
Monograph:
Debenham, J. (2002) Understanding Geodemographic Classification: Creating The Building Blocks For An Extension. Working Paper. School of Geography , University of Leeds.

School of Geography Working Paper 02/01

WORKING PAPER

02/01

UNDERSTANDING GEODEMOGRAPHIC CLASSIFICATION: CREATING THE *BUILDING BLOCKS* FOR AN EXTENSION

James Debenham

School of Geography

University of Leeds, Leeds, LS2 9JT

j.debenham@geog.leeds.ac.uk

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Despite the inclusion of some non-Census variables, the traditional proprietary geodemographic classification systems remain purely demand based and static. Nevertheless it has been shown that geodemographics can be extended using supply-side and change variables to create a classification system that measures small areas on the characteristics of the labour market and their propensity to change over time, in addition to the likely levels of affluence more commonly found in such systems (Debenham, Clarke and Stillwell 2001a; 2001b).

This paper presents a number of methods that will later be used to evaluate the success of adding these supply-side and change variables. A static demand classification is created in the style of traditional geodemographic system. Various techniques are then used to evaluate the robustness of the classification and to identify the most important cluster formative variables. Furthermore, this classification is benchmarked against an existing geodemographic system, Experían Ltd's GB MOSAIC system. It is hoped that this will show that the demand variables used here provide a suitable base to operationalise the theories behind extending geodemographics. In order to show how the evaluation techniques can be used to monitor the success of adding different supply-side and dynamic datasets, a suite of property transaction variables are added to the classification. The effect of these variables upon the robustness of the taxonomy and the importance of the individual variables is then displayed.

Key words: Geodemographics, evaluate, performance, cluster formative variables

UNDERSTANDING GEODEMOGRAPHIC CLASSIFICATION: CREATING THE *BUILDING BLOCKS* FOR AN EXTENSION

1 INTRODUCTION

The purpose of this document is to devise a series of methods that will evaluate the success of a progression of classifications that have been created to show how geodemographic taxonomies can be extended by adding supply-side and dynamic variables (Debenham, Clarke and Stillwell 2001a, 2001b). Despite the addition of some non-census variables, geodemographic classifications have remained purely demand based. Traditional systems pay no attention to the supply-side characteristics of the market that also vary spatially, and therefore might not necessarily fulfil the criteria of business need because no consideration is taken of the economic, social or environmental conditions that might influence the consumption of goods and services in an area, let alone how they might change (Ibid.).

However, the contribution of these new classifications to this field of study cannot be assessed without proper evaluation. Some of the geodemographic companies may have added non-census variables (credit ratings, county court judgements) but the impact of such additions upon the classification systems is unclear. It is important to understand which are the *key stock* variables that drive the segmentation in the clustering algorithm used to create the classification. Knowing this will help when choosing which variables to enter into subsequent classifications. Similarly, it is important to know which variables are having a neutral (or even negative) effect upon the segmentation. There is a tendency for some variables with extreme or poorly distributed values to create 'outliers' in multivariate taxonomic space. Such a situation can result in the dataset not being segmented optimally and a poorer classification being created.

The final classification (for a prototype see Debenham et al. 2001a) that includes demand and supply as well as dynamic variables will be created in a cumulative process with different elements of the regional dataset added piece by piece. After each intermediate classification has been created, some robust analysis is required to ascertain how well the segmentation works as a classification. Furthermore, it will be necessary to do some analysis to discover how each variable performs within each classification. The

identification of the strongest variables ensures proper understanding of the underlying socio-economic, demographic and temporal processes, while a knowledge of which variables are the least important will allow decisions to be made about whether they should be included at all in the final classification and whether such a system would be weaker or stronger for their inclusion

This document first describes the creation and evaluation of a static demand classification of postal sectors in Yorkshire and the Humber in the style of a traditional census-based geodemographic system. This classification will be used as a basis for further development of the concept of geodemographics by adding supply-side and change variables. However, the choice of which variables to use in a geodemographic system is a largely subjective exercise (Openshaw 1994) and it will be important to see how good the original choice has been. One important step will be the benchmarking of this classification against a real geodemographic system. Once this has been fully assessed, and the key driving variables identified, it will be shown how such a classification can be extended by adding further information, in this case a suite of variables describing the property market. The new classification will be described and then appraised to see how the new variables have affected the original classification.

The remainder of this document is divided into 7 sections. Section 2 will describe the creation of a static demand classification and present the associated pen portraits and cluster maps. Section 3 will conduct an evaluation of this classification, comparing it to GB MOSAIC, a commercial geodemographic system, and evaluating its robustness by assessing how well the postal sectors in the region have been allocated to clusters in the system. Section 4 will evaluate and analyse the contribution of the variables to the classification to try to highlight the most and least important. Three separate statistical measures will be employed to do this; an analysis of the ANOVA table created by SPSS during the implementation of the *K-means* clustering algorithm, some observations of the correlations between the variables and, finally, a principal component analysis. Then, using similar techniques, sections 5, 6 and 7 will show how the additional variables can be added to the classification and what effect they have had upon both the taxonomy itself and the performance of the individual variables. Finally, section 8 will present some conclusions.

2 CREATING THE STATIC DEMAND CLASSIFICATION

2.1 Variable selection and clustering.

An initial classification has been developed based on a set of demand side variables. The intention was for this first classification to be as similar to a commercial geodemographic system as possible in order to see how the addition of different types of data might transform the taxonomy. The matching process is made difficult because the companies that create and market geodemographic systems rarely release details of the variables they use in their classifications. Nevertheless, the 51 selected here are likely to be reasonably similar to those used currently. There were some elements of parsimony in the variable selection so as not to make the dataset too large. For instance, no ethnicity data has been used, nor data from the tables in the 1991 Small Area Statistics that cross reference variables, such as ethnicity and housing tenure. Nevertheless, the variables incorporated here are good proxy indicators for affluence and life stage. Table 2.1 shows the variables that were used. With the exception of variables 1 to 8 and 51, all the data have been extracted from the 1991 Small Area Statistics available from CASWEB at MIMAS. No Census data are available for current postal geographies so the SAS were obtained for enumeration districts and converted for 784 postal sectors in Yorkshire and the Humber using the look-up tables made available through the All Fields Postcode Directory (Simpson and Yu 2001). The age structure variables (1 to 8) were taken from mid-year estimates for 1999 in the Experían Postal Sector Data (ESRC/JISC agreement) also available at MIMAS. The unemployment rate (VAR51) is taken from the Computerised Claimant Count for July 1999, made available through NOMISWEB. The age structure and unemployment variables could have been extracted from the SAS but, as the intention of the this project is to improve upon traditional geodemographic classification, it is fitting to move away from using 10 year old Census data wherever possible.

The static demand classification was created using a *K-means* classification. The number of clusters (*K*) is specified before the clustering process starts. There is no optimal value for *K*; it depends entirely on the data being classified and the user's personal impression of how many typologies the segmentation should create.

