Office for National Statistics

Clustering similar local authorities in the UK, methodology

Methodology information for our clustering 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:
23 February 2024

Next release:
To be announced

Notice

27 March 2024

We have corrected an error in our analysis which resulted in the silhouette scores being slightly underestimated, this impacts tables 2 and 3 in this article. This error did not impact the clusters or any other model information. This was caused by a coding issue and we have corrected our approach for future releases.

Table of contents

1 . Overview of main changes

This article provides the methodology and quality information for our clustering analysis, which uses publicly available data produced by a range of sources from across government to group UK local authorities with similar characteristics and/or outcomes. This analysis provides an update to our 2023 article, Clustering local authorities against subnational indicators, and its accompanying methodology paper, Clustering local authorities against subnational indicators methodology. Where this article refers to the "previous analysis", this is the publication it refers to.

While our previous analysis focused on metrics highlighted in the Levelling Up the United Kingdom white paper, published on GOV.UK, and analysed similarities between local authorities in England, the current models utilise a wider range of data to identify broader similarities between local authorities across the whole of the UK. Additionally, in this release we have focused on improving our methodology for data imputation and transformation.

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 evaluating the impact of interventions. The results should not be viewed as being a judgement about the performance of a local authority.

This methodology article will provide additional information on the sourcing and transformation of the data, and the methods that we used to generate our results and ensure their quality. This methodology has been applied to our Clustering similar local authorities in the UK data tables, which outline the assigned cluster for each local authority for each model, the data used to build the models, statistical information relating to the models and the regional composition of each cluster.

2 . Data sources and geographical coverage

Our previous release exclusively sourced data from our Subnational Indicators Explorer tool. For this release, we have explored alternative metrics to those laid out in the Levelling Up white paper to enable this work to be UK wide and identify broader similarities between local authorities. While some of the data included in this analysis are still sourced directly from the December 2023 edition of the Subnational Indicators Explorer, we have also included data from a range of publicly available data sources across government.

The data included in this release are updated to 19 December 2023. Information on the data sources used, including caveats and notes, can be found in our data tables.

Innovative methods were used to address difficulties presented by expanding the geographical coverage. We prioritised using single-source data covering all UK local authorities, but for many of our data sources, local authorities within the devolved administrations (Northern Ireland, Scotland and Wales) were not included. In these cases, we worked with the devolved administrations to source comparable data for all countries of the UK to ensure full coverage for all metrics.

Where comparable data were not available, we dropped the metric from the models. This disproportionally affected areas under devolved governance (for example, health and education), which may be underrepresented in the models.

For instance, we initially sought to include data on unauthorised school absences. However, it was difficult to establish whether the definition of "unauthorised absence" was effectively comparable across countries. We also encountered coverage issues: England collects data for pupils aged 5 to 15 years and excludes part of the summer term for pupils in school Year 11, whereas Scotland's school absences data cover all pupils enrolled at school and are collected for the entirety of the school year. Finally, the latest available data for England were from the Autumn 2022 to Spring 2023 term, while for Scotland it was from the 2020 to 2021 school year. Because the 2020 to 2021 school year was atypical owing to coronavirus (COVID-19) restrictions, we decided we could not ensure data comparability across the UK for this metric. We plan on reviewing data exclusions in future iterations of this work.

When we identified definitionally coherent data from across the UK, we carried out distribution checks to ascertain whether there were substantial differences in distributions between the nations. We first ran the Shapiro-Wilkes test for normality on each country's data. If this highlighted that data were not normally distributed, we checked the mean and variance using T-tests and F-tests. Where distributions differed too drastically without a logical explanation, the data were removed from our models.

The country-level distribution for "Proportion of commuters travelling to work by bicycle/ by train/ on foot", which was sourced from the Census for England and Wales, the Census for Northern Ireland, and Scotland's Census, presented differences. However, we determined that these differences could be explained by variation in population density and service provision rather than differences in definitions or data-recording methods. After carrying out distribution checks, we worked with data owners from the devolved administrations to ensure the data could be included in our models.

Additionally, when the data were sourced separately for UK countries, we had to work around inconsistencies in temporal coverage. For example, Scotland's Census was carried out in 2022, compared with 2021 for England, Wales, and Northern Ireland. This meant that updated Scotland's Census data were not available at the right geography levels to be included in this analysis. Where possible we used 2021 mid-year population estimates; for metrics that could not be derived from mid-year population estimates, we reverted to 2011 Census. See the data tables for a detailed description of the different data sources used. We will look to include the updated Scotland's Census data in the next iteration of this work.

3 . Data transformations, standardisation and methods improvements

We used k-means clustering as a method for identifying and grouping similar data points within a dataset. Further information on how k-means clustering works can be found in the methodology article for our previous release.

