COVID 19 – Socio-Economic Factors in the UK

Thomas Fishwick

Abstract— In this report we look at the UK COVID-19 data at the local authority level and join it to the 2011 census data at the same level, along with geographic data at that level. We do this to see which, if any, factors influence COVID-19 cases, deaths and vaccine rates. We use a mixture of Seaborn plots in Python and Tableau graphs to analyse the data. We also try clustering the local authorities by the census data and then building ARIMA and XGB Regression models. Both models fail to map onto the underlying structure of the data, with ARIMA being worse. For the socio-economic factors which influence your COVID-19 infection likelihood we conclude that underlying health conditions, being younger and being economically inactive made you more likely to catch COVID-19, while the further past 70 you were the more unlikely you were to catch it. For the death data we conclude that there is not really enough data to pass an opinion. For the vaccination rate data there is a small effect between car ownership and vaccination rates, with areas with less car ownership having lower rates and higher car ownership areas having higher vaccine rates.

1 Problem Statement

Here we will be looking at the UK's 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 and its impact.

To solve this problem, we have the COVID-19 case, death and vaccine rates by UK region [2] from UKHSA which is the official source of COVID-19 data in the UK. The ONS estimated age breakdown by region (as of August 2021) [30], which is used as it is based upon the census data and estimated as to what has happened in the ten years between the census and 2021, as otherwise we would have to age the census data ourselves. COVID-19 cases by age and region. A portion of the 2011 Census data for England and Wales showing the shared/unshared dwellings, number of cars, long term health, ethnic breakdown, method travel to work, qualifications and residence type [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 all of 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.

2 State of the Art

In the Office of National Statistics paper [8] the authors look at the breakdowns of the COVID-19 deaths by different ethnic groups and by gender. The article links individuals' census and NHS records together (patient register and pandemic planning dataset, which are not publicly available) and looks at other health conditions the individuals might have. The authors were looking to get a risk factor for different ethnic groups indicating 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 as it links on the NHS number and date of birth. Those with a missing/invalid NHS number or date of birth were dropped from the analysis. Those individuals not present in the UK at the time of the 2011 census were also excluded. 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. We could not use their data as it was based on the USA's figures, 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] released in 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 their 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, supporting our idea of looking at the lower-level data. This paper assumed that COVID-19 would behave much the same as the theoretical virus merged with the early infection data.

3 Properties of the Data

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 UK Health Security Agency. 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 case data has the new cases by that date, new deaths within 28 days of a positive test by that date, new vaccine doses been given, the date and area code. The UK wide data has the total number of COVID-19 hospital cases at that time, but this figure is not available at the lower level.

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 local authorities. For the new merged districts, 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. Some of the columns are measured by the number of households and some of them are measured by the number of people (or at least those who replied to the census).

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

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).

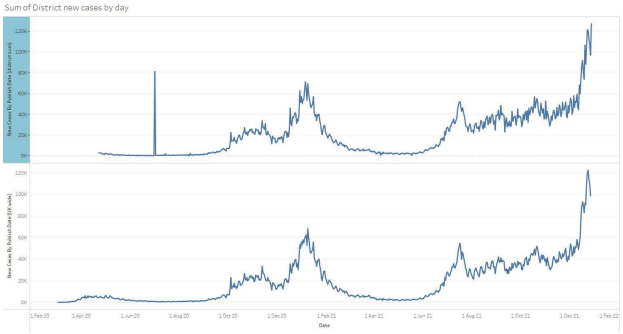

Fig. 1-District versus UK wide cases

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 others.

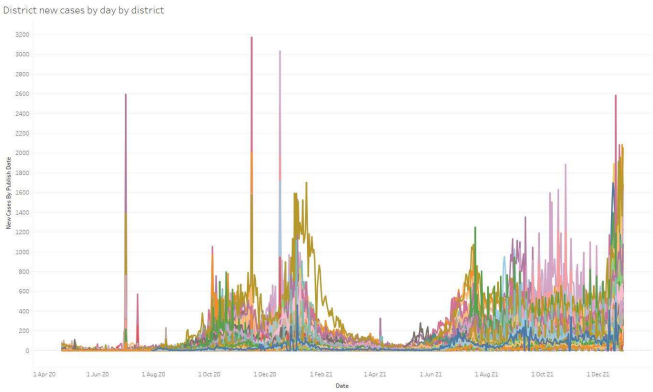

Fig. 2-District cases

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.

