Ecological analysis of Covid case rates

Greater Manchester

Map of Covid case rates by MSOA in Greater Manchester on week ending 4th October 2011


**Cartogram of Covid case rates in Greater Manchester by MSOA**

### Table of Contents

[Table of Contents 2](#_Toc86052840)

[Table of figures 2](#_Toc86052841)

[Spatial dataset 4](#_Toc86052842)

[Introducing GeoDa 7](#_Toc86052843)

[The GeoDa interface 7](#_Toc86052844)

[Mapping the data 8](#_Toc86052845)

[Exploring the data 10](#_Toc86052846)

[Spatial effects 11](#_Toc86052847)

[Exploring spatial clustering and sparsity 12](#_Toc86052848)

[Selecting the independent variables 13](#_Toc86052849)

[Multivariate analysis (linear regression) 14](#_Toc86052850)

### Tables and figures

[Table 1 – data dictionary 4](#_Toc86053004)

[Figure 1 – GeoDa interface 7](#_Toc86053005)

[Figure 2 – GeoDa initial load 8](#_Toc86053006)

[Figure 3 – quantile map 8](#_Toc86053007)

[Figure 4 – Covid case rates in October 2020 by MSOA in Greater Manchester using natural breaks 9](#_Toc86053008)

[Figure 5 – weights file creation – set file ID variable 11](#_Toc86053009)

[Figure 6 – space menu 12](#_Toc86053010)

[Figure 7 – regression model 1 14](#_Toc86053011)

[Figure 8 – classic regression for model 1 15](#_Toc86053012)

[Figure 9 – error regression model 1 16](#_Toc86053013)

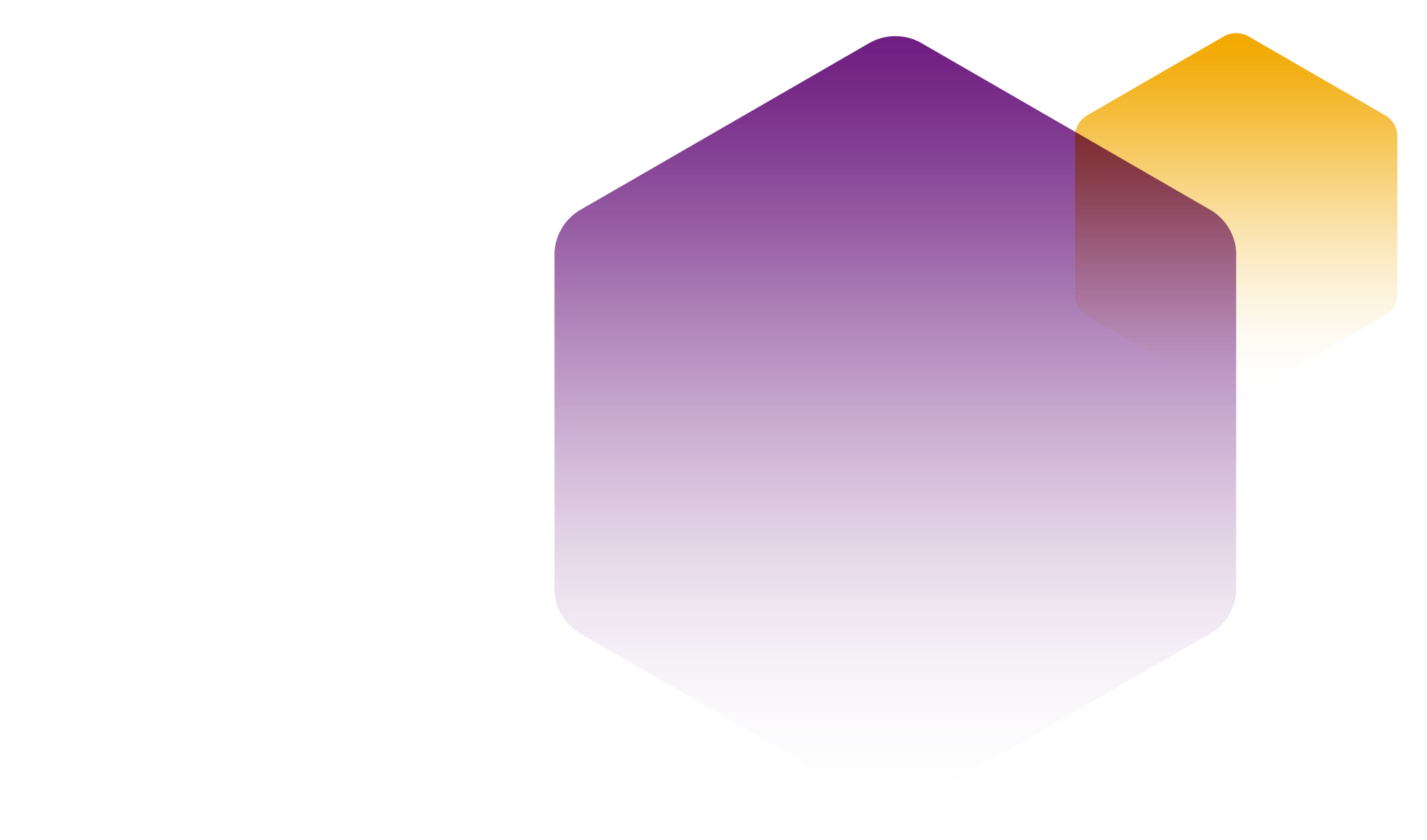
[Figure 10 – classic regression for model 2 17](#_Toc86053014)

