Office for National Statistics

Clustering similar local authorities and statistical nearest neighbours in the UK, methodology

Methodology information for our clustering and statistical nearest neighbours analysis, which groups UK local authorities with similar characteristics and outcomes.

Contact:
Subnational Methods for Dissemination team
subnational@ons.gov.uk
+44 1633 455135

Release date:
14 February 2025

Next release:
To be announced

Table of contents

1 . Overview of project

This article outlines the methodology and quality information for our clustering and statistical nearest neighbours analysis. It explains the improvements made since our Clustering similar local authorities in the UK, methodology (2024), using publicly available data from a range of government sources to create similar groups of local areas.

In this analysis, we have made the following changes to our approach:

- we have included a statistical nearest neighbours feature that provides an ordered list of the most similar areas to a given area
- counties are included in our analysis for the first time, within the statistical nearest neighbours feature (but are not included in the clustering)
- our variable selection has been improved, including replacing some correlated data with composite or proxy measures

The clusters of similar areas produced in this analysis can help policymakers identify local authorities that may be facing similar challenges to themselves and create control groups for assessing the impact of a policy intervention.

The new nearest neighbours feature offers several advantages for local users:

- flexible group sizes – by producing an ordered list of the most statistically similar areas, local authorities can tailor the size of their similar groups to meet specific needs
- improved insights – more focused groups of similar authorities lead to deeper insights as larger clusters include a broader range of areas, where some projects may require a smaller list of areas; by narrowing the focus, local authorities can gain a clearer understanding of shared challenges and opportunities
- collaboration – this approach promotes direct collaboration between local authorities; with a named list of the most similar areas, authorities can easily identify and engage with their counterparts on common issues

This methodology provides a summary of:

- our methods
- the changes to our approach
- our data transformation techniques
- our variable selection approach
- information on the quality of our results

For tables including all of the results from our models, some analysis of our clusters and other model information, see our accompanying data tables.

2 . Data gathering

Initial data gathering

The initial data gathering process involved a search for new and updated variables from a variety of sources. These sources included data from previous clustering projects, our Explore Local Statistics (ELS) service, NOMIS , and data portals from other government departments and UK nations.

For us to consider using the data, they had to satisfy the following conditions:

- already in the public domain
- available for all UK nations
- available at both lower-tier local authority (LTLA) and upper-tier local authority (UTLA) levels, or in a format where it can be accurately aggregated from LTLA to UTLA
- low levels of missingness
- relevant to local policy makers

We identified 47 variables that satisfied these conditions and entered our variable selection process.

Variable selection process

The variable selection process was an important step to ensure that the most informative and relevant variables were used in our analysis, ultimately simplifying the final models. This involved several techniques, each aimed at enhancing the quality of our output.

Correlation analysis

The purpose of our correlation analysis was to identify and remove highly correlated variables, to prevent certain variables from having an increased impact on the models. By examining the relationships between different variables, we aimed to ensure that each variable contributed unique information to the dataset. Variables with correlation coefficients above 85% were considered for removal from our models.

Variance analysis

We conducted variance analysis to identify the most informative variables. Variables with little variation across local authorities make a lower contribution to the final output as there is less of a pattern for the algorithm to identify. Low variance variables were considered for exclusion alongside the other requirements.

Principal component analysis (PCA) and loadings

We used principal component analysis (PCA) to identify the most informative components and reduce the total number of variables used in the final models. PCA condenses the variables into principal components, with the first 10 principal components explaining over 85% of the variance in the data. The loadings, which are a measure of the contribution of each variable to the principal components, were used to identify the most statistically significant variables, as these variables contribute to principal components 1 to 10. This reduction in the number of variables allowed us to focus on the most important aspects of the dataset. Variables like voter registration, which showed no contributions to the principal components of interest, were considered for exclusion as they did not substantially improve the explanatory power of the model. We used a combination of the results of our PCA and the correlations between variables to decide on variables for exclusion. If a variable is an important headline indicator that is relevant to the subnational context, we weighted our decision making towards including it in the model.