4 Analysis

4.1 Approach

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. This is to identify potential problems with the data sets.

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 districts 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. This is so that we can compare the census data to the case data and plot these on a map easily.

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 and seeing which factors affect the COVID-19 rates.

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. This is so that we have an easier to use variable to use in the forecasting process.

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.

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

We will also be looking at how the various COVID-19 statistics change over time by district.

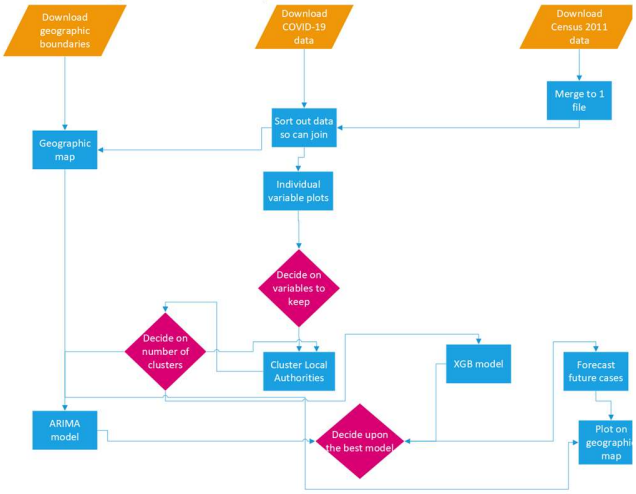

Fig. 3-Approach flowchart

4.2 Process

For the process of looking at the various columns and how they interact with the relative number of cases, we looked at the seaborn regression plot twice. Once for the absolute number of households/people and once for the relevant column divided by the total population. This was so that the effect of population on that variable could be removed. As most districts have similar numbers of people but some have much higher populations these were causing the absolute number graphs to have many values in a cluster on the left and one or two extreme values on the right-hand side causing the graphs to be very hard to read.

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 distorted by a few extreme points of data. Seeing the data for ourselves allows us to determine the level of the effect of the variable on the relative number of cases. 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 years old, with relatively higher numbers of 85-89 and 90+ year olds causing the relative rates to drop a lot more.

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 10 or fewer clusters seems to be preferable, then any number up to 100 is not too bad and then after that the performance got steadily worse as the number of clusters rose.

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. As the data does not trend around a central figure and the test-statistics are high we reject the stationary hypothesis.

For the process of running the Partial Autocorrelation Function and Autocorrelation Function graphs we used 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 meaning that we will using an integrated first order moving average flavour of ARIMA model. 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.

