Clustering similar local authorities and statistical nearest neighbours in the UK

What the clustering similar local authorities and statistical nearest neighbours statistics cover, how we produce them, and their quality and comparability. Includes definitions and latest, past and upcoming changes.

Contact:
Subnational Methods for Dissemination team
subnational@ons.gov.uk

Release date:
18 March 2026

Next release:
To be announced

Table of contents

1 . Overview

Our clustering and statistical nearest neighbours analysis uses machine learning techniques on an underlying dataset of 29 subnational indicators to identify groups of statistically similar subnational areas. We produce a global model that focuses on overall similarity, and two topic-specific models that investigate economic and demographic similarity.

Clustering models are provided for lower-tier local authorities (LTLA) only, while the statistical nearest neighbour lists also include upper-tier local authorities (UTLA) and combined authorities, as well as the LTLAs. Groups of similar areas have a range of applications, outlined in the Users and uses of these statistics section, including identifying suitable comparators for policy analysis.

The How we produce our statistics section covers the K-means and statistical nearest neighbours methods in detail. It also covers the transformation and aggregation techniques we have developed to ensure full geographical coverage of the chosen indicators.

In the Quality of the statistics section, we describe the steps taken to ensure that the input data feeding into our models is accurate. This section also covers our feature selection process. This is where we use principal component analysis (PCA), sensitivity analysis and correlations to ensure all indicators in our final model are having an appropriate impact on results and contributing unique information.

We publish an accompanying dataset, which includes cluster allocation, lists of nearest neighbours, data sources, and supplementary model information. The results are included on the Explore local statistics service through its "similar areas" feature. This is where users can view the clusters and neighbours on a map, and use the similar groups of areas for comparisons across more than 90 indicators. For expert users, we also publish a version of our modelling code on GitHub, which can be downloaded and adapted to meet specific user needs.

Further information on data gathering is available in the Where the data come from section of this release. Information on the specific indicators used is available in the data sources tab of the accompanying dataset.

These are official statistics in development. For more information, see the Quality of the statistics section.

2 . Latest changes to quality and methods

We updated this guide on 18 March 2026. Important changes to quality and methods include:

- the adding of statistical nearest neighbours for combined authorities
- the creation of separate lists of statistical nearest neighbours for LTLAs and UTLAs, which were previously combined; this has been changed to make a fair comparison between areas of similar size and administrative responsibilities
- the introduction of sensitivity analysis to our indicator selection process

For more information on latest, past and upcoming changes, go to Changes and their effects on comparability over time.

3 . What the statistics cover

This quality and methods guide explains the two ways we define similarity of local areas across the UK, clusters and statistical nearest neighbours. Clusters are groups of statistically similar areas that we produce for lower-tier local authorities (LTLA). Nearest neighbours provide an ordered list of neighbours for each area, based on their relative statistical similarity.

We produce lists of statistical nearest neighbours for LTLAs, upper-tier local authorities (UTLA) and combined authorities. All geographies included in the release are up to date as of 1 March 2026. We produce a global model that focuses on overall similarity, and two topic-specific models that investigate the economic and demographic similarity to give users the option to focus on those themes.

4 . Where the data come from

Initial data gathering

We gather initial data by searching for new and updated indicators from a variety of sources. These sources include the following:

- indicators from previous phases of this project
- our Explore local statistics (ELS) service
- Nomis
- data portals from other government departments
- the devolved governments

To be considered for inclusion in our analysis, data must:

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

We identified 33 UK indicators and 6 England only indicators that satisfied these conditions and entered our indicator selection process. These consist of 28 indicators from the Office for National Statistics (ONS) and devolved government, and 11 indicators from other government departments.

