CASA0012 Dissertation Book

Chenxi Zhao

CASA0012, MSc Spatial Data Science and Visualisation Dissertation

Supervisor: Dr Huanfa Chen

Repository: https://github.com/Audrey-chenxi/CASA_Dissertation_Chenxi-Zhao
This dissertation is submitted in part requirement for the
MSc (Or MRes) in the Centre for Advanced Spatial Analysis,
Bartlett Faculty of the Built Environment, UCL

Word count: 8,000+
2021-07-17

Abstract

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 xxx words in length

Contents

Appendix
List of Figures

List of Tables

Abbreviations

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 the start of mass quarantine and wearing of masks (Cheng et al., 2020). Later on it was found that there was also a strong correlation with some socio-demographic factors 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.

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 below the country. However, only by understanding the relevant factors in these small areas can we have a more accurate judgment on covid-19 and better reflect on 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. 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 countries with limited scarce resources allocate it? How does the government guarantee the people's health and basic life? How do companies face constant 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. (IMF, 2020) 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. 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 ranks second in the world,

accounting for 31% of the world's total. (Eurostat) The infection rate in the United Kingdom have always been of concern, because the cumulative number of cases in the United Kingdom is higher than 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.

In our opinion, 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 a 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 of time, 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)

2.3 To be added

Chapter 3
Data Source and Processing

3.1 Data Framework

The data mainly goes through four stages. First find the relevant data sets, including case rate and social demographic factors related data, MSOA and LSOA correspondence table, London shapfile. Secondly, after data processing, a complete data set is obtained, including dependent variables and 19 independent variables from

5 categories. The third and fourth steps are to process the verification data through different methods, and finally get the data results and visualization charts. The specific content of each part will be explained in the data and method section below.

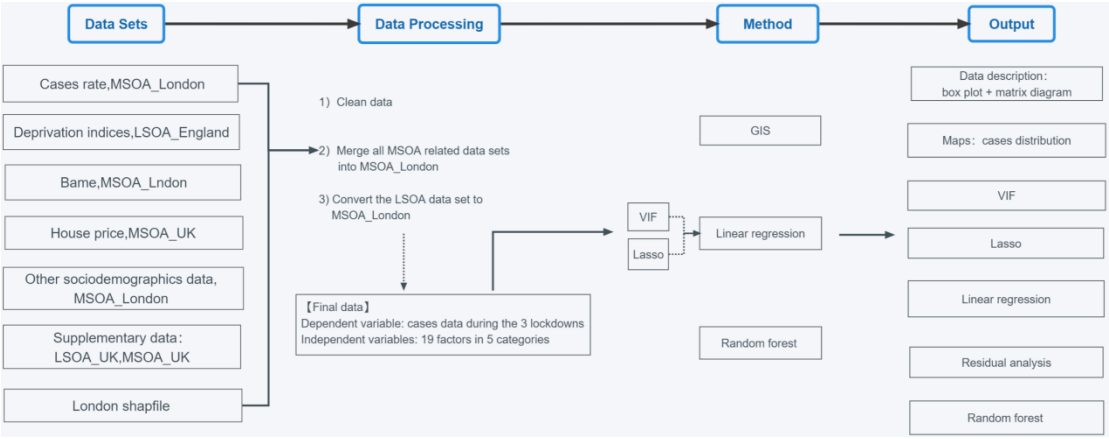


Figure1: Data Framework

3.2 Data Source and Processing

3.2.1 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. The data are selected from the relevant data of case rate and socio-demographic factors. Among them, the data of covid-19 uses the data of 2020, and the other data adopts 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 complicated, including MSOA and LSOA data, and the scope includes data from London, England and the United Kingdom. In terms of data processing, first clean each data set. Secondly, integrate all MSOA data sets into London MSOA level data. Then use the mapping relationship between LSOA and MSOA to convert LSOA related data into MSOA. Finally, merge the above data together.

3.2.2 Data explanation and description

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:

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

Figure2: Lockdown Timeline

In terms of independent variables, 19 factors in 5 categories are selected.

