Cities and Context: The Codification of Small Areas through Geodemographic Classification

Alex Singleton
Department of Geography and Planning, University of Liverpool, UK.
alex.singleton@liverpool.ac.uk

July 8, 2014

Abstract

Geodemographic classifications group small area geography into categories based on shared population and built environment characteristics. This process of "codification" aims to create a common language for the description of salient internal structure of places, and by extension, enable their comparison across geographic contexts. The typological study of areas is not a new phenomenon, and contemporary geodemographics emerged from research conducted in the 1970s that aimed at providing a new method of targeting deprivation relief funding within the city of Liverpool. This city level model was later extended for the national context, and became the antecedent of contemporary geodemographic classification. This paper explores the origins of geodemographics, to first illustrate that the coding of areas is not just a contemporary practice; and then extends this discussion to consider how methodological choices influence classification structure. Being open with such methods is argued as being essential for classifications to engender greater social responsibility.

Keywords: geodemographics,GIS.

1 Geodemographic Place Coding

Geodemographic analysis continues an extensive history of empirically driven models of urban socio-spatial structure, extending back to the 1920s and 30s human ecologists and then later, the large body of empirically driven work producing social area analysis models (Shevky & Williams 1949, Shevky & Bell 1955) for various urban locations (see Timms (1971, 56)). Representations created through such models attempted to reduce the complexities of population and built structure into meaningful and simplified typologies, giving order to multiple attributes about small areas (Abler et al. 1971). Some of the earliest published work on geodemographics were also described as social area analysis (Webber 1975) and focused on single cities (in this case, Liverpool, UK). It was only later that geodemographic techniques were expanded, to create classifications with national coverage (Webber & Craig 1978, Webber 1977). Such geodemographic systems were presented by Webber (1978) as a methodological solution for handling the highly dimensional 1971 UK Census:

> "What is needed is a solution which will pick out pattern from the detail, without loosing too much of the original information, and which will admit more detailed examination of parts of the pattern which become relevant to a particular issue or local area as and when required" Webber (1978, 275).

Webber (1978) also makes two further points: firstly, that geodemographics provide utility as a method of performing analysis on sparsely populated census variables which otherwise might suffer statistical unreliability at the local level. This has contemporary relevance in the context of the US, where the national census now only represents a limited number of questions, and is supplemented by more uncertain small area estimates from the American Community Survey (Singleton & Spielman 2013). Secondly, geodemographics were also argued as a useful framework within which non census indicators could be evaluated over time, and again, would be familiar to contemporary users of geodemographics, with examples of spatial policy evaluation (Batey et al. 2008) and small area population profiling (Singleton 2010*b*).

Although this early history concerned analysis in the public sector, during the 1980s geodemographics were adopted widely by the private sector as a tool for customer segmentation (Sleight 1997), as it was found that the grouping of areas into clusters showed strong correspondence with the consumption of certain product categories. This led to numerous commercial classifications being created, however, more recently, there has been a resurgence of interest in geodemographic applications within the public sector (Longley 2005). Although many geodemographic classifications are commercial, and as such have cost implications, within the UK, there have been a series of classifications built that correspond to the decennial release of the 1981 (Charlton et al. 1985), 1991(Blake & Openshaw 1995), 2001 (Vickers & Rees 2007) and 2011 Censuses (ref to be added).

Although of demonstrated utility (Harris et al. 2005, Singleton & Spielman 2013), geodemographics have been criticized as geographically over simplified (Twigg et al. 2000), or masking of diversity within small areas (Voas & Williamson 2001). However, there is evidence to suggest that geodemographic classifications perform well in comparison to more complex statistical models (Brunsdon et al. 2011). In the mid 1990s there was also extended critique of those negative images place-based marketing initiatives may elicit as part of a wider critique of GIS (Goss 1995, 2003). Uprichard et al. (2009) more recently raises concerns about the "automatic production of space" (Thrift & French 2002), through recursive, reiterative and transformative practices that are embedded within software.

Sociologists (along with numerous other social science discplines) have widely utilized classifications based on occupation (e.g. in the UK, the National Statistics Socio-Economic Classification –NS-SEC) to code individuals into occupational class based groupings / hierarchies. However, since the mid 2000s, interest has grown over the use of contemporary geodemographic classifications as part of research into the spatialization of class (Parker et al. 2007). Geodemographics have been argued as "emblematic of a significantly changing relationship between class and status" (Burrows & Gane 2006, 805), and within this context, geodemographics are seen as usefully encapsulating a wide range of social transactional data that are otherwise of restrictive access to academia, and additionally, also appearing to be engaging with a "rhetoric of sociological discourse" (Savage & Burrows 2007, 887), albeit arguably only at the level of cluster description. Other theoretical work has also made connections between how geodemographics fit within Bourdieu's field-capital theory (Tapp & Warren 2010). Most commercial geodemographic classifications are optimized on the basis of discriminating patterns of consumption (Webber 2007), which have been shown to have similar stratification by occupational group (Sivadas 1997); and as such, it is perhaps not unsurprising that parallels between these two classification approaches are drawn, despite their very different methodology.

2 Subjectivity and Classification Builder Preferences

A geodemographic is created using algorithms that aim to optimize the assignment of small areas into groups that offer the greatest similarity over a typically large set of attributes. However, such representations are explicitly linked to those methodological decisions taken in their construction. Such choices can be informed empirically, theoretically and more pragmatically based on the practitioner or collective of industry experience. As such, the process of geodemographic classification building is regularly described as both art and science (Harris et al. 2005).

The research presented in this chapter does not attempt to provide an evaluation of geodemographics relative to other techniques, nor does it aim to provide an exposition about the "best" method of building a geodemographic, or how this might be assessed. The empirical focus here is to explore how output geodemographic patterns can be sensitive to changes in methodological approach. Some potential options that a classification builder might take when building a geodemographic classification are outlined in the remainder of this section.

2.1 Geographic Extent

The choice of geographic extent impacts how similarity between areas are considered by clustering algorithms. Geographic extent selection has three impacts: firstly, by altering the statistical distributions of attributes, for example, the minimum, maximum and average values for each variable will change relative to the selected geographic extent. This alters the shape of the "attribute space" that is searched when a clustering algorithm is seeking an optimal partitioning of areas into groups. As such, it could be argued that classifications built for and from data about more localized extents will likely demonstrate greater sensitivity (Openshaw et al. 1980), and some have argued that national classification are not necessarily more complete relative to local models (Reibel & Regelson 2011). However, to some extent this also reflects a difference of view that geodemographics are seen as either method (e.g. application of clustering to uncover patterns) or tool (use of a classification system to illustrate patterns / contexts) (Singleton & Spielman 2013).

