Student number: 21185793

Does Time Series Data on COVID-19 Vaccination Rates Reveal Locational Patterns in Uptake?
An approach using time series clustering

Background

The coronavirus (COVID-19) pandemic in the UK has exposed sizeable differences in the uptake or intended uptake of vaccinations against COVID-19 between subpopulations in the UK. Vaccination rates in the UK were falling prior to the onset of COVID-19 (Lewis, 2020 and de Figueiredo, 2021), underscoring the need to identify and address low vaccine uptake in order to avert a public health crisis.

Much of the research on levels of uptake to date has focused on sociodemographic factors for regions as a whole. A study by Curtis et al. (2022) of 57.9 million primary care records in England found lower levels of first-dose uptake among ethnic minority and socially deprived populations in the first 100 days of the vaccine rollout. The study claimed that 96.2% of the eligible white population took up a vaccination versus only 68.3% of the eligible black population. The figures were 90.7% and 96.6% between the most and least socially deprived areas studied. These results are broadly supported by the longitudinal study conducted by Robertson et al. (2021) and the cross-sectional study conducted by Martin et al.

Despite the availability of spatial and temporal data on COVID-19 vaccine uptake, there appears to have been very little work in either area with the exception of the recent study by de Figueiredo, who found significant geographic and sociodemographic variations in vaccine uptake. This research proposes to address this gap by taking the parameters of the Curtis et al. study as a benchmark to answer a simpler question using time series data: **does time series data from the first 100 days of the vaccine rollout reveal locational patterns in uptake?**

Data

The data pertaining to COVID-19 vaccination rates used in this report was obtained from coronavirus.data.gov.uk, the official, UK government website for communicating data and insights on COVID-19 in the UK.

The site provides data for a range of geographic and administrative boundaries, including Local Authority, Middle and Lower Layer Super Output Area (MSOA and LSOA, respectively) and NHS Trust. However, the types and amounts of data provided are not consistent across reporting categories: for example, vaccination data at MSOA level is only provided as cumulative counts at the latest reporting date with no access to earlier data being made available.

Upper Tier Local Authority (UTLA) data on daily, first-dose vaccinations by vaccine date was used for this analysis. The available data covers 181 UTLAs in England and Scotland. No data for Wales or Northern Ireland was available.

The decision to use data at UTLA level was based on 3 considerations: (i) a full archive of daily data by vaccine dose was available; (ii) the granularity of information was thought sufficient to be able to show meaningful and interesting comparisons between areas; (iii) the dataset was not so large that it would impede the calculation of one of the distance matrices used in the hierarchical clustering procedure (see 'Choice of distance measure' under 'Methodology' for further details).

The time period analysed was chosen to match the Curtis et al. study: 100 days beginning 8 December 2020, the date of the first COVID-19 vaccination in the UK, and ending on 17 March 2021. Only first doses of the vaccine were considered, again to match Curtis et al.

The dataset therefore consisted of 181 time series, each with 100 daily observations of first-dose uptake. As a motivating example to illustrate the clear, observable differences in vaccine uptake between UTLAs, Figure 1 shows vaccine uptake for 5 UTLAs selected at random from the dataset. Visually, the series generated for the City of Bristol looks quite different to the one for East Dunbartonshire whilst Coventry and Cambridgeshire look more similar in nature.

Methodology

The analysis presented in this report uses time series clustering as a method for discerning similarities in vaccine uptake patterns between UTLAs. A time series is a sequence of values ordered by time. In the case of the data used herein, the time series is the number of vaccines delivered per day, ordered by date of delivery.

Clustering is an unsupervised data mining technique in which similar data is placed into related or homogeneous groups and time-series clustering can be defined as a special type of clustering (Aghabozorgi et al., 2015).

The following methodological choices were made concerning the clustering algorithm used in the analysis.

Scaling of data: UTLAs are not defined based on similar sized population groupings, unlike units such as MSOAs. To prevent larger UTLAs from influencing the clustering algorithm, min-max normalisation was applied to the vaccine uptake data to rebase all uptake values within a series to between 0 and 1.

