A Modified DBSCAN Clustering Method to Estimate Retail Centre Extent

Michalis Pavlis[1], Les Dolega[1], Alex Singleton[1]
[1]University of Liverpool, Department of Geography and Planning, Roxby Building, Liverpool L69 7ZT, United Kingdom

January 13, 2017

1. Introduction

Although it has been argued that depicting retail agglomerations at national scale, particularly accounting for more granular temporal shopping patterns is very challenging (Mackaness and Chaudhry 2011), the classification of shopping destinations and delineation of their spatial extent is essential to gain a better understanding of the relationship between use of retail space and changing consumer behaviour. A consistent and rigorous approach to defining town centre boundaries enables systematic metrics of retail centre morphology and performance to be actualised (Thurstain-Goodwin and Unwin 2000), alongside providing utility as input into many commonly implemented retail analytics tasks related to store location and demand estimation (Newing et al. 2015). The objective of this analysis was the development of a methodology that would enable the automated identification of retail agglomerations across Great Britain based on a national dataset of retail locations that was provided by the Local Data Company (LDC) through the ESRC Consumer Data Research Centre.

2. Available datasets

The national occupancy dataset that was collected by LDC during 2015, contained information regarding the current occupier and location for 529,062 retail locations across Great Britain (GB). Data pre-processing included removing locations that did not have building level location accuracy (18% of the locations) and duplicate locations. Vacant outlets were also removed given that they often occur as a failure of a particular retail setting and as such might indicate a potential change in extent morphology.

Supplementing the LDC retail data, additional information regarding the retail areas in GB were available from two sources. Firstly, reports produced by local authorities within GB, which even though contain rich information, can typically only be accessed in rendered pdf format. Given that only a small number of (qualitative) comparisons can be made against these sources without an extensive re-digitising, the reports were used for method selection and during the calibration process. Secondly, boundaries for the top 339 retail places in GB were acquired from the company Geolytix, and although they represent only a subset of the total retail boundaries, they nevertheless provide an additional and relatively large sample of retail areas suitable for comparison.

3. Evaluation of the candidate clustering methods

Cluster analysis is a collection of unsupervised learning methods that address the issue of grouping a set of objects based on similarity. It is a multivariate technique (multiple attributes of the phenomenon under investigation can be used), but in this study it is strictly spatial; utilizing only the locations of the retail units. This is an appropriate approach for the identification of retail agglomerations where the extent of the clusters is determined by spatial discontinuity in unit distribution (Dearden and Wilson 2011).

Five candidate clustering methods (DBSCAN, Kernel Density Estimation, K-means, Quality Threshold, Random Walk) were considered in this analysis and were compared in 8 case study areas

(Abertillery and Cardiff in Wales, Bristol, Clapham Junction, Winchester and Wolverhampton in England, Inverurie and Glasgow in Scotland) that were representative in terms of retail location density, size and retail centre morphology. Based on the results of the analysis the DBSCAN method was selected, however, this method is known to underperform in areas where the density is not uniform (Everitt et al. 2011). The reason being that the optimal DBSCAN parameter values (i.e. for the epsilon parameter which represents the radius that two points can be neighbours and the minimum points parameter which represents the minimum number of neighbours for every core point) depend on the point density of the study area. As such, we developed a refinement to the method which involves splitting the national-scale data into more homogeneous areas for separate treatment.

4. Development and application of a modified DBSCAN method

In the first step of the proposed methodology, a sparse graph representation of the spatial dataset is created based on a k-nearest neighbour matrix (where k is equal to the value of the minimum points parameter of DBSCAN) and the maximum distance constraint. The vertices of the graph are the locations that have at least one neighbour within the specified maximum distance. The next part of the methodology uses the Depth First Search algorithm to decompose the sparse graph to create more homogeneous (in terms of point density and distance between the retail units) subgraphs, under the condition that each subgraph has at least k vertices and that each location has at least one neighbour within the maximum distance. The vertices that are not part of any subgraph are removed as outliers.