1) People related factors (X1: NUK, X2: O65, X3: Bame) 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) added on this basis whether it is UK, because it determines Whether people can get the same social rights and benefits.
2) House related factors (X4: HP, X5: OO, X6: OML, X7: SR, X8: PR, X9: NR) are mainly divided into two parts: house price and tenure. Among them, due to the large difference in housing prices in London, the median is used here. 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) Covid-19 factors (X10: Density, X11: Geo, X12: IMD, X13: Health, X14: Income) Factors that have been confirmed to have a strong relationship with the covid-19 in other dissertations. It has been explained in detail in the lecture review.
4) Income-related factors (X15: AS, X16: Employment, X17: Education). According to the above lectures, we know that income will affect the situation of covid-19 to a certain extent. Therefore, we consider the skills, academic background and employment that will affect income as related variables.
5) Social related factors (X18: Crime and X19: Environment) represent the social security situation and living conditions of the area.

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

Figure3: Description of Variable

3.3 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 London MSOA 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.

In linear regression, two methods (VIF and Lasso) are used to deal with multicollinearity, and the confidence of P-value is discussed, and 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 machine learning, we try to continuously optimize the model by adjusting hyperparameters and cross-validation methods. In this process, the advantages and disadvantages of different methods were analyzed and compared, and different models and methods were viewed critically.

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:

$$VIF_k = \frac{1}{1 - R_k^2} \tag{1}$$

The value of VIF is greater than 1. The closer the VIF value is to 1, the lighter the multicollinearity, and vice versa. Those with a score greater than 5 are deemed more serious, and this variable will be deleted.

Least Absolute Shrinkage and Selection Operator (Lasso) is also one of the methods used to solve multicollinearity, but Lasso uses the L1 paradigm of the coefficient w (the L1 paradigm is the absolute value of the coefficient w) multiplied by the regularization coefficient α, so The expression of Lasso's loss function is:

$$J_L(w) = \frac{1}{2}\|y - Xw\|^2 + \lambda \sum |w_i| \tag{2}$$

As a regularization method, the basic idea of lasso regression is to minimize the sum of squared residuals. Thus, some regression coefficients that are strictly equal to 0 can be generated, and a model with strong explanatory power can be obtained. But it has a constraint that the sum of the absolute values of the regression coefficients is less than

a constant

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 vatiables):

$$y = X\beta + \varepsilon \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{3}$$

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:

$$\hat{\beta} = (X^T X)^{-1} X^T y = Hy \tag{4}$$

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) :

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{5}$$

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:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} \tag{6}$$

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. The insignificant outliers greater than 3 can be removed by linearity test, normality test, independence test, homoscedasticity test and other methods. If there are no obvious ones, they will be retained, because they will not affect much of the R^2.

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 trees. 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

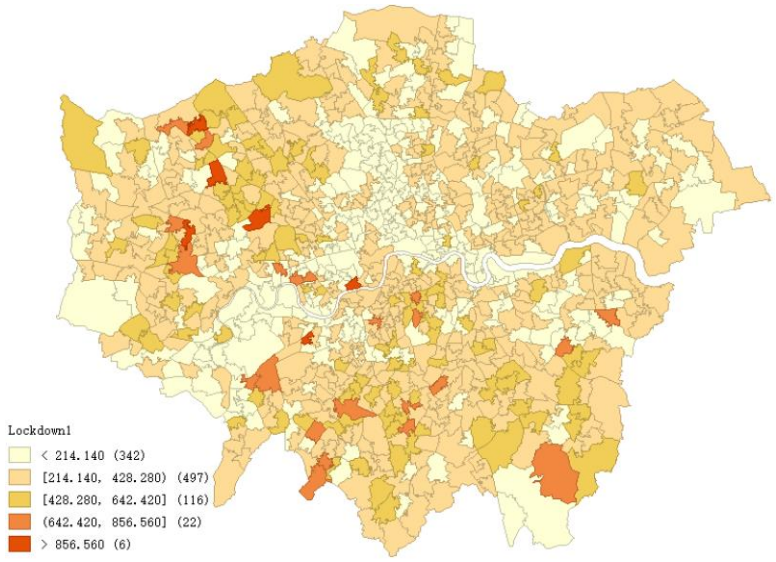
5.1 Covid-19 case rate under different lockdown periods