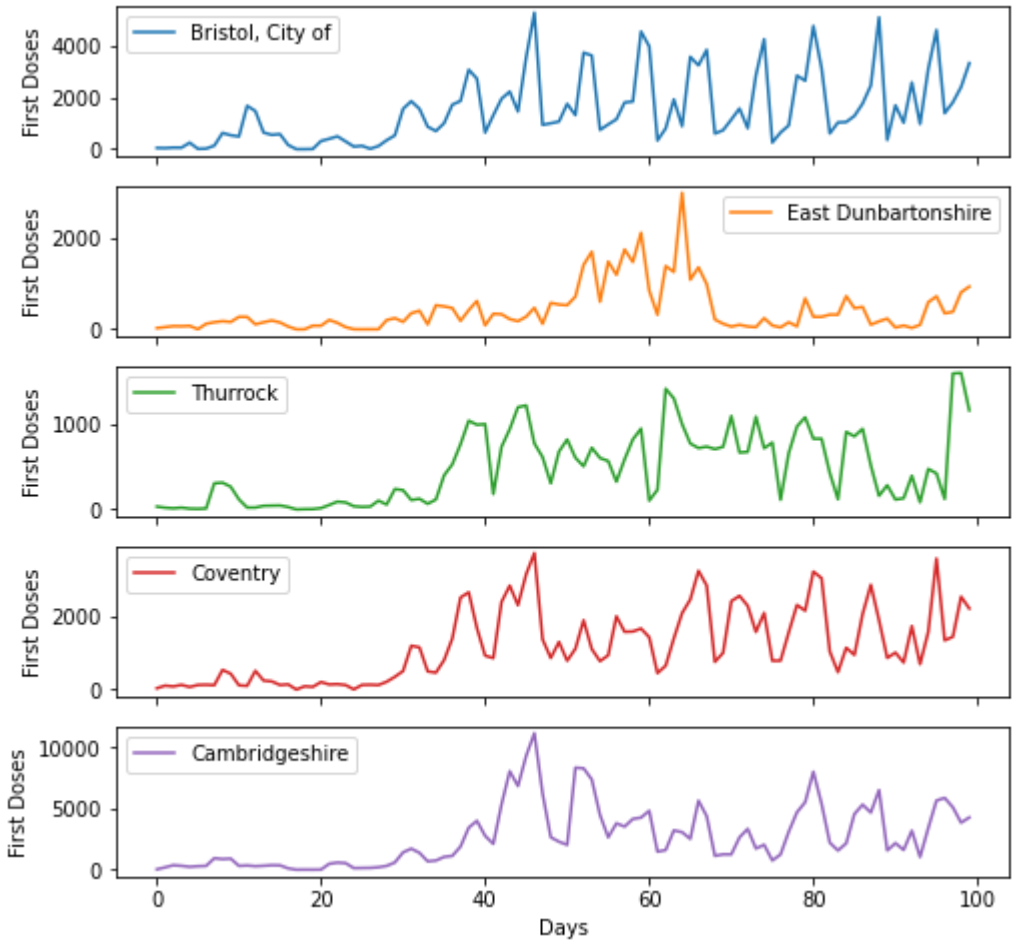

Fig. 1: Plots of first dose vaccine uptake for 5 UTLAs selected at random over the first 100 days of the COVID-19 vaccine rollout

Choice of clustering method: Agglomerative, hierarchical clustering was chosen based on its recognition as the most widely used hierarchical method (Maharaj et al.) and the fact that it is deterministic and therefore reproducible whereas partitioning methods are not (Özkoç, 2020). A hierarchical method was also preferred for use with time series data as the algorithm can incorporate non-Euclidean distances, as discussed below under 'Choice of distance measure'.

Choice of distance measure: The analysis employed two types of distance measure that were compared for effectiveness: Euclidean distance and Dynamic Time Warping (DTW).

Euclidean distances were used as they are among the most common distance measures used in cluster analysis (Everitt, 2011). Specialist texts on time series clustering such as Maharaj et al. cover methods using Euclidean distances and the measure also forms the basis of a large body of time series clustering research reviewed by Aghabozorgi et al.

