Office for National Statistics

Clustering local authorities against subnational indicators methodology

This methodology guide is intended to provide information on the data and method used on the article clustering local authorities against subnational indicators

Contact:
Will Haste
subnational@ons.gov.uk

Release date:
24 February 2023

Next release:
To be announced

Table of contents

1 . Overview of project

In our accompanying article, entitled Clustering local authorities against subnational indicators, England, we conducted analysis on data from the December version of the Subnational indicators explorer to group local authorities with similar characteristics. The explorer comprises a collection of publicly available data produced by a range of sources from across government. The latest data available on the explorer for each metric were included, which were all collected between 2018 and 2022.

The clusters from our analysis can be used by subnational policy makers to identify other local authorities that may be facing similar challenges, and to create control groups for intervention impact analysis. The results presented are not used to influence levelling up policy decisions and should not be viewed as being a judgement about the performance of a local authority. As the clustering methodology is complex, this article will provide additional information on the methods that we used to generate our results and ensure their quality.

2 . K-means clustering method

Clustering is a method of analysis that identifies groups and patterns within a dataset. Grouping data by the International Territorial Level (ITL)1 region the respondent lives in would be an example of basic clustering, however this may not result in the most similar clusters because of intra-regional variation. For instance, the demographic and socioeconomic characteristics of central Bristol would likely be different to areas of rural Cornwall, despite their sitting within the same ITL1 region.

We partitioned data by 2021 lower-tier local authority boundaries as this was the most common geography reported in our data sources. Using clustering to create groups of local authorities with a greater similarity means that those groups can be used in identifying complex patterns, predictive analysis and maximising the accuracy of conclusions from datasets by allowing similar areas with lower sample size to be analysed together.

We used k-means clustering (explained in this tutorial at 365datascience.com) in this analysis as a method for identifying and grouping similar data points within a dataset. The k-means algorithm functions by taking a chosen number (denoted by k) of random centroid points, and each individual data point is then assigned to its closest cluster. The total distance between points and their respective centroid is calculated and stored. The algorithm then updates the centroid points to be central within each cluster and the distances are calculated again, and new clusters formed. This process is continued until the centroid points no longer change and the total Euclidean distance (the length of a connecting line) between each point and their respective centroid is minimised.

The equation representing k-means clustering, where the algorithm minimises the objective function, is as follows:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2$$

- where J is the objective function to be minimised
- k is the number of clusters
- n is the number of data points
- x is the location of a specified data point
- c is the centroid point of the cluster

Silhouette scores (as explained in this article at towardsdatascience.com) were used to optimise the number of clusters in the clustering function, between 4 and 15 clusters. This allows us to be non-deterministic in our analysis by not prescribing the exact number of clusters. Scores range between negative 1 and 1, and we aimed for the scores to be as close to 1 as possible, as a higher score denotes more clearly distinguishable clusters. We intend to deliver improvements to the models in the future to increase these silhouette scores. The list below shows the silhouette scores across our models:

- Headline – 0.48
- Economic – 0.62
- Connectivity – 0.53
- Educational attainment – 0.21
- Skills – 0.43
- Health – 0.28
- Well-being – 0.25

We set the range of 4 to 15 clusters to ensure there were enough clusters for analysis, without having too many to make interpretation challenging. The formula for calculating silhouette scores is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- where s(i) is the silhouette score for the data point i.
- a(i) is the average distance between i and the other points within its cluster.
- b(i) is the average distance between i and all clusters to which i does not belong.

Limitations

Results are dependent on the initial “seed” values, which are the random centralised starting points for the algorithm. Choosing different seed values will result in differing clusters, so to mitigate this, we set the same randomised seed value for each of our clustering models.

The model is very sensitive to changes in the data. This means that simple operations such as standardising the data can result in a change in the number and distribution of the clusters. For this reason, it was important that all data used in the publication went through the same data cleaning and manipulation process before clustering alongside robust and rigorous quality assurance of results.

The K-means algorithm also cannot account for missing data. This means that when a geography did not have a value for one of the metrics used, that geography had to be dropped from the model. For instance, local authorities where recent boundary changes have occurred, such as Buckinghamshire, West and North Northamptonshire, could not be assigned to clusters because some raw data in our models did not account for boundary changes.