5 . How we produce the statistics

Brief overview of the main steps to produce these statistics

1. Identify indicators that satisfy the inclusion criteria.
2. Process indicators and apply transformations to remove missingness, creating the underlying data file (aggregation/weighting/imputation).
3. Quality check pre-processed indicators against the source data.
4. Prepare the quality assured data by limiting the effect of extreme outliers (winsorizing) and standardising the indicators.
5. Create initial clustering and nearest neighbour models.
6. Run principal component analysis (PCA), correlation and sensitivity analysis on initial models, and assess the impact of each indicator on the results.
7. Use the results of testing to decide which indicators to keep.
8. Rerun the clustering and nearest neighbour models with the new indicator list and recheck.
9. Check suitability of clusters and nearest neighbours.
10. Collate and quality assure the models and all supporting information.

K-means clustering

The K-means clustering method identifies and groups similar data points together within a dataset. The K-means algorithm works by starting with a chosen number (denoted by k) of random centroid points, then each individual data point is assigned to its closest centroid to form provisional clusters. The total distance between points and their centroid is calculated and stored. The algorithm then updates the centroid points to be central within each provisional cluster, the distances are calculated again, and new clusters formed. This process is repeated 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:

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

- 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

Our clustering approach

Our models include all UK lower-tier local authorities (LTLA), and we optimise the number of clusters between 4 and 15 for each of our models.

We have created one global model and two topic-specific models in this release. We produce the economic and demographic topic-specific models as they align well with the available input data and have established use cases. The economic and demographic models include only the indicators that align with those themes. The global model contains 22 indicators covering a variety of additional themes, including well-being, education and digital connectivity.

The global model should be used when a comprehensive view of similarity is required, and the topic-specific models are produced for analysis of policies that are relevant to those themes. These clusters can be used to identify broad similarities and investigate geographical patterns within the data.

For more information on the specific indicators included in each model, please see the definitions tab of our accompanying dataset. Summaries of the geographical and statistical features of the clusters from each of our models can be found on table 10.

Silhouette scores

Silhouette scores are a measure of how distinct the clusters are from each other. We calculate them by taking the ratio of the closeness of a point to all other points in its cluster, and all other points in the next-nearest cluster. We use these scores to assess the quality of our clustering models and to 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. Our silhouette scores range from 0.37 to 0.72 for the three clustering models we produce, showing that the clusters produced by our models are well defined. The most distinct results are produced by our global model.

Statistical nearest neighbours

The statistical nearest neighbours output provides an ordered list of the 20 most similar areas for any given lower-tier local authority (LTLA) or upper-tier local authority (UTLA). It also provides an ordered list of five most similar areas for each combined authority, across our three models. The nearest neighbours are limited to the same geography type as the area of interest, to ensure a fair comparison of areas with similar populations and legislative responsibilities. These lists, alongside the clusters, can be viewed through the "similar areas" feature of the Explore local statistics (ELS) service.

While the clusters are useful when a broad view of similarity is required, we produce the statistical nearest neighbour output in response to user demand for additional flexibility in identifying smaller groups of similar areas. This approach also allows us to include counties and combined authorities in our analysis, as there are too few of these areas to create a robust clustering model. These ordered lists are more suitable for projects that are looking to identify a small number of similar areas for direct comparison. An example of this is finding an area with a similar profile to use as a control to track the impact of a policy intervention over time.

We use Euclidean distance, a measure of the straight-line distance between two points in a multidimensional space to generate our statistical nearest neighbours lists. For each area, the Euclidean distance to all other areas, based on our underlying dataset, is calculated and these distances are stored. These values are then ordered and the areas with the lowest distances become our lists of statistical nearest neighbours. Euclidian distances are dimensionless and could be seen as similarity index scores instead of physical distances.

These distances are 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 (indicators)
- Ai and Bi are the indicator values of the points A and B

Our statistical nearest neighbours and clustering models are independent of one another, so the nearest neighbours for an area may not all be in the same cluster. This is expected because many areas will be situated at the edge of a cluster, and some of our clusters have a lower number of areas. Tests of how often this happens find an average of 77% of neighbours in our global model are within the same cluster as the area of interest, this figure is 78% and 84% for the economic and demographic models respectively. This shows that our neighbours lists are well aligned with our clustering models and further indicates the distinctness of our clustering results.