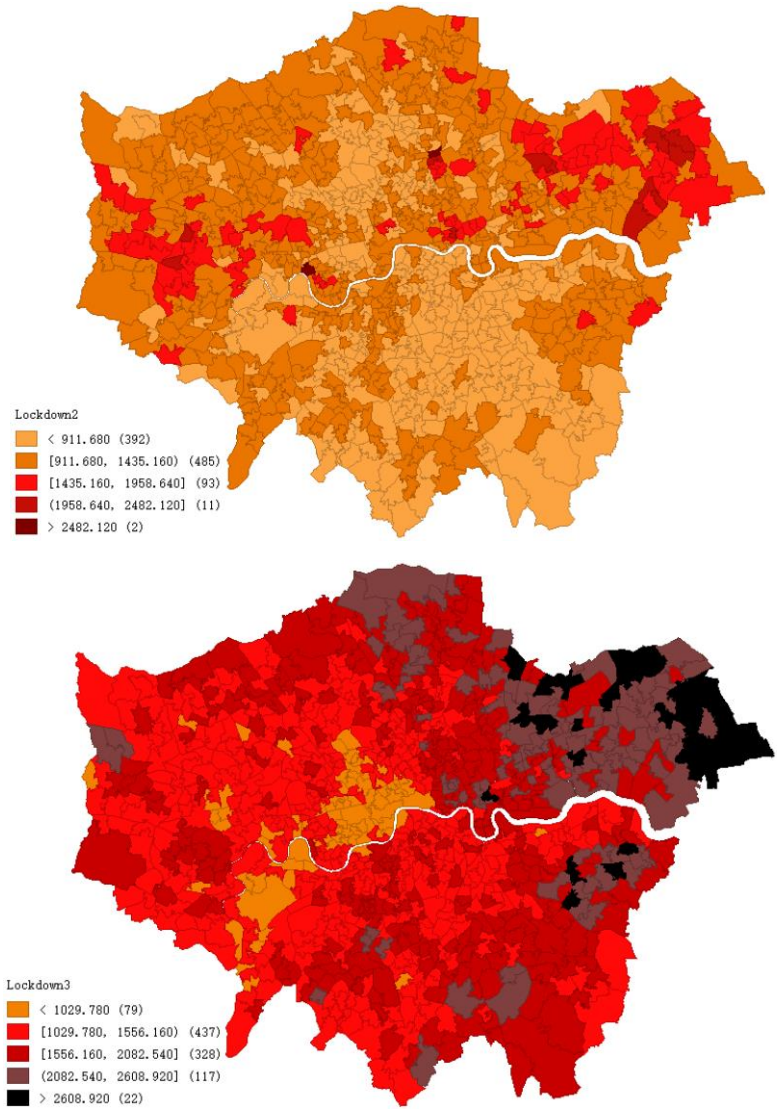
5.1.1 Visualization of covid-19 cases

5.1.2 Covid-19 case rate distribution map (Three lockdowns)

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 distribution of infection rates in different time periods will be supplemented)





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, multicollinearity is processed first, and the methods of VIF and Lasso are used here. When VIF was tested for the first time, it was found that the VIF

of the 6 independent variables such as X7: SR, X12: IMD was higher than 5. That means the multicollinearity is very strong and will affect the results of the model. Therefore, the 6 independent variables higher than 5 are deleted. When performing the second VIF verification, it was found that the VIF of X6: OML was 5.19, and it was also deleted since it was higher than the standard 5. Therefore, after the original 19 variables were tested for multicollinearity, 7 independent variables were deleted, and 12 independent variables remained. The following table shows the VIF value of the deleted features:

Variable	X7: SR	X12: IMD	X16: Employment	X15:AS	X1: NUK	X5: OO	X6: OML
VIF1	266.163	24.70	16.91	10.84	7.61	5.19	4.42
VIF2	-	-	-	-	-	-	5.19