DTW (Berndt and Clifford, 1994) is a mapping of points between 2 time series, T1 and T2 designed to minimise the pairwise Euclidean distance (Javed et al., 2020). It is preferred in some literature (Wang et al., 2012; He et al. 2018) due to the failure of Euclidean distances to recognise similarities between series when series are shifted horizontally relative to each other. Figure 2, below is a graphical representation of how DTW deals with series shifts. The downside of DTW is that it is computationally expensive and slow for large datasets (Maharaj et al.).

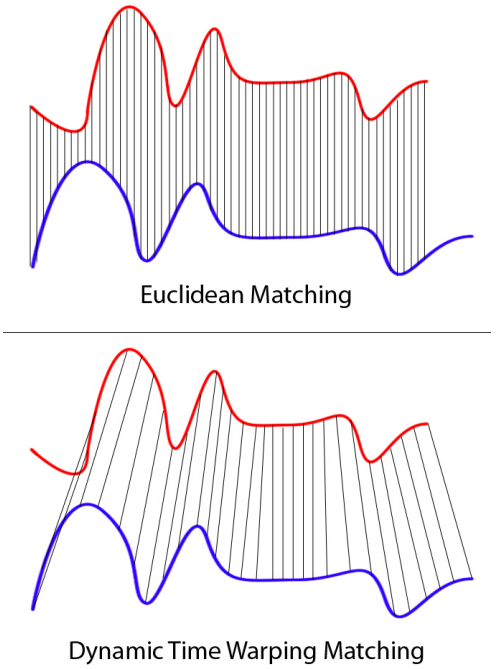

Fig. 2: Graphical representation of Euclidean and Dynamic Time Warping matching. Source: Wiki Commons: File:Euclidean_vs_DTW.jpg

Choice of linkage type: The analysis adopted the recommendations of Mather (1976) and tested a range of linkage methods, determining the best linkage through testing how closely the resulting dendrogram preserved the pairwise distances of the original data points using a measure known as the cophenetic correlation coefficient (Saraçli, 2013). A cophenetic correlation of very close to 1 indicates that the dendrogram is a good representation of the data.

Analysis

Figure 3, below summarises the results of testing combinations of distance and linkage measures and comparing the goodness-of-fit of the resulting dendrograms with the underlying data using the cophenetic correlation coefficient. The linkage types selected were those reviewed by Maharaj et al. Only certain linkages accept non-Euclidean distances, which explains why the combinations tested with DTW are smaller.

Distance Measure	Linkage Type	Cophenetic Correlation
Euclidean	Single	55.3
Euclidean	Complete	37.94
Euclidean	**Average**	**71.29**
Euclidean	Weighted	62.05
Euclidean	Centroid	64.33
Euclidean	Median	52.46
Euclidean	Ward	49.82
DTW	Single	29.5
DTW	Complete	37.15
DTW	Average	56.21

Fig. 3: Combinations of distance measure and linkage type tested

Fig. 3 shows that none of the combinations proved to be a particularly good fit for the data. The combination of Euclidean distance with average linkage had the highest cophenetic correlation coefficient and was selected as the basis of the subsequent agglomerative hierarchical cluster analysis.

Fig. 4 shows the resulting dendrogram. Even though the graphic is crowded and difficult to interpret due to the number of UTLAs in the dataset, it is evident from the heights of the branches that the clustering is poor. No branches joined until a height of c.1 in a tree that is only c.3.5 in height. Furthermore, the large cluster on the right, coloured light blue and occupying c.50% of the figure appears to be noise, based on the varying heights grouped together.

Selecting a cut point for the tree was difficult given the untidy nature of the clustering, which provided no clear cut points. The decision was taken to cut the tree at the 2.5 level, increasing the size of the noise cluster but reducing the number of singletons that would be generated should a cut point lower down the tree be selected. A total of 10 clusters were counted and this number was used in a Python routine to create a map of the clusters in list form that could then be mapped to the original dataset of vaccine uptake by UTLA.

