

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

Christopher Brunsdon¹ · Martin Charlton¹ ·
Janette E. Rigby¹

Received: 16 January 2016 / Accepted: 17 October 2016 /
Published online: 29 October 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Geodemographic classifications have progressed from manual classifications of areas through to complex, highly marketable products used in both the public and private sectors. As their production became commercialized, input variables moved beyond census variables to include other, often not publicly available datasets, and hence the resultant black-box approach increased in sophistication, but was less open to scrutiny. In the UK this was somewhat reversed with the production of the Output Area Classification (OAC) from the 2001 census. As an alternative approach, in this paper we demonstrate the production of a geodemographic classification for the Republic of Ireland, using a different approach to OAC, and extending the ethos of transparency and reproducibility.

Keywords Geodemographics · Ireland · Census · Algorithms · Reproducibility · Open source

Introduction

A geodemographic classification is essentially a grouping and labelling of geographical neighbourhoods, or other small areas, in terms of their social and economic characteristics. Attempts to classify areas from the characteristics of the people living there go back some way before computational approaches emerged, for example, Booth's 1903 maps of London and Rowntree's exploration of poverty in York in 1900. Some of these classifications have proved remarkably resilient: Orford et al. (2002) demonstrated that a poverty index constructed from Booth's survey and maps was a better

✉ Janette E. Rigby
jan.rigby@nuim.ie

¹ National Centre for Geocomputation, Maynooth University, Maynooth, Co Kildare, Ireland

predictor of mortality rates for the over 65s for the period 1991–1995 than some contemporary poverty measures.

Singleton and Longley (2009), Longley (2012) and more recently Singleton and Spielman (2014) have critiqued the development of the production of geodemographic systems. Longley outlines a rationale that the increasing differentiation in lifestyles has been used to justify the addition of a wide range of datasets to the census data variables routinely used to derive the classifications. He notes, however, that the additional sources can lack scientific validity, which therefore preclude the scientific scrutiny of the resultant classification. Regardless of academic critique, the geodemographics industry has proved remarkably resilient, including its recent UK diversification encouraging consumers to check their credit ratings.

As geodemographics of this type have moved towards classifying consumption and consumers, there have been parallel developments in the production of deprivation indices, which focus more on social need. The basic classifications produced in the 1980s and 1990s (e.g. Townsend, Carstairs) have also been subsumed into a deprivation industry using increasing numbers of indicators e.g. Index of Multiple Deprivation (England and Wales), Pobol HP Deprivation Index (Ireland). Again these can be used uncritically, and it is possible to see examples of these indicators used as predictors for outcomes which can be found in the morass of indicators used to construct the indices: something of a circular argument. We leave readers to identify these themselves. Applications through commercial products, and more worryingly academic research, have promoted a view that a single classification is suitable for an extraordinarily wide range of applications. Most existing classifications operate at a national scale. This can be problematic when the influence of a capital city does not follow the socio-economic ‘norms’ which are determined and applied, as illustrated for Greater London by Singleton and Longley (2015).

A classification is generally achieved by applying a clustering algorithm such as *k*-means (Hartigan and Wong 1979) to a dataset of social and demographic variables computed for each of the areas. A key reason to do this is that there may be links identified within these geodemographic classifications of areas and other processes. For example, Brunsdon et al. (2011) 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, <http://acorn.caci.co.uk>) - a system produced and sold by CACI.¹ In addition, the use of geodemographics has gained attention in the public sector (where it is gaining credibility as ‘social marketing’) for example to target areas for initiatives to encourage people to stop smoking (Tomintz et al. 2009). Indeed, the proliferation continues, so there are now products for specific sectors, e.g. ACORN Health and MOSAIC Health available, and in use, within the UK. Whilst geodemographics have been subject to some critical evaluations (e.g. Feng and Flowerdew 1999; Openshaw and Wymer 1994) and methodological enhancement in the academic literature,

¹ <http://acorn.caci.co.uk>

application elsewhere has been worryingly uncritical as the commercialization potential has been quickly identified and diffused.

There are commercial geodemographic segmentation products available in Ireland. These include Data Ireland's OGHAM product which classifies households into 34 'Lifestyle and Affluence' groups²; Experian Ireland's MOSAIC classification³; and Gamma's Inca segmentation system.⁴ There is no open source system currently available in the Republic of Ireland.

More recently attention has been focused on freely available geodemographic classifications, in particular the UK's Output Area Classification (OAC) system produced by Vickers et al. (2005) which 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, and laudable, characteristic of OAC is that information relating to the data and clustering method used is freely available (see also Singleton and Longley 2015; Spielman and Singleton 2015). This offers a number of advantages - it ensures that others are able to scrutinise the code, or adapt the approach for a different data set, different spatial units, or employ an alternative classification algorithm. 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 none of the dependent variables are included, and hence avoid the discovery of a misleading association.

It is in this spirit of availability and openness that the classification system discussed here has been created. The authors have produced an open geodemographic classification of the 2011 Irish Census, based on the Small Area areal units (CSO 2011) the genesis and production of which is the subject of this paper.

Methods

In Ireland, a population census is conducted every 5 years. The administrative unit designed for this was the Electoral Division (ED). However, by 2006 the populations of the EDs varied from 76 to 32,288, which had become problematic for data collection, reporting and subsequent analysis. For 2011 a new set of 'small area' units was commissioned, which resulted in the production of 18,488 small areas, with a median population of 240.

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:

- (i) Use of the Partitioning Around Medoids (PAM) cluster analysis algorithm (Kaufman and Rousseeuw 1987) instead of k-means. The algorithm is outlined in Appendix 1. This approach is proposed for a number of reasons. As its creators observe, the method is based on minimising an L_1 metric - that is, it chooses

² <http://www.dataireland.ie/Services/Pages/Consumer-Lists-OGHAM-Geo-demographic-Segmentation-Model>

³ <http://www.experian.ie/assets/marketing-services/brochures/MosaicRol%20Flipchart.pdf>

⁴ <http://www.gamma.ie/inca-irelands-neighbourhood-segmentation-index>

clusters based on minimising absolute distances from cluster centres (medoids), rather than squared distances, as k -means does. This makes it more robust to outliers. Less outlier-resistant approaches tend to assign unusual observations to single-item clusters, which is not desirable here. In addition, the approach defines each cluster in terms of a representative case - an observation from the data (in this case a Small Area) that typifies the cluster. This is very helpful when attempting to describe and interpret the characteristics of each cluster.

- (ii) Use of heat maps as an approach to interpreting the clusters – these are also to be made publicly available.
- (iii) 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. This also facilitates adaptation of the methodology e.g. using a different clustering method, different areal units, or updating with new data. A number of arguments for reproducibility in academic work are made, for example, by Peng (2009) and Laine et al. (2007). An overview of the approach is illustrated as a workflow in Fig. 1; specific details will now be considered.

Choice of Variables

The aim of the exercise was to create a general-purpose classification of the Small Areas in Ireland. This raises the question of an appropriate sets of variables to use as the basis for the classification. There appears to be little theoretical guidance, but one is cautioned against “the mindless approach in which numbers of variables ... easily culled from census volumes ... are picked over like cans on a rubbish tip” (Mather and Openshaw 1974, p.290). Early exercises in classification have useful suggestions for variables. Perusing the lists in Cullingford et al. (1975), Webber (1975), Webber (1977), Webber and Craig (1978) allow us to identify common themes. Appendix 1 in Webber (1977) contains a list of 40 variables, which form a basis for later work by Charlton et al. (1985). There are some choices in the earlier studies which point to social problems which are no longer pressing for example Cullingford et al. (1975) include indicators for shared bathrooms and outside toilets.

