COVID 19 – Socio-Economic Factors in the UK

Thomas Fishwick

**Abstract**—Put here a brief summary of your work: analysis task, data, approach, main findings. Length: up to 200 words.

# Problem Statement (250)

Here we will be looking at the COVID-19 statistics and comparing the spread of the virus in different areas and using the last census data to try to understand the various factors behind the spread of the virus.

To solve this problem, we have the COVID 19 case, death and Vaccine rates by UK region [2]. The ONS estimated age breakdown by region (as of August 2021). COVID 19 cases by age and region. A portion of the 2011 Census data showing the shared/unshared dwellings, number of cars, long term health and household deprivation [1]. Ideally, we would be using the 2021 census data, but it will not be released until 2023. We also have the geographic boundaries of the UK Local governments [11], so that we can plot this data onto a map.

We will look at the overall trends of the virus spread. Then look at various local risk factors and how they may interact with the spread, mortality and vaccination rates. Then look at whether we can make a predictive model for the virus spread at the Local authority level.

# State of the Art (500)

In the Office of National Statistics paper [8] the authors look at the breakdowns of the COVID 19 deaths by different ethnic groups and gender. The article links individuals’ census and NHS records (patient register and pandemic planning dataset) and looks at other health conditions the individuals might have. The authors were looking to get a risk factor for different ethnic groups for how likely they are to die of COVID 19. As we will not have access to census or medical records at the individual level, we can use their raw results or just use the ethic group data at the local authority level. They also break the model down by local authority district to account for any geographic variation. From this paper we have learnt that there are links between ethnic groups and COVID 19 mortality and we have a risk factor for these groups. They use various visualisations to show the differing rates of death between different ethnic groups. The paper makes some assumptions in linking NHS records to census data where there was not a direct link. For the district level data, they assume that the 2011 census records are still relevant.

Within this paper [9] from IEEE, the authors look at clustering US counties by various socio-economic factors and building time series forecasting models. They use various visualisations in their approach to decide upon how many clusters to use. They used GDP data and population breakdowns for US counties along with infection data. To cluster the counties, they used the k-means algorithm. They then compared ARIMA (auto-regressive integrated moving average) against Seasonal Trend Random Walk models to see which performed better, concluding that ARIMA was better. Their data is not usable due to being the wrong country, but their methods of normalising age and other data is very useable. One of their assumptions was that the 2019 socio-economic data could be used.

From one of SAGE’s reference papers [10] from February 2020 the authors were attempting to make a mathematical model to predict potential spreads of COVID 19 at the electoral ward level, based upon an earlier theoretical model. The authors use the 2011 census data and early infection data from the UK and China to predict potential infection rates in different parts of England and Wales. As almost two years’ worth of actual data is now available, we can use that rather than theoretical spread data. The authors have generated various visualisations of the virus transmission data. The authors separated the data out into the main regions of England and Wales, so we can see that our idea of looking at the lower-level data is valid. This paper assumed that COVID 19 would behave much the same as the theoretical virus merged with the early infection data.

# Properties of the Data (500)

Our COVID-19 case data [2] is at the local authority level (LTLA) from the results of PCR tests and positive lateral flow tests (which are reported, from 21/10/20). This data is collected from the various local authorities and then checked and published by Public Health England. The data itself is 8 columns by 244,442 rows from 13/3/20 to 28/12/21, with at least one row per day. There is an issue with the data for 1/7/20, which appears to be a correction for earlier data points. Generally, the aggregated local authority data matches the UK wide data, but does not face the same level of scrutiny as the UK wide data (being presented by the Prime Minister).

The data we have from the census [1] is at the same level, but some of the councils have been merged or split apart. Using Excel [4] we investigated the differences in councils, for the new merged councils we summed together their census data and for those splitting apart (Suffolk) we divided the data equally between them. We used Python’s Pandas library [5] to join up all of the census tables into one sheet. The census data was collected through questionnaires presented to every household in the UK. These were then aggregated by the Office of National Statistics. This dataset has 95 columns and 343 rows.

Finally, we have the local authority boundary data for 2020 from [11], which is from the ONS. This dataset has the various local authorities and their geographic properties.

The three datasets are joined together by their geography code. By plotting the cases over time using Tableau [12], we could see any obvious anomalies in the case data, such as the giant peak on 1/7/20 which isn’t replicated in the UK wide data (plotted underneath the data, in figure 1).