The second impact of switching from a national to constrained geographic extent is that the benefits of appending national surveys onto a classification are lost, unless adequate sample within the restricted extent can be extracted. Descriptive detail that could be obtained by appending such additional data potentially impacts the range of possible end user applications. These issues may however be minimized in the future as greater volumes of open data that can be partitioned into different geographic extents become available.

Finally, changes from the national extent impact the ability to use geodemographics as a measure for comparing places, and furthermore, can be expensive to maintain and update, an issue acute for the public sector (Webber 1980).

2.2 Scale, zones and input variables

The arrangement of areas into geodemographic clusters are impacted by the choice of zonal geography as this effects the calculation of summary values for input attributes. This is a prescient issue in statistical analysis involving aggregate geographic data, and is referred to as the modifiable areal unit problem (Openshaw 1984). Although of concern, Richard Webber, an expert on geodemographics noted "I have yet to come across any real world example of a conclusion being invalidly reached as a result of this hypothetical possibility" (cited in de Smith et al. 2009, :133). Nonetheless, sensitivity to this issue is required in selecting an "appropriate" geography and interpreting results derived at this selected scale. An "appropriate" zonal geography can be guided by a number of factors such as the availability of data inputs, the intended applications, stability of patterns over different scales or other motivations to provide more detailed classifications, such as leveraging competitive advantage.

Variable choices can be driven by multiple perspectives ranging from theories about what influences socio-spatial structure, empirical investigation of attribute influence on cluster formation, and pragmatic choices based on the experiences of the classification builder or the overarching purpose of the classification (e.g. general purpose versus bespoke - (see Singleton & Longley 2009*a*)). Precursors to geodemographics such a social area analysis (Shevky & Bell 1955) were constructed from a theory about the key drivers of small area differentiation and change, although, some have argued that these were ex post facto rationalization of earlier works featuring more ad-hoc choices (Timms 1971). Geodemographics were however established with a more applied focus. In one of the earliest national classifications Webber & Craig (1978, 6) notes "[a] general purpose classification should by definition, represent as wide as possible a variety of characteristics without over representing any particular aspect". Correlated attributes have a "weighting" effect that gives greater emphasis to such combined dimensions, thus potentially influencing cluster assignment. Inputs into this classification were organised around "Dimensions" not dissimilar to those presented in social area analysis models, and such typology of input attributes have also remained a feature of many present day classifications.

Knowledge about the input variables used to build geodemographics range in degrees of transparency. For many commercial classifications, the exact specification of inputs will be commercially sensitivive, and as such, will not typically be fully disclosed (Singleton & Longley 2009*b*). Conversely, in "open geodemographics" (Vickers & Rees 2007), a full specification of variables would normally be made, including links to where these data may be obtained in the public domain. For open geodemographic classifications, transparency requires that all data be publicly available, and as such, this could also restrict inputs to certain variable types where licences permit redistribution.

Finally, choice of variables are also related to the selection of scale or extent, given that each of which impacts whether or not certain attributes would be available to the classification builder, and how they may be almalagated (e.g. individual versus concatenated age ranges). For example, open data within one context may not be available in another, or, attributes available more universally, might be restricted in scale for a target area.

2.3 Measurement, weights and transformations

A classification can be built with attributes of numerous measurement types such as rates (e.g. percentages), averages, ratios, continuous measures (e.g. distance) or relative scores (e.g. index scores). The choice depends to an extent on the attributes of interest. For example, density would typically be presented as a ratio of population divided by area, whereas an example of a continuous measure might be the distance of an area to the coast or other feature of interest. However, the measurement of attributes using either rates or relative scores are more nuanced. The former relates to the expression of an attribute within an area on a standard scale, whereas the latter takes a rate for a given attribute, and then compares areas by the extent these deviate from the national average. Measurement types impact the range of values that an attribute can hold, for example, percentage scores range between 0 and 100, whereas other measures can hold a wider range of values, and such differences may alter the shape of output classification.

Historically, managing a large number of attributes when building geodemographic classifications was more difficult with restricted computing power limitations. Principal component analysis (PCA) was introduced as a method of reducing attribute dimensions (see Webber 1975), and also reducing the impact of correlated attributes (as PCA by definition comprise linearly uncorrelated variables). As computing power has increased, the necessity for PCA has been reduced, and given that PCA can remove non linear association between variables emergent within specific geographic contexts, some have argued against the use of PCA (Harris et al. 2005).

Weights can be added to attributes to increase their importance in a clustering solution, however, the choice of weights can be considered as subjective, and as such, have been avoided in a number of open geodemographics (Vickers & Rees 2007). Weighting does however see extensive use in commercial geodemographics, and has also been noted as a method to control unhelpful effects caused by highly skewed or otherwise problematic attributes (Harris et al. 2005).

Finally, prior to clustering, data standardization is required to ensure that all attributes are measured on the same scale, and as such, have the same influence on the final cluster solution. However, the exact methods chosen can either constrain or enhance the impact of outliers. For example, standardization with a z-score measures how far an attribute score is relative to the mean in standard deviation units, however, this can accentuate the effect of outliers. Other techniques such as the commonly used range standardization, redistributes attribute scores onto a fixed scale, typically 0-1, compressing outliers into this range, and suppressing their impact. Decisions on which techniques are appropriate are framed within classification builder views on whether they see outliers as an issue to correct, or as an interesting local pattern that is desirable to influence final cluster assignment. Such decisions will also be guided by practicality, given that outlier clusters will by definition be small in nature, and this may not be viewed as useful to feature to appear in a final typology.

2.4 Clustering methods

Clustering algorithms attempt to seek an optimal grouping of areas into clusters by maximizing some measure of within cluster homogeneity or between cluster heterogeneity. Methods of optimization vary between clustering approaches, however, choice of algorithm can influence the assignment of areas to into clusters. A further key decision must be made about how many clusters are desirable in a final solution. Such decisions are commonly guided by experience (Harris et al. 2005), however, can also be assessed empirically through analysis of divisions that "fit" the data most effectively. Common techniques include the use of "elbow criterion" measures (Vickers & Rees 2007) or methods such as silhouette plots (Adnan et al. 2010). A final consideration is whether the classification is to be hierarchical, and if so, whether these are to be built from the top down (most aggregate groups first), or bottom up (most disaggregate groups first).