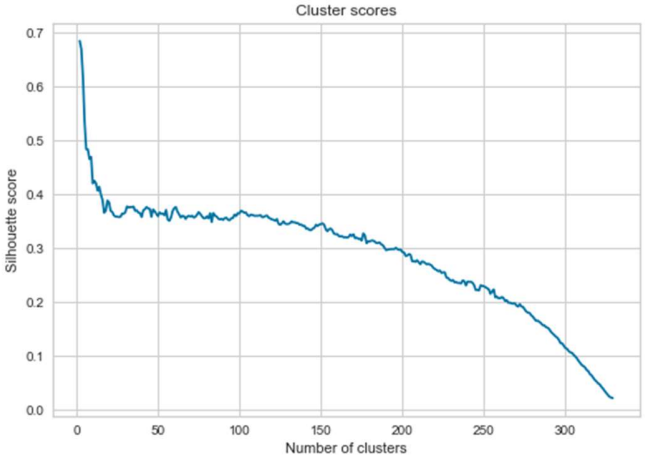

Fig. 4-Potential number of clusters versus average silhouette score

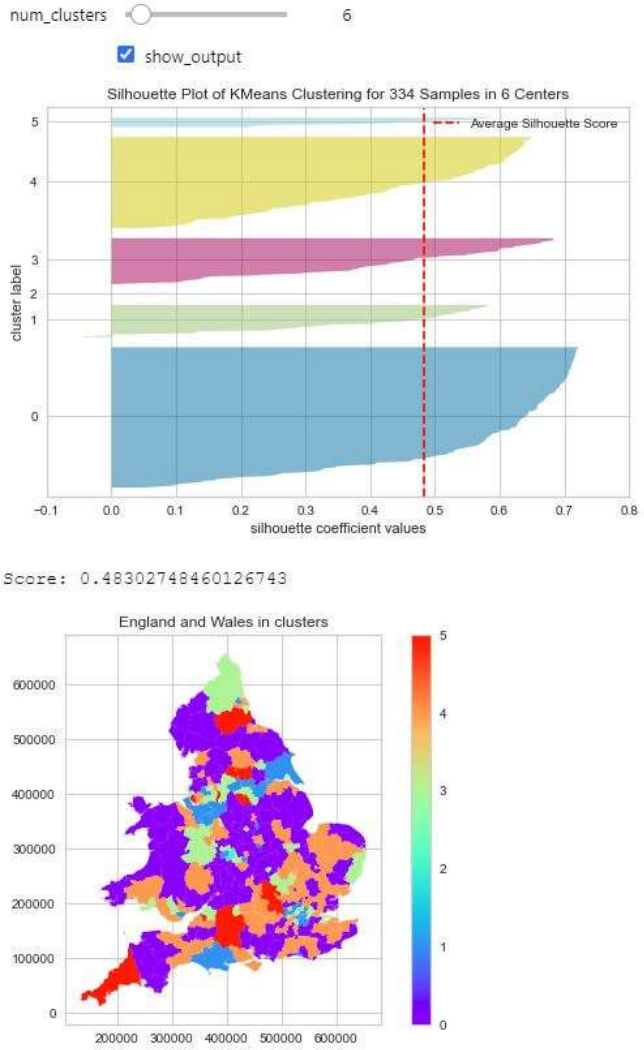

Fig. 5-Clustering

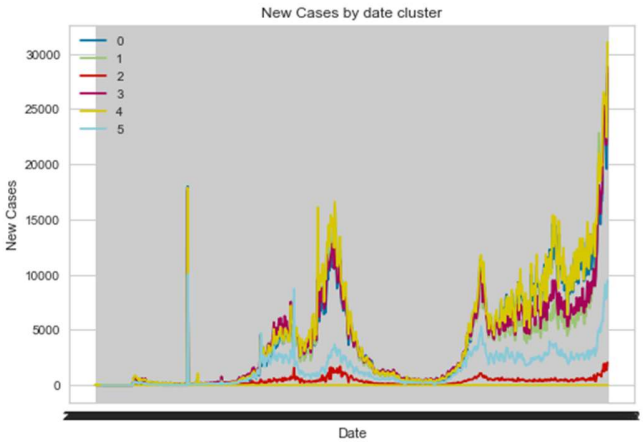

Fig. 6-Cluster Time-Series

In making the ARIMA models it does not appear that they are able to generalise to 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.

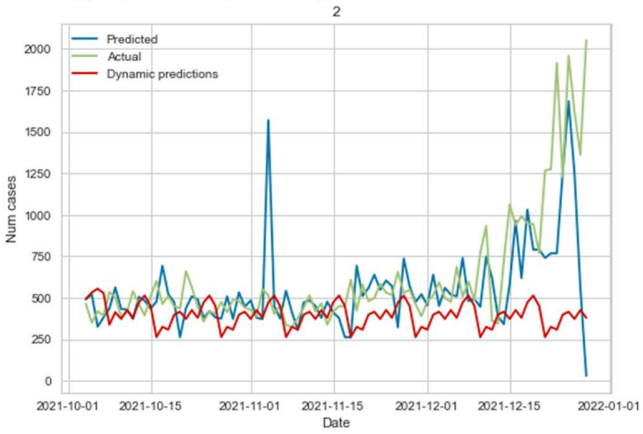

Fig. 7-XGB Cluster 2

In making the XGB regression model (figure 7) using [20] we can see that for cluster 2 (the best model) that using the actual previous day's case data the XGB model tracks fairly closely to the actual data. However, 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. XGB creates decision trees using boosting so that the next tree is trained on a mixture of the last trees mistakes and unseen data.