Table 2.1 The suite of demand variables

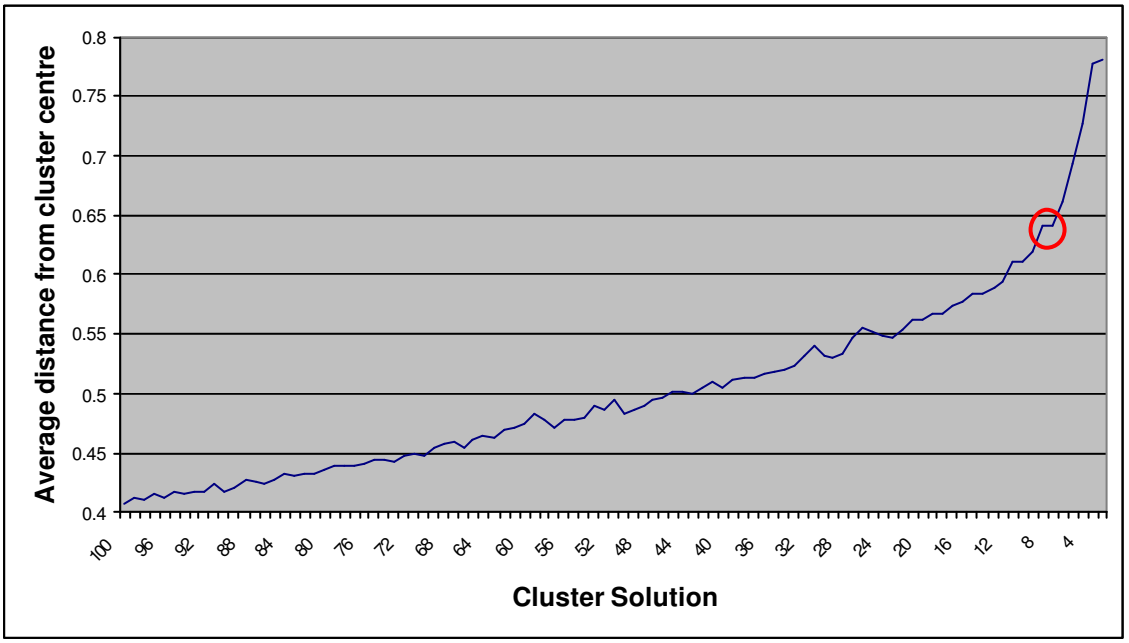
No.	Variable
VAR1.	Persons aged 0-4 (1999 Experían estimates)
VAR2.	Persons aged 5-14 (1999 Experían estimates)
VAR3.	Persons aged 15-24 (1999 Experían estimates)
VAR4.	Persons aged 25-44 (1999 Experían estimates)
VAR5.	Persons aged 45-64 (1999 Experían estimates)
VAR6.	Persons aged 65-74 (1999 Experían estimates)
VAR7.	Persons aged 75-84 (1999 Experían estimates)
VAR8.	Persons aged 85+ (1999 Experían estimates)
VAR9.	Total married population
VAR10.	Single population
VAR11.	Retired (pensioners)
VAR12.	Lone parents
VAR13.	Students (16+) in term-time addresses
VAR14.	Movers last year
VAR15.	Pensioner migrants
VAR16.	Home Owners
VAR17.	Mortgage owners
VAR18.	Privately rented
VAR19.	Rented from Housing Association, Local Authority or New Town
VAR20.	Detached
VAR21.	Semi-Detached
VAR22.	Terraced
VAR23.	Flats
VAR24.	Bedsits
VAR25.	No central heating
VAR26.	Lacking bath & shower
VAR27.	No car
VAR28.	2+ cars
VAR29.	Households > 1.5 persons per room
VAR30.	Households with > 7 rooms
VAR31.	No family household and owner occupied or privately rented
VAR32.	No family household and council rented
VAR33.	Married + cohabiting couple, no children and owner occupied or privately rented
VAR34.	Married + cohabiting couple, no children and council rented
VAR35.	Married + cohabiting couple, dependent children and owner occupied or privately rented
VAR36.	Married + cohabiting couple, dependent children and council rented
VAR37.	Households with two or more families and owner occupied or privately rented
VAR38.	Households with two or more families and council rented
VAR39.	Economically active residents aged 16+
VAR40.	Households with dependants
VAR41.	Self-employed
VAR42.	Households in Social Class I (Professional)
VAR43.	Households in Social Class II (Managerial & Technical)
VAR44.	Households in Social Class III (N) (Skilled Non-manual)
VAR45.	Households in Social Class III (M) (Skilled Manual)
VAR46.	Households in Social Class IV (Partly Skilled)
VAR47.	Households in Social Class V (Unskilled)
VAR48.	Workers with higher degrees
VAR49.	Workers with other qualifications
VAR50.	Persons with Long Term Limiting Illness
VAR51.	Unemployment (claimant count – July 1999) as proportion of working age population

One approach is to repeat the clustering process with different values of K to find the best results. Therefore the algorithm was run for values of $K = 2$ to $K = 100$. The optimum number of clusters can be detected by monitoring the distance of the cases from the cluster centre. For each clustering solution, the average distance of each case from its cluster

centre was computed and the results graphed. Figure 2.1 shows that the 'distance' indicator rises reasonably steadily until around the 87th procedure (13 clusters) when it starts to increment a little more sharply. An analysis of the graph might suggest that the 8-cluster solution (circled in red) might be the more appropriate as it is the final solution before a very rapid rise towards higher average distance values.

The optimal cluster membership can also be monitored by looking at the number of cases within each cluster. While we would neither expect nor want an equal number of cases in each cluster, the better segmentations are those that avoid having the majority of cases in one or two clusters and then a number of sparsely populated groups. Table 2.2 shows the cluster membership for the 8-cluster solution. There is a reasonable spread of values but it is not ideal due to there being only one zone in cluster 3 and seven in cluster 8. Normally this sort of return would see a user opt for a smaller cluster membership but an analysis of the cluster membership returns suggests that this problem is not avoided until the 5-cluster solution, by which time any valuable segmentation of the data set has been lost.

Figure 2.1 Monitoring the average distance from cluster centre with different values of *K*



This problem may be attributed to one of the failings of the *K-means* algorithm; data outliers can seriously affect the results by drawing the cluster centres away from their most favourable locations. Clearly, there are cases that are so far away from the others in

multivariate taxonomic space that the algorithm is forced to place them into their own cluster. Furthermore, because the schedule has defined a limit to *K*, the remaining cases must be attributed to a cluster that may be some distance away. This is why the average distance component in Figure 2.1 rises so sharply towards the end of the schedule.

There are two solutions to this problem. The first is to leave the cluster solution as it is and accept two crucial inadequacies; firstly that some cluster groups will be less valuable because they either contain single cases or too few observations to draw any relevant descriptions, and secondly that the remaining cases have not been optimally clustered. The second is to remove the offending cases and reclassify the remaining data.