3 Case Study - national versus local geodemographics

In this final section, two geodemographic classifications are compared, illustrating how from the same input data and methods, two different assignments of areas into clusters can be created on the basis of adjusting the geographic extent of the classification boundaries. The 2011 Office for National Statistics Output area

Table 1: 2011 Output Area Classification Input Variables

Domain	Sub Domain	Variables
Demographic	Age Structure	Age bands
	Family Structure	Marriage; children; dependent children
	Ethnicity	Ethnic Groups; Spoken English; EU V New EU
Housing	Composition	Density; communal establishment; student household; occupancy rating
	Type	Detached, semi, terrace, flats
	Tenure	Socially rented; private rented; owned or shared ownership
Socio-Economic	Health	Day-to-day activities limited a lot or a little; standardized illness ratio
	Employment	Unemployment; full time; part time
	Occupation	Occupation groups
	Education	Level 1; Level 2; Level 3; Level 4+
	Mobility	Car ownership; private transport; public transport; active transport

classification (OAC) will be used as the national classification, and the methodology repeated, however for the localized extent of Liverpool.

The full methodology for OAC 2011 is presented elsewhere (ref to be added). However, in brief: the input data for OAC are sourced entirely from the 2011 census, and are detailed in Table 1. Variables are organised around three domains; demographic, housing and socio-economics. These are then divided into a series of sub-domains comprised of a total of 60 variables. The input variables to OAC are all calculated as percentages against an appropriate denominator, with the exception of a standardized illness ratio and population density. Input data were selected on the basis of maintaining similarity to OAC 2001 Vickers & Rees (2007), but also exploiting some of those new variables added in the 2011 census. Such requirements were formulated after the outcome of a national consultation exercise delivered by the ONS [1] and extensive evaluation.

after the 2011 Census data were assembled and the attribute measures calculated, these were first standardized using an inverse hyperbolic sine function that transforms the attributes more closely to a normal distribution. It can be argued that more normally distributed input attributes assist clustering algorithms such as k-means given their optimization for finding spherical clusters, although, there is no statistical requirement for the data to be normally distributed, as might be the case with techniques such as regression analysis. Secondly, prior to clustering, all of the attributes were standardized onto a 0-1 scale using a range standardization method, thus ensuring that each variable had an equal influence on the clustering result. The K-means algorithm was then implemented to cluster the UK Output Areas and Small Areas (in Northern Ireland) into 8 initial clusters referred to as Super Groups. The data were then split by these clusters, and further divided into between 2 and 4 clusters, forming a second level called Groups and comprising 26 clusters in total. A final set of splits created a Sub Group level, comprising a total of 76 clusters. The nested hierarchy of OAC 2011 is shown in Table 2 and mapped for the UK and Liverpool in Figure 1 and Figure 2. Although the 8 Super Group clusters are visible in the UK map, within Liverpool, only seven clusters are present, excluding the predominantly rural Super Group "1 - Rural Residents".

[1] Details of the consultation exercise can be found `http://www.ons.gov.uk/ons/guide-method/geography/products/area-classifications/ns-area-classifications/new-uk-output-area-classification/index.html`