[Figure 11 – spatial lag model 2 19](#_Toc86053015)

Introduction

This workbook was developed for the *Placing the Pandemic* session at the National Centre for Research Methods e-festival on 27th October 2021.

The first section explains the dataset used for the workshop, how it was developed and what it represents. The second section introduces the GeoDa package which is used for the spatial analysis and how to produce maps and charts to explore the dataset. The third section covers the steps needed to carry out a linear regression that identifies and takes account of spatial autocorrelation.

­­­

# Spatial dataset

Description

The dataset was prepared using boundary datasets for mid layer super output areas (MSOAs). All variables after the geographical identities and counts of people and households reflect proportions. The Covid rates are calculated as the number of PCR cases testing positive in the first week of each month per 100,000 population. The remaining variables are the proportion of the relevant population in each category. The Covid case numbers were taken from the nationally published figures available at <https://www.coronavirus.data.gov.uk>. The population proportions are taken from 2011 census tables downloaded from [InFuse - access 2011 and 2001 UK Census data (mimas.ac.uk)](http://infuse.mimas.ac.uk/). The contents of the dataset are explained in Table 1.

##### Table 1 – data dictionary

| **Variable name** | **Description** |
| --- | --- |
| code | MSOA code |
| id | MSOA ID |
| name | MSOA name |
| la | Local authority |
| people | No of people |
| households | No of households |
| *Covid rates per 100,000 for each MSOA in the first week of each month* | |
| crateoct20 | Covid rate October 20 |
| cratenov20 | Covid rate November 20 |
| cratedec20 | Covid rate December 20 |
| cratejan21 | Covid rate January 21 |
| cratefeb21` | Covid rate February 21 |
| cratemar21 | Covid rate March 21 |
| crateapr21 | Covid rate April 21 |
| cratemay21 | Covid rate May 21 |
| cratejun21 | Covid rate June 21 |
| cratejul21 | Covid rate July 21 |
| crateaug21 | Covid rate August 21 |
| cratesep21 | Covid rate September 21 |
| crateoct21 | Covid rate October 21 |
| *Proportion in each age group* | |
| prage0-14 | Proportion aged 0-14 |
| prage15-29 | Proportion aged 15-29 |
| prage30-44 | Proportion aged 30-44 |
| prage45=59 | Proportion aged 45-59 |
| prage60-74 | Proportion aged 60-74 |
| prage75 | Proportion aged 75 or over |
| *Overcrowding indicator based on room standard – proportion overcrowded counted as (no households overcrowded by 1 room + (2\*no households overcrowded by two rooms))/no households in MSOA* | |
| ocind | Overcrowding indicator |
| *Occupational group* | |
| working | Working age population (proportion of total population) |
| health | Health or social care occupation |
| education | Education occupation |
| bluelight | Emergency services occupation |
| *Occupational social class* | |
| nssecprof | Professional |
| nssecint | Intermediate |
| nssecrou | Routine |
| nssecnw | No occupational social class |
| *Students* | |
| prstu | Students |
| *Ethnic group* | |
| prethwb | White British |
| prethwiri | White Irish |
| prethwoth | White other |
| prethmixwbc | Mixed white and black Caribbean |
| prethmixwba | Mixed white and black African |
| prethmixwa | Mixed white and Asian |
| prethmixoth | Mixed other |
| prethind | Indian |
| prethpak | Pakistani |
| prethban | Bangladeshi |
| prethchi | Chinese |
| prethasoth | Asian other |
| prethbafr | Black African |
| prethbcar | Black Caribbean |
| prethboth | Black other |
| prethoth | Other ethnic minority |
| *Mobility based on address 12 months ago* | |
| mobsame | Same address as 12 months ago |
| mobla | Moved in same area |
| mobuk | Moved within UK |
| mobint | Moved from overseas |
| mobout | Moved out of area |
| *Tenure of property* | |
| tenowno | Owned outright |
| tenownms | Owned with a mortgage or shared ownership |
| tensrent | Social rented |
| tenprent | Private rented |
| tenfree | Living rent free |