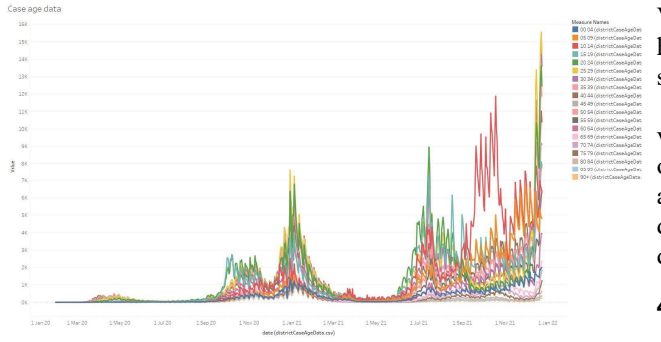

Fig. 8-Case age data

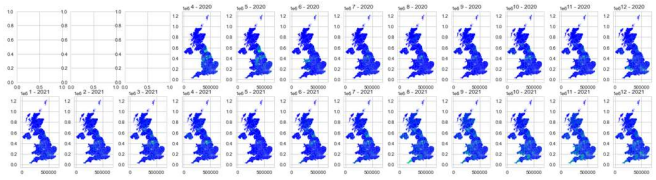

Fig. 9-Monthly Proportional case data

Looking at the case age data we can see that most of the age groups follow the same sort of trend, although from the age groups from 70 and above all have lower numbers than the other age groups (which may be because there are fewer people in those groups, they take more precautions or as they were vaccinated earlier than the younger population, that restricted the third wave's numbers). An outlier group are the 10–14-year-olds, who have a third wave of cases in September 2021 (when secondary schools opened up) and then they have a fourth wave with everyone else's third wave.

Looking at the month-by-month case data by Local Authority (fig. 9) we can see how the proportion of cases has changed between regions month by month (dark blue for lower cases going to green for higher proportional cases (with no data as white)). We can see that Birmingham and the area around Bradford most often have the highest proportion of cases (although the data was not captured for the earliest stages of the pandemic). Northumberland has one of the highest proportions of cases April-May 2020. In August to December 2021 Cornwall and London have the highest COVID case data.

From the map of where respondents mostly reported their health as either bad or very bad, we can see that Birmingham, County Durham, Cornwall, Bradford and Leeds all score very highly on these metrics and these counties also had higher relative COVID-19 rates per month (although not in the same proportions). These areas also have higher overall numbers of people proportionally to other local authorities in England and Wales.

None of the census variables have an especially strong relationship with the death figures, but this may be because the deaths were relatively evenly split by population (there is an abnormally high number of deaths in Birmingham compared to its number of cases (212k cases and 3117 deaths), with Leeds (162k cases and 1604 deaths) at the next highest number having a lower proportion.

For the first and second vaccine dose rates, none of the census variables appear to have any more than a negligible effect on the rate (most of the graphs are relatively flat lines). Very arguably areas with high rates of no cars in a household have lower rates of vaccine take up and one or more car rates show it going up, but these are very slight effects.

For plotting the death percentage against the census variables there is a suggestion that as the relative number of older people rises the death rate increases. This does agree with a lot of the news stories that older people were more likely to die of COVID-19, but we are talking about a half a percent difference in numbers.

4.3 Results

From this report we can see that the most important socio-economic factors involved in the rates of new cases of COVID-19 are age, underlying health conditions and total population in an area. As we can see in fig. 10 a lot of the places in fig. 9 with high COVID-19 rates have high numbers of people with bad health conditions.