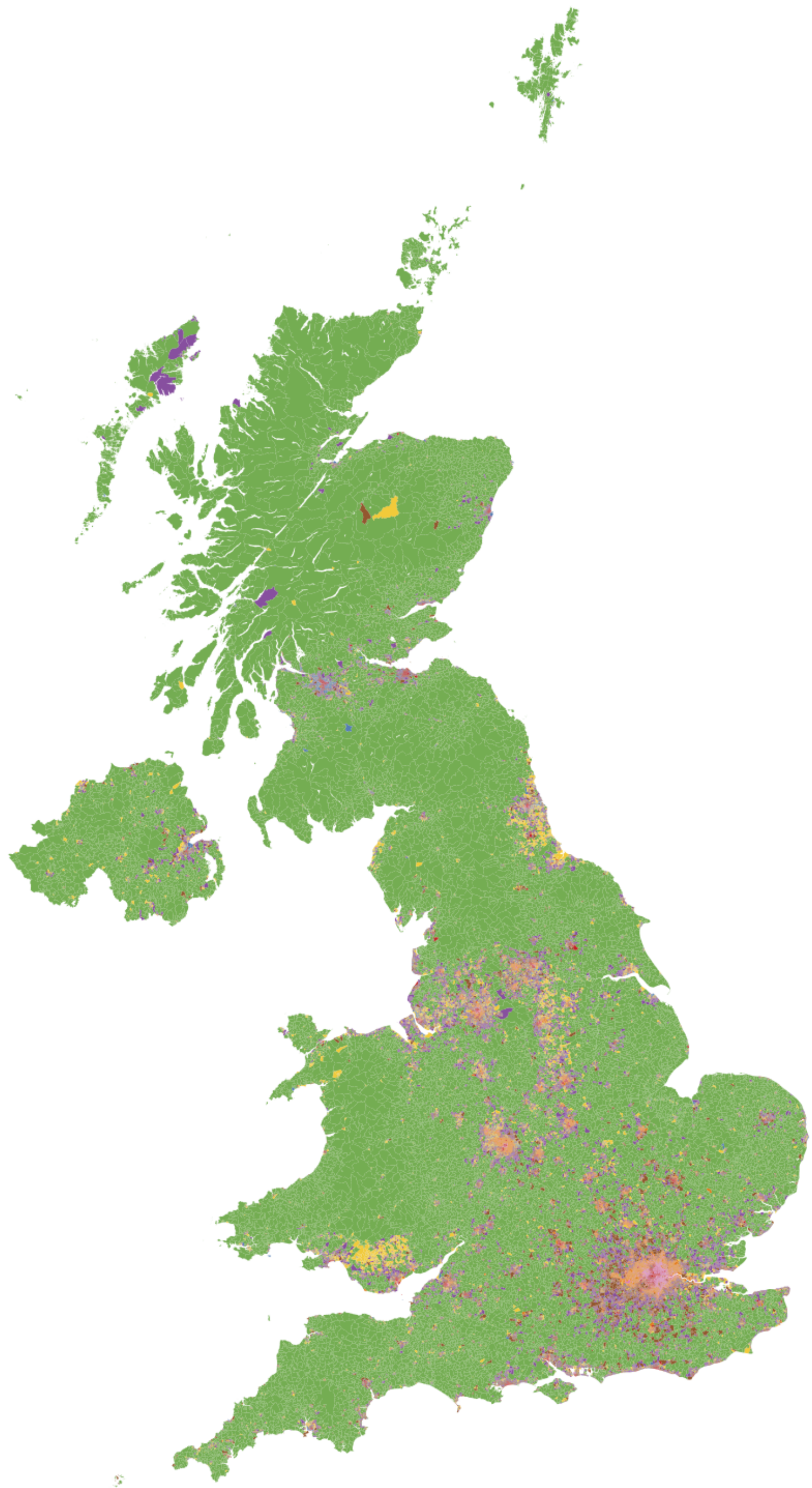

Figure 1: Super Group Level Output Area Classification - UK

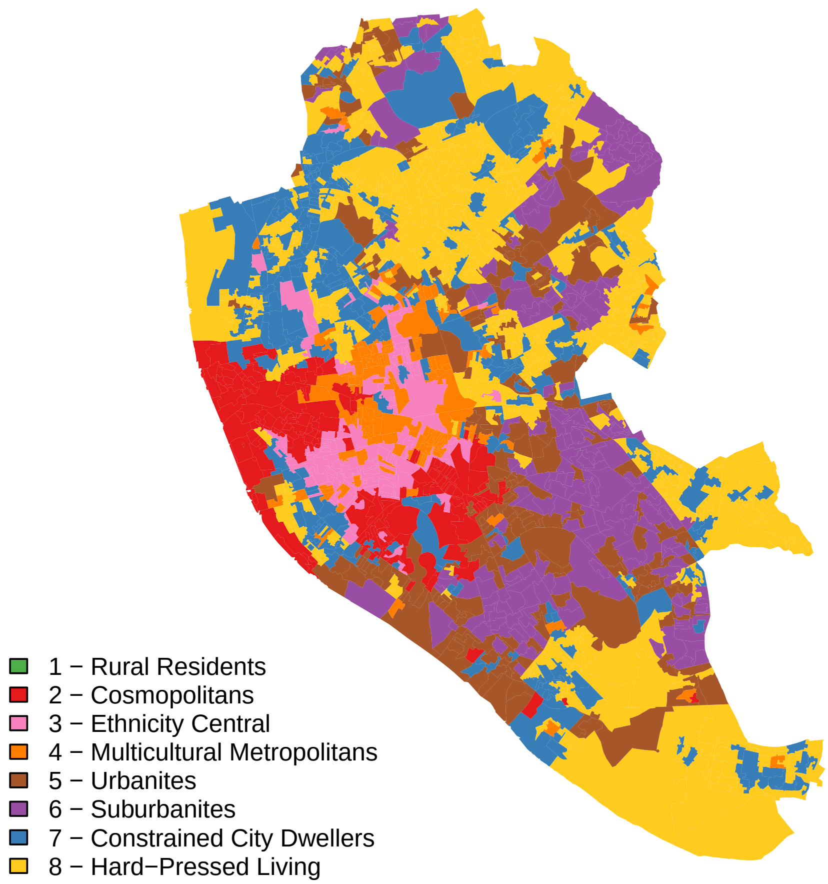

Figure 2: Super Group Level Output Area Classification - Liverpool

Table 2: The 2011 OAC Classification Hierarchy

Super Group	Group	Sub Group
1 - Rural Residents	1a - Farming Communities	1a1 - Rural Workers and Families
		1a2 - Established Farming Communities
		1a3 - Agricultural Communities
		1a4 - Older Farming Communities
	1b - Rural Tenants	1b1 - Rural Life
		1b2 - Rural White-Collar Workers
		1b3 - Ageing Rural Flat Tenants
	1c - Ageing Rural Dwellers	1c1 - Rural Employment and Retirees
		1c2 - Renting Rural Retirement
		1c3 - Detached Rural Retirement
2 - Cosmopolitans	2a - Students Around Campus	2a1 - Student Communal Living
		2a2 - Student Digs
		2a3 - Students and Professionals
	2b - Inner-City Students	2b1 - Students and Commuters
		2b2 - Multicultural Student Neighbourhoods
	2c - Comfortable Cosmopolitans	2c1 - Migrant Families
		2c2 - Migrant Commuters
		2c3 - Professional Service Cosmopolitans
	2d - Aspiring and Affluent	2d1 - Urban Cultural Mix
		2d2 - EU White-Collar Workers
		2d3 - Highly-Qualified Quaternary Workers
3 - Ethnicity Central	3a - Ethnic Family Life	3a1 - Established Renting Families
		3a2 - Young Families and Students
	3b - Endeavouring Ethnic Mix	3b1 - Striving Service Workers
		3b2 - Bangladeshi Mixed Employment
		3b3 - Multi-Ethnic Professional Service Workers
	3c - Ethnic Dynamics	3c1 - Constrained Neighbourhoods
		3c2 - Constrained Commuters
	3d - Aspirational Techies	3d1 - Established Tech Workers
		3d2 - Old EU Tech Workers
		3d3 - New EU Tech Workers
4 - Multicultural Metropolitans	4a - Rented Family Living	4a1 - Private Renting Young Families
		4a2 - Social Renting New Arrivals
		4a3 - Commuters with Young Families
	4b - Challenged Asian Terraces	4b1 - Asian Terraces and Flats
		4b2 - Pakistani Communities
	4c - Asian Traits	4c1 - Achieving Minorities
		4c2 - Multicultural New Arrivals
		4c3 - Inner City Ethnic Mix
5 - Urbanites	5a - Urban Professionals and Families	5a1 - White Professionals
		5a2 - Multi-Ethnic Professionals with Families
		5a3 - Families in Terraces and Flats
	5b - Ageing Urban Living	5b1 - Delayed Retirement
		5b2 - Communal Retirement
		5b3 - Self-Sufficient Retirement
6 - Suburbanites	6a - Suburban Achievers	6a1 - Indian Tech Achievers
		6a2 - Comfortable Suburbia
		6a3 - Detached Retirement Living
		6a4 - Ageing in Suburbia
	6b - Semi-Detached Suburbia	6b1 - Multi-Ethnic Suburbia
		6b2 - White Suburban Communities
		6b3 - Semi-Detached Ageing
		6b4 - Older Workers and Retirement
7 - Constrained City Dwellers	7a - Challenged Diversity	7a1 - Transitional Eastern European Neighbourhoods
		7a2 - Hampered Aspiration
		7a3 - Multi-Ethnic Hardship
	7b - Constrained Flat Dwellers	7b1 - Eastern European Communities
		7b2 - Deprived Neighbourhoods
		7b3 - Endeavouring Flat Dwellers
	7c - White Communities	7c1 - Challenged Transitionaries
		7c2 - Constrained Young Families
		7c3 - Outer City Hardship
	7d - Ageing City Dwellers	7d1 - Ageing Communities and Families
		7d2 - Retired Independent City Dwellers
		7d3 - Retired Communal City Dwellers
		7d4 - Retired City Hardship
8 - Hard-Pressed Living	8a - Industrious Communities	8a1 - Industrious Transitions
		8a2 - Industrious Hardship
	8b - Challenged Terraced Workers	8b1 - Deprived Blue-Collar Terraces
		8b2 - Hard-Pressed Rented Terraces
	8c - Hard-Pressed Ageing Workers	8c1 - Ageing Industrious Workers
		8c2 - Ageing Rural Industry Workers
		8c3 - Renting Hard-Pressed Workers
	8d - Migration and Churn	8d1 - Young Hard-Pressed Families
		8d2 - Hard-Pressed Ethnic Mix
		8d3 - Hard-Pressed European Settlers