We use Euclidean distance as the distance measure for nearest neighbours for consistency with the K-means clustering algorithm. We prioritise consistency because testing Manhattan distance as an alternative measure produces a similar pattern of agreement between neighbours and cluster allocations.

Data transformations

To ensure compatibility with our models, several transformations are applied to the data. These transformations are important to ensure that our data are as complete as possible, as the K-means clustering algorithm cannot process missing data. The transformations also prevent indicators with a larger scale from impacting the models too greatly.

Aggregation and weighted imputation to generate county and combined authority data

We use aggregation or weighted imputation techniques to create data for counties or combined authorities when they are not available from the source. Our aggregation process involves adding count values from the LTLAs, and sometimes recalculating rates or percentages from those count values, to create figures for the larger areas. Where aggregation is not possible, weighted imputation is used to estimate missing values based on the available data.

Our imputation method weights indicators based on population data. We choose the population source that best matches the sample of the indicator, using recent mid-year-population estimates. For example, the children in relative poverty source data do not include combined authorities, so we weight the figures for the underlying lower-tier local authorities by their child-age population. We then combine and divide by the total child-age population for each combined authority to create the estimates for our dataset.

Converting data from obsolete to current local authority boundaries

All of the indicators we use were published between 2021 and 2025, a period that includes several changes to local authority boundaries that are often reflected in the data. Data for obsolete local authority area boundaries need to be converted to the 2025 version of the boundaries to ensure that the new local authorities are not missing from our output. This is achieved through the same aggregation and weighted imputation approach as the county and combined authority data, allowing us to update the data to reflect current geographic boundaries more accurately.

Filling gaps in lower-tier local authority data with corresponding upper-tier local authority data

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

Winsorization to reduce the impact of outliers

Initial data checks show the presence of extreme values in some indicators. Unchecked, this can impact the model through the formation of very small clusters. After testing different methods, we chose to apply 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 results. We apply this approach to all indicators for consistency.

Standardisation of data

Before the data goes through the clustering algorithm, we standardise it to ensure consistency across all indicators. Standardisation involves transforming each indicator to have a mean of zero and a standard deviation of one. This step ensures that the unit or scale of an indicator does not change its impact on the model.

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

6 . Quality of the statistics

Statistical designation

These statistics are labelled as "official statistics in development". They are based on information from a range of government sources. We are developing how we collect the data and produce the statistics to improve their quality.

Once we have completed the developments, we will review the statistics with the Statistics Head of Profession.

If the statistics meet trustworthiness, quality and value standards based on user feedback, we will remove the "official statistics in development" label to publish under the "official statistics" label.

If they do not meet trustworthiness, quality and value standards, we will further develop them, and may stop producing them.

If they were "accredited official statistics" before the start of the developments, we will ask the Office for Statistics Regulation (OSR) to reassess and re-accredit them.

We will inform users of the outcome of our, and any OSR, review and any changes.

How we quality assure the data and statistics

As this project requires collation and processing of many indicators, it is important to conduct robust quality assurance after data processing to ensure that no errors occur at this stage. We do this by comparing our processed data against the source data, manually rebuilding any imputed results in a separate process to ensure accuracy. Additionally, all geography lookup files used in the code are checked to ensure accuracy and timeliness in all inputs.

We analyse the clusters and statistical nearest neighbours outputs to check that they align with our knowledge on subnational trends. Additionally, we conduct a manual check of the nearest neighbour for each area, ensuring that any results are logical, and investigating any unexpected outcomes.

Feature selection

We use principal component analysis (PCA) to identify the most informative components and reduce the total number of indicators used in the final models. This reduction in the number of indicators allows us to focus on the most important aspects of the dataset. We also analyse the variance of and correlations between the indicators, as pairs of correlated indicators can skew the models, while low variance indicators contribute less information.