Condensing variables

To simplify the underlying data, we combined some highly correlated variables into composite variables, allowing for their inclusion in the output.

- We combined age groups into a single-dependency ratio variable, which is defined as the proportion of non-working-age population (under 16 years and over 64 years) to the working-age population (16 years to 64 years); this condenses the three age-related variables into one variable that reflects how the population is distributed between working and non-working age groups.

- Similarly, we replaced male and female healthy life expectancy with one combined healthy life expectancy variable by working out the proportion of the population, by gender, in each area, and then weighting and combining the two variables using that proportion.

- We had in our dataset multiple variables representing different ethnicities. These ethnicity variables are highly correlated, so to simplify our analysis we decided to use the proportion of residents of white ethnicity as a proxy for overall ethnic diversity. This approach simplified our model while still including ethnic similarity as a factor within our similar groups.

Through our variable selection process, we reduced the number of variables in our underlying dataset from 47 to 31. This ensured that only relevant data that would have an impact on our models were included in our underlying data, reducing the noise that is introduced in the models by unnecessary data. Simplifying the inputs to our models in this way makes it easier to pick out important features and characteristics of similar areas, as less informative variables have been excluded.

Data transformations

To ensure compatibility with our models, several transformations were applied to the data. These transformations were important in ensuring that our data was as complete as possible, as the K-means clustering algorithm cannot process missing data, and in preventing variables with a larger scale from impacting the models too greatly.

Aggregation and weighted imputation for generating county data

When county-level data were not available from the source, we used aggregation or weighted imputation techniques to create the data. Our aggregation process involved adding count values from the smaller local authority districts, and sometimes recalculating rates or percentages from those count values, to create figures at the county level. Where aggregation was not possible, weighted imputation was used to estimate missing values based on the available data. Within our imputation method, variables were weighted based on single-year Census population data, which was subset for each variable using an age group relevant to the data being imputed. For example, our employment rate was weighted by population aged 16 to 64 years, as that is the age range that is covered by the data.

Conversion of data produced at obsolete local authority boundaries

Most of our variables were produced between 2021 and 2024, which is a period that included several changes to local authority boundaries, which are often reflected in the data. Data produced at obsolete local authority boundaries needed to be converted to the 2023 version of the boundaries to ensure that the new local authorities are not missing from our output. This was achieved through the same aggregation and weighted imputation approach as the county data, allowing us to update the data to reflect current geographic boundaries more accurately.

Filling gaps in lower-tier local authority (LTLA) data with corresponding upper-tier local authority (UTLA) data

The models used needed all values to be present at the LTLA level. In a few cases we filled gaps at the LTLA level with data from the UTLA level. This was only applied where there was low missingness in the LTLA data, or if there were no alternative similar variables available, to reduce the impact of the imputation.

Winsorization to reduce the impact of outliers

To reduce the influence of outliers on the dataset, we applied Winsorization to the data at the 1st and 99th percentile. This statistical technique involves capping extreme values at a given percentile to reduce their impact on the overall analysis. By Winsorizing the data, we reduce the impact of extreme values on our models and prevent very small clusters of areas from forming.

Standardisation of data

Before the data went through the clustering algorithm, we standardised it to ensure consistency across all variables. Standardisation involves transforming each variable to have a mean of zero and a standard deviation of one. This step ensures that the unit or scale of a variable does not change its impact on the model. Some variables vary in their directionality, for example a higher happiness score is a positive outcome, but a higher anxiety score is a negative outcome. K-means clustering does not take the outcome of data into account and measures pure similarity, so we did not have to account for the directionality of the data in our approach.

By applying these techniques, we created a complete, comparable dataset while mitigating against the susceptibility to extreme values and other limitations of the K-means clustering method.

3 . Model construction

K-means clustering

Once we had finalised our input data, we applied the K-means clustering method outlined in our previous published methodology. As with our previous analysis, these models include all UK LTLAs, and the number of clusters was optimised between 4 and 15 for each of our models.

We have created one global model and two topic-level models in this release. While our previous release had four topic-level models, we have selected only two for this release because the economic and demographic models have the best use cases and our dataset aligns more suitably to these topic areas.

