Identifying Cities Using Clustering

Author: Iurie Tarlev
Supervisor: Dr Nick McCullen

UNIVERSITY OF BATH

Introduction & Background

Urban boundaries in most countries are defined by either somewhat arbitrary administrative boundaries or population related metrics. In reality Identifying urban boundaries is a difficult and often subjective task, as the level of "urbanity" depends on various landscape (e.g. roads area) and socio-economic (e.g. income per household) characteristics.

Comnbination of computational techniques such as dimensionality reduction and clustering allow for identification of similarities between geographical regions based on aforementioned metrics. The aim of this study is assessment of those computational techniques applied for this task and comparison of the methods.

Key terms

Dimensionality reduction - technique used to reduce the number of dimensions in a multi-variate dataset into less dimensions, whilst preserving the similarities between the observations.

Clustering - "Unsupervised Machine Leaning" technique that allows identication of patterns in the data and groups the observations.

Methodology

Five different clustering algorithms were assessed on the data for Bath and North East Somerset data. Performance of clustering algorithms was evaluated using Silhouette Coefficients which measure compactness and separation of clusters. Performance of dimensionality reduction algorithms was assessed by comparing global and local clustering.

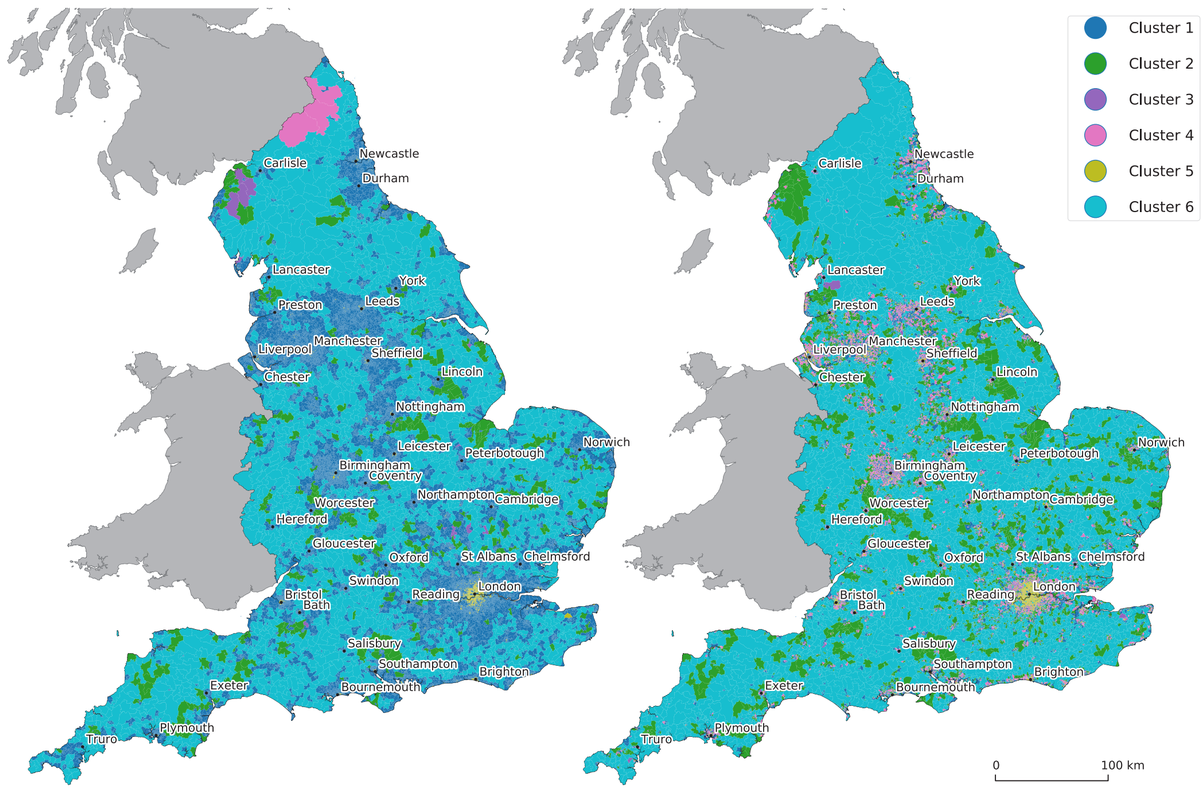
Dimensionality Reduction

Prior to clustering, dimensionality reduction was applied to reduce the number of dimensions in the data as well as de-correlate inter-related variables or increase visualisation performance. Two dimensionality reduction algorithms were used in this study: Principal Compoanent Analysis (PCA) and t-Distributed Stochastic Embedding (t-SNE).

The advantage of using PCA dimensionality reduction technique, is that inference back to the variables can be made, so it can be assessed what variables explain most of the data. Six dimensions outputted from PCA have been used for clustering as they seemed to explain about 75% of the total variation in the dataset.

In contrast, t-SNE is a more advanced non-linear and stochastic method for dimensionality reduction. It produces much richer visualisations and was found to be much better at preserving the local structure of the data.