The K-means algorithm cannot account for missing data: where a local authority has a null value for a single metric, the local authority cannot be included in the model. To maintain full UK coverage where possible, we imputed missing values in the data.

Imputation

Most metrics are defined at lower-tier (local authority district and unitary authority) level; however, male, and female healthy life expectancy 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. While this method has limitations, as imputation affects the geographical granularity of the data, we decided to include it because of the low availability of subnational health data with full UK coverage.

Where data for a local authority were absent from the original sources, the missing data had to be imputed or the local authority dropped from the model. The local authorities most affected by this issue were City of London and Isles of Scilly, which were missing from 3 and 7 metrics, respectively, because of their small sample sizes.

Because these local authorities were missing from a considerable proportion of metrics, City of London and Isles of Scilly are only included in the demographic and connectivity and sustainability topic-level models. The Orkney Islands are also not included in our health and well-being and global models because disclosive well-being values are suppressed in the source data.

In 2023 four new unitary authorities were formed: Cumberland, Westmorland and Furness, North Yorkshire, and Somerset, as described in our Local government restructuring methodology. These aggregate and replace 18 local authority districts (LADs). For this analysis, we use the latest available data, which for some metrics pre-date 2023 and reports figures for the, now inactive, 2022 LADs rather than the current unitary authorities.

As many of our metrics are rates, simply adding up the figures for the former LADs to get the current unitary authorities was not an option. To solve this issue, we devised a proportional imputation method. Since all the changes aggregated smaller LADs into a larger unitary authority, we used either census population by single year of age or area to account for different LAD population and area sizes and to calculate new "weighted" figures for the 2023 unitary authority (see Table 1). Deciding between census population or area, we selected what we considered to be the most appropriate denominator for each dataset.

Table 1: Data imputation for local authority boundary changes

2022 area name	2022 area code	Employment rate aged 16-64	Census 2021 population aged 16-64	Proportion of sum of populations	Employment rate * population proportion	2023 area name	2023 area code	Imputed value (sum of employment rate * population proportion)
Allerdale	E07000026	77.7	56700	0.34	26.71	E06000063	Cumberland	78.04
Carlisle	E07000028	81	67466	0.41	33.14			
Copeland	E07000029	73.6	40745	0.25	18.18			

Source: Office for National Statistics

Transformation and standardisation

Distribution analysis revealed that some metrics were skewed and displayed extreme values. K-means clustering is particularly sensitive to skewedness and extreme values in the data. Therefore, to reduce the impact of extreme values and prevent single clusters forming, we explored logarithmic and inverse hyperbolic sine (IHS) transformations and standardised using z-scores.

We calculated silhouette scores, as defined in this Silhouette Coefficient article published on the Medium website, for models using non-transformed, log and IHS transformed data to determine which approach led to more distinct 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. As shown by Table 2, the models using non-transformed data had significantly higher silhouette scores, meaning that the resulting clusters are more distinct. We therefore made the decision to avoid using data transformation techniques.

Table 2: Silhouette score by transformation technique

Model	Non-transformed	Logarithmic	Inverse hyperbolic Sine
Global	0.604	0.277	0.248
Economic	0.606	0.251	0.251
Demographic	0.696	0.343	0.356
Health/Wellbeing	0.304	0.307	0.343
Connectivity/Sustainability	0.550	0.308	0.307

Source: Office for National Statistics

Notes

1. Silhouette scores measure how distinctive the clusters in each model are. Scores range between -1 and 1 where a score of 1 represents clusters being as clearly distinguishable from each other as possible.

Initial data checks showed the presence of extreme values in some metrics. This led to the creation of very small clusters of local authorities in some of our initial models, which are less useful as control groups for analysis. This led us to devise an extreme values management approach. Owing to the issues created in the models by missingness outlined earlier, dropping extreme values from the affected metrics was not a feasible solution to this issue. We decided to cap extreme values at the 1st and 99th percentile values. Because typically only a small number of local authorities are major extreme values in the data, this approach has the advantage of bringing the extreme values in line with the rest of the data while maintaining their high or low value for the given metric, leaving most of the data unchanged. We applied this approach to all metrics for consistency, whether the data had extreme values or not.

Table 3: Silhouette score for percentile smoothed and non-percentile smoothed data

Model	Non-transformed	Percentile smoothing
Global	0.604	0.603
Economic	0.606	0.604
Demographic	0.696	0.697
Health/Wellbeing	0.304	0.308
Connectivity/Sustainability	0.550	0.546

Source: Office for National Statistics

Notes

1. Silhouette scores measure how distinctive the clusters in each model are. Scores range between -1 and 1 where a score of 1 represents clusters being as clearly distinguishable from each other as possible.