Table 2.2 Cluster membership return from the 8 cluster solution

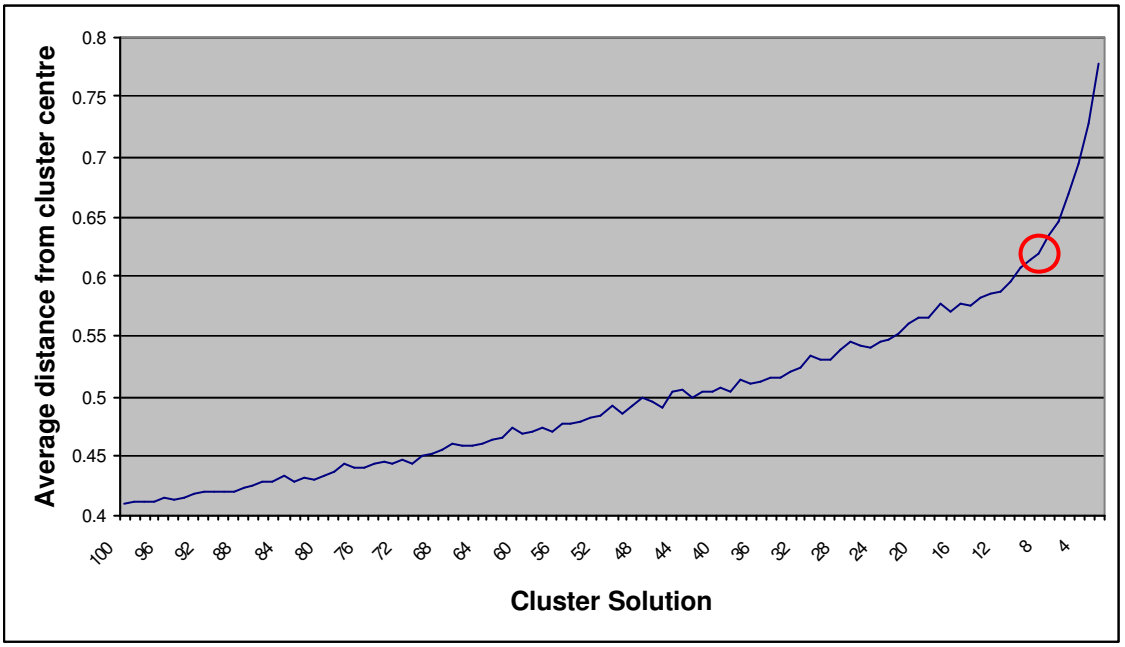
Number of Cases in each Cluster		
Cluster	1	263
	2	43
	3	**1** ←
	4	95
	5	35
	6	133
	7	207
	8	**7** ←
Valid		784
Missing		0

The zone that consistently occurs on its own throughout the different clustering schedules is LS1 8. With a population of just 11 it is liable to very extreme percentages which accounts for its unique position in the segmentation. Removing a zone does compromise the ideals of classification systems being totally comprehensive but as this zone has such a small population it is unlikely that its absence would affect the robustness of the static demand classification. Consequently the zone was removed. The 99 clustering solutions were run again and the results of graphing the average distance values displayed in Figure 2.2 while Table 2.3 shows the membership returns. It can be seen that the optimal number of clusters still appears to be 8 and the cluster memberships for this solution appear satisfactory. The removal of LS1 8 therefore appears to have been a sensible decision as there has been a small reduction in the distance indicator for the 8-cluster solution.

Table 2.3 Cluster membership returns for 8-cluster solution after the removal of LS1 8

Number of Cases in each Cluster		
Cluster	1	165
	2	45
	3	15
	4	201
	5	20
	6	112
	7	122
	8	103
Valid		783
Missing		1

Figure 2.2 Monitoring the distance values after the removal of LS1 8



By comparing the characteristics of the clusters, it is possible to determine their key features and build up a picture of the nature of the zones that fall into that category. Cluster labels and 'pen portraits' can then be derived. Pen portraits are small descriptive analyses of the clusters that draw upon their main identifiable characteristics. The proprietary systems use them to attach a real world context to the cluster labels and modify them to suit the particular application of geodemographics. Cluster evaluation for this system was performed using Z-scores for each cluster variable derived from calculating the standard deviations that occur above and below the global mean as follows:

$$Z_{Km} = (A_{Km} - B_m) / S_m \qquad (1)$$

where:

A_{Km} = mean of variable m in cluster k,

B_m = global mean of variable m,

S_m = global standard deviation for variable m.

Distinguishing variables will have a value that is larger than the global mean and the standard deviation categorisations offer an assessment of by how much. Typically one would examine 1, 2 and 3 standard deviations above and below the mean when evaluating the clusters.

The pen portraits of the 8 clusters created in the static demand classification are given in section 2.2 where the cluster geography is also mapped.

2.2 Cluster pen portraits for static demand classification

Figure 2.4 shows the geographic locations of the 8 clusters created by this classification of static demand data. The cluster pen portraits are now described below.

2.2.1 Cluster 1 *165 Zones*

The zones in this cluster are mostly urban although they are never found in the inner city areas of the bigger metropolitan districts. They are generally found outside the inner city areas of Leeds, Bradford and Sheffield and in the smaller towns of the region. There is a distinct belt running south-east from Leeds to Sheffield, incorporating Rothwell, Wakefield, Castleford, Pontefract, Barnsley, Doncaster and Rotherham, although there are also 'outlying' areas in Humberside and North Yorkshire.

Postal sectors in Cluster 1 are characterised by a high proportion of semi-detached houses and less large detached houses. Households are predominantly in social class III(M) with lower level of qualifications and higher degrees amongst the population. The proportion of households with mortgages is very high while privately rented housing is very low. Renting from the local authority is the same as the regional average. Therefore, the high

number of mortgages may suggest the purchase of council houses by residents. The majority of households are made up of established families, population mobility is low and the proportion of households with dependants suggests a mix of nuclear and extended families. This is corroborated by the higher proportions of both the younger and older age groups.

2.2.2 Cluster 2 *45 Zones*

The postal sectors in Cluster 2 are found mostly in inner city areas, if not in the city centres. Renting from the local authority or housing association is very common here and home ownership or mortgages are rare. The housing stock is largely made up of flats with a noticeable paucity of all other types. This would suggest high rise estates, an implication supported by this cluster's geographic location.

The population structure of these zones is a mixture of elderly and young adults (likely to be the older end of the 15-24 age group). The single and retired populations are both very high here; the married population noticeably lower. There are strong suggestions of lower levels of affluence. Unemployment is high and car ownership low. Limiting long term illness (LLTI) is high and economic activity is bw; it is likely that this is due to the influence of the elderly population. Few members of the population have any sort of qualifications and self-employment is particularly low. Despite this, population mobility is high, particularly pensioner migrants, suggesting a high turnover of both the elderly and young population.

2.2.3 Cluster 3 *15 Zones*

With just 15 zones, this cluster has the smallest membership and is heavily concentrated geographically in the very heart of the major urban areas. Students dominate these areas. Therefore there is a very high representation of the 15-24 age group in predominately privately rented accommodation. The population is mostly single, living in flats, terraced houses and bedsits that often lack central heating. There are very few semi-detached or detached houses and mortgage and home ownership is particularly low.

The population is very mobile as most students have a tendency to move every year. There is also strong evidence of recent graduates living in the area, as the proportion of the

population with higher degrees is greater. This is intuitive as graduates often decide to return to student areas when looking for or starting work due to cheaper accommodation and social links. Affluence levels in the area are hard to determine but it is likely that the non-student or non-graduate population will be quite poor. Unemployment is still at the regional average, car ownership low and there are larger numbers of households with more than 1.5 persons per room.

2.2.4 Cluster 4 *201 Zones*

This cluster has the highest membership and is found largely in the rural areas on the immediate outskirts of the urban areas and along an axis that follows the route of the A1 through North Yorkshire. The exception to this is a large belt running through North Lincolnshire and East Yorkshire.

There are strong suggestions of higher levels of affluence here. Household structure and tenure tends to be families (with and without children) in their own homes (mortgaged or owned outright); rented accommodation is less common. Houses tend to be large detached or semi-detached properties. Car ownership is high and few households have no access to a car at all. Self-employment and economic activity is higher here and there are a large number of households in social class II. Social classes I and III(N) are also common. Unemployment and LLTI are low. Population mobility is generally lower here, suggesting slightly more established families although a higher number of pensioner migrants also suggest retirement migration.