Fig. 1-District versus UK wide cases

On a similar plot (fig 2) we can see that some individual districts have anomalous spikes in cases (either due to not releasing figures over the weekend or reporting delays). We can see that there is some variation between the different districts in that some of them have different peaks to the others.



Fig. 2- District cases

By looking at the data in Excel we can see many days with zero cases for regions (27k in total) and other cells with null values.

To account for the noise in the case data we could move from a daily resolution to a weekly resolution.

# Analysis

## Approach (500)

To solve the problem of predicting COVID-19 cases at the district level we will be following the process set out in Figure 3.

Our first step will be obtaining the data and checking it for obvious discrepancies, using Tableau for quick visualisations.

As the districts have changed between the dates of the COVID-19 data and the Census data human reasoning is required to match up the merged and split counties between the datasets. The human needs to be presented with both lists of countries and where the computer cannot make a one-to-one join between the two datasets (using Excel’s vlookup function), the human needs to reason about what to do to make a join.

The next step is to create a calculated variable of the relative total number of cases. This is done by taking the total population (number of people 0-59 plus 60+) and dividing the sum of a region’s cases by that. This is then used as the y value in an array of Seaborn [13] regression plots with the various census columns used as the x values. The human analyst then needs to reason about which census columns to use further in the analysis based upon these and their correlation statistics. This is so that sensible columns are picked to help with the forecasting.

Once the relevant variables have been chosen the regions will be clustered based upon these factors. The human analyst will then have to decide upon the optimal number of clusters by using the Yellowbrick [14] silhouette plot, the seaborn plot of the clustered data, clusters on the geographic map and the silhouette scores for a range of cluster values. This is because although the computer could make this choice for the human analyst based upon the score, there is a balance to be struck between the number of clusters, the explanation of them and a sensible looking split.

The data will then need to be run through the Dicky-Fuller test to see if it is stationary or non-stationary time-series data so that the appropriate ARIMA model can be decided upon using Stats-models [15] AdFuller method.

An ARIMA model will then be constructed and tested upon the last few time periods of the data. This will be contrasted with an XGB (extreme gradient boost) regression model to see which is the better model. The human analyst will be deciding upon this using visualisation of both models predicted data against the actual data.

Finally, we will be predicting future cases and plotting these on the geographic map, then unpacking the clusters into regional predictions.