A subset of 1584 Output Areas were extracted for the extent of Liverpool, and inputs were created that mirrored the attributes, measures, transformation and standardization methods used for the OAC 2011 classification. Prior to clustering the Liverpool classification, a range of k values were considered for the initial Super Group level by plotting a total within sum of squares statistic for 2-12 cluster solutions. The purpose of this plot was to identify an “elbow criterion” which is a visual indication of where an appropriate cluster frequency might be set for Liverpool. As can be seen in Figure 3 there are no large decreases in the within sum of squares, and a minor moderation of the decrease around 7 or 8 clusters; which also mirrors similar patterns observered within UK OAC (ref to be added). As such, and to maintain comparability with how national OAC is represented within Liverpool, a 7 cluster solution was chosen.

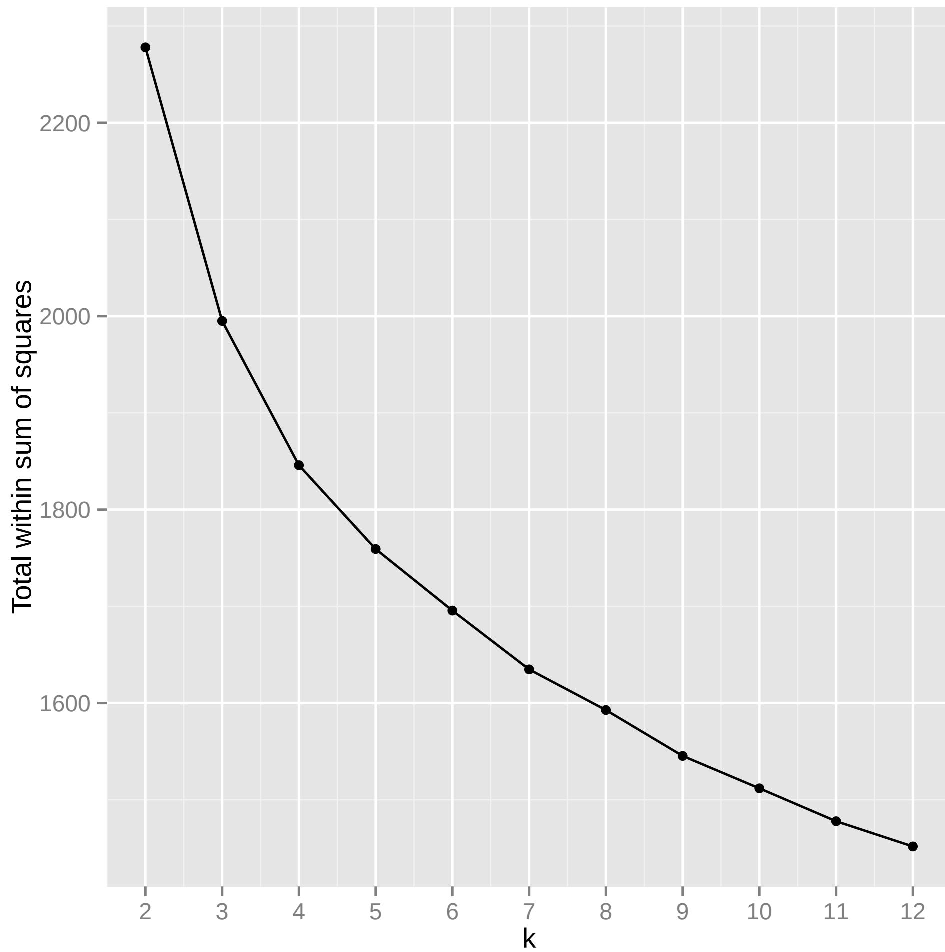

Figure 3: An “elbow criterion” plot used to consider an appropriate number of Super Group clusters in a Liverpool OAC

The next stage was to create the 7 cluster solution, and the k-means algorithm was run 10,000 times on the input data. This repetition is necessary as the initial starting conditions for k means are randomly allocated, and as such, a pool of outcomes must be generated in order to assess which result represents a best fit of the data. For full details of how the k-means algorithm makes an assignments of areas into clusters see (Harris et al. 2005), and for processes of optimization, see (Singleton & Longley 2009*a*). The final set of 7 clusters for Liverpool are shown in Figure 4. To contextualize these assignments, rates for input attributes within each cluster were compared with the Liverpool averages. From these scores, the labels and descriptions shown in Table 3 were formulated. Furthermore, the OA that were closest in attributes to their assigned cluster mean were identified, and a random postcode within these zones selected where an illustrative photograph was taken (see Figure 5).