Figure7: VIF of lockdown1

Secondly, the fitting of linear regression was performed three times in total. The first time is the fitting result of the original data. The second time is the fitting result after removing the multicollinearity, and the third time is the residual analysis based on the second time. It is found that the overall residual distribution is relatively uniform, and there is no obvious non-linear relationship. However, there is a partial separation of the group value, where outliers with residuals greater than 3 are removed. Then the third fitting is performed. The result is shown in the figure:

Regression	First	Multicollinearity processed	Residuals processed
Number of samples	972	972	959
Number of independent variables	19	12	12
R-squared	0.181	0.154	0.186
Adj.R-squared	0.165	0.144	0.176
F-statistic	11.11	14.59	18.04
Prob(F-statistic)	1.16e-30	2.42e-28	2.27e-35

Figure8: Linear regression comparison in lockdown1

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 reslut 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.

From the perspective of P-value, X10: Density and X14: Income 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 X9: NR, X13: Health and X17: Education 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 with case rate.

From the perspective of coefficients, when all the independent variables take 0, the predicted value of the dependent variable is 297.174. People, income, and covid-19

related factors can affect the infection rate more than other factors. Specifically, the coefficients worth paying attention to are X2: O65 (256.1346), X3: Bame (254.0734), X14: Income (-68.8467), X11: Geo (48.621) and X18: Crime (-34.0174). This shows that with the other independent variables remaining constant, each increase of O65, Bame and Geo by one unit may increase the case rate by 256.134, 254.0734 and 48.621 units. Income and Crime are negatively correlated. When other independent variables remain the same, each increase of one unit of them may reduce the case rate by -68.8467 and -34.0174 units. X4: The coefficient of HP is extremely low, indicating that the impact on the dependent variable is small. The specific regression results are as follows:

Dimensions	Name	Coefficient	Std error	T-value	P-value
	(Intercept)	297.1742	35.843	8.291	0.000
People	X2: O65	256.1346	113.505	2.257	0.024
	X3:Bame	254.0734	29.727	8.547	0.000
House	X4: HP	-9.87e-05	2.79e-05	-3.541	0.000
	X8: PR	-3.3659	0.633	-5.317	0.000
	X9: NR	3.7471	2.18	1.719	0.086
Covid-19	X10: Density	-0.0106	0.095	-0.112	0.911
	X11: Geo	48.621	11.726	4.147	0.000
	X13: Health	25.8371	14.374	1.798	0.073
Income	X14: Income	-68.8467	165.173	-0.417	0.677
	X17: Education	-1.5223	0.807	-1.888	0.059
Society	X18: Crime	-34.0174	14.601	-2.33	0.02
	X19: Environment	1.7739	0.792	2.239	0.025

Figure9: Linear regression result in lockdown1

5.2.1.2 Regression Tree and Random Forest

Use regression trees and random forests to verify and predict the above results.

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 final prediction result is 0.2043. At the same time, when n_estimators is 200, the random forest has the best prediction result with a score of 0.242.

In the regression tree, the most important feature is the income-related X15:AS of 0.352. The importance of AS in the random forest is also the highest at 0.125. It is worth noting that the overall importance of covid-19 related features is very high at 0.308. The specific values are shown in the table:

Regression Tree			Random Forest		
Core index	best_depth : 4		**Core index**	n_estimators: 200	
Score	0.2043		**Score**	0.242	
People: 0.17022	X1: NUK	0	**People: 0.18632**	X1: NUK	0.04053
	X2: O65	0.04703		X2: O65	0.06879
	X3:Bame	0.12319		X3:Bame	0.077
House: 0.21014	X4: HP	0.05972	**House: 0.28521**	X4: HP	0.06309
	X5: OO	0.04864		X5: OO	0.04632
	X6: OML	0		X6: OML	0.03663
	X7: SR	0.07865		X7: SR	0.03555
	X8: PR	0		X8: PR	0.049
	X9: NR	0.02313		X9: NR	0.05462
Covid-19: 0.15365	X10: Density	0	**Covid-19: 0.30832**	X10: Density	0.06571
	X11: Geo	0.15365		X11: Geo	0.07309
	X12: IMD	0		X12: IMD	0.0269
	X13: Health	0		X13: Health	0.04262
Income: 0.35213	X14: Income	0	**Income: 0.22981**	X14: Income	0.01849
	X15: AS	0.35213		X15: AS	0.12488
	X16: Employment	0		X16: Employment	0.03138
	X17: Education	0		X17: Education	0.05506
Society: 0.11386	X18: Crime	0.08772	**Society: 0.09034**	X18: Crime	0.0445
	X19: Environment	0.02614		X19: Environment	0.04584