Given that the spatial extent of each subgraph depends on the connectivity and number of points within an area, each subgraph can represent a town centre, a city centre or even a metropolitan region. DBSCAN, however, assumes that the epsilon value is a representative indicator of the local density. To fulfil that assumption, in the third step of the methodology, DBSCAN is iteratively applied for each subgraph and the cluster that has density (as estimated by the local epsilon, i.e. the 95^{th} percentile of the 4-nearest neighbours' distances within each cluster) closer to the overall epsilon value is selected and extracted from the subgraph. Following the extraction of a single cluster, a new sparse graph representation of the remaining locations is created (by recalculating the k-nearest neighbour matrix), which is then further decomposed using the Depth First Search algorithm, and for each (more homogeneous) subgraph the DBSCAN method is iteratively applied until no cluster can be formed. This process is summarised in Figure 1.

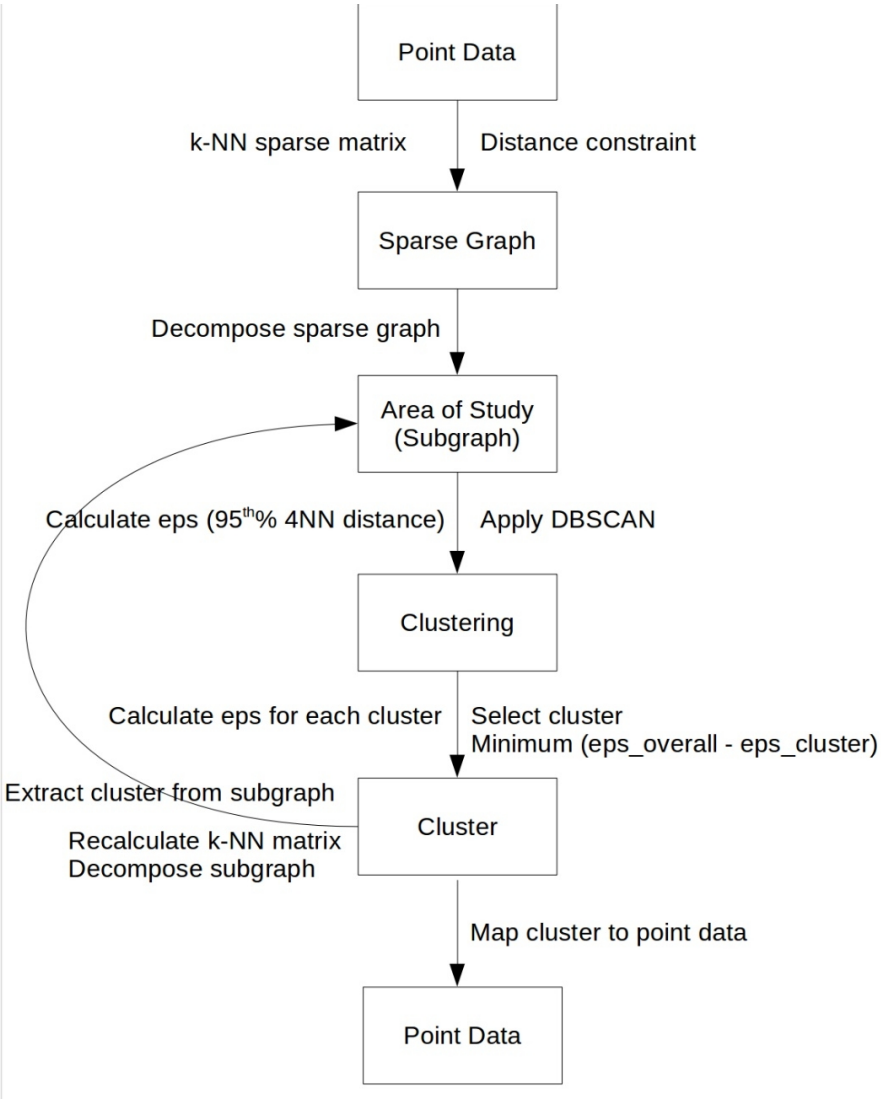

Figure 1. The process of the suggested modified DBSCAN methodology

5. Comparative verification of the results

The results derived with this new method were compared to data on retail centre extents supplied by the company Geolytix. The comparison was based on two metrics, the n-ary relation between the two datasets and the proportion of points within the Geolytix polygons.

