An Open Source Geodemographic Classification of Small Areas In the Republic of Ireland

Chris Brunsdon, Martin Charlton, Jan Rigby

National Centre for Geocomputation
National University of Ireland, Maynooth
Christopher.Brunsdon@nuim.ie

Background and Motivation

A geodemographic classification is essentially a grouping of geographical neighbourhoods, or other small areas, in terms of their social and economic characteristics. The classification is generally achieved by applying a *clustering algorithm* such as *k-means* [1] to a data set of social and demographic variables (such as the unemployment rate) computed for each of the areas. A key reason to do this is that there may be links identified with these geodemographic classifications of areas and other processes. For example, Brunsdon *et al* [2] use geodemographic approaches to predict participation in higher education in the UK. Another influential motivation is that there are many commercial and marketing applications of geodemographics - for example identifying which particular neighbourhood groups are most likely to yield customers for certain products, so that marketing campaigns can then target these areas. These kinds of application have led to several commercially available geodemographic classifications - one such example being A Classification of Residential Neighbourhoods (ACORN) [3] - a system produced and sold by CACI. In addition, the use of geodemographics has gained attention in the public sector – where it is sometimes referred to as 'social marketing' – for example to target areas for initiatives to encourage people to stop smoking [4].

More recently attention has been focussed on freely available geodemographic classifications, in particular the UK's Output Area Classification (OAC) [5] system produced by Vickers, Rees and Birkin provides a geodemographic classification based on the 2001 UK Census. The focus here arguably moves away from market research and towards social applications - and a notable characteristic of OAC is that information relating to the data and clustering method used is freely available. This offers a number of advantages - it ensures that others are able to scrutinise the code, or adapt the approach so that a different data set, different spatial units, or an alternative classification algorithm could be used. In addition, many studies involve analysing the linkage between exogenous dependent variables and the geodemographic groups - however it is necessary to know which variables were used to determine the groups to ensure the dependent variable is not included - and a misleading association is discovered.

It is in this spirit of availability and openness that the classification system discussed here has been created. The authors intend to produce an open geodemographic classification of the 2011 Irish Census, based on the *Small Area* areal units [6].

Further Details

As the Irish Census differs from the UK ONS census in the questions asked, and the size and geography of the underlying population, our process of clustering and analysis differs from OAC - but the intention of producing an open and freely available area classification remains. Some of the features unique to our approach are:

- Use of the Partitioning Around Medoids (PAM) cluster analysis algorithm [7] instead of *k-means*.
- Use of heat maps as an approach to interpreting the clusters – these are also to be made publicly available
- Use of a *reproducible research* approach - so that in addition to providing a public description of the analytical techniques and variables, the actual code and data will be made available, allowing third parties to reproduce the exact results. A number of arguments for reproducibility in academic work are made, for example, by Peng [8] and Laine *et. al.* [9].

Preliminary Results

The PAM approach was applied to principal components of a number of variables derived from the 2011 Irish Census. A characteristic of this approach is that it is more robust to outlying cases than *k-means* and less inclined to produce a classifications with very small numbers of cases. For brevity, full listings of the variables and the formulae used to compute them are omitted from this abstract. However the variable names, and their association with the clusters identified by the PAM algorithm are listed in the heatmap in figure 1.