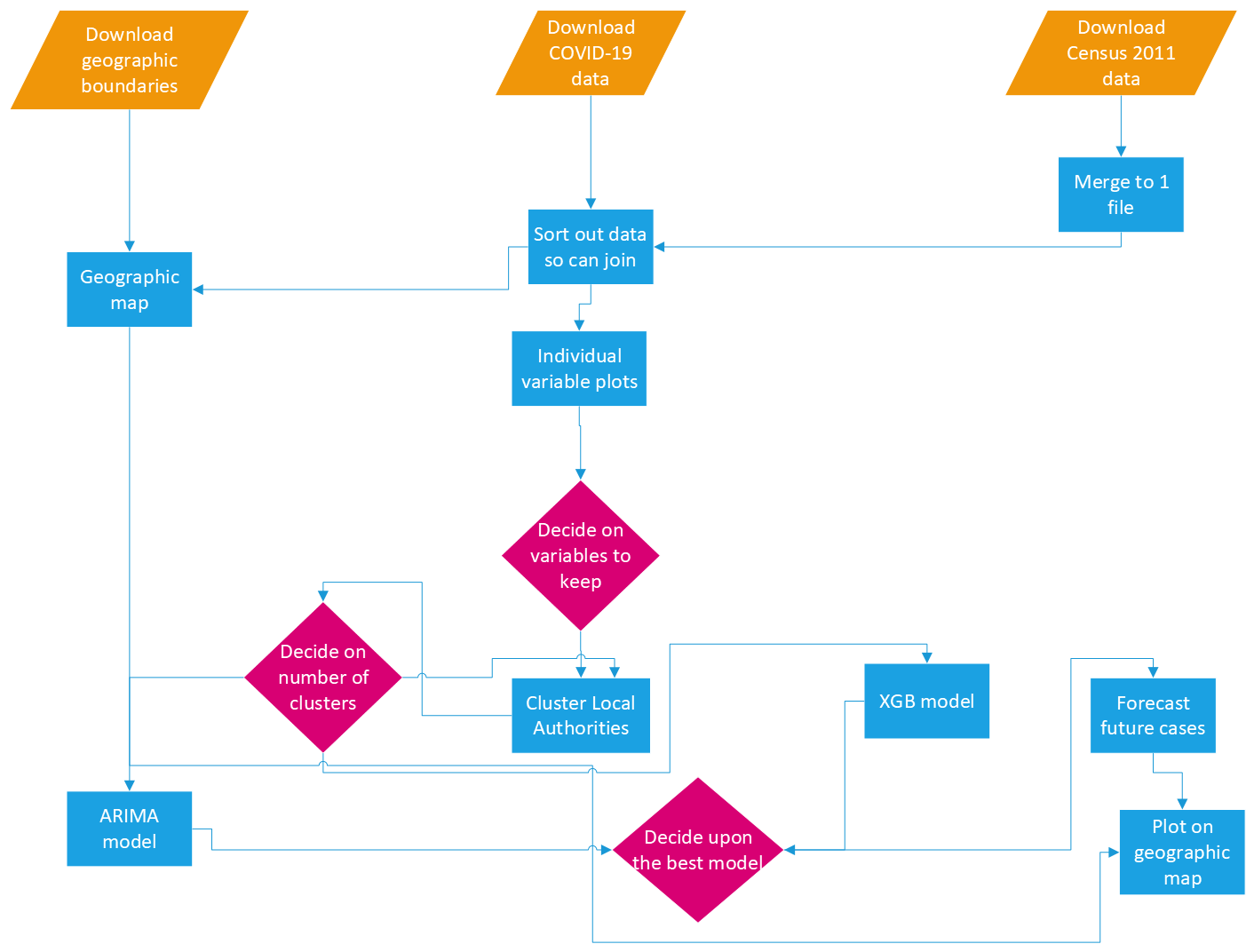


Fig. -Approach flowchart

## Process (1500)

For the process of looking at the various columns and how they interact with the relative number of cases, it was very quickly noticed that there was a problem. Almost every number no matter what it was, when it increased it meant that there was a higher relative number of cases. As unfortunately the total population was a confounder, as there were more people to catch and spread COVID-19. Dividing each column by the total population and then running it through seaborn’s regression plot function we could see the actual effect of that variable. Having the computer make the decision on the correlation between one of the variables and the relative number of cases would be a quick way of doing things, but the correlation figure can be easily misled by a few extreme points of data. Seeing the data for ourselves allows for more control. Interesting preliminary results from this analysis showed that the relative number of people in each ethnic group appeared to have no effect on the relative number of cases. Unsurprisingly the more people with lower health rankings the higher the number of cases. Those areas with more people with no qualifications had higher COVID-19 rates, but only the level 4 qualifications had an effect of lowering the rates of COVID-19. For the age groups there was the surprising effect that the younger the relative population the higher the rate of COVID-19, but this effect gradually settled down and then started reversing at 45-49 (or 55-59 if you account for the fact that the census is 10 years old).

In the process of clustering up the data, which we did as a way of reducing the total number of dimensions for the overall model, the human analyst had to decide upon the total number of dimensions. We filtered out Scotland and Northern Ireland as we do not have census data for them. Using Sci-Kit Learn’s [16] K-Means algorithm “which seeks an optimal partition of the data by minimising the sum-of-squared-error criterion” [17], we began clustering up the data. We used the relative numbers for our chosen interesting columns (in the appendix), except for the two population figures where we used the absolute numbers. So that the human analyst didn’t have to look at all 333 potential numbers of clusters (which would have been rather time consuming), the computer was set the task of looking at each number and calculating the average silhouette score, these were then put into a graph so that the analyst could look at the regions with higher scores (Fig 4). As you can see from the graph below 10 clusters seems to be preferable, then any number up to 100 is not too bad and then after that the performance got steadily worse.



Fig 4. Potential number of clusters versus average silhouette score

So that the human analyst could make the final decision we used Yellowbrick’s [14] silhouette visualizer and GeoPandas [18] to show the cluster’s scores and the clusters’ locations on the map. A balance needed to be struck between using too many clusters and them becoming meaningless and using too few clusters and them not being able to capture the essence of the underlying data. In the end we decided upon six clusters. Its average score is 48% and the area seem to be split out between very rural, rural, suburban, urban and somewhere in between rural and suburban (plus another cluster invisible on the map and silhouette plot) as you can see in figure 5.