Vickers et al. (2005) describe the genesis of the UK’s 2001 Output Area Classification (OAC), and for the 2011 OAC exercise the Office for National Statistics (2015a) demonstrates that a similar methodology and variable choice was

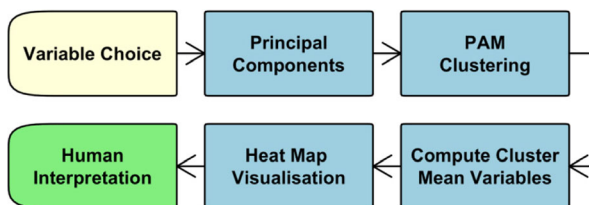


Fig. 1 Workflow of the approach to geodemographic classification. *Blue, sharp-cornered boxes* indicate computational (and therefore reproducible) activities

used for both classifications. To enable some element of international comparison we have attempted to identify a parallel set of variables from the Irish Census. It should be noted that harmonisation of the definitions is not always straightforward. Not only are there differences in the definitions of the indicators, there can be differences in the derivation of the data: in the UK a *de jure* count is the basis, and in Ireland, the basis is the *de facto* population. The essential difference is that the UK counts the usually resident population on census night (so visitors are returned to their place of usual residence) but in Ireland the count is of the population present on census night.

There is an additional challenge in that a comparable set of areal units is required. It has been known since the early 1930s that there are scale effects on the correlation structure of variables for modifiable areal units (Gehlke and Biehl 1934). Ordnance Survey Ireland commissioned a set of 'Small Areas' which have a common and consistent definition, and are comparable in population size and spatial scale with the Output Areas used in the United Kingdom. The Small Areas are the finest grained spatial units for which 2011 Census of Population data are available in Ireland. It is well established that graded grains make the finest units (Homepride, 1990, personal communication).

Choices in Clustering

There are a number of additional choices concerning the clustering methodology. An approach which has been used widely, and also in many of the commercial classification systems is to provide a hierarchical classification. This may have only two layers (for example that of Charlton et al. 1985), although the UK 2001 and 2011 OACs have a three layer classification. This raises the question of methodology. One approach is to cluster the individual spatial units into a moderate number of groups. There is little theory to guide the analyst on the choice of the number of clusters, although a scree plot is often used to guide the selection of the number of groups. With datasets of several 10s or 100s of thousands of cases, the number that this process yields may be inconveniently large (60 or 70 will not conveniently list on a single page, although it might encapsulate some dimensions of the social structure of the study area). A small number of groups can be arrived at by using a hierarchical clustering procedure (Ward's method (1963) is a frequent choice for this) using the cluster centroids (the mean vector for each cluster) as data. Vickers et al. (2005) describe a process where the hierarchy is created from the top down. The initial classification of the Output Areas was into 8 'Supergroups'. The members of each Supergroup are then classified into 2, 3, or 4 Groups, and finally the members of each Group are classified into Subgroups. Thus there are 76 Subgroups and 26 Groups in the 2011 OAC (ONS 2015a).

We have followed the general workflow described in Charlton et al. (1985). The data are subjected to an orthogonalising transform, and the component scores from this are used as the basis for a non-hierarchical classification into 18 clusters. The cluster centroids are then grouped into 8 classes using complete linkage. There are two differences: we do not scale the component scores by their eigenvalues and we use a non-parametric clustering algorithm. The first component has an eigenvalue of 9.78 - there are about 10 variables which are measurement of a single dimension, so to scale the scores by 9.78 would give undue prominence to this component in the clustering and undo the effect of the component transformation.

Interpreting the Clusters

One area of effort which should not be overlooked is that of attempting to characterise the members of the clusters. This tends to be an activity dominated by the subjective tendencies of the persons responsible. The OAC Pen Portraits are typical as outputs (ONS 2015b): each cluster is given a short verbal description of its most notable characteristics and then provided with a two or three word short title. As an example, 'Renting rural retirement' Output Areas are members of the 'Ageing rural dweller' group, which itself is a member of the 'Rural residents' Supergroup. It is difficult not to confuse the characteristics of the areas with the characteristics of the residents of those areas through ecological fallacy, and the short names can have a tendency towards stereotyping e.g. 'Ageing juveniles'.

How do we identify the characteristics of a group? One technique is to compare the values of the mean vector for a cluster with the population values for the same variables. If the value of variable p for a cluster lies solidly in the upper or lower tail of the global distribution of values for that variable, then it may be taken as 'characteristic' in some sense. We can compute the mean vectors, and tabulate those variables in each cluster whose mean values are in the tails of their global equivalents. However, this asks of the analyst that he or she makes comparison not only between clusters but also within clusters. ONS (2015c) provides radial plots showing the relationship between the 60 variables of the classification for every individual Supergroup, Group and Subgroup with the global values. One hundred and ten pages of plots become daunting after a time, and leads us to consider whether the richness of the data reduction can be encapsulated in a single graphic.

The R heatmap is one solution which is appropriate for our case. The rows represent the variables, and the columns the clusters. The stronger the shade of green that a cell is coloured, the more positively characteristic is the variable of that cluster; the strong the shade of brown, then vice versa. The heatmap also clusters the variables which are related among the clusters, and also clusters the cluster centroids themselves, providing a second tier in a hierarchy. Inspection of Fig. 2 suggests a transition from more 'rural' clusters on the left part of the diagram to 'urban' on the right.

Additionally the cluster locations can be mapped - whether the plots are static or not depends on the application. In the exercise described in this paper, the dataset consists of 18,488 spatial units. However, there is additional information we can add to aid interpretation of the plots. Recall that cluster membership for k -means requires the spatial unit to be closest to the centroid of the cluster to which is assigned - there is a distance between the mean vector and the vector of values for the spatial unit in question. It would be unfortunate to examine a sample of spatial units which are not representative of the cluster under scrutiny. The values of the membership distance should also be mapped, either as a choropleth map, or in some other mode (for example Wood et al.'s (2012) 'sketchy' approach).