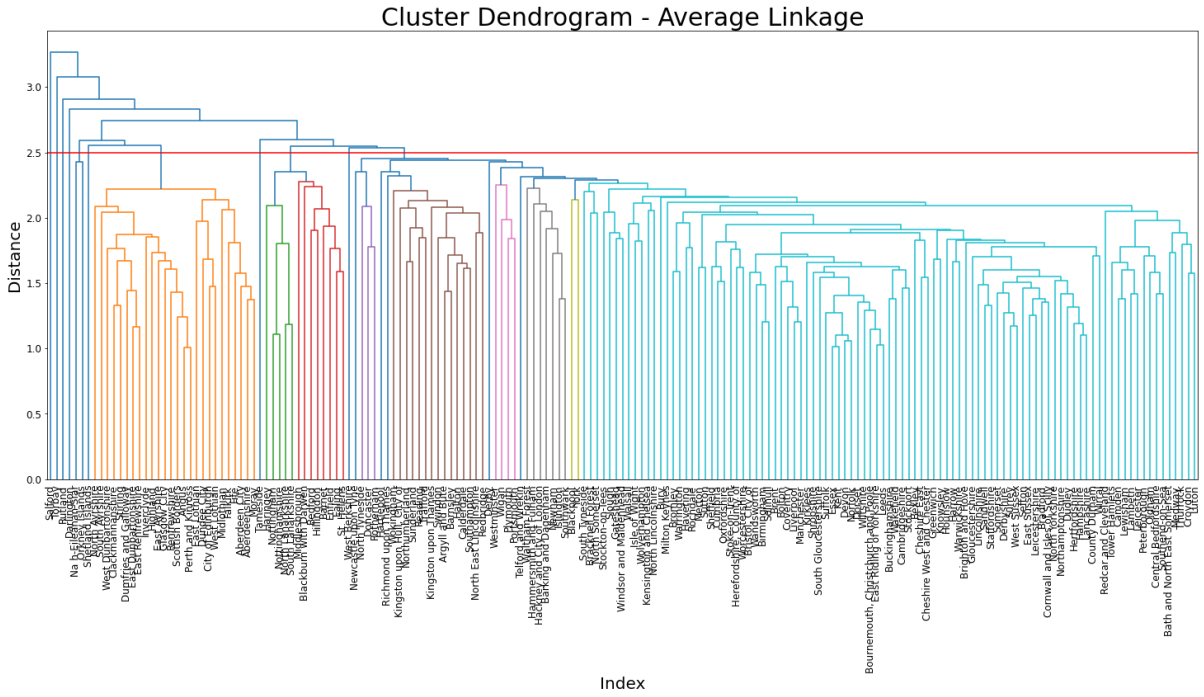

Fig. 4: Dendrogram of an agglomerative hierarchical clustering of the dataset using Euclidean distance and average linkage

'Clustermaps' or hierarchically clustered heat maps of vaccine uptake created from the mapping of clusters to UTLAs reveal some interesting patterning. As shown in figure 5, one of the clusters contains exclusively Scottish UTLAs with a distinct uptake pattern across all UTLAs in the cluster: uptake was initially low until c. day 55 of the rollout, at which point activity increased before subsiding again around day 70.

Another clustermap of interest (not shown here due to space constraints but available to view via the github link) contains 13, geographically dispersed UTLAs: Barnet, Ealing, Enfield, Haringey and Hillingdon in London; Bedfordshire in the East of England; Nottingham and Nottinghamshire in the Midlands; Blackburn with Darwen and St Helens in the North-West; Middlesbrough in the North-East and North and South Lanarkshire in Scotland. As the clustermap indicates, uptake in these 13 UTLAs was broadly low until day 40 of the rollout, at which point uptake levels increased and remained at increased levels until the end of the series.