The purpose of such descriptive material is to give a very brief overview of the “typical” characteristics of the clusters. Although was not the case here, such processes of labeling are often completed by a wider review group rather than an individual. For the 2011 OAC, this involved consultation and approval of names and descriptions by the ONS. An alternative method of validation of both the cluster assignment,

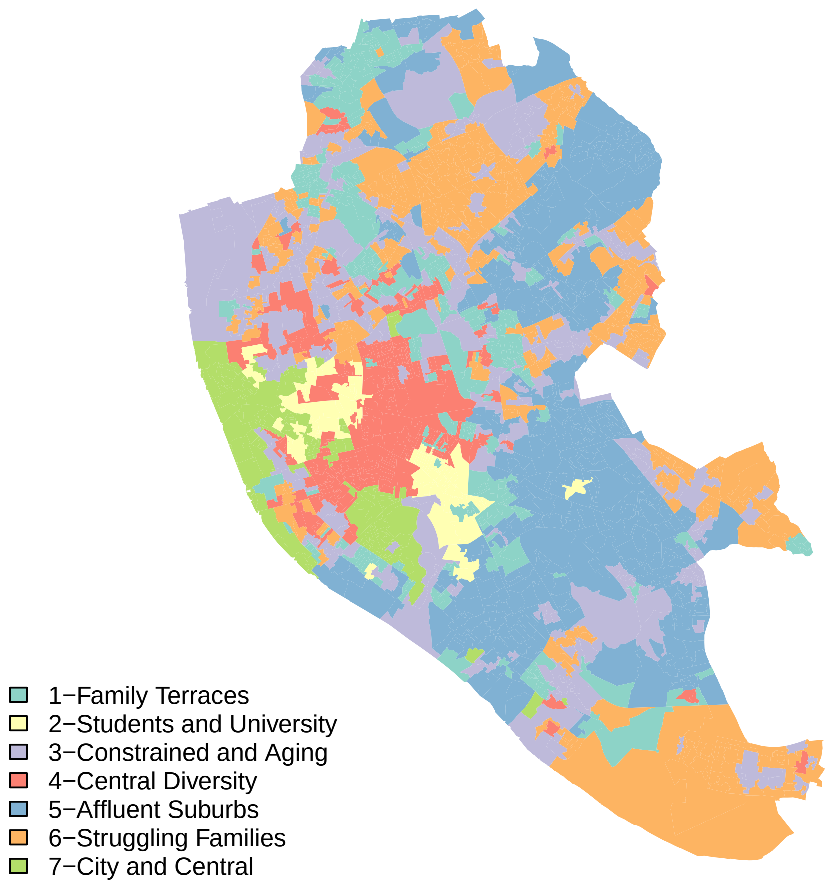

Figure 4: Super Group Level Liverpool Output Area Classification

and the descriptive interpretation was illustrated by Longley & Singleton (2009) who used an online public consultation portal to gather feedback on the classification. Such systems give the general public a method of responding to assignments, and this feedback could be incorporated into revised classifications.

Although the colours are not comparable, if the arrangement of areas into clusters between the Liverpool OAC in Figure 4 and the subset of the 2011 UK OAC for Liverpool (see Figure 2) are compared, the overall patterns are broadly similar, although, in the Liverpool OAC there is a greater degree of spatial autocorrelation (less “noise”). Such effects would likely occur because the optimization process in building the local classification forms clusters in relation to local attribute means rather than those of the UK. The impact is that the resulting clusters fit the data better for their locality.

Table 3: Liverpool OAC labels and brief descriptions

Super Group	Brief Description
1 - Family Terraces	Within these predominantly terraced areas, there are many families with young children, however, fewer ethnic minorities than the Liverpool average. Most property is owner occupied or rented from the private sector.
2 - Students and University	The majority of students studying in higher education live within these areas in shared accommodation, typically rented from the private sector.
3 - Constrained and Aging	These areas have a high concentration of elderly residents and others living in constrained circumstances. There are higher than Liverpool average rates of divorce, and also unemployment. Many of the property are flats which are rented from the social sector.
4 - Central Diversity	These centrally located areas have high ethnic diversity. There are many families within these areas with young children, although higher than the Liverpool average rates of divorce. Unemployment within these areas is high, and those in work tend to work in low level service occupations.
5 - Affluent Suburbs	These affluent suburban areas feature larger detached and semi-detached houses, many of which are owner occupied. Residents are typically well qualified and in the latter stages of successful careers in the public sector, finance or education. Families who have had children are old enough to be no longer dependent.
6 - Struggling Families	Families within these areas typically have young children and live in terraced housing rented from the social sector. There are high levels of unemployment in these areas, however those in work typically have blue collar occupations.
7 - City and Central	These central areas are occupied typically by young professionals, with high ethnic diversity, and particularly high rates of wider EU residents. Many residents within these areas are single and living in flats rented from the private sector, are well qualified and work in white collar occupations.

(a) 1 - Family Terraces - E00034061 - L13 2AY, Colwyn Road.

(b) 2 - Students and University - E00176614 - L15 3LE, Borrowdale Road.

(c) 3 - Constrained and Aging - E00034483 - L25 5LL - Halewood Place.