Details of the Analysis

The PAM approach (Kaufman and Rousseeuw 1987, 1990) was applied to principal components of a number of variables derived from the 2011 Irish Census. This approach detects clusters by identifying a set of medoids (typical cases for each cluster)

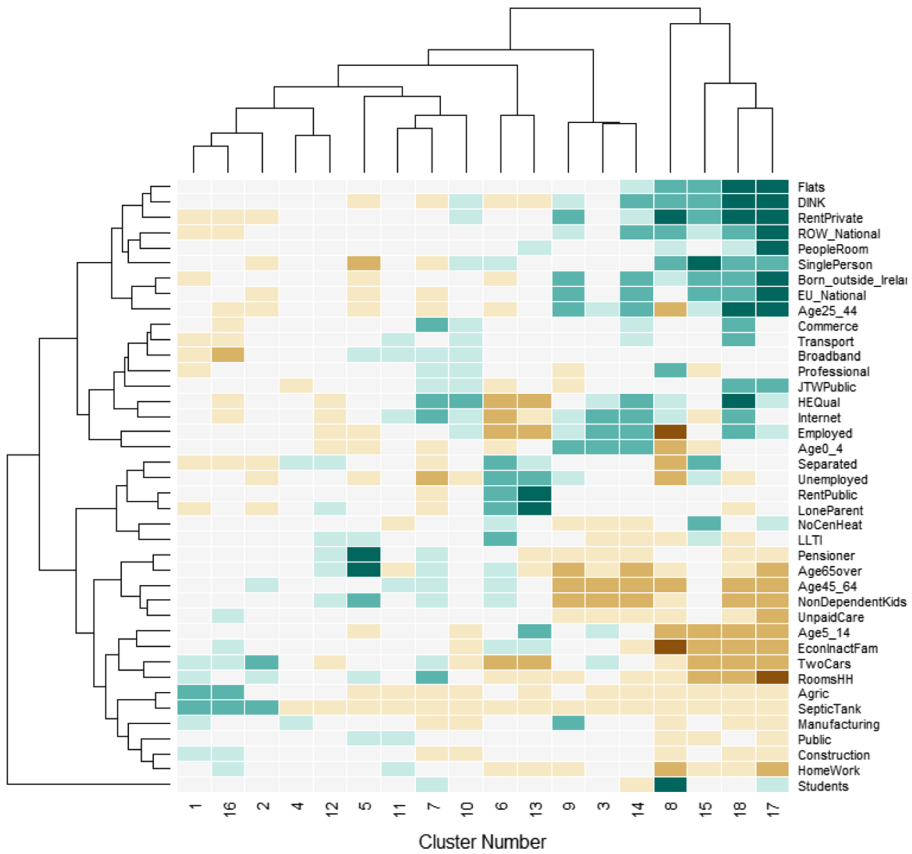


Fig. 2 Heatmap of PAM cluster characteristics

and assigning the other observations to clusters on the basis of the closest medoid. Here, ‘closest’ can be defined flexibly. All that is required is a distance matrix for the n observations being clustered. There are no constraints on how this distance may be defined - here it is defined by treating the first j principal components as coordinates in Euclidean space and computing distances on that basis.⁵ A characteristic of this approach is that it attempts to minimise sums of absolute distances rather than squared distances (as is the case with k -means) and as a consequence it is more robust to outlying cases than and less inclined to produce classifications with very small numbers of cases – this sometimes being a consequence of the effect of outliers in k -means. Full listings of the variables and the code used to compute them (in the case of derived variables) can be found by visiting the ‘Rpubs’ web site describing this procedure (<http://rpubs.com/chrisbrunsdon/14988>).

A justification of the use of principal components is that both PAM and k -means clustering make use of the idea of distance between different locations, in terms of the variables associated with them. The distances are defined in an m -dimensional space, where m is the number of variables measured. If the raw variables are used (even if

⁵ However, given the adoption a paradigm of open and reproducible code, it should be possible for others sufficiently motivated to experiment with other definitions of distance.

standardized to have zero mean and a variance of one), pairs of correlated variables tend to have similar values – with an effect of increasing the weighting on the underlying cause driving both variables. The issue with this is that unintentional over-representation of certain correlated groups of variables (for example by having a large number of age-category variables) will have the effect of creating a spuriously high emphasis of this group on the distance metric. However, by a principal components transformation (turning the m variables to the corresponding principal components) overcomes this problem, as the components are uncorrelated. The components effectively represent a set of independent underlying factors ‘driving’ the data – but each factor is allocated precisely one dimension, so the problem of unintentional over-representation is addressed.

An additional issue for clustering is the computation of the distances. If we are using Euclidean distances, then the angles between the axes of the multidimensional space must be $\pi/2$. If a pair of variables is correlated with correlation ρ , then the angle between the axes is given by $\cos^{-1}(\rho)$. With higher correlations, ρ approaches 1, and the influence of one of the variables should disappear from the calculation. If we assume that ρ is zero in such cases, we give one of the variables in question an undue influence in the clustering process. A principal components transform yields variables with orthogonal axes.

Results

Overview of Clusters

Although the PCA approach is helpful in the reliable formation of clusters, to interpret the clusters once assigned, it is then helpful to return to the original variables. For each cluster, the cluster mean of each variable is computed, and the relative values of these are shown in the heatmap of Fig. 2 above.

Here, the blue-green shaded elements correspond to higher average values of a variable within a cluster, compared to the Irish national average. In contrast, the brown values correspond to low values. The clusters were then subjected to a hierarchical cluster analysis – that due to Ward (1963) to attempt to identify similar clusters. The resultant dendrogram is shown on the x-axis of the heatmap; this also drives the ordering of the categories on the axis (Fig. 3). Similarly variables that are associated by being linked with similar profiles of clusters are also subject to Ward’s hierarchical clustering, with a dendrogram as seen against the y-axis, and again their ordering is determined by the dendrogram. The dendrograms convey information not only about the structure of the clusters, but also the degree of difference in the splits between groups. Divides higher up in the tree are based on greater differences. Thus, for example, the split between the group 8 and the remainder is based on the greatest level of difference which may be seen since the highest branch in the tree represents this division.

Descriptions of the clusters appear in Appendix 2. The dendrogram has suggested a higher order grouping. Clusters 1, 16 and 2 form a very coherent ‘rural’ group. This is not only suggested by their positions on the heatmap, but also by the relatively low population density (see Fig. 6 below), The septic tank variable is a strong discriminator for this group. Agricultural employment is a strong feature

of groups 1 and 16, with 16 forming the more remote rural communities - the noticeable lack of broadband connectivity reinforces this interpretation. While a ‘rural’ cluster, members of 2 are closer to main settlements, and are characterised as reasonably well-off older residents.

Apart from the ‘Students’ group which stands out markedly in character from the other groups, the others may all be sub-divided. For example, cluster 16 is in the broad ‘rural’ group but is characterized as having particularly low broadband uptake. Similarly, in the ‘Struggling’ group, group 6 is characterized as having a higher level of Limiting Long-Term Illness (LLTI) than the other member of this broad group (group 13).

Geographical Pattern

Although all of the clusters may also be mapped, here just one (corresponding to the ‘Students’ category above, in the Dublin area) is shown as Fig. 4 below.

As a first pass verification, the highlighted areas in Fig. 4 correspond to the locations of universities and halls of residence in Dublin.