Figure10: The result of RT and RF in lockdown1

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, the fitting of the tree times linear regression is also performed. The first time is the fitting of the original data. The second time is the fitting after removing the multicollinear variables. 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 LR model. After processing 19 variables, the remaining 12 independent variables. After the residual analysis is performed, the third fitting is performed. In this process, 9 outliers with residuals greater than 3 are removed. The result is shown in the figure:

Regression	First	Multicollinearity processed	Residuals processed
Number of samples	983	983	974
Number of independent variables	19	12	12
R-squared	0.261	0.163	0.164
Adj.R-squared	0.246	0.153	0.153
F-statistic	17.90	15.74	15.65
Prob(F-statistic)	3.87e-51	9.56e-31	1.50e-30

Figure11: Linear regression comparison in lockdown2

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 0.186 and 0.176. 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 X4: HP, X9: NR 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 1050.4358. The coefficient of X14: Income is surprisingly high at -2677.65. This means that with other independent variables remaining constant, each additional unit of income may reduce the case rate at lockdown2 by 2677.65 units. At the same time, the variables X3:Bame and X2:O65 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 X4: HP is also very small, which is the same as lockdown1, indicating that house prices in these two time periods have a small impact on the infection rate. The specific regression results are as follows:

Dimensions	Name	Coefficient	Std error	T-value	P-value
	(Intercept)	1050.4358	77.385	13.574	0.000
People	X2: O65	166.5387	246.626	0.675	0.500
	X3:Bame	517.3545	64.656	8.002	0.000
House	X4: HP	2.063e-05	6e-05	0.344	0.731
	X8: PR	-1.2581	1.372	-0.917	0.359
	X9: NR	-1.8452	4.703	-0.392	0.695
Covid-19	X10: Density	-0.4595	0.204	-2.249	0.025
	X11: Geo	-56.9937	25.387	-2.245	0.025
	X13: Health	36.0244	30.815	1.169	0.243
Income	X14: Income	-2677.6513	361.096	-7.415	0.000
	X17: Education	15.4932	1.812	8.549	0.000
Society	X18: Crime	13.5358	31.178	0.434	0.664
	X19: Environment	-1.4238	1.712	-0.831	0.406

Figure12: Linear regression result in lockdown2

5.2.2.2 Regression Tree and Random Forest

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 200, the score is 0.235. As in lockdown1, the most important feature is also X15: AS, where the importance is 0.18. And it is found that the importance of house-related features is generally high, with a total of 0.254.

Regression Tree			Random Forest		
Core index	best_depth : 2		**Core index**	n_estimators: 200	
Score	0.06877		**Score**	0.23514	
People: 0.09375	X1: NUK	0	**People: 0.21693**	X1: NUK	0.07648
	X2: O65	0		X2: O65	0.05223
	X3:Bame	0.09375		X3:Bame	0.08822
House: 0.36865	X4: HP	0	**House: 0.25423**	X4: HP	0.0424
	X5: OO	0.04864		X5: OO	0.04359
	X6: OML	0		X6: OML	0.03182
	X7: SR	0.32001		X7: SR	0.069
	X8: PR	0		X8: PR	0.0396
	X9: NR	0		X9: NR	0.02782
Covid-19: 0	X10: Density	0	**Covid-19: 0.15877**	X10: Density	0.0565
	X11: Geo	0		X11: Geo	0.04372
	X12: IMD	0		X12: IMD	0.02462
	X13: Health	0		X13: Health	0.03393
Income: 0.58624	X14: Income	0	**Income: 0.294**	X14: Income	0.02251
	X15: AS	0.58624		X15: AS	0.18131
	X16: Employment	0		X16: Employment	0.04566
	X17: Education	0		X17: Education	0.04452
Society: 0	X18: Crime	0	**Society: 0.07607**	X18: Crime	0.04233
	X19: Environment	0		X19: Environment	0.03374