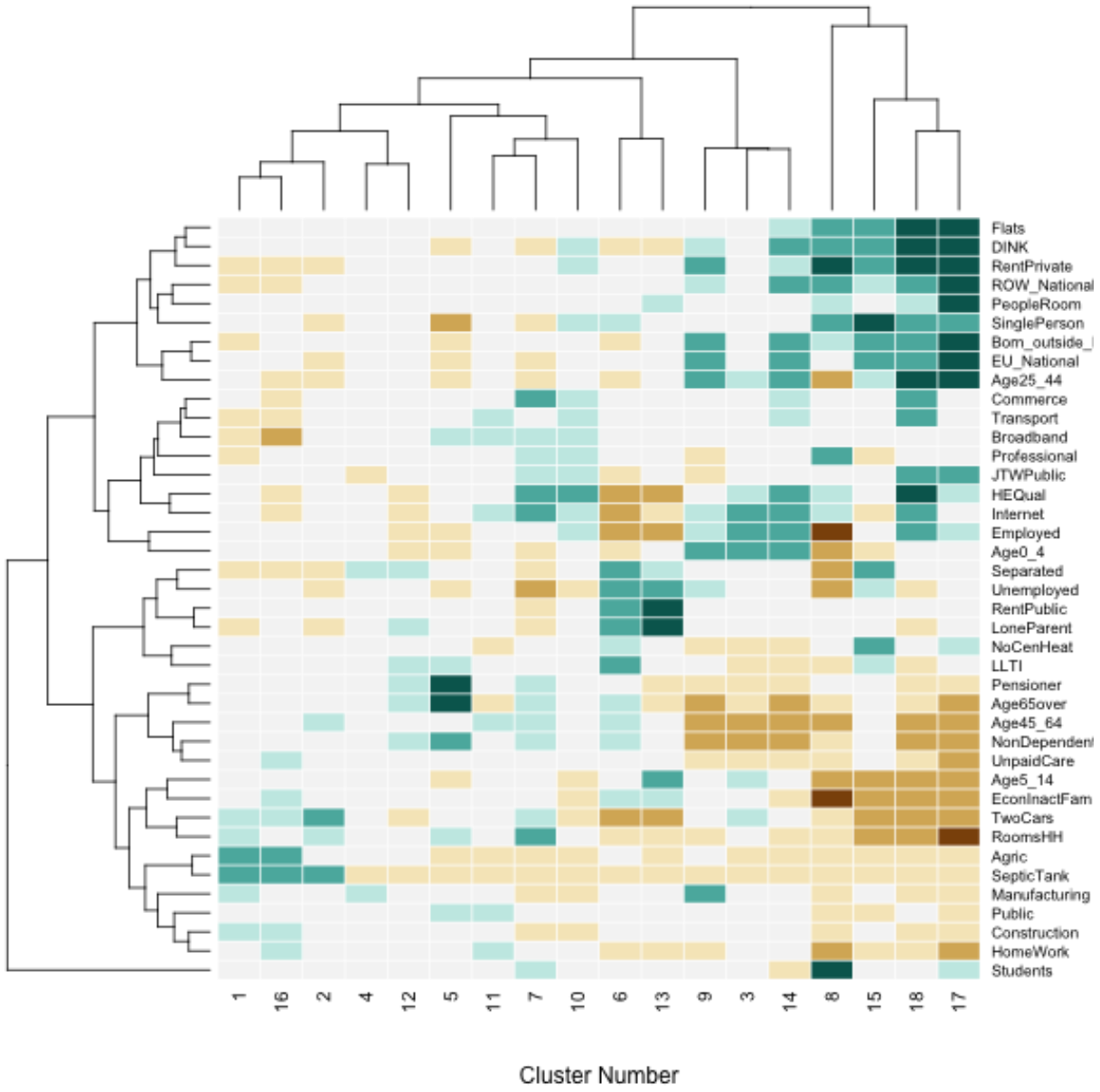

Figure 1 – Heatmap of PAM Cluster Characteristics

Here, the blue shaded elements correspond to higher average values of a variable within a cluster, compared to the Irish national average. The brown values correspond to low values. The clusters were then subjected to a hierarchical cluster analysis (Ward's method) to attempt to identify similar clusters. The resultant dendrogram is shown on the x-axis of the heatmap. Similarly variables that are associated by being linked with similar profiles of clusters are also subject to hierarchical clustering, with a dendrogram as seen against the y-axis.

At this stage, a full naming of the clusters has not been carried out. However using the dendrogram, a broad, higher level nomenclature can be suggested, as shown superimposed on the heatmap in figure 2.