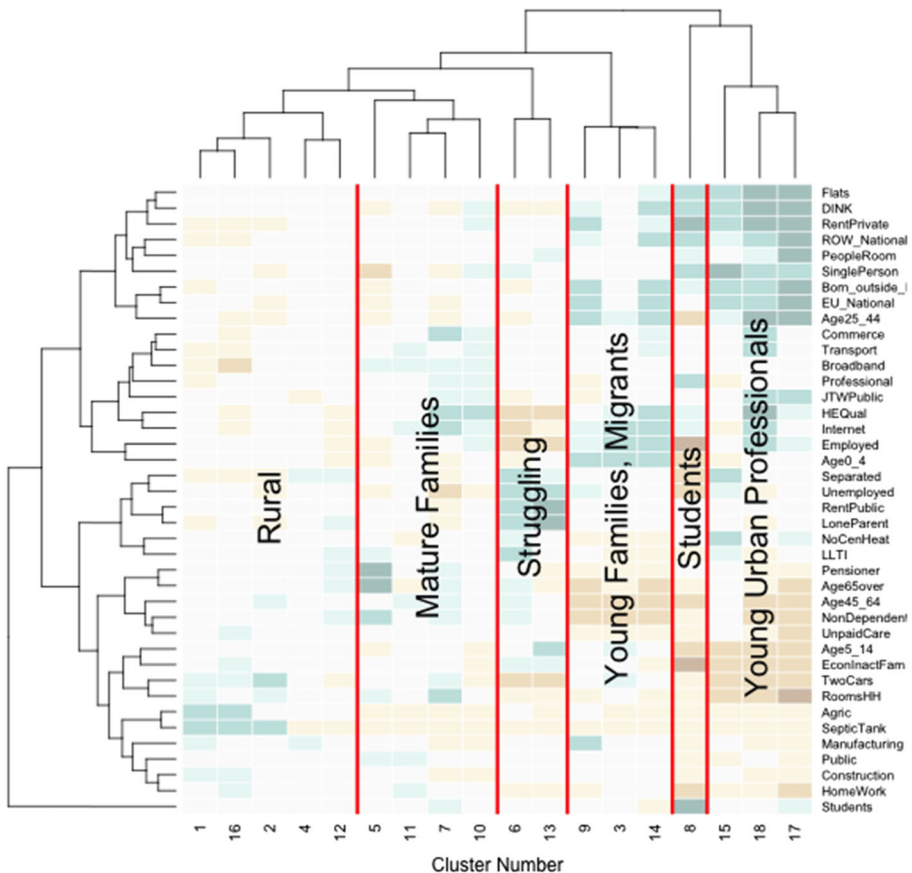


Fig. 3 Broad-scale cluster naming

We can examine the variation in population density between the clusters as a further clue to geographical pattern. A histogram of population density at Small Area level reveals a long right tail. If we log the density (in residents per hectare) then we obtain the histogram in Fig. 5:

The red line represents a density estimate. Notice the bi-modal nature of the distribution - the modes corresponding to 'more rural' on the left and 'more urban on the right'. The minimum between the modes corresponds to a population density of about 2.21 residents per hectare.

This allows us to create a boxplot of population density by cluster, with the boxes in dendrogram order, and width proportional to the square root of the number of objects in each cluster (Fig. 6). The design of the Small Areas has the constraint that they should

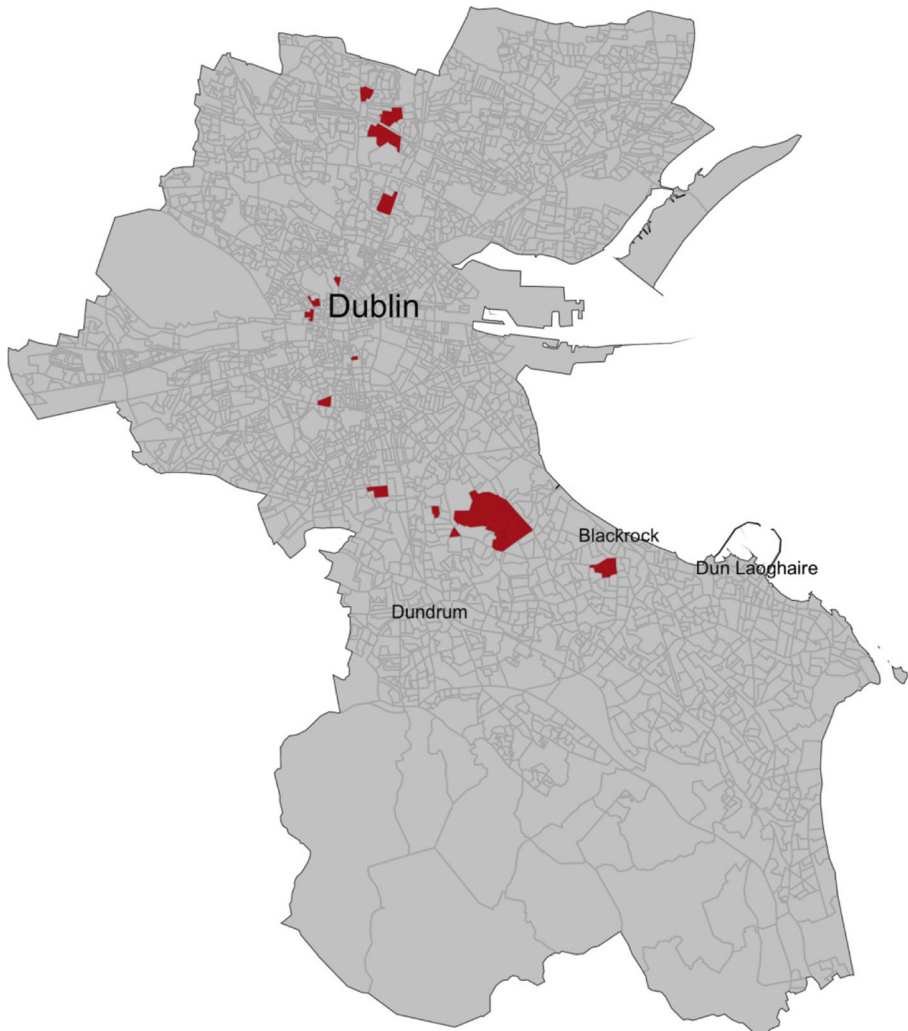


Fig. 4 'Student' Small Areas in the Dublin Region (Dublin and Dun Laoghaire/Rathdown)

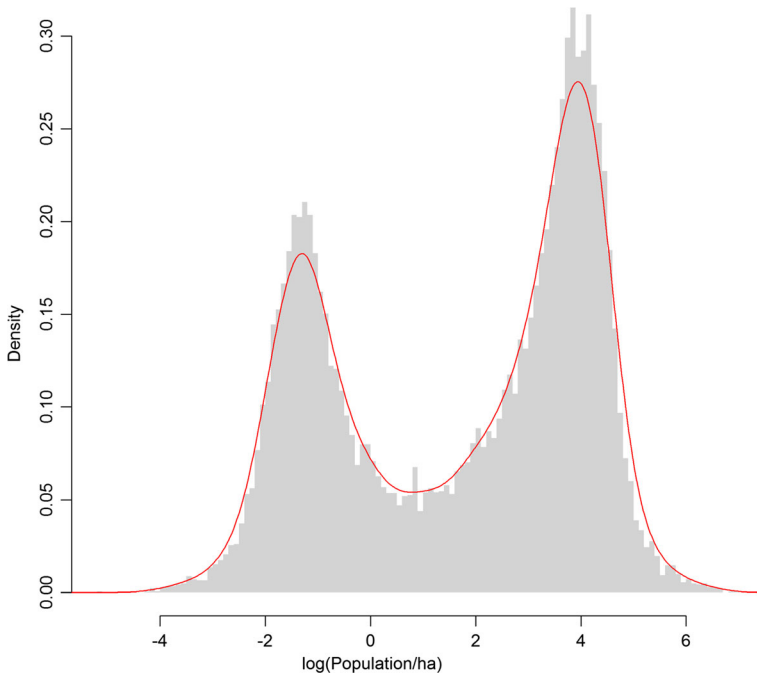


Fig. 5 Population density at Small Area level

have similar populations, so the box widths are also approximately proportional to the population in each cluster.