2.2.5 Cluster 5 *20 Zones*

Postal sectors in this cluster are only found in the very centre of the major urban areas. The only exception to this is YO11 2, the centre of Scarborough. The population is very mobile, including pensioner migrants, and mostly live in flats. There are very few other housing types. Like Cluster 2, there is a mixture of young and old and the population is generally single but this time in privately rented accommodation rather than from the council. Higher degrees and other qualifications are common here and most households are in Social Class II, with very few in the lowest groups. Car ownership is low but this is not necessarily a sign of less affluence here because of the central location. Unemployment is

mostly the same as the regional average. There is a strong suggestion that these areas may be gentrified.

2.2.6 Cluster 6 *112 Zones*

This cluster is generally found in the inner city areas of the larger cities, stretching to the outskirts but not the suburbs. Cluster 6 also dominates the postal sectors in the towns between Leeds and Sheffield that do not fall into cluster 1.

It is likely that the population here is less affluent. Renting from the local authority is particularly common here and mortgaging is rare. Most households are in social class III(M), IV or V. Unemployment is above the regional average and LLTI is high. The housing stock is predominately semi-detached or terraced. Car ownership is low and there are few workers with any qualifications. Population mobility and economic activity are low. Most households tend to be families with younger or teenage children although there are larger numbers of lone parents here.

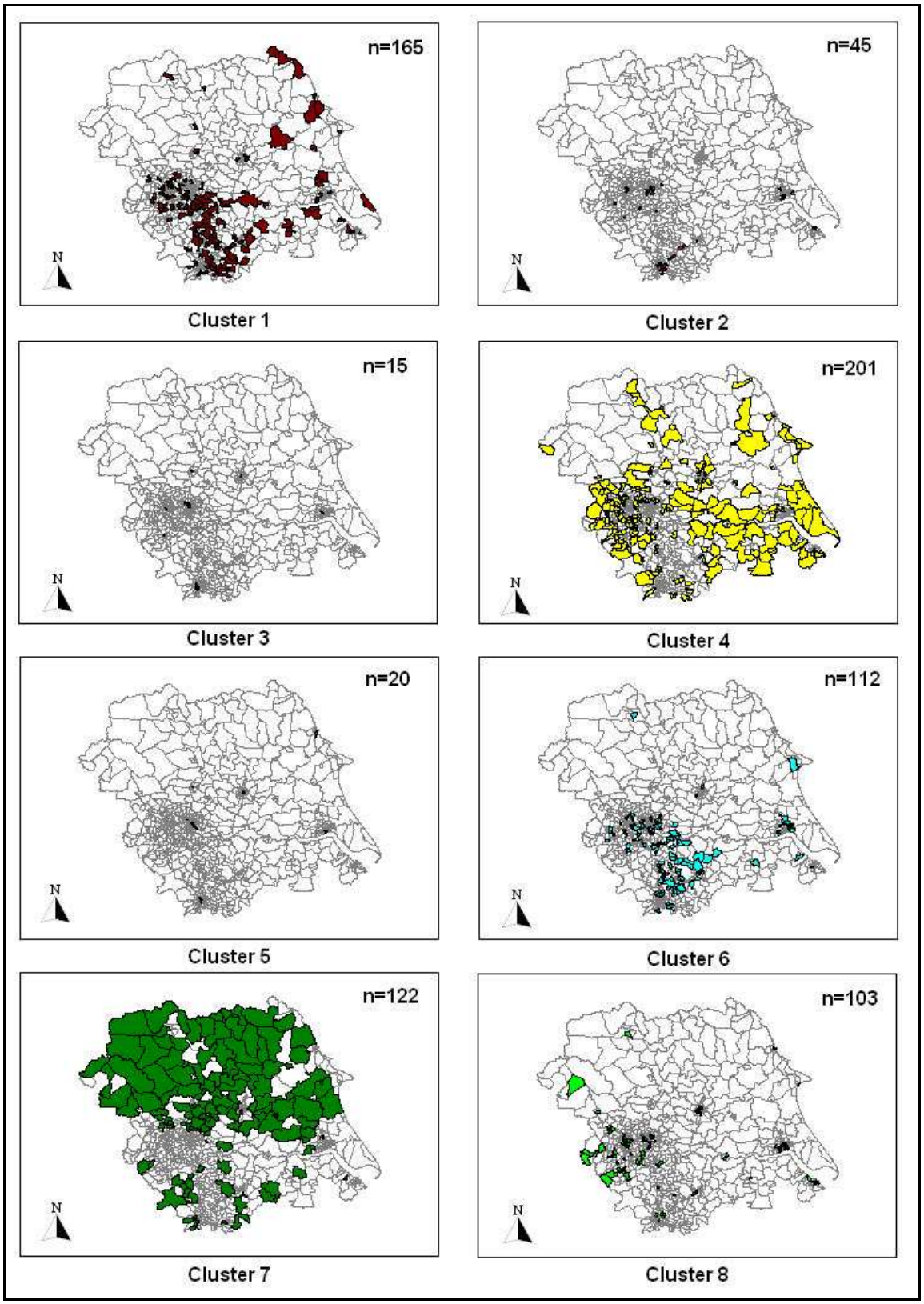
2.2.7 Cluster 7 *122 Zones*

These postal sectors are more rural than those in Cluster 4. They show similar signs of affluence; high car ownership, larger houses, large numbers of households in social classes I and II. However, the population is slightly older with much more established families. There tend to be less young families, particularly the 0-4 age group. Despite this, there are less of the very elderly and retired population; pensioner migrants are fewer, suggesting that there is less retirement migration here than cluster 4. Population mobility in general is lower.

2.2.8 Cluster 8 *103 Zones*

The Cluster 8 postal sectors are predominately found in inner city areas. They are dominated by terraced housing that often lack basic amenities and can be over-crowded. Other housing types are particularly rare here, especially semi-detached housing. Nevertheless there is a mixture of tenure and household types although there tend to be more younger families and fewer in the middle-aged groups. Furthermore, where renting occurs it tends to be under a private arrangement rather than from the council. Unemployment is high and there is a tendency to move.

Figure 2.3 The 8 Clusters of the static demand classification



3 EVALUATING THE CLASSIFICATION

There are no prescribed methods for assessing the 'performance' or accuracy of a classification system. Assessments of how good a geodemographic classification is can be as subjective as some of the decisions made to create it in the first place (choice of variables, number of clusters). There are no statistical methods for testing that the dataset has been classified optimally by the clustering algorithm. Only intuitive comments can be made about how the classification looks to the end user and how sensible the clusters look.

In fact, the static demand classification looks reasonable in the context of Yorkshire and the Humber. There is a clear distinction between the rural and urban zones and little overlap between the two, in other words there are very few urban zones that are placed in predominantly rural clusters. Furthermore, the areas that are perceived to be prosperous and those generally held to be less affluent are preserved in the cluster characteristics. The age structure patterns suggested by the cluster geography are also reasonable. There are occasional anomalies, as one might expect, but the significance of these can be downplayed if one abandons the notion that the pen portrait must exactly describe each zone within the cluster and instead accept that it only describes a number of possible similarities within the cluster.