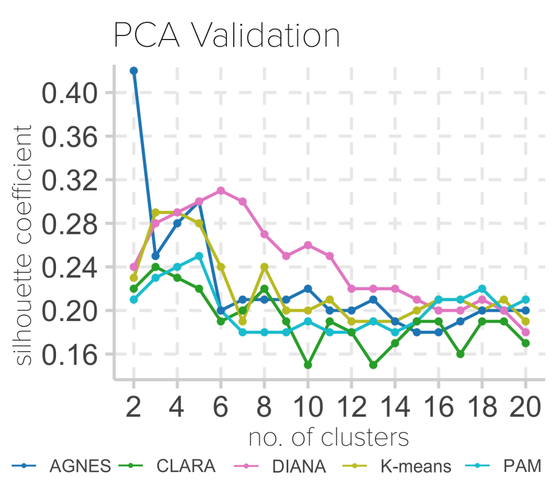
The maps below demonstrate t-SNE's ability to preserve local structure, as multiple types of urban areas have been detected, unlike in the PCA's dataset, where only two types of urban areas have been detected: urban (cluster 1) and super-urban (cluster 5), with several outliers spotted.



DIANA Cluster Map based on PCA and t-SNE reduced datasets respectively

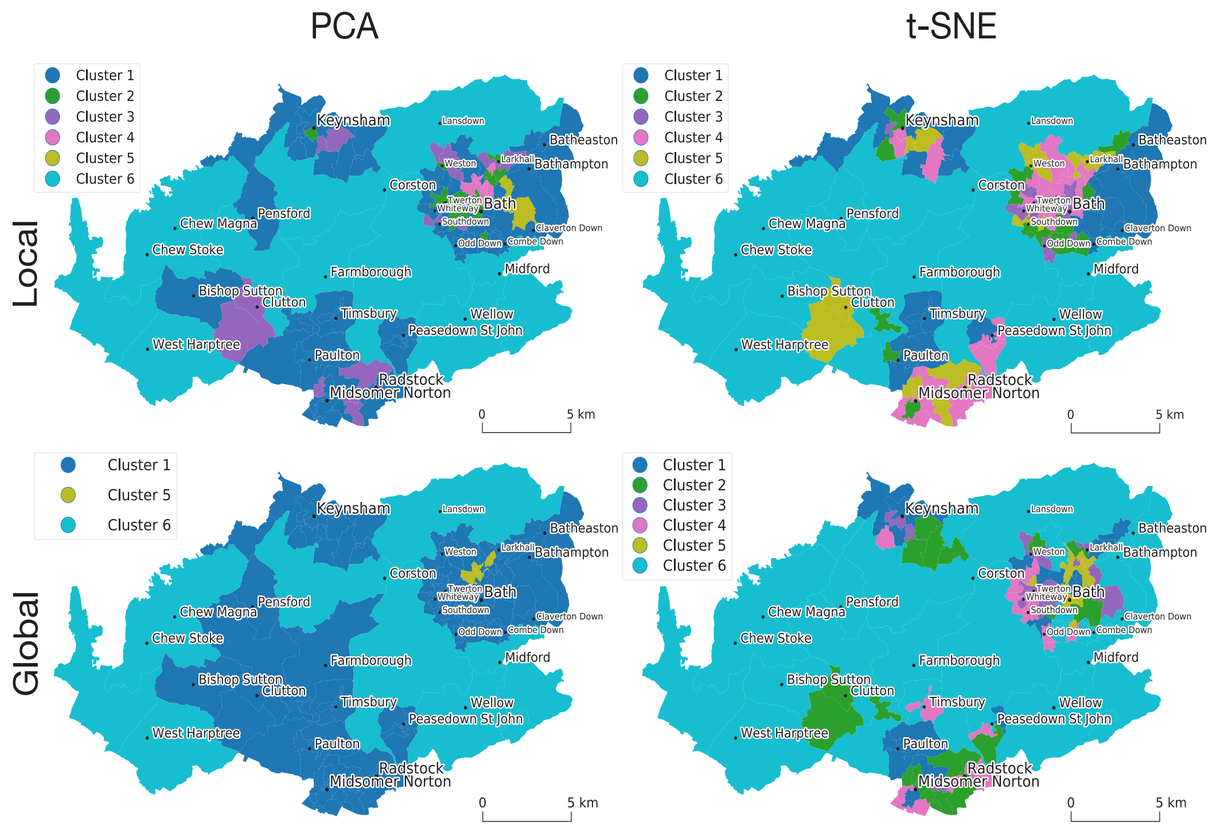
Clustering algorithms assessment

PCA (Principal Component Analysis) dimensionality reduced dataset was used for clustering algorithms assessment as it is a stable non-probabilistic method of dimensionality reduction, unlike t-SNE. As shown in the validation graph, DIANA clustering algorithm was considered to have the best sillhouette coefficient, based on 6 clusters, therefore considered optimal for this analysis.



Clustering based on global vs local variance

A comparison of clustering quality has been done between clustering on global (England) and local (Bath and North East Somerset) dataset. It was further confirmed that t-SNE is less dependent on the global variance and manages to preserve local as well as global similarities, as can be seen in the cluster map plots shown below.



Global vs Local Clustering on PCA and t-SNE reduced dataeset

Conclusion

Using Bath and North East Somerset dataset it was found that DIANA clustering algorithm has a comparatively better performance. This performance assessment is based on silhouette coefficients measure and on the PCA dimensionality reduced dataset. t-SNE dimesionality algorithm has better ability of preserving both local and global structure of the data, however this comes at great computational cost and inability to infer back to the input variables.

The recommendation of this research is to use t-SNE when global as well as local structure of the data is very important and sufficient computational resources are available, whilst using PCA when only global structure needs preserving. DIANA clustering is recommended for use as the primary clustering method.