We can also conclude that you cannot predict the next day's COVID-19 rates from the previous day or the moving average alone. This may be because these factors alone cannot tell you the direction of the case numbers or because of the different variants of COVID-19 bringing their own new difficulties (differing infection and survival rates).

For the death statistics, these are harder to tie back to census figures due to the fortunately low death rate (2.6% in Rother at the highest, to 0.5% in Moray at the lowest and an average of about 1%).

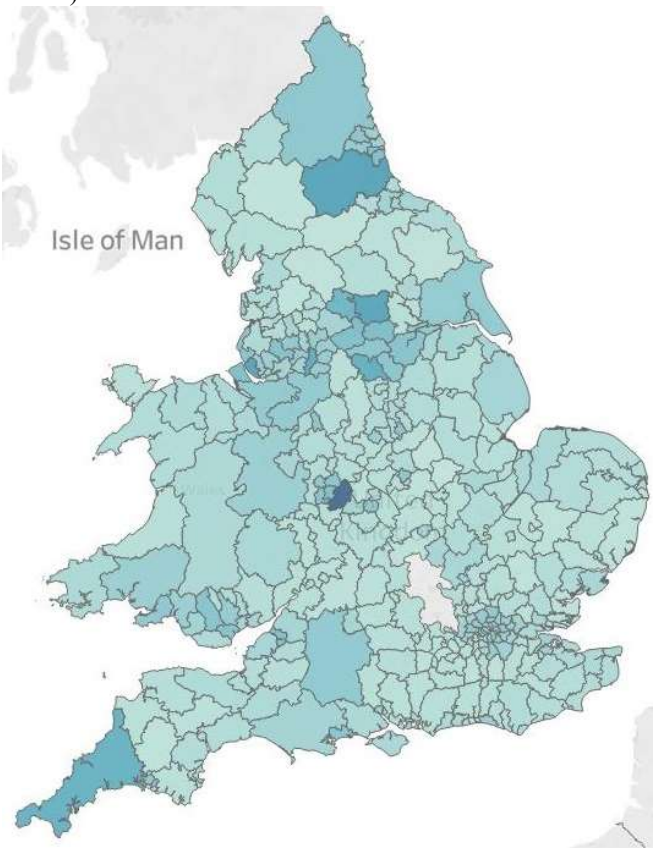

Fig. 10-Bad health by Local Authority (darker blue means more)

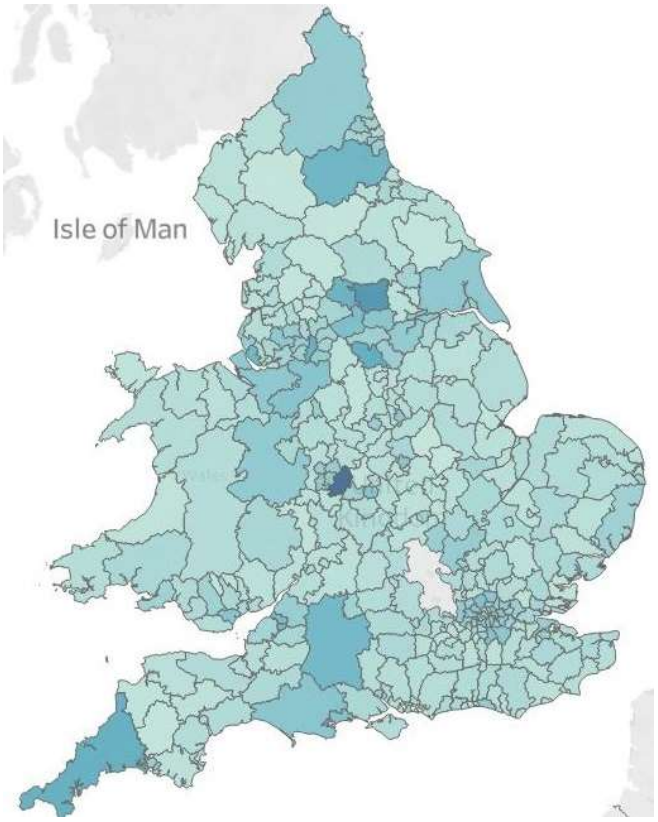

Fig. 11-Population by Local Authority

With the vaccine statistics these are rather uniform across the board with most places having a rate between 60-80% and, apart from the cars/no cars in a household, very little in terms of relationships back to the census figures.