The pecked red line in the boxplot corresponds to the ‘more rural’/‘more urban’ threshold identified in the histogram above. Clusters 1, 16 and 2 are predominantly rural in location, with cluster 16 having the lowest general level of population density - remoter rural locations. The rest of the clusters are more ‘urban’ in character, although there are some notable left tails. Clusters 8, 15, 18, and 17 are almost exclusively urban. Clusters 5, 11, 6, and 13 have some noticeable outliers in rural areas. Cluster 2, which being predominantly ‘rural’ has a noticeable right tail of ‘urban’ locations.

An initial attempt at naming and characterising the clusters is provided in the Appendix. One possibility, given the open nature of this classification, may be to provide access to the heat maps and geographical maps relating to the clusters on the internet, and use some kind of crowd-sourced approach to cluster naming.

Clusters and Deprivation

It is sometime considered that a single deprivation index will suffice to encapsulate social variation in the population. We examine the Kelly-Teljeur 2011 score (Kelly and Teljeur 2013) for the clusters in the different groups. Positive numbers on the Kelly-Teljeur Index indicate higher levels of deprivation. A multiple boxplot is shown in Fig. 7:

The boxplots are ordered corresponding their cluster’s position on the dendrogram in the heatmap, and Groups are alternately shaded light grey and white to aid interpretation. The pecked red line represents the median score. Of the clusters in Group A, many

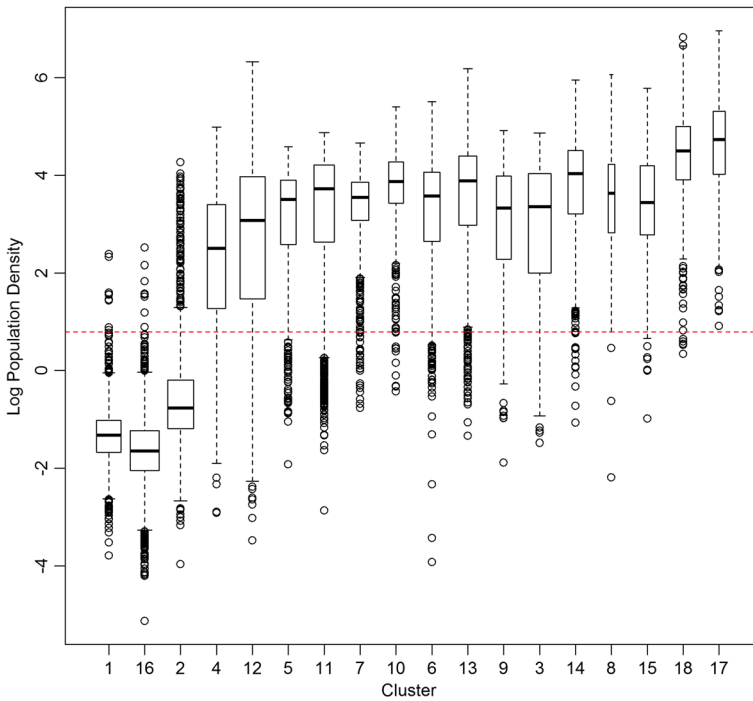


Fig. 6 Population density by cluster

of the Small Areas in each cluster are below the median, although cluster 16, the ‘remoter rural communities’ has a noticeable tail of deprived Small Areas. Group B, ‘Urban poor’ shows evidence of deprivation, particularly in cluster 12, and in the case of Group E, ‘Struggling Urban Peripheries’, there is strong evidence of deprivation. Of Group H, ‘City diversity’, cluster 15, ‘Stressed inner city singles’ shows some evidence of deprivation, as does Cluster 17. It would be a mistake to suggest that the Classification and Deprivation Index are substitutes for one-another. The index is the result of a spatially uniform Principal Components Analysis, so each value on the index has arisen via the same spatial process. By contrast, there is a different relationship between the variables within each cluster within each cluster; this is clear from the heatmap.

The Overall Pattern

There are 18,488 Small Areas. To show the complete picture on a map in the Irish National Grid or Irish Transverse Mercator projections would result in spatial units less than 1 pixel wide in the urban areas. One solution lies in a cartogram transform, and we have used Brunsdon’s *getcart* package, available on Github.⁶ This resulting cartogram gives more prominence to urban areas, but it is still difficult to see the detail (Fig. 8):

We can also examine the structure of the groups in Dublin - the same colours have been used. This map is based on the Irish National Grid projection (Fig. 9).

⁶ Available at <https://github.com/chrisbrunsdon/getcart>

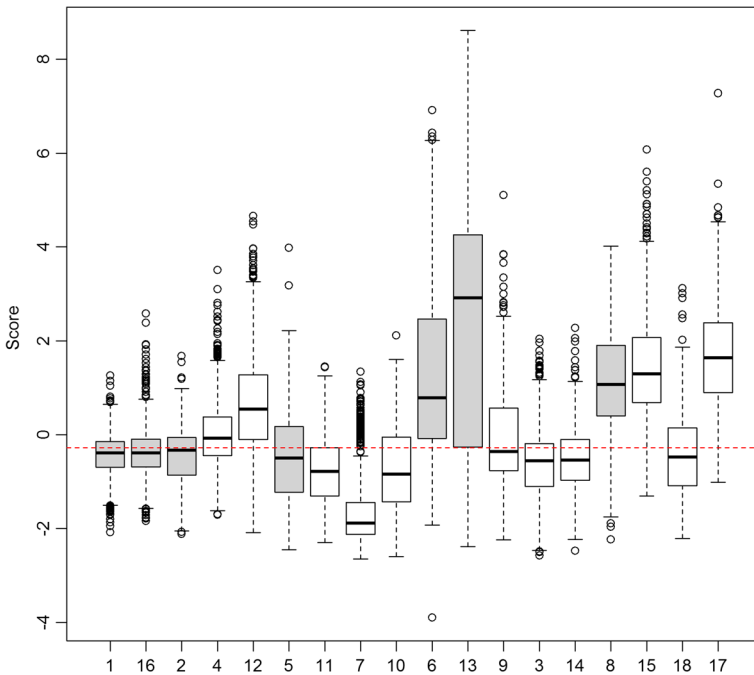


Fig. 7 Kelly-Teljeur deprivation scores by cluster

Discussion and Conclusion: Applying Reproducible Research

This paper has outlined a methodology for providing a classification of Irish Small Areas based on publicly available census data and cluster analysis techniques, similar in intention to the UK’s OAC classifications. A distinct feature is that not only is the data publicly available, but also the code used to carry out the analysis. Thus, as well as a classification that may be of use in its own right, this could also be a springboard to alternative classification schemes created by modifying this code. For example, a scheme for a different set of spatial units, or one adding some extra variables could be created relatively easily by modifying this code, and possibly supplying some extra or alternative data.

There are a number of other advantages to this approach, one of which is a kind of future-proofing. Should an alternative approach to cluster analysis be proposed at some future point that is more reliable, more robust or simply more appropriate for geographical data, then this could be easily ‘grafted’ into the existing analysis template, and results compared to the current classification. More generally the need for reproducibility is becoming apparent. Data analyses are becoming more complex and make use of computationally intensive techniques, but with this there is a danger that the analyses become opaque, and characteristics that could play a key role in interpreting the outputs can become hidden inside a ‘black box’. Reproducibility admits the possibility to view these analyses critically, which is often possible if only general details are outlined in a written summary as part of an academic paper or an official report.