The global model incudes all 31 variables while the economic and demographic models include only the variables that align with those themes. The global model should be used when a comprehensive view of similarity is required, and the topic-level models are produced for analysis of policies that are relevant to those themes.

Silhouette scores are a measure of how distinct the clusters are from each other as defined in our previous Clustering local authorities against subnational indicators article, were used to assess the quality of our clustering models and optimise the number of clusters produced. Silhouette scores range from negative 1 to 1 with a higher silhouette score indicating that the clusters are more well defined. For the three clustering models produced, the silhouette scores ranged from 0.58 to 0.70, which is an improvement on the silhouette scores of our previous analysis, which ranged from 0.31 to 0.70.

The following list contains the variables used for the economic and demographic models, while the remaining variables are unique to the global model.

Economic model

- Employment rate (percentage)
- Gross value added per hour worked (£)
- Workers employed in construction or manufacturing (percentage)
- Median house price (£)
- Business births (percentage)
- High-growth businesses (percentage)
- Children in relative poverty (percentage)
- Housing completions (per 10,000 population)
- Gross median weekly pay (£)

Demographic model

- Proportion of residents of white ethnicity (percentage)
- Population change from 2011 to 2021 (percentage)
- Residents who are not religious (percentage)
- Dependency ratio
- Population density (residents per square km)

Global model only variables

- Residents with level 3 or above qualifications (percentage)
- Residents with no qualifications (percentage)
- Population who smoke cigarettes (percentage)
- Healthy life expectancy (years)
- Households that are renting privately (percentage)
- Guest nights stayed (per 10,000 population)
- Personal well-being (0 to 10 scale)
- 4G or 5G coverage (percentage)
- Gigabit capable broadband availability (percentage)
- CO2 emissions (per capita)
- Mean domestic electricity consumption (KwH per meter)
- Museums (per 100,000 population)
- Supermarkets (per 10,000 population)

For more information on these variables, please see the definitions tab of our accompanying data tables.

Statistical nearest neighbours

The clusters produced by our K-means algorithm are useful for identifying geographical patterns and broad similarities. However, we received feedback that access to smaller, more concentrated groups of similar areas would be valuable to our local government users. In response to this feedback, we have added a statistical nearest neighbours output to our approach, which provides an ordered list of the 20 most similar areas for any given area across our three models.

Using the Euclidean distance, we have calculated how close these local authorities are to each other. These distances are dimensionless and could be seen as similarity index scores instead of physical distances. To ensure that our list of nearest neighbours logically belong together or share common characteristics, we needed to set a distance threshold. This value defines how far apart local authorities can be to consider them "similar" and hence help us to determine the optimum number of neighbours. We used a pairwise distance histogram method to determine the distance threshold.

We generated a histogram of pairwise distances to visualise the distribution of the distance measures. With this, we can understand the natural spread of the data and identify a distance threshold on the histogram. Typically, setting a smaller distance threshold will mean closer or more similar areas while larger threshold indicates points that are farther apart or less similar. We chose our distance threshold within the range of the most frequent distances as this would ensure that the neighbours are not too far apart.

The Euclidean distance is a measure of the straight-line distance between two points in a multidimensional space and has been calculated using the following formula:

Given two points A = (A1, A2, ..., An) and B = (B1, B2, ..., Bn) in an n-dimensional feature space, the Euclidean distance d between these two points is calculated as:

$$d(A,\ B) = \sqrt{\sum_{i=1}^{n} (A_i - \ B_i)^2}$$

Where:

- n is the number of dimensions (variables).
- Ai and Bi are the variable values of the points A and B.

For each area, the distance to other areas has been computed by iterating through the dataset and calculating the distance from the area of interest to every other area. Our statistical nearest neighbours output includes both LTLAs and UTLAs, providing a county-level measure of similarity for the first time. This presents an issue as some of the LTLAs nest within the UTLAs. To prevent this leading to the statistical nearest neighbours for a county just being the LTLAs contained within it, we remove any geographies that cover the same area from the list of neighbours for the area of interest. A list of 20 areas ranked in order of closeness is now available through the "similar areas" feature of the Explore Local Statistics (ELS) service.