# Introducing GeoDa

A free spatial analysis package developed by the University of Chicago

GeoDa supports a range of spatial analysis – univariate, bivariate and multivariate. Note that when looking at variables with multiple categories you will need to include them in your preparatory and final analysis. This will be illustrated as the program is introduced.

You need to install the GeoDa program, download the dataset and the boundary files from the links provided below and put them in a folder which you can access.

## The GeoDa interface

GeoDa provides a menu interface – this guidance is based on version 1.4.6. There might be some changes to the way the screens appear depending on the version you are running. GeoDa opens a small window for each action you ask it to carry out. You can maximise or change the size of these if they are not clear.

When you start GeoDa a small toolbar is displayed.

##### Figure 1 – GeoDa interface

Graphical user interface, text, application, Word

Description automatically generated

The first task is to load the boundary data into GeoDa. Select <File><Open Shapefile> and find the boundary file you want to use. In the examples I have used *GM\_MSOA\_2011\_ctgm* which holds the cartograms adjusted to reflect the number of people living in an area. Once you have done this a second window shows a map as shown in Figure 2.

##### Figure 2 – GeoDa initial load

Graphical user interface, text, application, Word

Description automatically generated

## Mapping the data

The map menu command allows you to visualise the variables in the dataset using a number of different methods of generating the category definitions. In the first example we will look at a map based on quantiles – four equal sized groups.

Select <Map><Quantile Map> and <Click>. You can change the number of groups at this stage. Stay with four and then you need to select the variable to map. Use the Covid rate for October 2020 *crateoct20* and a map like the one in Figure 3 should be displayed.

##### Figure 3 – quantile map

Graphical user interface, application, Word

Description automatically generated

The map shows Covid case rates per 100,000 people by MSOA in Greater Manchester together with the values and the number of areas.

The use of equal numbers of areas (quantiles) means that the difference in values between one level and another may be very close. An alternative approach is to use <Natural Breaks> which tries to identify real differences. Try using this approach for *crateoct20* with four categories. This produces a significantly different picture of the distribution of Covid case rates as shown in Figure 4.

##### Figure 4 – Covid case rates in October 2020 by MSOA in Greater Manchester using natural breaks

Graphical user interface, application, Word

Description automatically generated

The category ranges have changed significantly and you can see the effect on the apparent distribution with the majority between 103 and 520 cases per 100,000 and one area with rates of over 4,000 cases per 100,000.

**Note:** GeoDa allows us to quickly explore data and make decisions about which variables to include but for most publication purposes you are advised to use a GIS package where you have more control over the quality of the output.

## Exploring the data

The <Explore> menu allows you to produce univariate and bivariate analysis of the variables. The histogram and box plot allow you to see the distribution of the numeric variables. Have a look at the distribution of Covid rates for two different months. What do they suggest?

You can also look at the relationship between two variables. Have a look at the relationship between Covid rates in Oct 2020 and the proportion of students in the population. What does this chart suggest?

# Spatial effects

We are using GeoDa to allow us to account for spatial effects – to identify the extent to which places around the areas we are examining share similar characteristics. To do this we need to calculate values for variables for each output area and its surrounding areas. GeoDa allows us to calculate these neighbourhood weights. Select <Tools><Weights><Create> and the following window are displayed

##### Figure 5 – weights file creation – set file ID variable

Graphical user interface, text, application

Description automatically generated

Click <Add ID Variable> and select <OK> in the popup window which suggests the name for the new variable. Now set the way the weights are calculated by checking the Queen contiguity box. Click the <Create> button and the weights file will be created. You are asked to save it – this means that in future you can load the file without going through these steps.

## Exploring spatial clustering and sparsity

The creation of the weights file allows us to produce maps which show where specific values are clustered.