As part of our feature selection process, we also conduct a sensitivity analysis on our models. This involves removing one indicator at a time and analysing the level of change. Where an indicator has an outsized impact on the models, or a large negative impact on the silhouette score, it is also considered for removal. Using these techniques, we removed a total of 10 unsuitable indicators from our models, simplifying and improving the final output.

Strengths and limitations

Strengths

- The inclusion of both K-means clusters and statistical nearest neighbours supports different uses; clusters allow users to view groups with broad similarities and investigate geographical trends, while statistical nearest neighbours provide individual close comparators for policy analysis.

- The visualisations published through the Explore local statistics (ELS) service, alongside our output, provide an interactive way for a range of users to explore similarity and view subnational trends; the global, demographic and economic models are all used as comparator groups on the service.

- With the extensive data gathering across government, we can include high-quality indicators across a wide range of topic areas; this allows us to find nuanced links between areas that more limited models would miss, through analysing more dimensions of similarity.

Limitations

- We do not have direct control over the quality of all source data; however, we use publicly available data and take steps to assess the quality and suitability of each source.

- The clustering and statistical nearest neighbours output is designed as a snapshot, and should not be viewed as comparable over time with previous versions of the output.

- The random initialisation of the centroids in the K-means clustering algorithm can lead to instability in the results and sub-optimal solutions; to mitigate this, our models are each initialized 10,000 times and the result with the best silhouette score is chosen.

European Statistical System Quality Dimensions

The Office for National Statistics (ONS) has developed Guidelines for measuring statistical quality, based on the five European Statistical System (ESS) Quality Dimensions. These are:

- relevance
- accuracy and reliability
- timeliness and punctuality
- comparability and coherence
- accessibility and clarity

We have integrated these considerations into the guide.

7 . Changes and their effects on comparability over time

Latest changes

This output is not designed to be comparable over time. Improvements have been made over the course of development of this methodology, and we advise using the latest version to benefit from the latest methods and input data. View the Related links section of this guide to find our previous methodology articles.

The latest changes were made on 18 March 2026 with the release of this guide.

We included combined authorities within the statistical nearest neighbours output

These were processed as a separate model including additional (England only) input datasets. This meets a user need from combined authorities to identify suitable comparator areas for state of the region reporting.

We provided separate lists of statistical nearest neighbours for upper-tier and lower-tier local authorities

This was to ensure the comparison of areas of similar sizes and administrative responsibilities.

We further improved the indicator selection process

We did this using sensitivity analysis to identify indicators that are having a disproportionate impact on the results.

Past changes

The following changes were made to the methodology on 14 February 2025.

We added the statistical nearest neighbours output

This was to provide an ordered list of the most similar areas to each given area. This added flexibility for users on the number of similar areas to use.

We included counties within the statistical nearest neighbours output

Our early models did not include county councils. We added counties to the nearest neighbour output in response to user demand.

We improved the indicator selection process

This includes replacing some correlated data with composite or proxy measures.

The following changes were made to the methodology on 23 February 2024.

We introduced winsorization on all model inputs

This was to avoid extreme outliers in some indicators influencing the K-means clustering model.

We expanded the model to cover the whole of the UK

Our initial proof-of-concept covered England only. We expanded our models to cover the UK to meet user need.

Upcoming changes

We are pausing further developments to our clustering similar local authorities and statistical nearest neighbours methods, in line with recent Office for National Statistics (ONS) prioritisation decisions.

This guide will be updated if changes are made in the future.

8 . Comparability and coherence with other statistics producers

The clustering and nearest neighbour methods we use are all established statistical and machine learning techniques. As we share our code and methods, other government analysts apply our approach to their own areas of interest, so similar outputs may be seen across government.

Some private producers offer flexible nearest neighbours models that allow users to select their own metrics and create bespoke models. While this allows the user to customise outputs, our feature selection process ensures that all indicators included in our models are appropriate, and are contributing unique information. Models that have not been through that selection process are more likely to be skewed by correlated data.