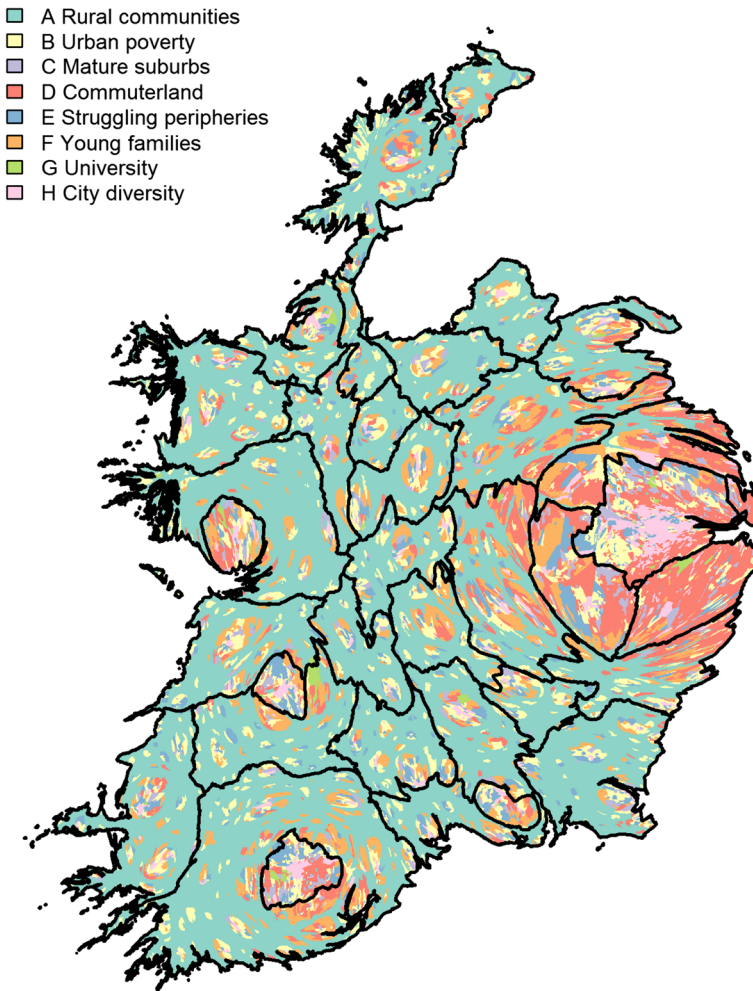


Fig. 8 Cartogram showing groups of each Small Area in Ireland

In arguing for increased reproducible activity in the social sciences, we need to examine work which permits reproducibility, though that may not have been the original aim. One example is Kelly and Teljeur's Deprivation Index (2013). Appendix II of their report contains descriptions and definitions of the indicators they use in sufficient detail to allow a researcher to return to their analysis and arrive, with a high degree of reliability, at their results. Whether this would be possible for, say, Cullingford et al.'s (1975) work is debatable: it took place 40 years ago; the software appears to be bespoke code; the data have been obtained from a variety of sources, and were available on a now obsolete medium: magnetic tape.

In order to ensure reproducibility, as defined earlier in this paper, 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 by RStudio, a tool designed to facilitate reproducible

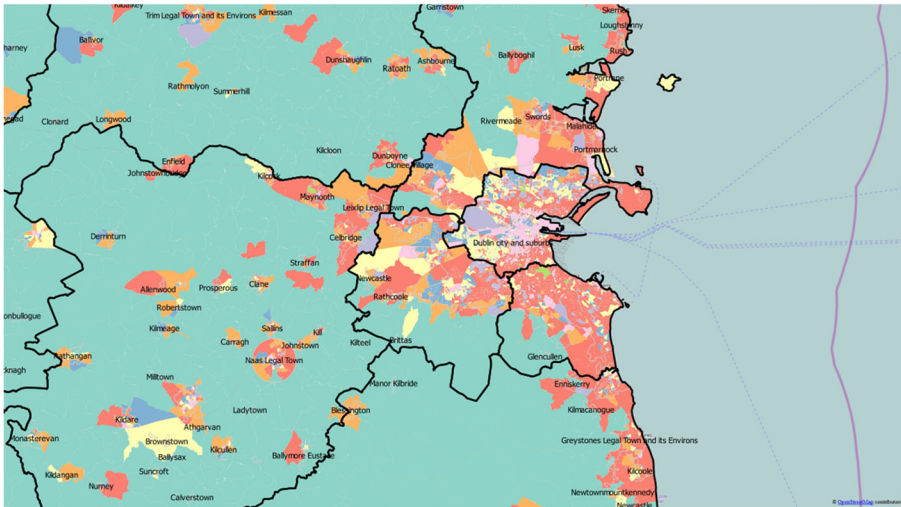


Fig. 9 The four counties of Dublin (Dublin City, Fingal, South Dublin, and Dún-Laoghaire-Rathdown) as well as parts of Louth, Meath, Kildare, and Wicklow

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.⁷

A recent innovation in Ireland has been the launch of postcodes. The Eircode is a service operated by a private company. Information on the structure, design and analytical capabilities offered by the codes is limited, although Irish residents were being informed of their Eircodes in summer 2015. The postcode is an address based code - this resolves to an individual residential address point and building level for commercial address points. The codes have seven characters organised as 3 characters-space-4 characters. The first, the 'routing key' refers to area of varying sizes which appear to be unsuitable for analytical purposes (that for Mallow, P51, is over 100 km from west to east, and is cut into two parts by the area for Fermoy, P61). The last four characters are organised randomly from a 25 character alphanumeric set. Whilst this gives in theory 390,625 combinations, a subset is used to remove the possibility of unsuitable words appearing. Adjacent addresses have entirely dissimilar Eircodes.

Hence the location of the address cannot be inferred from the code. The user is required to access a database, known as the Eircode Address Database (ECAD), which contains, inter alia, the small area code for each Eircode. There is a complex pricing structure for Eircodes which contrasts strongly with the open availability of both Census of Population data and digital boundaries for the Small Areas. However, in time, acceptance and wider use of the Eircode may open opportunities for market analysis. We observe that the once closed Postal Address File in the UK is now freely available, as are tables linking Postcodes and Output Area codes - this has allowed the development of a range of private and public analytical possibilities, and it is to be hoped that Ireland may benefit from similar developments in future.

Acknowledgments We gratefully acknowledge the helpful comments of two anonymous referees on an earlier version of this paper.

⁷ http://www.rstudio.com/ide/docs/authoring/using_markdown

Appendix 1: Kaufmann and Rousseeuw's Partitioning Around Medoids (PAM) Algorithm

The following is the algorithm for PAM proposed by Kaufman and Rousseeuw (1987).

1. Build Phase:
 - a. Compute a distance (dissimilarity) matrix for all of the cases to be clustered.
 - b. Given an initial set of k representative cases assign each other case to a cluster based on which representative case is closest.
 - c. Compute the sum of absolute distances (dissimilarities) between each case and the representative case and its cluster. This is the objective function to be minimised.
2. Swap Phase:
 - a. For each representative case and each remaining case, evaluate the objective function obtained if the pair are swapped (ie the remaining case becomes a cluster representative and vice versa).
 - b. Note the swap obtaining the smallest value of the objective function - if this is not lower than the previous value, return the previous set of representative cases. Otherwise return to **2.a** and repeat the swapping phase.