Nevertheless, it is still possible to conduct some analysis of the overall strength of the classification using other information available to us. The first way of assessing the performance of this classification is to benchmark it against a real geodemographic classification. This will show whether the variables selected have been segmented into sensible looking clusters. If we know that the clusters are similar to an existing system then it can be assumed that we have a firm base from which to start extending the geodemographic classification with new data.

Secondly, using the measure of distance from its cluster centre, we can investigate how well the postal sectors fit with their clusters. In theory, the lower the distance value the more optimal the fit of case to cluster. Large distance values suggest the presence of outliers which may draw the cluster centres away from their optimal location, thus affecting the performance of the system.

3.1 Comparing the Static Demand Classification with GB MOSAIC

The static demand classification is, to all intents and purposes, a traditional geodemographic classification, in that it is almost entirely made up of data from the Census. As this classification is to be used as a springboard for further, more advanced classifications it is prudent to see how it compares to a real geodemographic classification. Experían's GB MOSAIC system has been chosen as the benchmark here because of the data made available in the Experían Postal Sector Dataset. GB MOSAIC classifies all the enumeration districts or output areas in Great Britain into 52 distinct Lifestyle Types which are aggregated into 12 Groups. As there are 8 clusters in the static demand classification it makes sense to compare them to the 12 Groups listed in Table 3.1.

Table 3.1 GB MOSIAC Groups

Group	Label	n	Group	Label	n
A	High Income Families	71	**G**	Town Houses & Flats	40
B	Suburban Semis	52	**H**	Stylish Singles	30
C	Blue Collar owners	205	**I**	Independent Elders	14
D	Low Rise Council	145	**J**	Mortgaged Families	18
E	Council Flats	22	**K**	Country Dwellers	92
F	Victorian Low Status	93	(**L**	Institutional Areas	n/a)

Source: Experían (2001)

Only 11 of the 12 Groups are included here as the "special EDs" that are grouped in Group L – *Institutional Areas* – are not included in the regional information system used to make the static demand classification.

GB MOSAIC is a postcode level geodemographic system and this immediately presents problems for a comparison with a postal sector level classification. However, the Experían Postal Sector Dataset gives an estimate of the number of households in each MOSAIC lifestyle Type in each postal sector. The number of households in each Group was obtained by aggregating these types up and each postal sector in Yorkshire and the Humber was assigned to a group on the basis of which was most prevalent. The number of zones assigned to each group is given in Table 3.1.

The 'pen portraits' of the GB MOSIAC Groups are given in Appendix 1. These have been shortened to miss out the consumer behaviour information but still contain some of the

sociological information that Experían has attached to its geodemographic classification, such as lifestyle aspirations and leisure choices.

The keywords from the GB MOSIAC groups can be used to compare them to the pen portraits from the static demand Clusters to see if it is possible to link them. Once this has been done, the basic geography of the Clusters and Groups can be mapped to see how they compare. It is hoped that the static demand Clusters and MOSAIC Groups that are most similar in descriptive appearance will also be similar on the ground.

3.1.1 Descriptive appearance of MOSAIC Groups

The key words from the 11 MOSAIC Groups may be defined as follows:

Group A – High Income Families: Affluent, 2 car households; large, owner occupied housing; older children; high levels of educational and professional qualifications; few pensioners.

Group B – Suburban Semis: Middle aged families with children; middle class, managerial etc.; owner occupied.

Group C – Blue Collar Owners: Skilled manual workers; former council houses sold to tenants; semi-detached housing; families, few single people.

Group D – Low Rise Council: Renting from Local Authority; lower wages; unemployment; smaller houses; middle aged or older population.

Group E – Council Flats: High and mid-rise flats and "overspill" estates; low incomes, pensioners; long term sick; unemployed; single parents; low qualifications; low car ownership.

Group F – Victorian Low Status: Young families and childless elderly; terraced housing; private renting; some evidence of gentrification

Group G – Town Houses & Flats: Middle income; smaller, younger families; flats and converted large houses.

Group H – Stylish Singles: Young professionals and students; city centre locations; high levels of qualifications.

Group I – Independent Elders: Owner occupied housing and privately rented flats; aged but fit and active population.

Group J – Mortgaged Families: Younger families with mortgages; also high numbers of younger singles and childless couples.

Group K – Country Dwellers: Rural neighbourhoods; varying levels of affluence; high car ownership.

There are bound to be some significant differences between the two systems. GB MOSAIC is a national classification and the clusters are created on national averages which may be different to the regional ones used to create the static demand classification. For instance,

areas that are singled out as having a high car ownership for Yorkshire and the Humber may not be so significant when looking on a national scale. Furthermore, the only GB MOSAIC Group not to be immediately identified with the any of the static demand clusters is Group I *Independent Elders*. This is not surprising as the pen portrait in Appendix 1 suggests that it is quite specific to certain areas of the country, particularly the South Coast. Nevertheless, 14 zones in the region have been placed into this Group, most of which are found on the North Yorkshire coast around Scarborough and Whitby.

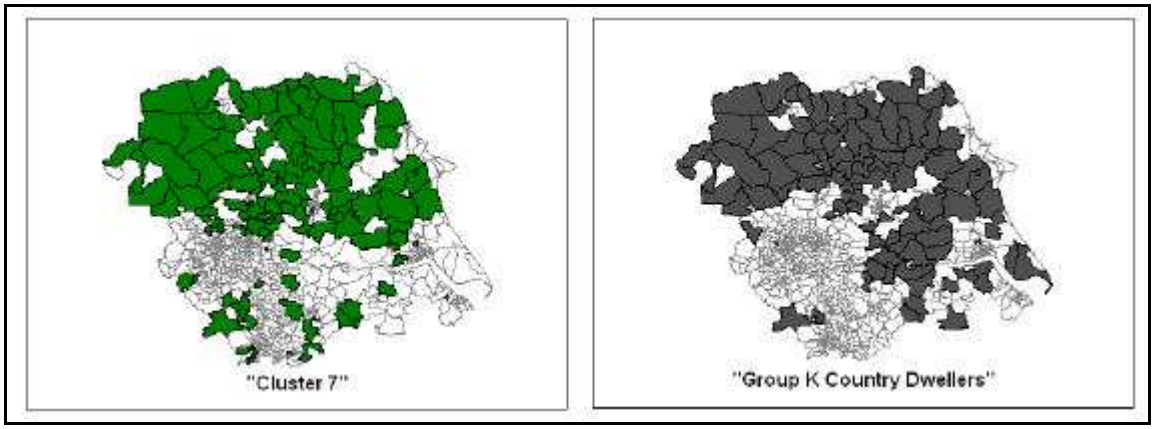
Nevertheless the pen portraits suggest a number of similarities between the two systems. For instance, there are immediate descriptive similarities between Group E and Cluster 2, Group H and Clusters 3 and 5 and Group K and Cluster 7. However, the similarities between the systems can really only be ascertained by mapping the clusters and groups in conjunction with the descriptive comparisons.

3.1.2 Mapping and comparing the Clusters and Groups

Cluster 7 with Group K

As mentioned before, the most immediate similarity in the cluster descriptions is between GB MOSAIC Group K *Country Dwellers* and the static demand classification Cluster 7 that Section 2.2.7 identifies as being predominately rural. Figure 3.1 shows that their geographical distribution is also quite similar. There are some differences, principally the predominance of Group K around Selby and the head of the Humber estuary but otherwise the patterns are consistent.