Figure13: The result of RT and RF in lockdown2

5.2.3 The result of the third lockdown

5.2.3.1 Linear regression

In the linear regressions fitting in lockdown3, the first time has the highest degree of fit, but because there are many factors of multicollinearity, it cannot be used as a final reslut. The independent variable that removes multicollinearity here is also greater than 5, and is consistent with lockdown1 and 2. 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 figure:

Regression	First	Multicollinearity processed	Residuals processed
Number of samples	983	983	979
Number of independent variables	19	12	12
R-squared	0.499	0.301	0.302
Adj.R-squared	0.490	0.292	0.293
F-statistic	50.58	34.74	34.76
Prob(F-statistic)	3.05e-130	2.87e-67	2.91e-67

Figure14: Linear regression comparison in lockdown3

Specifically, the performance of P-value is better than that of lockdown1 and lockdown2. Among them, X11: Geo and X13: Health 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, X14: Income is -3573.85, and X2: O65 is -934.49. This means that under the condition that other characteristics remain unchanged, each unit reduction of Income and O65 may increase the case rate by 3573.85 and 934.49 units. At the same time, X4: HP is still the variable with the smallest coefficient of -0.0003. The specific regression results are as follows:

Dimensions	Name	Coefficient	Std error	T-value	P-value
	(Intercept)	2460.2062	103.824	23.696	0.000
People	X2: O65	-934.4918	331.587	-2.818	0.005
	X3:Bame	189.2974	86.656	2.184	0.029
House	X4: HP	-0.0003	8.05E-05	-4.242	0.000
	X8: PR	-7.6214	1.838	-4.146	0.000
	X9: NR	-19.0532	6.32	-3.015	0.003
Covid-19	X10: Density	-0.5753	0.275	-2.093	0.037
	X11: Geo	-47.694	34.082	-1.399	0.162
	X13: Health	57.8922	41.337	1.401	0.162
Income	X14: Income	-3573.8497	482.273	-7.41	0.000
	X17: Education	17.2098	2.361	7.289	0.000
Society	X18: Crime	136.3015	41.858	3.256	0.001
	X19: Environment	-4.6989	2.302	-2.041	0.042

Figure15: Linear regression result in lockdown2

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 0.33 (best_depth is 4) and 0.45 (n_estimators is 150). It shows that there is a certain non-linear relationship in this data. X15: The independent variables related to AS and house are still the most important features. See the table below for specific data:

Regression Tree			Random Forest		
Core index	best_depth : 4		**Core index**	n_estimators: 150	
Score	0.32507		**Score**	0.44775	
People: 0.04994	X1: NUK	0.02502	**People: 0.18193**	X1: NUK	0.10355
	X2: O65	0		X2: O65	0.03583
	X3:Bame	0.02492		X3:Bame	0.04255
House: 0.33661	X4: HP	0.09374	**House: 0.30312**	X4: HP	0.0568
	X5: OO	0.1777		X5: OO	0.07669
	X6: OML	0		X6: OML	0.04699
	X7: SR	0.04754		X7: SR	0.03537
	X8: PR	0		X8: PR	0.04936
	X9: NR	0.01763		X9: NR	0.03791
Covid-19: 0.0648	X10: Density	0	**Covid-19: 0.13147**	X10: Density	0.04068
	X11: Geo	0.0648		X11: Geo	0.04618
	X12: IMD	0		X12: IMD	0.0192
	X13: Health	0		X13: Health	0.02541
Income: 0.52513	X14: Income	0	**Income: 0.3128**	X14: Income	0.01935
	X15: AS	0.52513		X15: AS	0.24486
	X16: Employment	0		X16: Employment	0.02359
	X17: Education	0		X17: Education	0.025
Society: 0.02352	X18: Crime	0	**Society: 0.07065**	X18: Crime	0.0417
	X19: Environment	0.02352		X19: Environment	0.02895

