Project report

Paula Cordero Encinar

September 2023

1 Introduction

Characterizing the structures of cities, towns and neighbourhoods constitutes an important task since the identification of similar cities can promote sharing respective experiences. In particular, it is interesting to obtain clusters of similar locations at different resolution levels.

The specific aim of this project is:

1. To classify the hexagons (H3: Uber's Hexagonal Hierarchical Spatial Index) of resolution 8 or 7 covering the UK into 6 classes that range from totally rural to completely urban.
2. To analyse the performance of the classification in terms of different metrics and visualize the output of the clustering.

However, this objective is not without its challenges, such as the determination of the variables to be used for the analysis, as well as the selection of the specific algorithm to be employed or the lack of a ground truth or benchmark to test our results.

2 Literature review

Different works in the literature have addressed similar problems of place classification not only restricted to the urban-rural distinction. However, all the reviewed works perform the classification at a lower granularity, e.g. cities or villages, or census tracts, but none of them using the hexagonal grid system H3.

Some papers use demographic and economic characteristics to classify their areas of interest (Morton et al., 2018). Others base their classification in characteristics of their street or road network. For example, da F. Costa and Tokuda (2022) use the following five features of the street network: average of the vertex degrees; standard deviation of the vertex degree; standard deviation of the vertex local transitivity (allows to measure the interconnectivity around each node); dispersion of the point locations (considers geographical distribution of vertices) and standard deviation of the vertex accessibility (vertex accessibility generalizes the concept of vertex degree, it quantifies the frequency of the visits to each node of self-avoiding random walks departing from all other nodes h steps before, where $h = 2$ in the mentioned paper). It might be interesting to build on this work by carrying out the analysis separately for walking streets and for roads, and incorporate some extra features as the number of major (principal) roads or in terms of streets the number of streets with a significant number of facilities (supermarkets, shopping areas, hospitals). While Boeing (2019) uses OpenStreetMap data to obtain the street network of a city and characterize it using the entropy of street bearings in weighted and unweighted network models, along with each city's typical street segment length, average circuity, average node degree (which was used before too), and the network's proportions of four-way intersections and dead-ends. Other works in this line is the ones of Varoudis and Penn (2021).

Alternatively some analyses employ landuse data from Open Street Map in the form of statistics or directly from images using computer vision techniques (Dobesova, 2020). It is important to mention that many works use a combination of these features to capture more information about the instances they are trying to classify.

Lastly, the Office for Nationa Statistics (ONS) also provides documentation of the methodology they use for the urban-rural classification at the local authority level, where they perform a 3-fold and 6-fold classification (`https://www.ons.gov.uk/methodology/geography/geographicalproducts/ruralurbanclas`

sifications/2011ruralurbanclassification). Their idea is to classify each settlement using the population density at different radii from the centre of the settlement and then aggregate them to obtain the corresponding class at the local authority level.

3 Methodology

As mentioned above, the main objective is to perform a 6-fold classification of the resolution 8 or 7 hexagons covering UK. The project is structured in two parts: the first one deals with the selection of variables to perform the clustering and the second one focuses on the clustering algorithm itself. The overall workflow is summarised in Figure 1.

3.1 Data extraction and feature selection

For the project, we have decided to combine demographic data from the Office for National Statistics (ONS) and Scotland Census, together with geospatial data from Open Street Map (OSM). It is important to note that while OSM geospatial data can be obtained directly at the hexagon level, demographic data cannot. The latter are available for different census geographical areas defined by the ONS. To calculate the values of the different attributes at the granularity we want, we need to interpolate the data. Simple interpolation could be done using the the function `tobler.area_weighted.area_interpolate` from the `tobler` library. However, this does not lead to sensible results because there might be hexagons in the census areas with no inhabitants and a uniform interpolation would assign a non-zero value to these hexagons for the different attributes. We have therefore decided to also use data from a weighted interpolation, where the weights are given by the proportion of residential area in the hexagon of interest.