(d) 4 - Central Diversity - E00176732 - L7 2PT - Stamford Street

(e) 5 - Affluent Suburbs - E00033295 - L12 3HB - Blackmoor Drive

(f) 6 - Struggling Families - E00034134 - L11 7BG - Faversham Road

(g) 7 - City and Central - E00033032 - L17 8UG - Parkfield Road

Figure 5: Liverpool OAC Super Groups

This comparison can be extended by cross-tabulating the assignment of OA in the two classifications. These are presented as percentage scores in Table 4. A number of interesting trends are highlighted, the first is that the OAC Super Group "2-Cosmopolitans" which represents the gentrified core of most large cities in the UK, is split within Liverpool OAC into a cluster with similar characteristics "7-City and Central", and a further cluster that represents many of the student areas ("2-Students and University"). Such areas are not necessarily as concentrated or extensive in other urban areas of the UK. OAC Super Groups maintaining similarity to those in the Liverpool classification include "3-Ethnicity Central" and "6-Suburbanites" with 80.4% and 99.5% similarity respectively. The UK OAC Super Group "4-Multicultural Metropolitans" maintains broad similarity to the Liverpool OAC Super Group "4-Central Diversity", although, some OA are reassigned into "1-Family 2- Terraces" which have lower ethnic diversity and "2-Students and University" which although ethnically diverse, have different age profiles and many more residents in full time education. Similarly, the UK OAC Super Group “5-Urbanites” is split into two between the less affluent “1-Family Terraces” and “5-Affluent Suburbs”. The Super Group “7-Constrained City Dwellers” maintains most similarity to the Liverpool OAC Super Group “3-Constrained and Aging” (67.7%), however OA are also reassigned into “1-Family Terraces” (15.8%) and “6-Struggling Families” (8.2%). The Super Group “8-Hard-Pressed Living” has the majority of OA assigned to “6-Struggling Families” (60.6%), however, other OA are assigned into areas that although are less affluent, have either more elderly residents (“3-Constrained and Aging”; 9.3%) or younger families (“1-Family Terraces”;23.5%). There are also some assignments into the most affluent Super Group in Liverpool (“5-Affluent Suburbs”). This latter difference is interesting as “8-Hard-Pressed Living” might be a cluster where use could be envisioned in public sector targeting of resources - for example - university widening participation or health care initiatives. However, in the context of Liverpool, 6.6% of these areas are classified as “5-Affluent Suburbs” when examined with the city focused classification.

	1-Family Terraces	2-Students and University	3-Constrained and Aging	4-Central Diversity	5-Affluent Suburbs	6-Struggling Families	7-City and Central
2-Cosmopolitans	7.2	34.3	1.8	1.8	0.0	0.0	54.8
3-Ethnicity Central	0.9	3.7	0.0	80.4	0.0	0.0	15.0
4-Multicultural Metropolitans	15.0	8.8	0.0	69.9	1.8	4.4	0.0
5-Urbanites	41.7	0.5	5.9	0.0	51.5	0.0	0.5
6-Suburbanites	0.0	0.5	0.0	0.0	99.5	0.0	0.0
7-Constrained City Dwellers	15.8	0.0	67.7	7.6	0.3	8.2	0.3
8-Hard-Pressed Living	23.5	0.0	9.3	0.0	6.6	60.6	0.0

Table 4: Percentage of OA assigned to OAC Super Groups (rows) and Liverpool OAC Super Groups (columns)

4 Discussion and Conclusions

This chapter has provided an overview of how geodemographic classification emerged as a method of describing the characteristics of areas from rich multidimensional census data. The use of contemporary geodemographics are widespread in the public and private sectors, and effectively code people and the places in which they live into aggregate groupings based on shared attribute similarities. As a representational method, details of reality are balanced in favour of generalization, with the aim of providing a model that has utility in aiding understanding about how places are structured, or, used as a component of area based targeting strategies. Such codification is informed firstly by those choices made when compiling the classification, and secondly, by the choice of labels and descriptive materials associated with the output typology to provide context. As such, there are no “correct” or “true” geodemographic representations, and between classifications these organise a variety of different granular geographies into aggregate typologies of varying characteristics.

Methodological decisions that a classification builder might take when they build a geodemographic vary, and some typical choices were reviewed, alongside discussion of their likely impacts. A comparison of all possible methods and their combinations would run the length of many doctoral theses, and as such, an illustrative case study was selected to focus on the impact one specific methodological decision, the geographic extent of the classification. In this comparison, the national classification OAC 2011 was mapped for the extent of Liverpool. The methodology used to create this classification was then repeated to derive a new classification, however, with cluster optimisation restricted to the geographic extent of Liverpool. The impact of this single decision resulted in a classification which arguably represents the geography of Liverpool more appropriately, given that the clusters were optimised based on a constrained geographic area, and as such, do not have to account for the wider variance of a UK dataset. Reassignments from 2011 OAC into the Liverpool classification were considered, and highlighted local socio-spatial structures which either deviate or are similar to national patterns.

It is important to differentiate between geodemographic method, that is the process by which a classification can be built and a geodemographic system, which are those classifications pre-compiled and often integrated into software coding solutions that can be applied to a range of applications. In this chapter, a classification system is compared with an implementation of a method, and the results indicate that, and perhaps unsurprisingly, a bespoke classification (in this case optimised for local context) offers a potentially more effective representation than the generic geodemographic system. The purpose here is not to make the case for general purpose versus bespoke classifications, as such arguments have been rehearsed since the inception of geodemographics (see: (Openshaw et al. 1980)), and are discussed and evaluated elsewhere (Singleton 2010*a*). In a geodemographic system, the aim is to provide the "best" representation for a wide range of purposes. For example, a commercial classification may find utility in the retail, the automotive or insurance sectors; however, is not designed specifically for any of these application areas[2]. Whereas geodemographic methods aim to provide contextual structure for a given application or locality. Given the divergent aims and objectives of geodemographic systems and methods, the exact choices about how a classification will be created become application specific. For example, the UK OAC 2011 required input attributes that would be available in all counties of the UK, and as such, ignores those attributes that might only be available only within specific countries. Examples could include the input of Welsh language variables in Wales, or within England, attributes about second home ownership.

As illustrated by the case study presented in this chapter, choice of methods can impact the output representation, and as such, it is critical when building a geodemographic to be open and transparent about methodological specification, and present a clear rationalle about why these decisions were taken. This is of particular importance for applications in the public sector where life chances might be apportioned through those decisions informed by geodemographics (Singleton & Longley 2009*b*). Such methodological clarity engenders greater scientific rigor, as methods are more open to scrutiny, testing and reproduction. Arguably, best practice in this regard is to embed code, data and written interpretations, and place these within the public domain, for example, utilising public code sharing repositories such as github[3]. Furthermore, and as argued elsewhere (Longley & Singleton 2009), mechanisms that enable end users to be empowered to give feedback about classification reliability should also be encouraged.

Building geodemographics employs scientific methods of data reduction to provide summary measures of the characteristics of typically small area geography. The art of building geodemographics relates to methodological choices and their justifications, which are typically guided by classification builder expertise. Given this subjectivity, there are no "best" solutions, although some classifications may perform better for certain applications, either serendipitously, or by design such as with a bespoke classification. Given their prevalence of use, it is argued here, that for geodemographics to attain greater social responsibility, all aspects of the build process should be placed within the public domain, and additionally, mechanisms should be enabled to provide end users (either those who are coded or who are coding) with the ability to give feedback on the quality of assignments.

5 Bibliography

References

Abler, R., Adams, J. S. & Gould, P. (1971), *Spatial Organisation: The Geographer's View of the World*, Prentice Hall, New Jersey.