Figure 3.1 Comparing static demand Cluster 7 with GB MOSAIC Group K

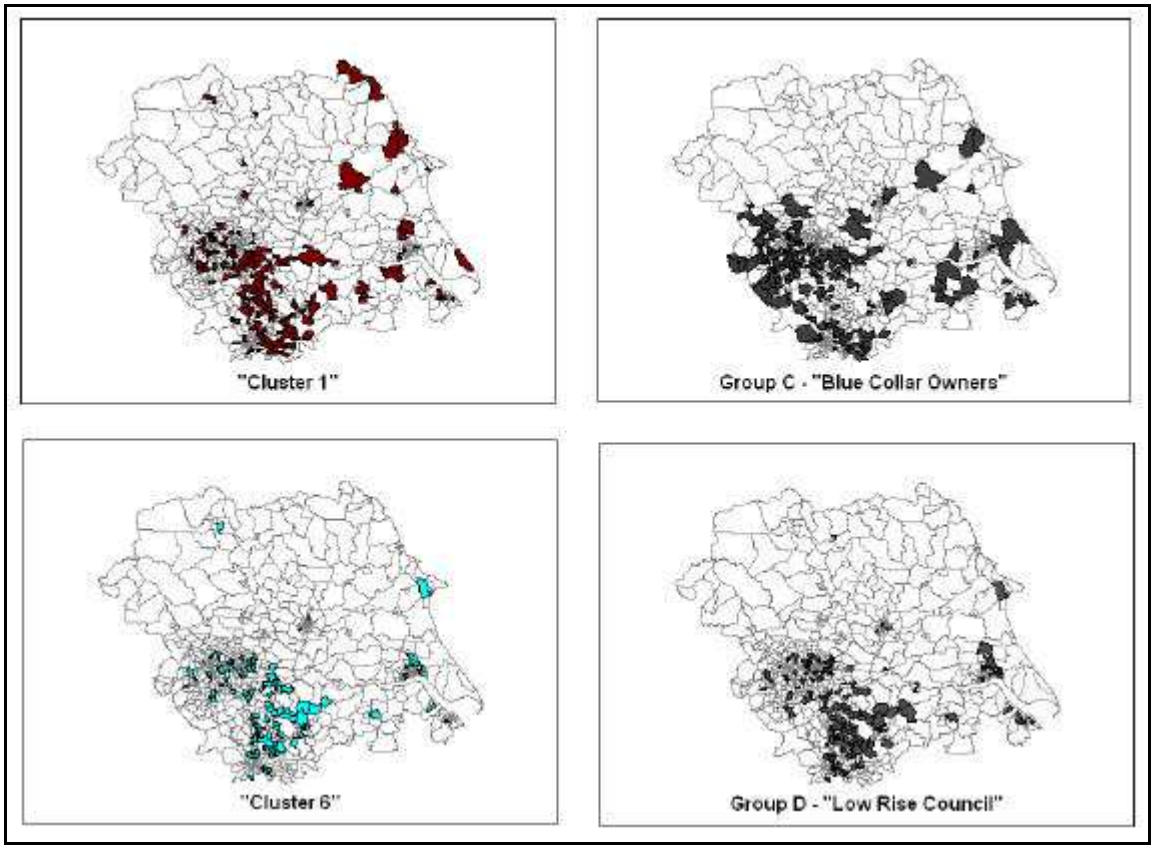


Group K map from Experían Postal Sector Dataset (ESRC/JISC Agreement)

Cluster 4 in the static demand classification accounts for most of the zones allocated to Group K but missed by Cluster 7.

Cluster 1 with Group C and Cluster 6 with Group D

Figure 3.2 Comparing static demand Cluster 1 with GB MOSAIC Group C and Cluster 6 with Group D



Group C & D maps from Experían Postal Sector Dataset (ESRC/JISC Agreement)

In descriptive appearance Cluster 1 seems most similar to Group C *Blue Collar Owners*. Both are made up predominately of skilled manual workers in semi-detached housing with established families making up the majority of the population. The most striking similarity is that the claim in the static demand classification pen portrait that these zones may be made up of former council houses now sold to tenants is echoed in the Group C pen portrait. Figure 3.2 shows these two typologies mapped and it is fair to say that there is distinct geographical similarity in addition to the descriptive one. There are more zones in Group C than Cluster 1 (206 versus 165) suggesting overlaps with other zones but nevertheless the basic geography of Cluster 1 is replicated in MOSAIC Group C. Group C

has less representation in North Yorkshire than Cluster 1 and is more prevalent in the far west of the region to the south of Bradford but otherwise the match is good, particularly in the areas to the south of Leeds and surrounding Sheffield. The disappointing aspect is that the characteristic belt of zones in Cluster 1 in the static demand classification that links Leeds with Sheffield is not replicated in GB MOSAIC.

A number of these zones are allocated to Group D *Low Rise Council.* The description of this group suggests that it is quite similar to Cluster 6 in the static demand classification. Essentially these areas are comprised of less affluent households that rent from the Local Authority. Figure 3.2 shows that the geographies of these two typologies are again similar. Once more there are more zones in the MOSAIC Group than there are in the static demand cluster (145 against 112) but the overall distribution looks similar, particularly the concentration in the south east of the region around Sheffield and Rotherham.

There is likely to be some co-habitation between households of these typologies. The principal difference between Cluster 1 and Cluster 6 in the static demand classification and Group C and D in GB MOSAIC is that the former type is made up of households who are likely to have bought their former council houses and the latter is made up of those who, because they may be unemployed or in a job which does not provide sufficient wages, have not done so. These households tend to exist along side each other and this may be why there is some overlap between all four types displayed in Figure 3.2. The aggregation of units up to postal sector will have blurred the distinctions further. If one combines both static demand clusters and both MOSAIC groups into two 'super-groups' then we can suggest that there is a distinct geographical similarity although the static demand classification still fails to identify the zones picked out by MOSAIC in the west of the region