From the census data, we have computed the following features:

- **Population density** derived from the total number of inhabitants in the `age` collection of the `h3r8data` database or calculated by linearly interpolating data in the `populationdensity` collection of the `gb_nationalstatistics` database at the hexagon resolution 7 level.
- **Average age** obtained either from the `age` collection of the `h3r8data` database or linearly interpolated at the hexagon resolution 7 level using the `age` collection of the `gb_nationalstatistics` database.
- **Average household size** obtained from the `housholdsize` collection of the `h3r8data` database or linearly interpolated at the hexagon resolution 7 level using the `householdsize` collection of the `gb_nationalstatistics` database.

In the case where we are performing a linear or uniform interpolation to calculate values at the hexagon resolution 7 granularity, the census data used is obtained from the `gb_nationalstatistics` database where data is available for different types of census tracts. It is important to highlight that for England there is no age data available at the outputs area level so we have used data from local authority districts as it is available for the 3 ONS collections we are interested in. In the case of Scotland, we have information about all the variables we are interested in at the outputs area level, therefore we have use that for the interpolation. We have employed resolution 7 to reduce the computation times when interpolating the data.

It is also worth mentioning some data issues, that would need further exploration. The collections used from the `h3r8data` database that is already interpolated using weights at the hexagon level have different lengths which mean that not all resolution 8 hexagon IDs are present in all the databases. To address this issue, we have only performed the clustering using the hexagons which are in the intersection of the collections used which results in some gaps in the final map with the clustering output.

While from the OSM data stored in the `h3r8data` or `h3r7data` databases we have extracted:

- **Landuse data** taken directly from the collection of the same name.
- **Length of tertiary and residential roads** from the `roads` collection. Tertiary roads are those connecting smaller settlements, and within large settlements, roads connecting local centres. Whereas residential roads primarily provide access to dwellings, with no connecting function between settlements.

3.2 Clustering algorithm

We have decided to perform the clustering algorithm in two steps, this is because we originally wanted to perform a three-fold classification into rural, medium and urban areas, however some of the clusters derived from the initial classification cover large areas that are diverse and additional sub-classification is needed. It is important to remark that different clustering algorithms were tested including k-means, DBSCAN (density-based spatial clustering of applications with noise) and hierarchical clustering. Based on their performances using different clustering metrics (including Silhouette score, Davies-Bouldin index and inertia) and given that the number of clusters is determined in advance, we decided to use the k-means algorithm.

- **1st step: rural-middle-urban**. We start by performing a three-level clustering using census variables: population density, average age and average household size. In addition, we analyse the density distribution for each variable used in the algorithm based on the output label and the value of different clustering performance metrics.
- **2nd step: subclassification of rural and middle**. Looking at the results of the previous step, we observe that while the urban class is limited to large urban centres, the middle and rural classes are more extensive. This is one of the reasons that motivated us to further classify the latter two categories. We decided to distinguish three different classes within the rural category and two within the middle category. In this case, the variables selected for the clustering algorithm are a mixture of demographic and geospatial variables. Specifically, population density, area of residential land use type and length of residential and tertiary roads. As in the previous step, we have also done a post-analysis of the clustering output.

The post-analysis of the clustering results in both steps could be used for variable selection when other characteristics are available, as we can test which set of variables is best in terms of performance metrics.

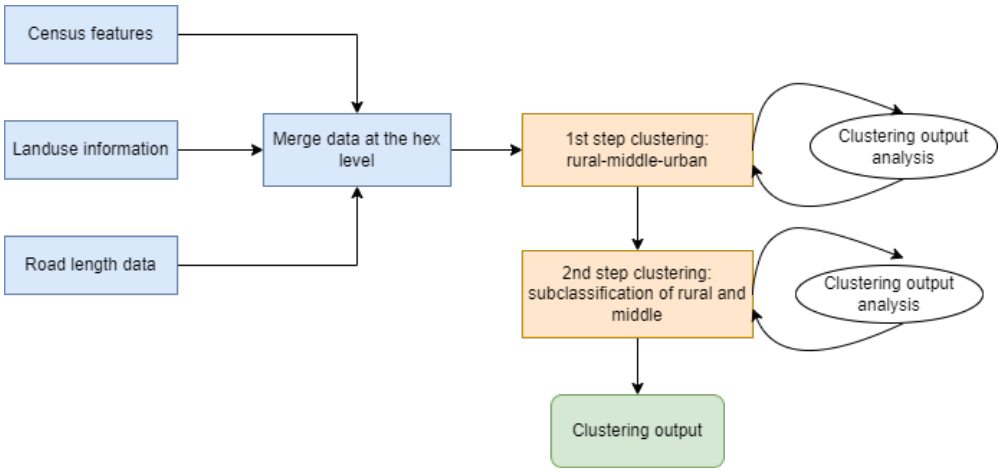

Figure 1: Overall workflow of the implementation.