The n-ary relation returns a score where the higher the number of clusters that had one-to-one relation with the clusters identified by Geolytix the better the relation. There were 274 spatial intersections between the two datasets, out of which 250 were one-to-one.

Summary values of the spatial distribution of the clustered locations within the Geolytix boundaries are shown in Table 1. On average (based on the median value) almost 90% of the clustered points were within the Geolytix boundaries.

Table 1. Summary values describing the spatial distribution of the clustered locations within the Geolytix boundaries.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
0.015	75.53	89.31	77.06	94.94	100.00

6. Conclusion

The objective of this analysis was to develop a clustering method that would facilitate the identification of retail agglomerations across a national extent and that could be updated over time. For this purpose, five of the most frequently used clustering methods were compared within 8 representative locations across Great Britain. The DBSCAN method was selected on the basis that it provided the most accurate representation of those retail areas relative to formal definitions; it was faster to produce a clustering solution and also easier to calibrate input parameter values.

However, in order to address a well-known issue that DBSCAN does not cope well in areas of varying densities, the DBSCAN method was adapted so that it could be iteratively applied within smaller more homogeneous sites that were created using a k-NN sparse graph representation of the retail locations. Each selected retail cluster was created by the DBSCAN algorithm with an epsilon value that was representative of the local point density.

The clusters produced were comparable to those retail areas designated by the local authorities for the sample areas of study, and in some cases, were more accurate when compared to the traditional DBSCAN method. In addition, the identified clusters were in most areas similar in terms of spatial extent to those produced by the Geolytix company using alternative dataset and methodology. Furthermore, the output of this analysis provides a better spatial coverage and option for automated update in comparison to the existing DCLG town centre boundaries. Given that the DCLG boundaries were widely used by academics, local authorities and private organizations across the country it can be anticipated that these results will prove to be valuable for research and analysis.

7. Acknowledgements

The authors would like to acknowledge funding provided by the ESRC.

8. Biography

Michalis Pavlis is a data analyst at the University of Liverpool. His work interestes lie in the area of spatial data analysis, retail and geodemographic analysis.

Les is currently a Lecturer in Geographic Data Science within the Department of Geography and Planning at the University of Liverpool. His main research interests lie in the area of retail geography with a focus on town centres performance, their spatial complexity and implementation of consumer-related (big) data.

Alex Singleton is a Professor of Geographic Information Science at the University of Liverpool and has research interests concerning an informed critique of the ways in which geodemographic methods can be refined for effective yet ethical use in public resource allocation applications.

9. References

Dearden J. and A. Wilson (2011). A Framework for Exploring Urban Retail Discontinuities. *Geographical Analysis*, 43(2), 172-187.

Everitt B S, S Landau, M Leese and D Stahl (2011). Cluster Analysis. 5th ed. Chichester, UK Wiley.

Mackaness W A. and Chaudhry O Z (2011). Automatic Classification of Retail Spaces from a Large Scale Topographc Database. *Transactions in GIS*, 15(3), 291-307

Newing A, G P Clarke and M Clarke (2015). Developing and Applying a Disaggregated Retail Location Model with Extended Retail Demand Estimations. *Geographical Analysis*, 47, 219-239.

Thurstain-Goodwin M and Unwin D (2000). Defining and delineating the central areas of towns for statistical monitoring using continuous surface representations. *Transaction in GIS*, 4 (4), 305-317.