demographic factors in the three lockdowns in London in 2020. Three methods of linear regression, regression tree and random forest are used to compare and analyze the results. Python, R, GIS are used as tools for data processing and visualization.

In 2020, the new cases per 100,000 population in London have a certain relationship with socio-demographic factors. First of all, in the three lockdowns, the overall fit is increasing. That means with the development of covid-19, these independent variables can better explain the case rate. Secondly, the factors with strong correlation are mainly concentrated in adult skills score, income score and demographic-related factors (proportion of people over 65 and proportion of bame). And among all the factors, adult skills score has the strongest correlation. Third, there are different priorities in different methods. Income score is the most important feature in linear regression, and adult skills score is the most important feature in regression tree and random forest. At the same time, with the change of time, in the linear regression model, the independent variable of income score has become more and more important, and the coefficient is much higher than other characteristics.

This article emphasizes the importance of multiple variables, multiple models, and multiple time periods for studying covid-19 cases and has made new discoveries. Through these findings, we realize the importance of multi-dimensional analysis. Firstly, among all the factors, adult skills score has the strongest correlation with cases. However, this factor is rarely discussed in existing papers. Secondly, the core factors obtained by different models are also different. Finally, by comparing the results in time and space, the development trend conclusions of covid-19 can be obtained.

Although in the process of fighting against covid-19, we cannot avoid the infection, but the extent and scope of the infection can be controlled by our efforts and methods (Public Health England, 2021). By combining different models and methods, we can see that income score, the proportion of people over 65, the bame ratio, and the adult skill score have a significant impact on new cases per 100,000 population. We have seen the government and society make certain efforts. For example, give priority to vaccinations for people over 65 years old (COVID-19 vaccination first phase priority groups, 2021). The westminster area will provide senior citizens with daily necessities, health checks and chat services every week (Age UK Westminster-COVID-19 Response, 2021).

But according to this result, we still have some new expectations. For one, we hope to strengthen the improvement of adult skills. In the results, we know that adult skills score has the strongest correlation with cases. However, this factor has been ignored in most current studies, so there is no good relevant policy at present. Although under

normal circumstances, few adults use online learning as a training method (OECD, 2019). But in this special period of covid-19, governments and institutions can provide more online learning courses to help adults improve their skills and abilities. Let those who do not have the advantage in social competition, such as the disabled and people in remote areas, can also master certain skills systematically and formally. Secondly, we would prefer a specific area or group of people could get some help, such as bame. At present, the policies we see are to increase publicity through bame institutions or the media, but have not some very substantial help (Public Health England, 2020). We think we can give them some specific services. For example, regularly provide protective products such as masks and disinfectants. Or when they feel that they have symptoms of infection, they can give priority to nucleic acid testing.

It has been a year and a half since the discovery of covid-19, which shows that fighting the epidemic is still a long-term thing. Although the current situation in London is under control, it is still a long-term and arduous challenge that requires continuous efforts from all walks of life. Let everyone work together to help specific groups and regions reduce the infection rate, so that everyone has the right to live a healthy life.

Appendix

Appendix A:

The Summary List of All Data Source Employed in this Research

Appendix B

Repository: https://github.com/Audrey-chenxi/CASA_Dissertation_Chenxi-Zhao

Bibliography

Ageuk.org.uk. 2021. *Age UK Westminster - COVID-19 Response*. [online] Available at: <<https://www.ageuk.org.uk/westminster/our-services/covid-19-response/>> [Accessed 19 August 2021].

Andrew Gregory, H., 2021. *Inside the commuter 'Covid triangle', 1 in 16 are now infected*. [online] Thetimes.co.uk. Available at: <<https://www.thetimes.co.uk/article/inside-the-commuter-covid-triangle-1-in-16-are-now-infected-q09sngcnk>> [Accessed 13 July 2021].

Bray, I., Gibson, A. and White, J., 2020. Coronavirus disease 2019 mortality: a multivariate ecological analysis in relation to ethnicity, population density, obesity, deprivation and pollution. *Public Health*, 185, pp.261-263.

Cheng, V., Wong, S., Chuang, V., So, S., Chen, J., Sridhar, S., To, K., Chan, J., Hung, I., Ho, P. and Yuen, K., 2020. The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2. *Journal of Infection*, 81(1), pp.107-114.

Dong, E., Du, H. and Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), pp.533-534.

GOV.UK. 2021. *COVID-19 vaccination first phase priority groups*. [online] Available at: <<https://www.gov.uk/government/publications/covid-19-vaccination-care-home-and-healthcare-settings-posters/covid-19-vaccination-first-phase-priority-groups>> [Accessed 19 August 2021].