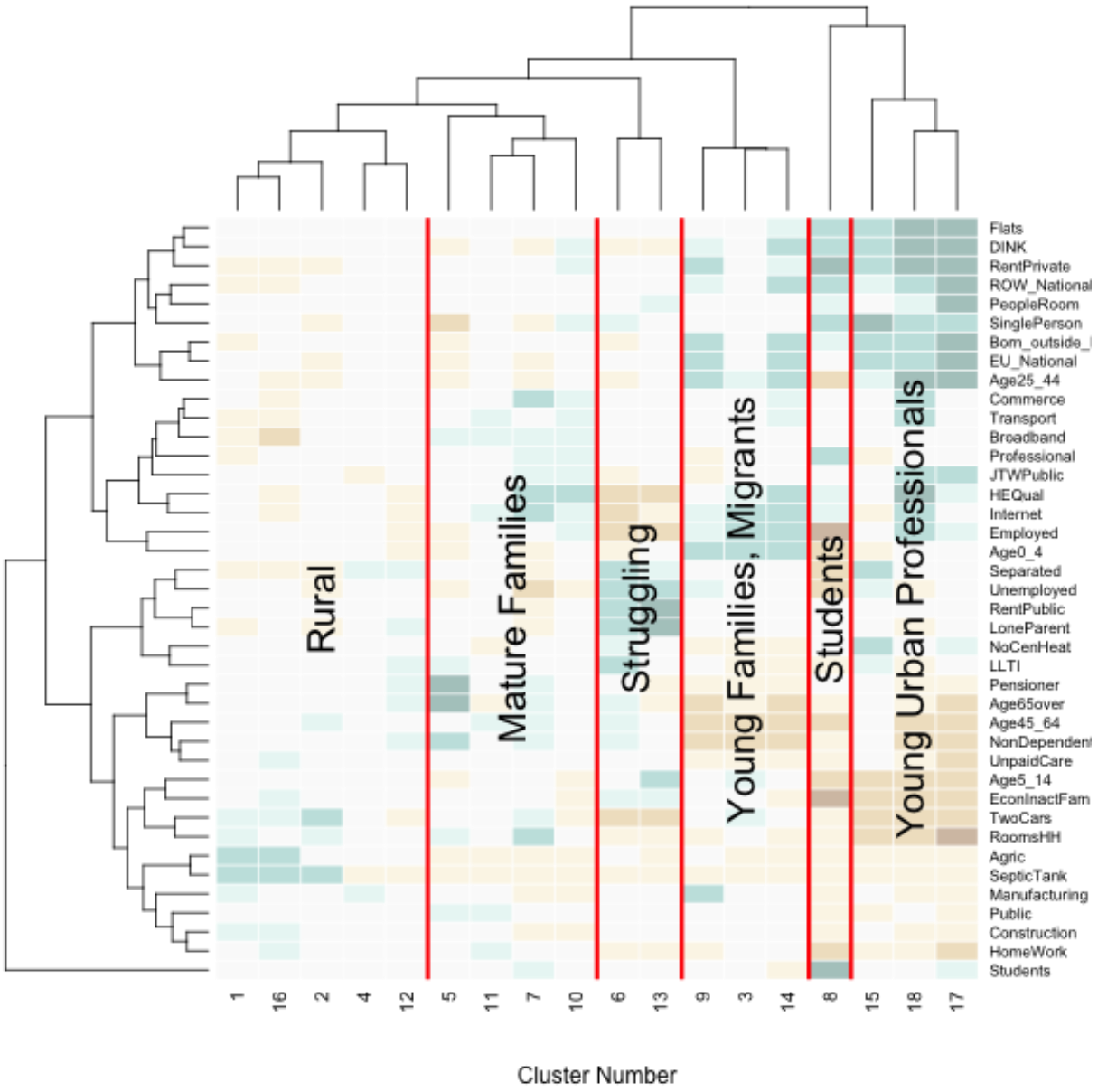

Figure 2 – Broad-scale Cluster Naming

Although all of the clusters may also be mapped, here just one (corresponding to the 'Students' category above, in the Dublin area) will be shown.

Figure 3. Map Showing 'Student' Small Areas in the Dublin Region

As a 'quick and dirty' verification, the highlighted areas correspond to the locations of universities and halls of residence in Dublin – see Figure 3. However, as yet the naming of the remaining individual clusters requires further work. One possibility, given the open nature of this classification, may be to provide access to the heatmaps and geographical maps relating to the clusters on the internet, and use some kind of crowd-sourced approach to cluster naming.

Details of Computation: Applying Reproducible Research

In order to ensure reproducibility, as defined earlier in this abstract, a web-based document outlining the analysis is provided at the web site: http://rpubs.com/chrisbrunsdon/11732 . This document contains all of the R code

executed to obtain the classification, and information about data sources. The document was produced using RMarkdown [10] – a tool designed to facilitate reproducible research, by storing documents with embedded R code, so that reporting of results and the code used to obtain the results are integrated in the same document.

References

[1] HARTIGAN, J. A. AND WONG, M. A., 1979. *A K-means clustering algorithm*. Applied Statistics 28, 100–108.

[2] BRUNSDON, C,. SINGLETON, A. D., LONGLEY, P. A ., ASHBY, D., 2011. *Predicting Participation in Higher Education: a Comparative Evaluation of the Performance of Geodemographic Classifications.* Journal of the Royal Statistical Society: Series A (Statistics in Society) 174: 17-30

[3] ACORN - The Smarter Consumer Classification: http://acorn.caci.co.uk

[4] TOMINTZ, M.N., CLARKE, G.P., AND RIGBY, J.E., 2009. *Planning the Location of Stop Smoking Services at the Local Level: A Geographic Analysis.* The Journal of Smoking Cessation 4:61-73.

[5] VICKERS D., REES, P. AND BIRKIN, M., 2005, *Creating the National Classification of Census Output Areas: Data, Methods and Results* Working Paper 05/2, School of Geography, University of Leeds

[6] CENSUS 2011 Boundary Files – CSO – Central Statistics Office http://www.cso.ie/en/census/census2011boundaryfiles/

[7] KAUFMAN, L. AND ROUSSEEUW, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis* John Wiley and Sons, New York.

[8] PENG, R., 2009. *Reproducible Research and Biostatistics*. Biostatistics 10, 405-408

[9] LAINE, C., GOODMAN, S. N., GRISWOLD, M. E. AND SOX, H. C., 2007. *Reproducible research: moving toward research the public can really trust.*. Annals of Internal Medicine 146, 450–453.

[10] RSTUDIO: http://www.rstudio.com/ide/docs/authoring/using_markdown

Biographies

Chris Brunsdon, Martin Charlton and Jan Rigby are economic migrants working in Maynooth, Co. Kildare, Republic of Ireland.