Conclusion

The results from the cluster analysis are limited due to the clustering only representing the underlying data to a moderate degree but the few, well-defined clusters obtained do appear to be able to affirmatively answer the research question that time series

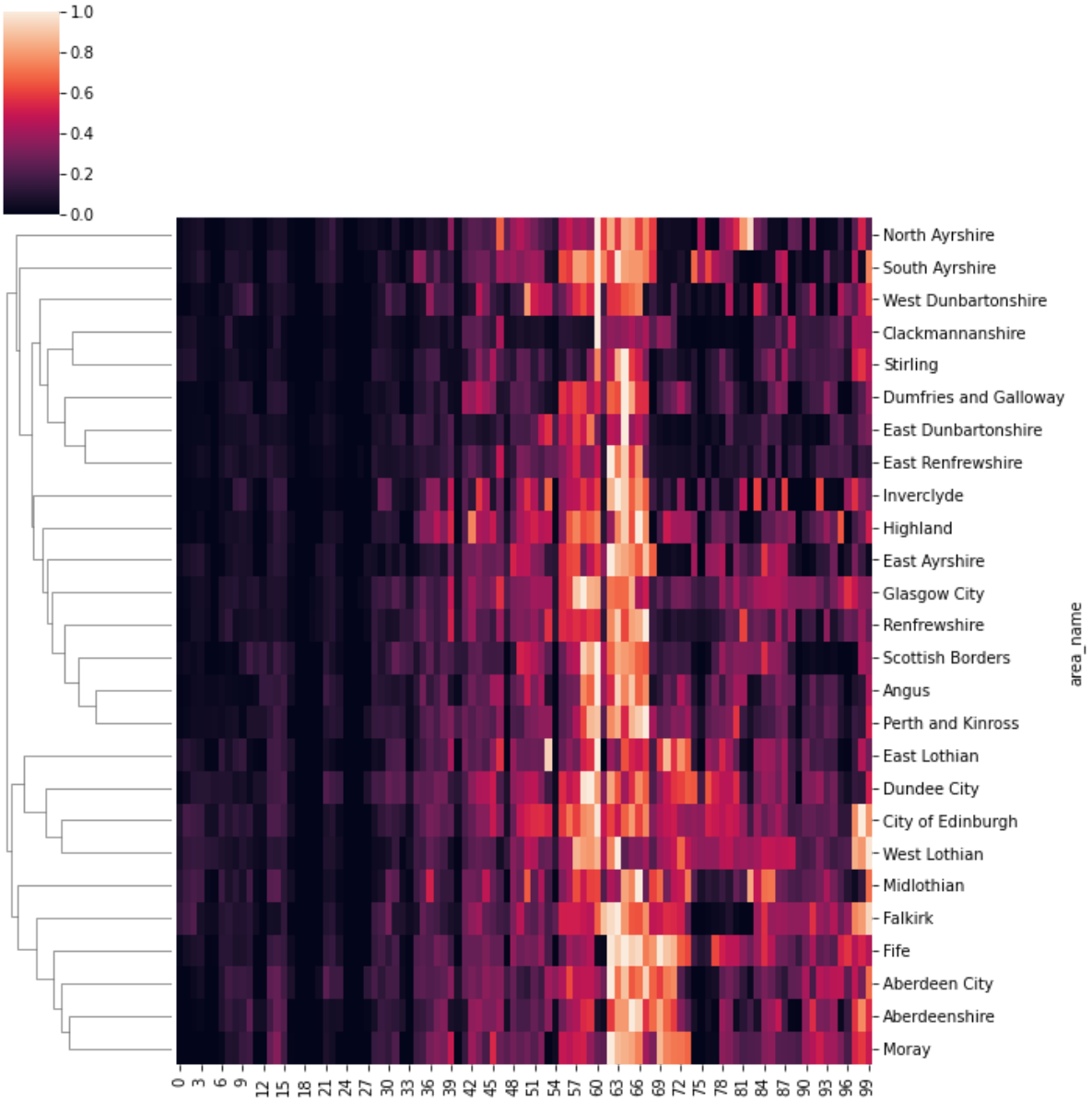

Fig. 5: Clustermap showing an exclusively Scottish cluster in which all UTLAs can be seen to have very similar patterns of first-dose uptake

Conclusion (continued)

(cont.), data from the first 100 days of the rollout does reveal locational patterns in uptake. An important insight is that mainland Scotland behaved remarkably cohesively in terms of uptake of first doses in the period and that peak uptake activity occurred during a narrow window of c.2 weeks. Further investigation would be required to determine whether these effects can be attributed to public health policy or sociodemographic factors but the information could serve public health authorities planning future vaccination campaigns.