4 Results

We now present the results from the post-analysis of the clustering outputs. Figure 2 shows the density distribution of the variables used in the first step of the clustering based on the output label, which distinguishes between urban, middle and urban, when using linear interpolation of the census data. We observe that population density increases in areas classified as urban, while the average age is, on average, lower in urban areas, which is consistent with our subjective knowledge.

On the other hand, Figure 3 shows the clustering output, where each category is represented in a different color. This is the result when using uniformly interpolated census data. We note that the main urban centres are coloured in blue, indicating the most urban category, while the red tones represent the middle class.

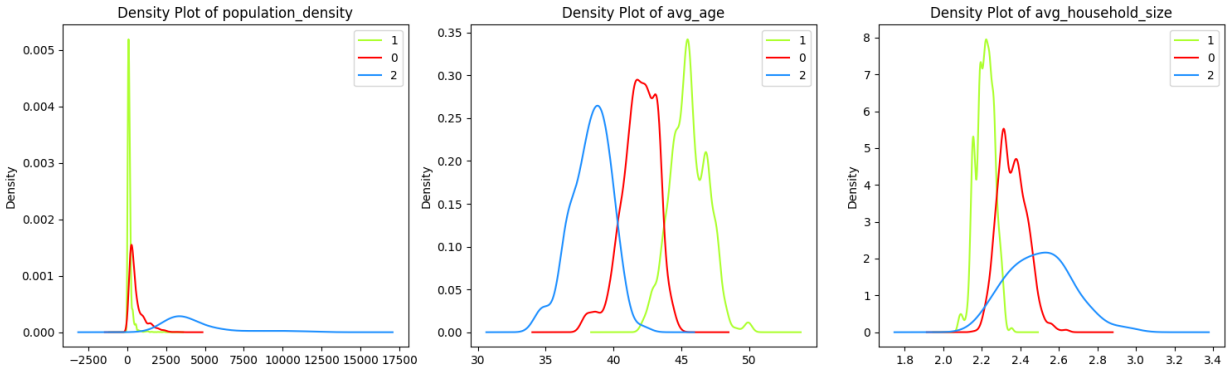

Figure 2: Density distribution of the variables used in the first step of the clustering for each label.