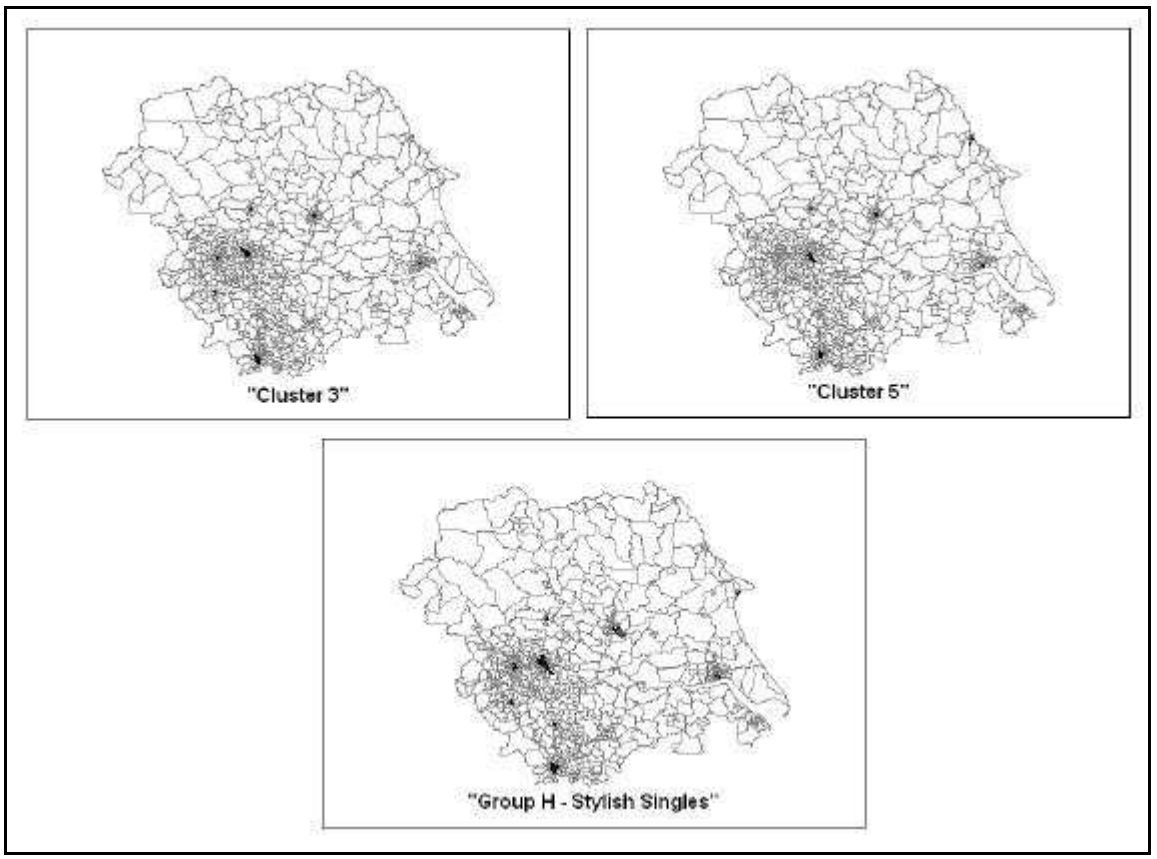
Cluster 3 and Cluster 5 with Group H

In the static demand classification, Clusters 3 and 5 are identified as being dominated by young adults. Cluster 3 is characterised by students and recent graduates while Cluster 5 seems to contain young professionals in city centre flats and developments. These groups seem to correspond with GB MOSAIC Group H *Stylish Singles*. All three have large proportions of young adults with high levels of educational attainment. The pen portraits

for Group H indicates high proportions of students and recent graduates and this links it to Cluster 3 while the younger professional element of the MOSAIC Group identifies it with Cluster 5. It is interesting that the static demand classification has distinguished between these two elements although Cluster 5 also highlights large numbers of elderly residents which may have distinguished it from the younger populations in Cluster 3.

Mapping these three typologies in Figure 3.3 shows a high degree of geographical similarity. Cluster 3 and 5 combine to account for all the central urban zones that fall in MOSAIC Group H. The only exception to this is in York where none of the postal sectors in the very centre are allocated to Group H but instead a zone on the outskirts. This is more likely to be due to a problem with the method of allocating zones to MOSAIC groups on the basis of majority group membership rather than a problem with the segmentation of the static demand classification. Otherwise the differences are minor.

Figure 3.3 Comparing static demand Clusters 3 and 5 with GB MOSAIC Group H



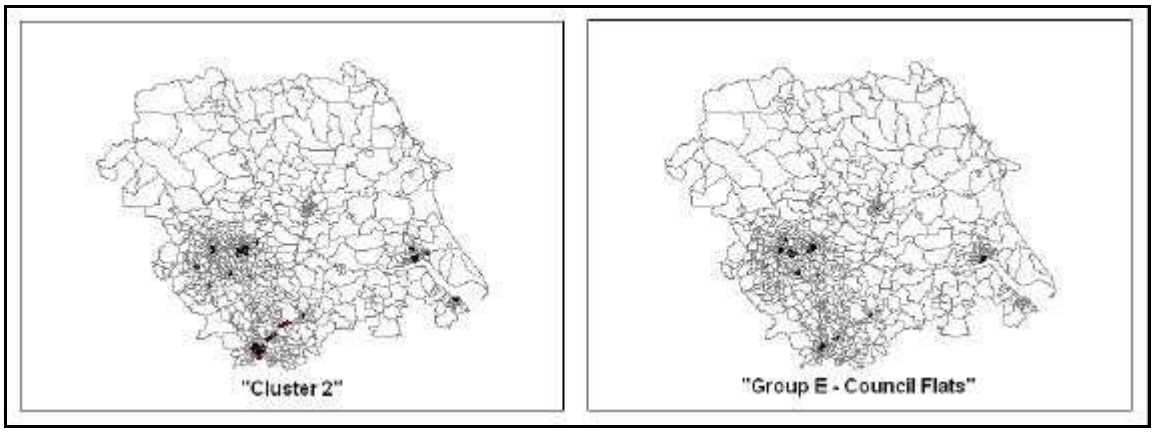
Group H map from Experían Postal Sector Dataset (ESRC/JISC Agreement)

Cluster 2 with Group E

Cluster 2 and MOSAIC Group E *Council Flats* share many similarities, not least the implication that they are high rise council estates. The majority of the housing stock is flats and the most common tenure is rental from the Local Authority. Both have high levels of unemployment and long term illness and high numbers of pensioners and younger adults with no qualifications.

Figure 3.4 suggests that the geographical similarities between the two typologies are as close as the descriptive ones. Cluster 2 has twice as many zones as Group E (45 to 22) and this suggests that it also has similarities with other MOSAIC types. Many of the zones in Cluster 2 that are not also Group E are instead found in Group D, thus suggesting that Cluster 2 also has low rise council estates as well as high rise. As the static demand classification only has 8 clusters rather than 11 groups, it is likely that these distinctions become blurred, especially as high rise and low rise council estates are frequently located close to one another.

Figure 3.4 Comparing static demand Cluster 2 with GB MOSAIC Group E



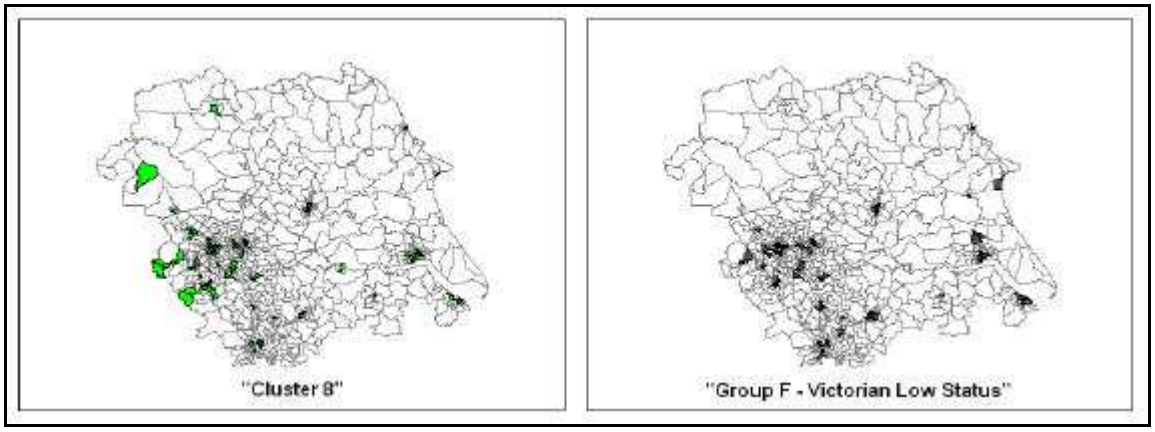
Group E map from Experían Postal Sector Dataset (ESRC/JISC Agreement)

Cluster 8 with Group F

Descriptively, Cluster 8 and GB MOSAIC Group F *Victorian Low Status* appear similar because of the dominance of terraced housing. As the pen portrait for Cluster 8 points out, other types of housing are rare here. Both have a predominance of younger families and

where renting occurs it tends to be private. The GB MOSAIC description hints towards the possibility of gentrification but looking at the locations of the static demand classification Cluster this would seem unlikely in Yorkshire and the Humber; they tend to be depressed inner city areas. The gentrified areas are located in Clusters 3 and 5.