Gray, J. and Woodall, W., 1994. The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis. *The American Statistician*, 48(2), pp.111-113.

Gupta, M., Wahl, B., Adhikari, B. *et al.* The need for COVID-19 research in low- and middle-income countries. *glob health res policy* 5, 33 (2020).
<https://doi.org/10.1186/s41256-020-00159-y>

International Monetary Fund (2020). "World Economic Outlook," International Monetary Fund. April.
<https://www.imf.org/en/Publications/WEO/Issues/2020/04/14/weo-april-2020>

Journal of Mathematical and Statistical Psychology, 28 (1), pp. 51–62. doi: 10.1111/j.2044-

Kakkar, D., Dunphy, D. and Raza, D., 2021. *Ethnicity profiles of COVID-19 admissions and outcomes*.

Lloyd, C., 2015. Spatial scale and small area population statistics for England and Wales. *International Journal of Geographical Information Science*, 30(6), pp.1187-1206.

MAXWELL, A. E. (1975). 'LIMITATIONS ON THE USE OF THE MULTIPLE LINEAR REGRESSION MODEL'. *British*

OECD (2019), Dashboard on priorities for adult learning - OECD, <http://www.oecd.org/employment/skills-and-work/adult-learning/dashboard.htm> (accessed on 16 April 2020).

Palmer, B., 2021. *Chart of the week: Covid-19 kills people in the most deprived areas at double the rate of those in the most affluent*. [online] The Nuffield Trust. Available at: <<https://www.nuffieldtrust.org.uk/resource/chart-of-the-week-covid-19-kills-the-most-deprived-at-double-the-rate-of-affluent-people-like-other-conditions>> [Accessed 13 July 2021].

Pan, J., Yao, Y., Liu, Z., Meng, X., Ji, J., Qiu, Y., Wang, W., Zhang, L., Wang, W. and Kan, H., 2021. Warmer weather unlikely to reduce the COVID-19 transmission: An ecological study in 202 locations in 8 countries. *Science of The Total Environment*, 753, p.142272.

Public Health England, 2020. *Beyond the data: Understanding the impact of COVID-19 on BAME groups*. [online] Wellington: Public Health England, p.4-49. Available at: <<https://www.gov.uk/government/publications/covid-19-understanding-the-impact-on-bame-communities>> [Accessed 21 August 2021].

Publichealth.hscni.net. 2021. [online] Available at: <https://www.publichealth.hscni.net/sites/default/files/2021-02/Infection_Prevention_and_Control_Guidance_January_2021.pdf> [Accessed 19 August 2021].

Schellekens, I., 2021. *COVID-19 is a developing country pandemic*. [online] Brookings. Available at: <<https://www.brookings.edu/blog/future-development/2021/05/27/covid-19-is-a-developing-country-pandemic/>> [Accessed 13 July 2021].

Schellekens, P. and Sourrouille, D., 2021. *COVID-19 Mortality in Rich and Poor Countries : A Tale of Two Pandemics?*. World bank group, pp.2-16.

Shrestha, N., 2020. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, 8(2), pp.39-42.

Statistics.laerd.com. 2021. Mean, Mode and Median - Measures of Central Tendency - When to use with Different Types of Variable and Skewed Distributions | Laerd Statistics. [online] Available at: <<https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>> [Accessed 26 July 2021].

Un.org. 2021. *How COVID-19 is changing the world: a statistical perspective (Volume II) | Population Division*. [online] Available at: <<https://www.un.org/development/desa/pd/news/how-covid-19-changing-world-statistical-perspective-volume-ii>> [Accessed 13 July 2021].

Van Rens, T. and Oswald, A., 2021. Age-Based Policy in the Context of the Covid-19 Pandemic: How Common are MultiGenerational Households?. [online] Ideas.repec.org. Available at: <<https://ideas.repec.org/p/cge/wacage/522.html>> [Accessed 26 July 2021].

Voltes-Dorta, A. and Sánchez-Medina, A. (2020) ‘Drivers of Airbnb prices according to property/room type, season and location: A regression approach’, *Journal of Hospitality and Tourism Management*, 45, pp. 266–275. doi: 10.1016/j.jhtm.2020.08.015.

Weir, G. and Oakley, M., 2020. NEIGHBOURHOOD LEVEL COVID-19 MORTALITY IN LONDON. Technical Paper for London’s Poverty Profile. Truist for London, pp.3-5.

Worldometer (2020). “Coronavirus Countries Where COVID-19 Has Spread,” Worldometer. Updated daily. Accessed May 27.

World Health Organization COVID-19 situation. 2020. World Health Organization, p.113.8317.1975.tb00547.x.