Our statistical nearest neighbours approach is a separate process to our clustering models, so the nearest neighbours for an area may not all be in the same cluster. As many areas will be situated at the edge of a cluster, and some of our clusters have a lower number of areas, it is expected that there will be some cases of nearest neighbours being in separate clusters. We tested how often this happens and found that a median of 85% of neighbours in our global model were within the same cluster as the area of interest, this figure was 79% for the economic model and 89% for our demographic model. We tested using Manhattan distance as the distance measure and found that the consistency of neighbours with clusters was roughly the same as when Euclidean distance was used, we chose to continue with Euclidean distance as it is the distance measure used within the K-means clustering algorithm. We also validated the neighbours with a manual check of the closest neighbours for each model, ensuring that all closest neighbours had a similar data profile.

The 10 nearest neighbours to Newport, Fareham and Darlington, are shown in the following list. Each neighbour is given alongside its geographical code.

Newport (W06000022)

- Wakefield, E08000036
- Barnsley, E08000016
- Peterborough, E06000031
- Tameside, E08000008
- Leeds, E08000035
- Rotherham, E08000018
- St. Helens, E08000013
- Medway, E06000035
- North Tyneside, E08000022
- Wirral, E08000015

Fareham (E07000087)

- North Somerset, E06000024
- Solihull, E08000029
- Bromsgrove, E07000234
- Chelmsford, E07000070
- Lichfield, E07000194
- Eastleigh, E07000086
- Cheshire East, E06000049
- North Hertfordshire, E07000099
- South Gloucestershire, E06000025
- East Dunbartonshire, S12000045

Darlington (E06000005)

- Wirral, E08000015
- Newcastle-under-Lyme, E07000195
- Gateshead, E08000037
- Sefton, E08000014
- Swansea, W06000011
- Staffordshire, E10000028
- Vale of Glamorgan, W06000014
- Redcar and Cleveland, E06000003
- St. Helens, E08000013
- Stockton-on-Tees, E06000004

Uses and users

- The clusters created in our analysis can be used as geographical groupings by central government and other analysts who are interested in how their variable of interest interacts with subnational similarity.
- The topic-level models are designed to be used by policy analysts, when the policy they are investigating is expected to have an economic or demographic impact.
- Local government users can use the clusters and nearest neighbours produced here to identify areas that may be facing similar challenges to themselves, or to identify comparators to measure the impact of a policy intervention.
- Academic researchers can analyse the spatial patterns of our clusters to explore the reasons for regional trends or use the groups to construct counterfactuals to use as synthetic controls for policy impact analysis.
- The enquiring citizen will be able to use our output through the Explore local statistics service, where it is located in the "similar areas" section of the "explore local indicators" feature, to explore similarities between their area and others.

4 . Strengths and Limitations

Strengths

- The inclusion of both the K-means and the statistical nearest neighbours approach means that users can choose the size of similar groups to better meet individual needs, increasing the flexibility of our output.
- The visualisations published via the ELS service, alongside our output, provide an interactive way for a range of users to explore similarity and view subnational trends, with the demographic model used as one of the comparator groups on the service.
- With the vast amount of work and data gathering across government, we were able to include several variables, which were not available to us previously, such as housing completions and access to museums and supermarkets; accounting for these extra factors allows us to present a more nuanced view of similarity.

Limitations

- While our goal was to ensure all variables were updated to the same time period, we encountered some differences; as our data are compiled from various government departments, the latest updates of some of the data used span a period of approximately three years, which was a necessary decision as it allowed us to include data from more sources and to create a more complete underlying dataset.
- The reliability of our results is dependent on the accuracy and completeness of the underlying dataset; as we have used publicly available data, we are reliant on the quality checks from the data producers.

5 . Cite this methodology

Office for National Statistics (ONS), released 14 February 2025, ONS website, methodology, Clustering similar local authorities and statistical nearest neighbours in the UK, methodology