Appendix 2 Cluster Descriptions

The clusters are presented in the order they appear in the dendrogram. There is a cultural aspect to the naming of the clusters, so that while 'Struggling Blue Collar' might have some meaning within a UK context, the term 'Blue Collar' has little meaning in the Irish context and might be replaced with a more suitable term.

Group	Group Name	Cluster	Population
A	Rural communities	A1: Rural settlements	440,752
		A16: Remoter rural settlements	538,779
		A2: mature rural communities	521,590
B	Urban poor	B4: Struggling urban workers	221,719
		B12: Hard-pressed urban communities	412,338
C	Mature suburbs	C5: Retired and mature suburbs	165,646
D	Commuterland	D11: Comfortable commuter settlements	347,272
		D7: Affluent suburbs	242,613
		D10: Affluent commuterland	144,191
E	Struggling peripheries	E6: Older urban singles	169,269
		E13: Single parent urban deprivation	279,366
F	Young families	F9: Migrant service workers	159,915

Group	Group Name	Cluster	Population
		F3: Affluent young families with children	441,170
		F14: Affluent migrant communities	166,861
G	University	G8: University communities	28,673
H	City centre diversity	H15: Stressed inner city singles	94,016
		H18: Inner area affluent renters	122,531
		H17: City migrant renters	91,551

Group A: rural communities

Cluster 1: rural settlements

Very high proportions of residents with septic tanks, and working in agriculture. High proportions of construction and manufacturing workers, two cars households, and large dwellings. Low proportions of lone parents, separated persons, professional employees, broadband connections, transport workers, foreign migrants and private renters.

This cluster is not strongly spatially concentrated, although 17 % of the residents of this cluster are to be found in Cork County. More than 20 % of the residents of the following counties are in this cluster: Carlow, Cavan, Limerick, Monaghan, South Tipperary, and Waterford.

Cluster 16: remoter rural settlements

Very large proportions of residents with septic tanks, and working in agriculture. Large proportions of students, home workers, two car households, economically inactive families, older residents, and very low proportions of households with broadband connections. Low proportions of separated persons, internet connections, qualified workers, transport and commerce works, young adults, private renters and foreign migrants.

More than 20 % of the residents of the following counties are in this cluster: Cavan, Clare, Donegal, Galway, Kerry, Kilkenny, Laois, Leitrim, Longford, Mayo, Monaghan, North and South Tipperary, Offaly, Roscommon, Sligo, and Wexford.

Cluster 2: mature rural communities

Very high proportions of households with septic tanks and two cars. High proportions of large houses and older adults. Low unemployment, low proportions of lone parents, separated persons, young adults, private rents and immigrants.

This cluster is not strongly spatially concentrated, although 24.5 % of the residents of this cluster are to be found in Cork County and Galway County. More than 20 % of the residents of the following counties are in this cluster: Galway, Mayo, Roscommon and Sligo.

Group B: Urban poor

Cluster 4: Struggling urban workers

High proportions of workers in manufacturing and separated persons. Low proportions of households with septic tanks, and workers using public transport.

This cluster is not strongly spatially concentrated, although 21.1 % of the residents of this cluster are to be found in Cork County (10.1), Mayo (6.0) and Clare (5.1).

Cluster 12: Hard-pressed urban communities

High proportions of households with non-dependent children, older adults. Pensioners, families with long term health problems, lone parents, and separate persons. Low proportions of households with septic tanks, two cars, and internet connectivity; as well as low proportions of young children, employed and qualified residents.

About 24.6 % of the residents of this cluster can be found in Dublin City, with 6.9 % in Cork, and 5.7 % in Donegal.

Group C: Mature suburbs

Cluster 5 retired and mature suburbs

Very high proportions of older pensioners. High proportions of public sector workers, large houses, households with health problems, and households broadband connectivity. Low proportions of workers in agriculture, and households with septic tanks, children, young children, employed residents, young adults, migrants, and apartments. Very low proportions of single person households.

This cluster has a strong spatial concentration. 52.8 % of the residents are to be found in Dublin City (23.1), South Dublin (15.1), Cork City (9.1), and Dún Laoghaire-Rathdown (5.1).

Group D: Commuterland

Cluster 11 comfortably off commuter settlements

High proportions of transport and distribution works, broadband and internet connectivity, older adults, and public sector workers. Low proportions of houses without central heating, pensioners, workers in agriculture and houses with septic tanks.

This cluster is strongly spatially concentrated, with 78.0 % of its residents living in Fingal (19.0), South Dublin (18.3), Dublin City (12.1), Meath County (7.2), Wicklow (6.2), and Cork County (5.7). More than 20 % of the residents of the following counties are in this cluster: Fingal and South Dublin.

Cluster 7 affluent suburbs

Very high proportions of workers in commerce, qualified workers, households with internet connectivity and large houses. High proportions of professional workers,

households with broadband connectivity, older adults and pensioners, two car households and students living at home. Very low unemployment levels.

This cluster is strongly spatially concentrated in Dublin with 72.8 % of the residents in the four counties of Dún Laoghaire-Rathdown (36.2), Fingal (13.2), Dublin City (13.0) and South Dublin (10.4). More than 20 % of the residents of the following counties are in this cluster: Dún Laoghaire-Rathdown (42.6 % of the county's residents)

Cluster 10 affluent commuterland

Very high proportions of qualified workers. High proportions of couple without children, apartments, single person households, professional workers, workers in commerce and transport, broadband and internet connectivity, employment rates and public transport users.

This cluster is strongly spatially concentrated, with 82.4 % of its residents living in Dublin City (52.1), Dún Laoghaire-Rathdown (18.9), Fingal (5.7) and Galway City (5.7).

Group E: Struggling urban peripheries

Cluster 6: Older urban singles

Very high proportions of lone parents, public renters, unemployed, separated and in poor health. High proportions of single persons, households without central heating, older adults and pensioners, and economically inactive. Very low proportions of two car households, qualified worked, employed residents, and households with internet connectivity.

This cluster is not strongly spatially concentrated: 35 % of the residents of this cluster are to be found in Cork City (9.1), Limerick City (7.2), Wexford (6.7) and South Tipperary (6.3). More than 20 % of the residents of the following counties are in this cluster: Limerick

Cluster 13: Single parent urban deprivation

Extremely high proportions of public renters and lone parents. Very high proportions of young children, unemployed workers. High proportions of economically inactive families, separated people, and overcrowded households.

Very much a Dublin phenomenon. 20.4 % of the residents of Fingal are in this cluster. Some 22.5 % of the residents of this cluster are in Dublin City, and 15 % in South Dublin.

Group F: Young families

Cluster 9: Migrant service workers

Very high proportions of private renters, young adults, EU migrants, young children, and workers in manufacturing. High proportions of couples without children, non-EU migrants, employed workers, internet-connected households, Very low proportions of older adults and retirees, and small houses.

This cluster is not strongly spatially concentrated, although 24.7 % of its residents live in Cork County (13.1), Wexford (6.0) and Galway County (5.6).

Cluster 3: Affluent young families with children

High proportions of, employed workers, and internet-connected households and young children. High proportions of young adults, qualified workers, children and two car households. Very low proportions of older adults.