Finally, our results aim to map which places have similar outcomes based on the chosen metrics, however these data do not provide an indication of how much each local authority has affected the clusters formed.

3 . Data sources and geographical coverage

The data used for this analysis were taken from metrics included in our subnational indicators explorer. Information on the data sources used, including caveats and notes, can be found in our accompanying Subnational indicators dataset (December 2022 edition).

Although some metrics on the explorer were available for Scotland, Wales and Northern Ireland, we have used England-only data in this publication to have the same consistent geographical coverage across our models. We are working with these devolved nations to improve the coverage of our analysis. Ongoing work to improve the range of geographies that our statistics are available for will reduce the impact of this in future iterations of this work but may also impact the clusters in these results.

4 . Model construction

K-means clustering was used on the data, to group similar local authorities for the purpose of analysis. Once clusters were defined, we analysed the common characteristics of local authorities in each cluster.

Headline Model

Although our article is not official government policy, the data used from the subnational indicators explorer are in line with the headline and supporting metrics for the respective missions from the Levelling Up the United Kingdom: missions and metrics technical annex. Where possible, we aligned our headline model with the headline metrics outlined in the annex. To create a multidimensional model that groups local authorities across all themes, we chose one headline metric from each theme where available. Including the same number of metrics from each theme in the headline model ensured that the clusters were not weighted towards a particular topic area. Where multiple headline metrics could have been selected for a theme, we consulted with cross-government topic experts on which metric was most appropriate to include.

Rationale for selecting one metric from multiple headline metrics

Economic theme: Gross value added (GVA) per hour worked was chosen as it is recognised as an important metric for productivity. It is also a measure that is used extensively by subnational policy makers and, as such, there was high stakeholder demand for its inclusion.

Transport connectivity theme: Average travel time to employment centre with 500 to 4,999 jobs by public transport or walking was recommended as the metric to include by topic experts. It was also found to be the most representative metric for this theme when compared with analysis combining figures for all modes of transportation.

Digital connectivity theme: The lack of subnational variation in 4G coverage means that access to gigabit-capable broadband is a better representation of the subnational disparities regarding digital connectivity.

Well-being theme: There is a precedent for using life satisfaction to represent wider personal well-being such as the Origins of Happiness paper by the London School of Economics, therefore we aligned our analysis accordingly.

For the skills theme, we could not use the headline metric of 19 years and over further education and skills achievements, because our data represent a count rather than a rate. There is a wide range in the total population of local authorities and using a count would result in the outcome being weighted towards the local authorities with higher populations. We used apprenticeship achievements as an alternative to the headline metric.

Additionally, healthy life expectancy data is only available disaggregated by sex at local authority level. To fill this data gap, local authority level figures for male and female life expectancy were weighted to the UK level proportion of men and women (49% and 51%, respectively), and then combined. We used the national figure to weight these results rather than local authority level sex proportions as timely data for this were not available.

The metrics included in the headline model were therefore as follows:

- Economic theme – GVA per hour worked
- Transport connectivity theme – average travel time by public transport or walking to the nearest large employment centre
- Digital connectivity theme – percentage of premises with gigabit-capable broadband
- Education theme – percentage of pupils reaching expected KS2 levels by the end of primary school
- Skills theme – apprenticeship achievements per 100,000 population
- Health theme – healthy life expectancy
- Well-being theme – life satisfaction

We will continue to explore the impact of selecting different metrics to represent the important themes in future work.

Individual theme models

Alongside the headline model, we also produced models based on individual themes; economy, connectivity, education, skills, health and well-being.

Data modifications

Most metrics are defined at lower tier (local authority district and unitary authority) level; however, some are only reported at upper tier (county and unitary authority) level. We imputed lower tier data by setting it to be the same as the upper tier data for all missing lower tier local authorities within an upper tier local authority. We will continue to work with data owners in the future to ensure that data are available at the most granular level possible to align with the Government Statistical Service subnational data strategy.