Other methodological improvements

The K-means clustering algorithm works by choosing a starting centroid, called a seed. This can either be specified or chosen randomly, and a different seed may lead to different clusters being formed. In our previous release, we set the same randomised seed value for all our models. In this analysis, the seed is randomly chosen, but the clustering algorithm is set to run 10,000 times and select the best-performing clusters in terms of inertia. This is a measure of the distance of points within clusters from the cluster centroid, as explained in this Clustering with K-means article, published on the Medium website.

4 . Model construction

Once we identified a list of metrics that conformed to our geographical needs we ran correlation analysis, using Pearson's correlation, to identify highly correlated metrics. High correlation was not an exclusion criterion for our data because K-means clustering can handle correlated variables in the same model. Correlation analysis was used to identify variables for further investigation, as we were concerned that very high correlation between metrics that measure similar concepts would double the "weight" of that concept in the model, disproportionally affecting the results. As a result of this analysis, we dropped the metric "feeling that the things done in life are worthwhile", which was highly correlated with another well-being metric (life satisfaction), and country of birth, which was highly correlated with "proportion of population that belongs to a religion" and ethnic group.

Global Model

The global model includes all 34 metrics and aims to identify local authorities that are similar across several topic areas. These metrics are listed under the individual theme models below, and information on the data sources used can be found in the accompanying data tables. The aim of this model is to provide the widest possible overall view of similarity between local authorities. The resulting clusters are best suited for use by local authorities seeking to identify other local authorities that may be facing similar challenges, and to create control groups for evaluating the impact of interventions.

Theme models:

Alongside the global model, we also produced models based on subsets of metrics (full list and data sources available in the accompanying data tables) relating to individual themes; economy, demographic, health and well-being, and connectivity and sustainability, which local authorities can use to assess similarity relevant to specific policy areas. These are:

Economic model:

- employment rate for 16- to 64-year-olds
- gross disposable household income per head
- gross value added per hour worked
- median house price
- percentage of businesses born in 2022
- percentage of children in relative poverty
- percentage of workers employed in the construction sector
- percentage of workers employed in the manufacturing sector
- percentage of workers employed in the services sector
- rate of businesses per 10,000 people

Demographic model:

- population change from 2011 to 2021 (percentage)
- population density (per square km)
- proportion of population aged 15 years and under
- proportion of population aged 16 to 64 years
- proportion of population aged 65 years and over
- proportion of population of Asian or Asian British ethnicity
- proportion of population of Black, Black British, Caribbean or African ethnicity
- proportion of population of Mixed or multiple ethnic groups
- proportion of population of White ethnicity
- proportion of population that does not belong to a religion
- proportion of population that belongs to a religion

Health and well-being model:

- anxiety
- female healthy life expectancy
- happiness
- life satisfaction
- male healthy life expectancy
- proportion of adults that currently smoke cigarettes

Connectivity and sustainability model:

- CO2 Emissions (per square km)
- domestic mean electricity consumption in 2021 (kwh per meter)
- gigabit capable broadband coverage (percentage)
- proportion of commuters travelling to work by bicycle
- proportion of commuters travelling to work on foot
- proportion of commuters travelling to work by train

Global model only metrics

- proportion of the population aged 16 to 64 years with level 3 or above qualifications

5 . Limitations and further work

Limitations

Our results aim to map which places have similar characteristics and/or 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.

In our previous release we carried out lookup analysis where we explored the demographic and geographic composition of the clusters, which we only replicated for ITL 1 (regions) in this update. In this release, where possible, we included this information directly in the models (population density, age). However, some lookups were not available with UK coverage (IMD, rural/urban classification, coastal towns). In these cases, we were unable to include these metrics in the models or carry out lookup analysis using the data.

Further work

In the future, we aim to continue working with devolved administration colleagues and UK-wide/coherence colleagues within the Office for National Statistics (ONS) to expand the availability of definitionally coherent data for the whole of the UK. We will consider whether we can produce similar analysis at the upper-tier local authority geography, depending on user needs.

We aim to include visualisations of the clusters and the capability to compare with similar local authorities as part of our Explore Subnational Statistics service.

Additionally, we will work to include the UK-wide datasets we have sourced into our subnational dissemination products, such as our Subnational indicators explorer webpage, where appropriate.

Finally, we will share the methodological insights derived from the clustering improvements we have implemented and the data imputation process with colleagues across government. We will share the code used in this analysis via our GitHub repository shortly.

If you are interested in the code used to produce this analysis or have any feedback on this methodology article or on the accompanying data tables, please feel free to contact us at subnational@ons.gov.uk.

6 . Cite this methodology

Office for National Statistics (ONS), released 23 February 2024, ONS website, methodology, Clustering similar local authorities in the UK methodology

.