Some 53.8 % of the residents of this cluster are to be found in Cork County (14.6), Meath (10.1), Kildare (10.0), Fingal (0.6) and South Dublin (9.5). More than 20 % of the residents of the following counties are in this cluster: Kildare and Meath.

Cluster 14: Affluent migrant communities

Very high proportions of couple without children, non-EU migrants, EU migrants, young adults, qualified, employed workers, internet connected households, young children. Very low proportions of older adults, pensioners.

This cluster is strongly spatially concentrated; 82.3 % of its residents live in Fingal (33.5), South Dublin (14.75), Kildare (9.3), Meath (6.9), Dublin City (6.7), Cork County (5.8) and Dún Laoghaire-Rathdown (5.2). More than 20 % of the residents of the following counties are in this cluster: Fingal.

Group G: University communities

Cluster 8: University communities

Extremely high proportions of students and private rented households. Very high proportions of apartments, couples without children, non-EU born, single person households, and professional workers. High proportions of Apartment households, private renters, residents with qualifications, and households with internet connectivity. Low proportions of residents with health problems, residents over 65, older adults, children, and families with non-dependent children, unpaid carers, large dwellings, and employed residents. Rural characteristics including agricultural workers and houses with septic tanks are absent.

This cluster is strongly spatially concentrated. Some 73.7 % of the residents of this cluster live in Limerick County (15.2), Galway City (14.6), Cork City (14.4), Dublin City (14.0), Dún Laoghaire-Rathdown (9.0), and Sligo County (6.4).

Group H: City centre diversity

Cluster 15: Stressed inner city singles

Extremely high proportions of single person households. Very high proportions of apartment households, couples without children, both EU and non-EU migrants, separated persons, and households without central heating. High proportions of crowded households, young adults, unemployed workers and residents with health problems. Very low proportions of older adults and retirees two-car households and large dwellings.

Some 22.4 % of the residents of this cluster live in Dublin City, and 18.5 % in Cork City and Cork county.

Cluster 18: Inner area affluent flat renters

Extremely high proportions of apartment dwellers, couples without children, private renters, young adults, qualified workers, Very high proportions of EU and non-EU migrants, single person households, works in commerce and transport, public transport users, households with internet connectivity, and employed workers. High proportions of crowded households, Very low proportions of older adults, young children, two-car households and large dwellings.

This cluster is strongly spatially concentrated, with 81.6 % of its residents living in the three counties of Dublin City (55.6), Dún Laoghaire-Rathdown (14.2), and Fingal (11.8). Much of the areas are in the city centre.

Cluster 17: City migrant apartments

Extremely high proportions of apartment dwellers, couple without children, private renters, EU and non-EU migrants, and young adults. Very high proportions of single person households, qualified workers, households with no central heating and households with students. Very low proportions of older adults and retirees, carers, children two-car households, home workers. Extremely low proportions of large dwellings.

This cluster is strongly spatially concentrated, with 75.0 % of its residents living in Dublin City (62.6), Cork City (7.2) and South Dublin (5.3).

References

- 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.
- Charlton, M., Openshaw, S., & Wymer, C. (1985). Some new classifications of census enumeration districts in Britain: a poor man's ACORN. *Journal of Economic and Social Measurement*, 13, 69–98.
- CSO (2011) Census 2011 Boundary Files – Central Statistics Office <http://www.cso.ie/en/census/census2011/boundaryfiles/>
- Cullingford, D., Flynn, P., & Webber, R. (1975). *Liverpool Social Area Analysis (Interim Report)*, PRAG technical papers TP9. London: Centre for Environmental Studies.
- Feng, Z., & Flowerdew, R. (1999). The use of fuzzy classification to improve geodemographic targeting. In B. Gittings (Ed.), *Innovations in GIS 6* (pp. 133–144). London: Taylor & Francis.
- Gehlke, C. E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A), 169–170.
- Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28, 100–108.
- Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L_1 -norm and related methods* (pp. 405–416). Delft University of Technology: Faculty of Mathematics and Informatics.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. New York: Wiley.
- Kelly, A., & Teljeur, C. (2013). *The National Deprivation Index for health and health services research - update 2013*. Dublin: Small Area Health Research Unit.
- Laine, C., Goodman, S. N., Griswold, M. E., & Sox, H. C. (2007). Reproducible research: moving toward research the public can really trust. *Annals of Internal Medicine*, 146, 450–453.

- Longley, P. (2012). Geodemographics and the practices of geographic information science. *IJGIS*, 26(12), 2227–2237.
- Mather, P. M., & Openshaw, S. (1974). Multivariate methods and geographical data. *The Statistician: Journal of the Institute of Statisticians*, 23(3/4), 283–308.
- Office for National Statistics (2015a) Methodology Note for the 2011 Area classification for Output Areas. <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/methodology-and-variables/methodology-oa.pdf>
- Office for National Statistics (2015b) Pen portraits for the 2011 classification of Output Areas <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/pen-portraits-and-radial-plots/index.html>
- Office for National Statistics (2015c) Radial plots for the 2011 classification of Output Areas <http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/ns-2011-area-classifications/pen-portraits-and-radial-plots/radial-plots-oa.pdf>
- Openshaw, S., & Wymer, C. (1994). Classifying and regionalizing census data. In *Census Users Handbook* (pp. 239–270). Cambridge: GeoInformation International.
- Orford, S., Dorling, D., Mitchell, R., Shaw, M., & Davey Smith, G. (2002). Life and death of the people of London: a historical GIS of Charles Booth's inquiry. *Health and Place*, 8(1), 25–35.
- Peng, R. (2009). Reproducible research and biostatistics. *Biostatistics*, 10, 405–408.
- Singleton, A. D., & Longley, P. (2009). Geodemographics, visualisation and social networks. *Applied Geography*, 29, 289–298.
- Singleton, A. D., & Longley, P. (2015). The internal structure of greater London: a comparison of national and regional demographic models. *Geo: Geography and Environment*, 2, 69–87.
- Singleton, A. D., & Spielman, S. F. (2014). The past, present and future of geodemographic research in the United States and United Kingdom. *Professional Geographer*, 66(4), 558–567.
- Spielman, S., & Singleton, A. (2015). Studying neighborhoods using uncertain data from the American community survey: a contextual approach. *Annals of the Association of American Geographers*, 105, 1003–1025.
- Tomintz, M. N., Clarke, G. P., & Rigby, J. E. (2009). Planning the location of stop smoking Services at the Local Level: a geographic analysis. *Journal of Smoking Cessation*, 4, 61–73.
- Vickers D., Rees, P., & 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
- Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.
- Webber, R. J. (1975). *Liverpool Social Area Study 1971 Data: final report, PRAG technical papers TP14*. London: Centre for Environmental Studies.
- Webber, R. J. (1977). The National Classification of residential neighbourhoods: an introduction to the classification of wards and parishes. In *PRAG technical paper TP23*. London: Centre for Environmental Studies.
- Webber, R., & Craig, J. (1978). *Socio-economic classification of local authority areas. Office of Population Censuses and Surveys: studies on medical and population subjects no. 35*. London: Her Majesty's Stationery Office.
- Wood, J., Isenberg, P., Isenberg, T., Dykes, J., Boukhelifa, N., & Slingsby, A. (2012). Sketchy rendering for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 18(9), 2749–2758.