Select the top menu <Space> which gives a number of options as shown in Figure 6.

##### Figure 6 – space menu

Graphical user interface, text, application, Word

Description automatically generated

Select <Univariate Moran’s I> and then select one of the Covid rate variables. A scatterplot of the value of the variable in each areal unit against the value of the variable for that areal unit and those around it together with a value for Moran’s I is displayed. What does this suggest?

*Hint: From the scatterplot is there a relationship? Is it positive or negative? From the value of Moran’s I (possible values -1 to 1) How strong is the relationship?*

We can also look at the local indicators of spatial association. Select <Univariate Local Moran’s I>. You can now choose three types of optional output

**A cluster map** shows clustering of high (red) and low (blue) areal units for this variable Pink areas are high value areas surrounded by low value areas, light blue areas are low value areas surrounded by high values. Other areal units are left white.

**A significance map** shows the statistical significance of the areal units shaded on the cluster map.

**A Moran scatterplot** shows the value of the variable in each areal unit against the value of the variable for that areal unit and those around it together with a value for Moran’s I

What do these maps and chart suggest?

## Selecting the independent variables

You can explore the descriptive statistics for potential independent variables using the techniques outlined previously. In the examples we will look at two models.

1. Coronavirus rate in October 2020 ~ age + proportion students
2. Coronavirus rate in September 2021 ~ age + proportion students + tenure

Have a look at the dependent and independent variables in the 2 models.

## Multivariate analysis (linear regression)

**Coronavirus rate in October 2020 ~ age + proportion students**

To create the first model we use <Methods><Regression> and then select the dependent and independent variables. For the age category we need to omit one of the variables which then acts as a reference category. In the first example I have omitted age 30-44. I have also included the weights file created earlier. Figure 7 shows the selection window.

##### Figure 7 – regression model 1

Graphical user interface, text, application

Description automatically generated

First run the Classic model (click <Run>) and the output as shown in Figure 8 is shown.

##### Figure 8 – classic regression for model 1

Graphical user interface, text, application

Description automatically generated

Adjusted r2=0.63

For each 1% increase in coefficient

Case rate = Case rate + (effect / 100)

Significance

Lag and error diagnostics significant – check robust diagnostics. Suggests an error model

The error model can be selected by clicking on the <error> button in the Regression window. Figure 9 shows the results. Have a look at the change to the adjusted r2 value and the coefficients.

##### Figure 9 – error regression model 1

Graphical user interface, text, application, email

Description automatically generated

**Coronavirus rate in September 2021 ~ age + proportion students + tenure**

To start a new model, press the <Reset> button and put in your dependent and independent variables. For the age and tenure categories we need to omit one of the variables which then acts as a reference category. I have age 30-44 and owns outright. I have also included the weights file created earlier. Figure 10 shows the output from the classic regression.

##### Figure 10 – classic regression for model 2

Graphical user interface, application, Word

Description automatically generated

Have a look at the adjusted r2, coefficients and significance and note what the model suggests.

The Lagrange multiplier diagnostics show that both lag and error are significant, but the robust tests suggest we should use spatial lag to take account of spatial autocorrelation. You can do this by selecting the <Lag> button and clicking <Run>. The output is shown in figure 10.

##### Figure 11 – spatial lag model 2

Graphical user interface, application, Word

Description automatically generated

Have a look at how the r2, coefficient and significances have changed by taking account of spatial lag.

Thank you very much

Nigel de Noronha [nigel.denoronha@manchester.ac.uk](mailto:nigel.denoronha@manchester.ac.uk)

Dave Rawnsley david.rawnsley@jisc.ac.uk

[www.ukdataservice.ac.uk](http://www.ukdataservice.ac.uk)

[help@ukdataservice.ac.uk](mailto:help@ukdataservice.ac.uk)

+44 (0) 1206 872143

We are supported by the University of Essex, the University of Manchester, UKRI through the Economic and Social Research Council, and Jisc.

32%

Text box

Velloria estiorpor modigendae dolum aspici nestrumetur, sim quiditatur mos aut eaqui sapid quibusa ndessi voluptaqui tet rerupta dellitatur sed ullamen duciis exeratur ad ea.

32%

Text box

Velloria estiorpor modigendae dolum aspici nestrumetur, sim quiditatur mos aut eaqui sapid quibusa ndessi voluptaqui tet rerupta dellitatur sed ullamen duciis exeratur ad ea.