Figure16: The result of RT and RF in lockdown2

5 Discussion

5.1 Discussion on the results of the three lockdowns

1）比较每次 lockdown 三种方法的结果（r 方，LR 的系数，RT/RF 的重要性）

2）比较 LR 的三次 lockdown 结果（r 方 系数 p-value）

5.2 Results of advantages and limitations .

问题：

1） LR 还要加些什么？总感觉少些什么
2） Case 走势图+地图分布分析到什么程度
3） Lasso 如何分析
4） Discussion/conlusion 方向

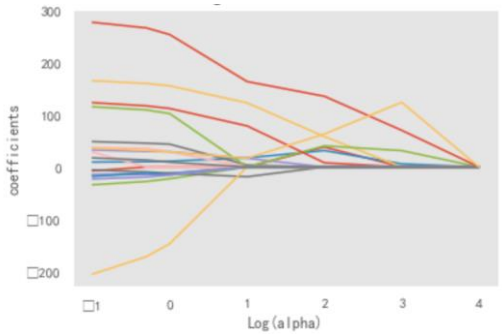
```
Lasso(max_iter=100000000.0, normalize=True)
```

```
lasso_model2.score(X=predictors_data2, y=response_data2)
```

```
0.1311153103004562
```

```
# print(lasso_model.coef_)
df_coef_lasso2 = pd.DataFrame({"var":predictors_data2.columns.v
print(df_coef_lasso2)
```

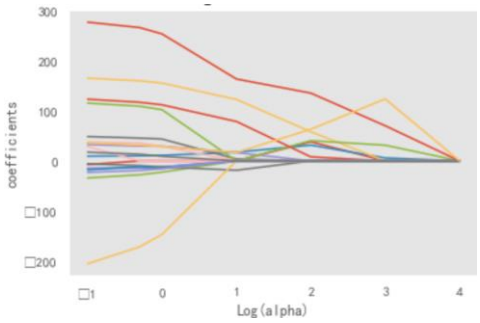
	var	coef
0	NUK	0.0000
1	O65	0.0000
2	Bame	10.9974
3	HP	-0.0000
4	OO	0.0000
5	OML	0.0000
6	SR	-1.9216
7	PR	0.0000
8	NP	0.0000
9	Density	-0.0000
10	Geo	-0.0000
11	IMD	-0.0000
12	Health	-0.0000
13	Income	-0.0000
14	AS	897.4315
15	Employment	-0.0000
16	Education	0.0000
17	Crime	-0.0000
18	Environment	-0.0000




```
0.1311153103004562
```

```
# print(lasso_model.coef_)
df_coef_lasso2 = pd.DataFrame({"var":predictors_data2.colu
print(df_coef_lasso2)
```

	var	coef
0	NUK	0.0000
1	O65	0.0000
2	Bame	10.9974
3	HP	-0.0000
4	OO	0.0000
5	OML	0.0000
6	SR	-1.9216
7	PR	0.0000
8	NP	0.0000
9	Density	-0.0000
10	Geo	-0.0000
11	IMD	-0.0000
12	Health	-0.0000
13	Income	-0.0000
14	AS	897.4315
15	Employment	-0.0000
16	Education	0.0000
17	Crime	-0.0000
18	Environment	-0.0000




```
0.3616163679131458
```

```
# print(lasso_model.coef_)
df_coef_lasso3 = pd.DataFrame({"var":predictors_data3.colum
print(df_coef_lasso3)
```

	var	coef
0	NUK	-5.0536
1	O65	-0.0000
2	Bame	0.0000
3	HP	-0.0000
4	OO	0.0000
5	OML	2.6639
6	SR	-3.9784
7	PR	-0.0000
8	NP	-15.6170
9	Density	-0.0000
10	Geo	-0.0000
11	IMD	-0.0000
12	Health	0.0000
13	Income	-0.0000
14	AS	1,783.1685
15	Employment	-0.0000
16	Education	-0.0000
17	Crime	0.0000
18	Environment	-0.0000