For the process of running the Dicky-Fuller test (used to test if time-series data is stationary or non-stationary (tending around a number or unbound)), we used statsmodels [15] adfuller test. From the test statistics we can see that cluster 2 is the most stationary and 3 is the least stationary. From the p-values we can accept that the data is stationary. As the case data (as of 28/12/21) is still trending upwards as we can see in figure 6 (and we are assuming that they will fall at some point) this is likely throwing off the test statistics.

Fig. 5-Clustering

For the process of running the Partial Autocorrelation Function and Autocorrelation Function graphs using the relevant statsmodel functions. We get a series of graphs for the clusters, taking the first difference we can see that the ACF graph cuts off after the first lag and the PACF tails off. Using the pmdarima [19] package we can run through all of the time series to confirm our analysis, the computer disagrees for clusters 2 and 3, but their numbers are not that different to the others. So, we will be disregarding the computer’s answer.

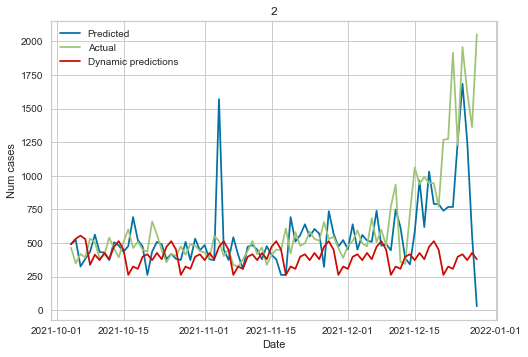
In making the ARIMA models it does not appear that they are able to generalise to be able the test data (4/10/21 – 28/12/21). Effectively deleting the erroneous records for 1/7/20 improves upon the fit, but the model failed to predict anything other than a straight line for the number of cases. Using the weekly data does not improve this and the residual fit is only moderately improved by getting rid of the over-correction on the 1/7/20’s case data. So, it appears that the ARIMA model is not appropriate for predicting the number of new cases.

Fig. 7-XGB Cluster 2

Fig. 6-Cluster Time-Series

In making the XGB regression model (figure 7) we can that for cluster 2 (the best model) that using the actual previous day’s case data the XGB model tracks fairly close to the actual data. But when the model uses its own previous prediction as input data it very rapidly loses track with the actual data. The other clusters generally predict an almost straight line for the number of cases.

## Results (200)

First paragraph...

Following paragraphs...

*<200 words, <=2 images*

# Critical reflection (500)

First paragraph...

Following paragraphs...

*<500 words*

Table of word counts

|  |  |
| --- | --- |
| Problem statement | 183 |
| State of the art | 463 |
| Properties of the data | 387 |
| Analysis: Approach | 428 |
| Analysis: Process | 1500 |
| Analysis: Results | 200 |
| Critical reflection | 500 |

References

The list below provides examples of formatting references.

1. Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency (2017): 2011 Census aggregate data. UK Data Service (Edition: February 2017). DOI: http://dx.doi.org/10.5257/census/aggregate-2011-2
2. Gov.UK Coronavirus. ‘Cases in the UK | Coronavirus in the UK’. HTML, 2021. <https://coronavirus.data.gov.uk/details/cases?areaType=overview&areaName=United%20Kingdom>.
3. Office for National Statistics (2011). 2011 Census: boundary data (England and Wales) [data collection]. UK Data Service. SN:5819 UKBORDERS: Digitised Boundary Data, 1840- and Postcode Directories, 1980-. <http://discover.ukdataservice.ac.uk/catalogue/?sn=5819&type=Data%20catalogue>, Retrieved from <http://census.ukdataservice.ac.uk/get-data/boundary-data.aspx>. Contains public sector information licensed under the Open Government Licence v3.
4. ‘Microsoft Excel Spreadsheet Software | Microsoft 365’. Accessed 30 December 2021. <https://www.microsoft.com/en-us/microsoft-365/excel>.
5. ‘Pandas - Python Data Analysis Library’. Accessed 30 December 2021. <https://pandas.pydata.org/>.
6. Hilton, Joe, and Matt J. Keeling. ‘Estimation of Country-Level Basic Reproductive Ratios for Novel Coronavirus (COVID-19) Using Synthetic Contact Matrices’, 27 February 2020. <https://doi.org/10.1101/2020.02.26.20028167>.
7. Klepac, Petra, Adam J. Kucharski, Andrew JK Conlan, Stephen Kissler, Maria L. Tang, Hannah Fry, and Julia R. Gog. ‘Contacts in Context: Large-Scale Setting-Specific Social Mixing Matrices from the BBC Pandemic Project’, 5 March 2020. <https://doi.org/10.1101/2020.02.16.20023754>.
8. Larsen, Tim, Matt Bosworth, and Vahé Nafilyan. ‘Updating Ethnic Contrasts in Deaths Involving the Coronavirus (COVID-19), England: 24 January 2020 to 31 March 2021’. Updating ethnic contrasts in deaths involving the coronavirus (COVID-19), England - Office of National Statistics, 2021. <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/updatingethniccontrastsindeathsinvolvingthecoronaviruscovid19englandandwales/24january2020to31march2021>.
9. Lucic, Michael C., Hakim Ghazzai, Carlo Lipizzi, and Yehia Massoud. ‘Integrating County-Level Socioeconomic Data for COVID-19 Forecasting in the United States’. IEEE Open Journal of Engineering in Medicine and Biology 2 (2021): 235–48. <https://doi.org/10.1109/OJEMB.2021.3096135>.
10. Danon, Leon, Ellen Brooks-Pollock, Mick Bailey, and Matt Keeling. ‘A Spatial Model of CoVID-19 Transmission in England and Wales: Early Spread and Peak Timing’, 14 February 2020. <https://doi.org/10.1101/2020.02.12.20022566>.
11. ONS. ‘Local Authority Districts (May 2021) UK BFE’. Accessed 1 January 2022. <https://geoportal.statistics.gov.uk/datasets/ons::local-authority-districts-may-2021-uk-bfe/about>.
12. Tableau Software. ‘Tableau: Business Intelligence and Analytics Software’. Tableau. Accessed 1 January 2022. <https://www.tableau.com/node/62770>.
13. Waskom, Michael. ‘Seaborn: Statistical Data Visualization’. Journal of Open Source Software 6, no. 60 (6 April 2021): 3021. <https://doi.org/10.21105/joss.03021>.
14. scikit-yb developers. ‘Yellowbrick: Machine Learning Visualization — Yellowbrick v1.3.Post1 Documentation’, 2019. <https://www.scikit-yb.org/en/latest/index.html>.
15. Perktold, Josef, Skipper Seabold, and Jonathon Taylor. ‘Introduction — Statsmodels’, 2019. <https://www.statsmodels.org/dev/index.html>.
16. Cournapeau, David, Matthieu Brucher, Fabian Pedregosa, Gael Varoquaux, Gramfort Alexandre, and Vincent Michel. ‘Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.0.2 Documentation’, 2010. <https://scikit-learn.org/stable/>.
17. Xu, Rui, and Donald Wunsch II. ‘Partitional Clustering’. In Clustering, 63–110. 68: John Wiley & Sons, Ltd, 2008. <https://doi.org/10.1002/9780470382776.ch4>.
18. Van den Bossche, Joris, Nick Eubank, Kelsey Jordahl, Martin Fleischmann, James McBride, Chris Holdgraf, and Philipp Kats. ‘GeoPandas 0.8.2 — GeoPandas 0.8.2 Documentation’. Accessed 4 January 2022. <https://geopandas.org/en/v0.8.2/index.html>.
19. Smith, Taylor. Pmdarima: Python’s Forecast::Auto.Arima Equivalent (version 1.8.4). MacOS, Microsoft :: Windows, POSIX, Unix, C, Python, 2017. <http://alkaline-ml.com/pmdarima>.

**Appendix**

Interesting columns:

* economic\_activity\_economically\_active\_unemployed
* economic\_activity\_economically\_active\_full-time\_student
* economic\_activity\_economically\_inactive\_total
* general\_health\_bad\_health
* general\_health\_very\_bad\_health
* qualification\_no\_qualifications
* qualification\_level\_4\_qualifications\_and\_above
* residence\_type\_communal\_establishments\_with\_persons\_sleeping\_rough
* 60+
* 00\_59
* method\_of\_travel\_to\_work\_work\_mainly\_at\_or\_from\_home
* method\_of\_travel\_to\_work\_taxi
* method\_of\_travel\_to\_work\_bus\_minibus\_or\_coach
* method\_of\_travel\_to\_work\_motorcycle\_scooter\_or\_moped
* method\_of\_travel\_to\_work\_passenger\_in\_a\_car\_or\_van
* method\_of\_travel\_to\_work\_on\_foot