The analysis could be re-run using a longer time series and combining data from doses 2, 3 and booster shots to assess whether a larger number of data points results in an improved clustering.

A code repository containing the Jupyter notebooks used for analysis and the raw data sets can be accessed via: https://github.com/MissMaya/casa0007_cwk3

Word count: 1,794 words

References

Aghabozorgi, S., Shirkhorshidi, A.S. & Wah, T.Y., 2015. Time-series clustering – A decade review. Information Systems. Available at: https://www.sciencedirect.com/science/article/pii/S0306437915000733 [Accessed January 18, 2022].

Berndt, D. & Clifford, J., 1994. Using dynamic time warping to find patterns in time series. Association for the Advancement of Artificial Intelligence . Available at: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf [Accessed January 25, 2022].

Curtis, H.J. et al., 2022. Trends and clinical characteristics of COVID-19 vaccine recipients: A federated analysis of 57.9 million patients' primary care records in situ using OpenSAFELY. British Journal of General Practice. Available at: https://bjgp.org/content/72/714/e51 [Accessed January 16, 2022].

de Figueiredo, A., 2021. Forecasting sub-national trends in COVID-19 vaccine uptake in the UK. medRxiv. Available at: https://www.medrxiv.org/content/10.1101/2020.12.17.20248382v2.full-text [Accessed January 17, 2022].

Everitt, B., 2011. Cluster analysis, Chichester, West Sussex, U.K.: Wiley.

He, L., Agard, B. & Trépanier, M., 2018. A classification of public transit users with smart card data based on time series distance metrics and a Hierarchical Clustering Method. Taylor & Francis. Available at: https://www.tandfonline.com/doi/abs/10.1080/23249935.2018.1479722 [Accessed January 25, 2022].

Javed, A., Lee, B.S. & Rizzo, D.M., 2020. A benchmark study on time series clustering. Machine Learning with Applications. Available at: https://www.sciencedirect.com/science/article/pii/S2666827020300013 [Accessed January 18, 2022].

Lewis, P., 2020. Falling vaccination rates: The case of the MMR jab. www.researchbriefings.files.parliament.uk. Available at: https://researchbriefings.files.parliament.uk/documents/LLN-2020-0033/LLN-2020-0033.pdf [Accessed January 20, 2022].

Maharaj, E.A., D'Urso, P. & Caiado, J., 2021. Time series clustering and classification, Boca Raton: Chapman & Hall/CRC.

Martin, C.A. et al., 2021. SARS-COV-2 vaccine uptake in a multi-ethnic UK healthcare workforce: A cross-sectional study. PLOS Medicine. Available at: https://journals.plos.org/plosmedicine/article?id=10.1371%2Fjournal.pmed.1003823 [Accessed January 20, 2022].

Mather, P.M., 1976. Computational methods of multivariate analysis in physical geography, London: Wiley.

Özkoç, E.E., 2020. Clustering of time-series data. IntechOpen. Available at: https://www.intechopen.com/chapters/65772 [Accessed January 18, 2022].

Robertson, E. et al., 2021. Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study. Brain, Behavior, and Immunity. Available at: https://www.sciencedirect.com/science/article/pii/S0889159121001100 [Accessed January 20, 2022].

Saraçli, S., Doğan , N. & Doğan , I., 2013. Comparison of hierarchical cluster analysis methods by cophenetic correlation - journal of inequalities and applications. SpringerOpen. Available at: https://journalofinequalitiesandapplications.springeropen.com/articles/10.1186/1029-242X-2013-203 [Accessed January 20, 2022].

Wang, X. et al., 2012. Experimental comparison of representation methods and distance measures for time series data - Data Mining and Knowledge Discovery. SpringerLink. Available at: https://link.springer.com/article/10.1007/s10618-012-0250-5 [Accessed January 19, 2022].