Adnan, M., Longley, P. A., Singleton, A. D. & Brunsdon, C. (2010), 'Towards real-time geodemographics: Clustering algorithm performance for large multidimensional spatial databases: Towards real-time geodemographics', *Transactions in GIS* **14**(3), 283–297.

Batey, P., Brown, P. & Pemberton, S. (2008), 'Methods for the spatial targeting of urban policy in the UK: a comparative analysis', *Applied Spatial Analysis and Policy* **1**(2), 117–132.

Blake, M. & Openshaw, S. (1995), 'Selecting variables for small area classifications of 1991 UK census data', *School of Geography, University of Leeds Working Paper* **95/5**.

[2]Many commercial geodemographic companies in addition to general purpose classifications also offer systems that have been tailored to markets. For example, CACI produce "Financial Acorn" - `http://www.caci.co.uk/integrated-marketing/data-products/financialacorn`.

[3]`https://github.com/`

Brunsdon, C., Longley, P., Singleton, A. & Ashby, D. I. (2011), 'Predicting participation in higher education: a comparative evaluation of the performance of geodemographic classifications', *Journal of the Royal Statistical Society, Series A.* **174**(1), 17–30.

Burrows, R. & Gane, N. (2006), 'Geodemographics, software and class', *Sociology* **40**(5), 793–812.

Charlton, M. E., 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**, 69–96.

de Smith, M., Goodchild, M. F. & Longley, P. (2009), *Geospatial Analysis*, 3 edn, Matador, Leicester.

Goss, J. (1995), Marketing the new marketing. the strategic discourse of geodemographic information systems, *in* J. Pickles, ed., 'Ground Truth', Guildford Press, New York, pp. 130–170.

Goss, J. (2003), The instrumental rationality of geodemographic systems, *in* D. Clarke, ed., 'The Consumption Reader', Routledge, New York. Published: Paperback.

Harris, R. J., Sleight, P. & Webber, R. J. (2005), *Geodemographics, GIS and Neighbourhood Targeting*, Wiley, London.

Longley, P. (2005), 'Geographical information systems: a renaissance of geodemographics for public service delivery', *Progress in Human Geography* **29**(1), 57–63.

Longley, P. & Singleton, A. (2009), 'Classification through consultation: Public views of the geography of the e-society.', *International Journal of Geographical Information Science* **23**(6), 737–763.

Openshaw, S. (1984), *The Modifiable Areal Unit Problem*, CATMOG 38 Geoabstracts, Norwich.

Openshaw, S., Cullingford, D. & Gillard, A. (1980), 'A critique of the national classifications of OPCS/PRAG', *Town Planning Review* **51**(4), 421.

Parker, S., Uprichard, E. & Burrows, R. (2007), 'Class places and place classes geodemographics and the spatialization of class', *Information, Communication & Society* **10**(6), 902–921.

Reibel, M. & Regelson, M. (2011), 'Neighborhood racial and ethnic change: The time dimension in segregation', *Urban Geography* **32**, 360–382.

Savage, M. & Burrows, R. (2007), 'The coming crisis of empirical sociology', *Sociology* **41**(5), 885–899.

Shevky, E. & Bell, W. (1955), *Social Area Analysis*, Stanford University Press, California.

Shevky, E. & Williams, M. (1949), *The Social Areas of Los Angeles*, University of California Press, Berkley.

Singleton, A. (2010*a*), *Educational opportunity: the geography of access to higher education*, International population studies, Ashgate, Farham.

Singleton, A. (2010*b*), 'The geodemographics of educational progression and their implications for widening participation in higher education', *Environment and Planning A* **42**(11), 1560–2580.

Singleton, A. & Longley, P. (2009*a*), 'Creating open source geodemographics - refining a national classification of census output areas for applications in higher education.', *Papers in Regional Science* **88**(3), 643–666.

Singleton, A. & Longley, P. (2009*b*), 'Geodemographics, visualisation, and social networks in applied geography', *Applied Geography* **2009**(29(3)), 289–298.

Singleton, A. & Spielman, S. (2013), 'The past, present and future of geodemographic research in the united states and united kingdom', *Professional Geographer* **In press**.

Sivadas, E. (1997), 'A preliminary examination of the continuing significance of social class to marketing: a geodemographic replication', *Journal of Consumer Marketing* **14**(6), 463–479.

Sleight, P. (1997), *Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business*, NTC Publications, Henley-on-Thames.

Tapp, A. & Warren, S. (2010), 'Field-capital theory and its implications for marketing', *European Journal of Marketing* **44**(1/2), 200–222.

Thrift, N. & French, S. (2002), 'The automatic production of space', *Transactions of the Institute of British Geographers* **27**(3), 309–335.

Timms, Duncan, W. (1971), *The Urban Mosaic: Towards a Theory of Residential Differentiation*, Cambridge University Press, Cambridge.

Twigg, L., Moon, G. & Jones, K. (2000), 'Predicting small-area health-related behaviour: a comparison of smoking and drinking indicators', *Social Science & Medicine* **50**(7-8), 1109–1120.

Uprichard, E., Burrows, R. & Parker, S. (2009), 'Geodemographic code and the production of space', *Environment and Planning A* **41**(12), 2823–2835.

Vickers, D. & Rees, P. (2007), 'Creating the UK national statistics 2001 output area classification', *Journal of the Royal Statistical Society. Series A. Statistics in society* **170**(2), 379.

Voas, D. & Williamson, P. (2001), 'The diversity of diversity: a critique of geodemographic classification', *Area* **33**(1), 63âĂŞ–76.

Webber, R. (1975), *Liverpool Social Area Study, 1971 Data: PRAG Technical Paper 14.*, Studies, Centre for Environmental.

Webber, R. & Craig, J. (1978), *Socio-Economic Classifications of Local Authority Areas (Studies on Medical and Population Subjects)*, London.

Webber, R. J. (1977), *An Introduction to the National Classification of Wards and Parishes*, number 23, London.

Webber, R. J. (1978), 'Making the most of the census for strategic analysis', *The Town Planning Review* **49**(3), 274–284.

Webber, R. J. (1980), 'A response to the critique of the national classifications of OPCS/PRAG', **51**(4), 440–450.

Webber, R. J. (2007), 'The metropolitan habitus: Its manifestations, locations, and consumption profiles', *Environment and Planning A* **39**(1), 182–207.