Using Census data from across the UK, ONS Area classifications are produced at various levels of geography, including local authority district. While similar clustering techniques to this publication are used, each product has its own advantages. The 2021 Area Classification for local authority districts will be the best source for when in-depth analysis of demographic similarities between areas is required. However, this publication is more appropriate when a broader range of metrics or more timely data are required.

9 . Users and uses of these statistics

Our clustering and nearest neighbours outputs are used by policy influencers at all levels of government. Users access our outputs in different ways. These are directly from the dataset, indirectly through the functionality on our Explore local statistics service or by applying our methodology to their own data using our published code.

Several stakeholders have used our statistics.

- Regional and local government bodies have used our outputs to set realistic benchmarks for acceptable service provision, identifying comparators for analysis and tracking progress over time.
- Combined authority users plan to use our new output in their State of the Region reporting, enabling comparative assessment and providing economic evidence analysis to support policy making.
- Local government users have found the clusters and statistical nearest neighbours available through the Explore local statistics service helpful when communicating evidence to non-specialists.
- Analytical expert teams across local, regional and central government have adapted our published code to create their own similarity models for specific policy needs.

10 . Definitions

K-means clustering

K-means clustering is an unsupervised machine learning technique that we use to group statistically similar areas into clusters, based on a range of indicators.

Lower-tier local authorities (LTLA)

Lower-tier local authorities provide a range of local services. At the latest update of this guide, there are 296 lower-tier local authorities in England, made up of 164 non-metropolitan districts, 63 unitary authorities, 36 metropolitan districts and 33 London boroughs (including City of London). In Wales there are 22 unitary authorities. In Scotland there are 32 council areas. In Northern Ireland there are 11 local government districts.

Upper-tier local authorities (UTLA)

Upper-tier local authorities provide a range of local services. At the latest update of this guide there are 153 upper-tier local authorities in England made up of 63 unitary authorities, 36 metropolitan districts, 33 London boroughs (including City of London) and 21 counties. In Wales there are 22 unitary authorities. In Scotland there are 32 council areas. In Northern Ireland there are 11 local government districts.

Combined authorities

According to the Local Government Association, a combined authority is a legal body set up using national legislation that enables a group of two or more councils to collaborate and take collective decisions across council boundaries.

There are 15 combined authorities in England at the latest update of this guide. Combined authorities are larger areas that currently comprise between 2 and 17 lower-tier local authorities.

11 . Related links

Explore local statistics - ONS
Web page | Updated regularly
Find, compare and visualise statistics about places in the United Kingdom. This page presents the nearest neighbours and clusters alongside over 90 indicators where our outputs enhance exploration and visualisation options for users.

Subnational Statistics and Analysis; Clustering and Nearest Neighbours
Web page | Last updated March 2025
Public version of the code used to create the Office for National Statistics (ONS) clustering and statistical nearest neighbours analysis. This can be adapted by users to run their own analysis using our methods.

Clustering similar local authorities and statistical nearest neighbours in the UK, methodology
Methodology | Released 14 February 2025
Methodology information for our clustering and statistical nearest neighbours analysis, which groups UK local authorities with similar characteristics and outcomes.

This is the third of three methodology publications we produced during development. The quality and methods guide published on 18 March 2026 replaces all three historic publications and become the definitive guide from this date.

Clustering similar local authorities in the UK, methodology
Methodology | Released 23 February 2024
Methodology information for our clustering analysis, which groups UK local authorities with similar characteristics and outcomes.

This is the second of three methodology publications we produced during development. The quality and methods guide published on 18 March 2026 replaces all three historic documents and become the definitive guide from this date.

Clustering local authorities against subnational indicators, England
Methodology | Released 24 February 2023
Provided for reference only. This is a link to the first of three methodology documents we produced during development. The quality and methods guide published on 18 March 2026 replaces all three historic documents and become the definitive guide going forward.