5 Critical Reflection

In this project we often had to rely on critical thinking rather than blindly trusting in the computer. For example, by using the regression plots along with the correlation figure we could see whether or not the correlation was influenced by one or two outliers (which it quite often was).

Another part which required a lot of human interaction was linking the various local authorities together, as the government changed these between the census, the geography file and the COVID-19 data. These needed human reasoning to link the datasets together.

For clustering the data, the computer was able to divide the data into different clusters, but it needed human reasoning to decide upon the appropriate number of clusters.

In the ARIMA model the computer's fit had managed to miss the underlying structure and interpreted that as noise. This was mostly because the data did not have enough information for the model to learn from it.

In a similar manner the XGB Regressor was able to follow the structure of the data when it was relying upon the actual previous days data, but fell over when it had to work out this data for itself. We had also tried a Support Vector Regression model, which also failed to map to the data.

Something that would have made the analysis easier and more efficient is if the COVID-19 data had been available in a way that was linked to the census (much as the paper in [8] where the authors joined it). As apart from the age data, which was released publicly, we were joining by area to ten-year-old data and assuming that the population statistics had not changed in that time (for the population age data we were assuming that the ONS forecast was correct). Another limitation was that the local authority data had many artificial peaks from when various local authorities submitted data late (they usually missed out the weekend leaving a peak on a Monday).

A limitation on the government end for the death data was that this was death for any reason within 28 days of a positive COVID test result. So, if you died of COVID but had not taken a test you presumably were not counted, but being killed by something else after a positive test would count.

Software that would have made the task easier would have been something that could more easily work with space-time data. For the monthly geographic data, we ended up creating an animation to watch the time transition (although we find the 3D plots hard to read).

Another piece of software that would have been nice would be a sci-kit learn style API for the ARIMA algorithms. As the Statsmodel implementation was difficult to get used to.

Finally, it would have been reassuring if the Local Authority and UK wide data followed the same pattern when the Local data was summed up, as it was difficult to completely trust the data without that.

Table of word counts

Problem statement	250
State of the art	500
Properties of the data	498
Analysis: Approach	500
Analysis: Process	1476
Analysis: Results	194
Critical reflection	493

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.
[20] xgboost developers. 'XGBoost Documentation — Xgboost 1.5.1 Documentation', 2021. https://xgboost.readthedocs.io/en/stable/.
[21] Fishwick, Thomas. 'SL477/VA-Coursework: Visual Analytics Coursework'. GitHub, 2022. https://github.com/SL477/VA-Coursework.
[22] Bitfuul. 'Split Your Image Online for Free without Any Limits', 2021. https://bitfuul.com/split-image.
[23] Imgflip. 'Imgflip - Create and Share Awesome Images', 2022. https://imgflip.com/.
[24] Hunter, John, Darren Dale, Eric Firing, Michael Droettboom, and Matplotlib development team. 'Matplotlib — Visualization with Python', 2012. https://matplotlib.org/.
[25] Project Jupyter. 'Ipywidgets', 2017. https://ipywidgets.readthedocs.io/en/latest/.
[26] Project Jupyter. 'Jupyter-Labs', 2014. https://jupyter.org.
[27] Rossum, Guido van. 'Python'. Python.org, 1991. https://www.python.org/.
[28] R Core Team. 'R: The R Project for Statistical Computing', 1993. https://www.r-project.org/.
[29] Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 'Dplyr', 2018. https://dplyr.tidyverse.org/.
[30] ONS. 'Population Dataset - Excluding MSOAs'. Gov.Uk, 05 2021. https://coronavirus.data.gov.uk/downloads/supplements/ONS-population_2021-08-05.csv.
[31] Fulton, James. 'ARIMA Models in Python'. ARIMA Models in Python, 2021. https://app.datacamp.com/learn/courses/arima-models-in-python.

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