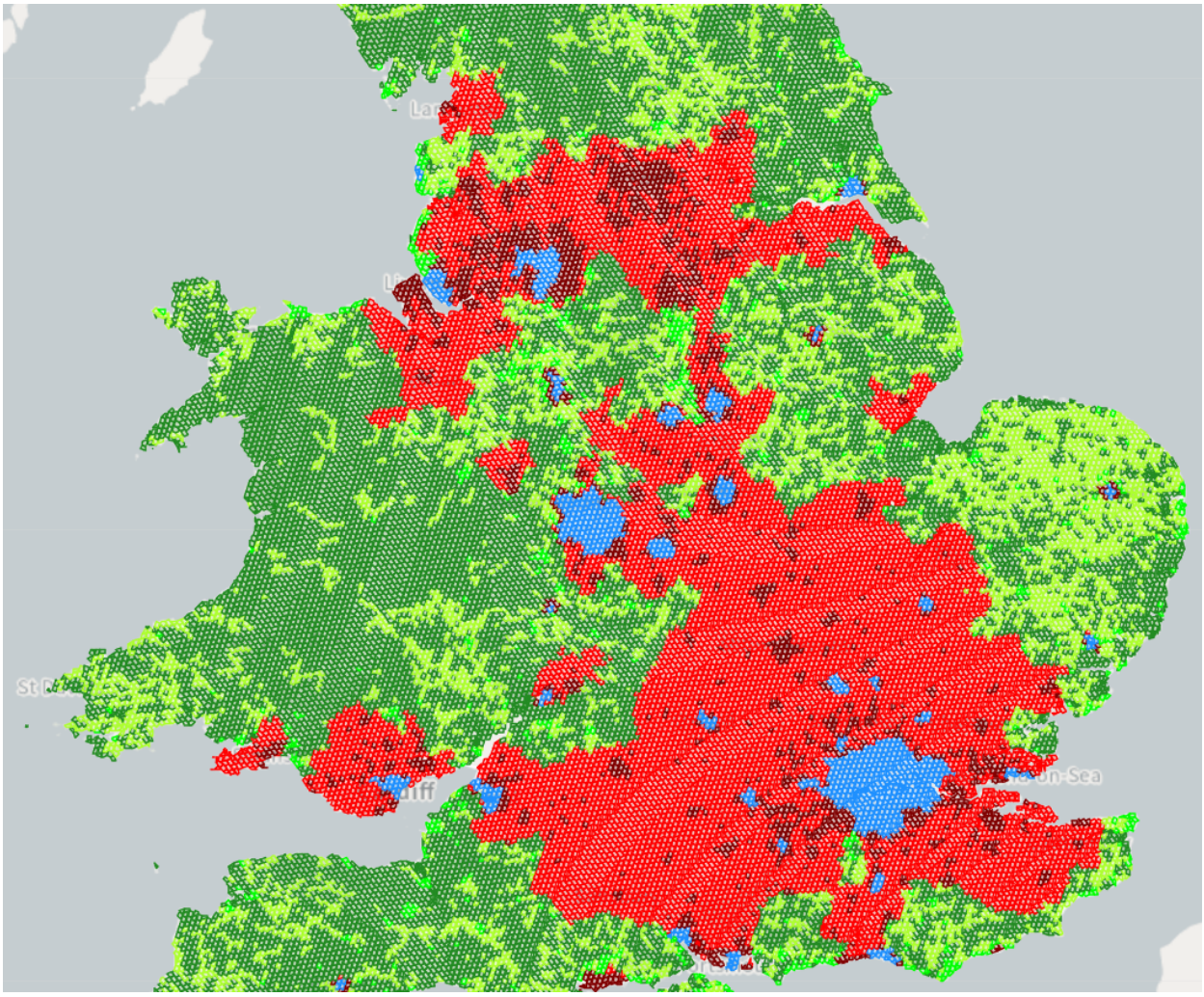

Figure 3: Result of clustering with 6 categories when using uniformly interpolated census data in the first clustering step..

5 Conclusions

This report summarises the algorithm to extract features from the data and perform the clustering. There are several lines of future work to improve this project.

Firstly, it will be interesting to manually label a certain region to assess the performance of the algorithm based on our subjective knowledge. In addition, once other variables are calculated at the hexagon level, we can use additional variables such as the number of workers in each industry or the number of people with

different levels of education for the clustering. To choose which model is better we can select the one with the best value for one of the performance metrics mentioned above.

It will also be interesting to establish a prototype of each cluster in order to understand its main characteristics. This will help to make future decisions regarding each cluster.

References

Boeing, G. (2019), 'Urban spatial order: street network orientation, configuration, and entropy', *Applied Network Science* **4**(1), 1–19.

da F. Costa, L. and Tokuda, E. K. (2022), 'A similarity approach to cities and features', *The European physical journal. B, Condensed matter physics* **95**(9).

Dobesova, Z. (2020), 'Experiment in finding look-alike european cities using urban atlas data', *ISPRS International Journal of Geo-Information* **9**(6).

Morton, C., Anable, J., Yeboah, G. and Cottrill, C. (2018), 'The spatial pattern of demand in the early market for electric vehicles: Evidence from the united kingdom', *Journal of Transport Geography* **72**, 119–130.

Varoudis, T. and Penn, A. (2021), Spectral clustering and integration: The inner dynamics of computational geometry and spatial morphology, *in* S. Eloy, D. Leite Viana, F. Morais and J. Vieira Vaz, eds, 'Formal Methods in Architecture', Springer International Publishing, Cham, pp. 243–250.