Median values for each metric in a cluster were calculated as well as an overall median for each metric for all local authorities included in the analysis. Cluster titles were named by comparing the medians for each cluster against the overall median for all clusters for each metric. For models where the clusters are relatively “linear” (meaning that all metrics in a cluster are above, or all are below the median), the cluster title compares them with that median. For more complex models (for example, where some metrics in a cluster may be above and some may be below the median), the words “higher” and “lower” have been used for the particular metrics that best highlight the variation between clusters.

5 . Analysis of clusters

Lookups

We analysed socio-spatial and economic trends within each cluster’s local authorities. This allows us to pick out characteristics of similar local authorities. We were not able to map all local authorities into each of the lookups because of data availability. This section provides information on the lookups used in the analysis.

Rural/urban classification

The 2011 rural/urban classification was used. The 2011 classification was updated to the 2021 Local Authority boundaries by using more granular geographies, such as output areas, to construct the new boundaries. We used a simplified version of the classification, where the sample of local authorities is split into three categories based on their rurality:

- predominantly rural
- urban with significant rural
- predominantly urban

Index of multiple deprivation

The Index of multiple deprivation (IMD) is a composite, multidimensional measure of deprivation. The link provides information on the variables considered in the IMD and what proportion of the final weighted measure they comprise. For the analysis here, the IMD rank of each local authority grouping has been used to split local authorities into quintiles. This was done to ease drawing conclusions from the data, while maintaining a statistically robust sample size of local authorities in each category.

International Territorial Level (ITL) 1 region

The 9 English ITL1 regions were used to analyse the geographic distribution of each cluster.

Population density

The population density figures we have used were taken from the 2021 census. Quintiles, each with a similar number of local authorities, were defined and used to analyse the relationship between population density and subnational indicators. Urban rural classification also measures population density, so a close relationship with that lookup exists.

These groups are as follows:

- less than 185 people per kilometres (km) squared
- 185 to 440 people per km squared
- 441 to 1073 people per km squared
- 1074 to 2776 people per km squared
- 2777 or more people per km squared

Median age

The data for the median age of everyone living in a local authority was taken from 2021 census publications. Local authorities were again split into three equal categories based on their median age. The median age was used as a measure to reduce the impact of outliers, a small local authority with a normal population but a large retirement home or school may have a less representative mean age. The median age bands used were as follows:

- aged 40 years or younger
- 41 to 44 years
- 45 years and older

Coastal towns

A binary marker was generated to indicate the presence of coastal towns within each local authority according to our Coastal towns in England and Wales dataset.

Correlations

As an addition to the lookup analysis, we processed correlations between each clustering model and the lookup values. The clustering categories are considered as nominal data as we cannot definitively say that one is higher than the other because of their multidimensional nature. As a result, we used Cramer's V correlation test to measure the strength of the correlation between clusters and lookups. As we are working with nominal data this measure does not provide us with information about the direction of the association, but a measure of the similarity of the distributions.

The results of the Cramer's V tests are presented in our accompanying dataset to show which of the clusters are most correlated with which lookups. While not a measure of causation, this analysis does indicate which of the lookup topics have the strongest statistical association with each cluster and can provide valuable information for further analysis that focuses on causation.

6 . Further work

Some of the maps used to visualise clusters contain "white spaces". This is a result of inconsistency in the geographies that the data were provided at or missing data for one or more of the metrics at that geography. While the visualisations presented in our article use the most recent 2021 local authority boundaries, some of the data used to generate the clusters were produced using older boundaries. A weighting methodology to assign old local authority data to the new boundaries is being developed as a priority for future analysis.

We will also work with devolved administration colleagues to expand the coverage of the data to the whole of the UK. We will also explore using different types of clustering models, as well as building some of the lookups analysis included in this publication directly into models. Furthermore, we will explore conducting this analysis across multiple time periods, to analyse how the clusters change over time.

If you have any feedback on what you would like to see included in future versions of this analysis, feel free to email your suggestions to us at subnational@ons.gov.uk.

7 . Cite this methodology

Office for National Statistics (ONS), released 24 February 2023, ONS website, content type, Clustering local authorities against subnational indicators methodology