Figure 3.5 Comparing static demand Cluster 8 with GB MOSAIC Group F



Group F map from Experían Postal Sector Dataset (ESRC/JISC Agreement)

Geographically the similarities are evident. Both typologies have roughly similar number of zones, 93 in Group F and 103 in Cluster 8. The static demand cluster has more of a presence in some peripheral areas, such as the small congregation of zones in the west of the region but otherwise both types have a similar inner city pattern. Some of the zones placed in Cluster 8 to the west of Bradford are found instead in GB MOSAIC Group C (Blue Collar Owners). This may be due to such areas being highly mixed, containing large numbers of households in different MOASIC Groups and the aggregation process has placed such zones into one Group when they could be identified in another.

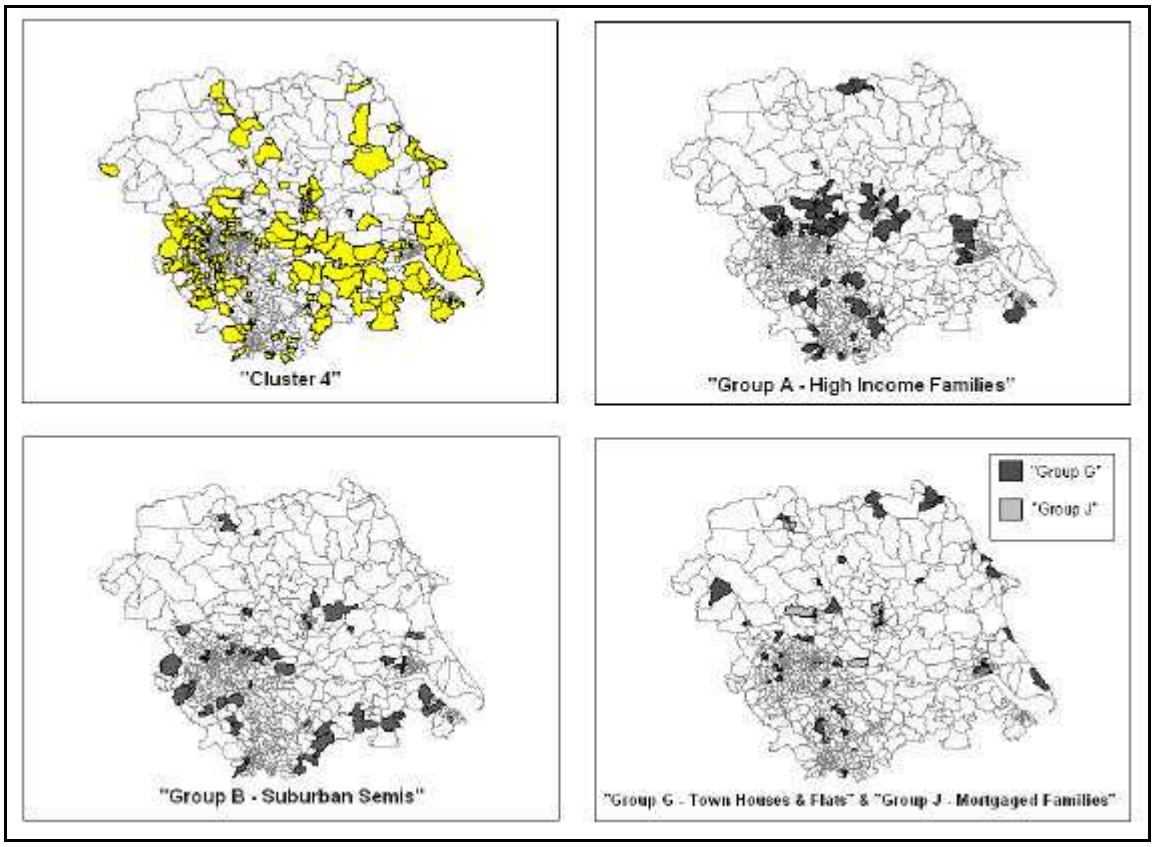
Cluster 4 with Groups A, B, G and J

Cluster 4 has the highest membership of all the clusters in the static demand classification with 201 zones. It is therefore likely that it will be associated with more than one MOSAIC Group. The cluster descriptions suggest that it can be associated with 4 MOSAIC Groups; Group A *High Income Families*, Group B *Suburban Semis*, Group G *Town Houses and Flats* and Group J *Mortgaged Families*. Each Group bears some descriptive similarity to Cluster 4. The most obvious are Groups A and B with their emphasis on higher affluence levels, established families, larger houses (detached or semi-detached) high car ownership

and managerial or professional occupations. However the characteristics of Groups G and J, with "middle income", possibly slightly younger families.

Figure 3.6 shows that the geographical similarities are not so implicit. Many of the zones in Group A *High Income Families* are more comparable with the static demand Cluster 7. This is not discouraging as the description of Cluster 7 indicates that it has a high proportion of very affluent households thus suggesting that Cluster 7 has as varied a mix of rural affluence as its GB MOSAIC counterpart (Group K) claims to have. Nevertheless, there are some similarities, particularly to the north of Bradford, around York and rural parts of north-east Lincolnshire around Grimsby.

Figure 3.6 Comparing static demand Cluster 4 with GB MOSAIC Groups A, B, G and J



Groups A, B, G and J maps from Experían Postal Sector Dataset (ESRC/JISC Agreement)

The geography of Cluster 4 is more similar to Group B, particularly in the West Yorkshire conurbation. This is encouraging. Furthermore, many of the zones placed in MOSAIC

Groups G and J also seem to show some similarities with Cluster although this is by no means exclusive.

Cluster 4 seems to show geographical similarities with many MOSAIC Groups. There is a clear concentration of zones to the south west of the Leeds-Bradford conurbation, which, according to GB MOSAIC, are *Blue Collar Owners* (Figure 3.2). Although the immediate impression is that this is a poor match a little more observation suggests that it may be intuitive. One of the principal characteristics of Cluster 4 is home ownership, hence the links with Group C.

3.1.3 Summary

The level of geographical association between the static demand classification (SDC) and GB MOSAIC can be summarised in Table 3.2. The number of zones that appear in the right combination of clusters and groups is given as a percentage of the number of zones in both the static demand classification Cluster and the GB MOSAIC Group. 438 (56%) postal sectors fall into the correct combination of Clusters and Groups. The highest level of association in terms of the static demand classification is the 83% of postal sectors in Cluster 6 that are also in MOSAIC group D. The smaller percentages are generally found in the Clusters that have the lowest membership. The exception to this is the level of association between Cluster 4 and the four MOSAIC groups. Section 3.1.2 highlighted that this association was weak and the table corroborates this with just 40% of postal sectors matching the criteria of membership in both classifications. The association appears even worse when calculated as a percentage of the number of zones in the corresponding GB MOSAIC Groups.

Nevertheless, the static demand classification shows a number of strong similarities with the commercial system. Although there are some differences, the overall geography of affluence and deprivation is maintained between both of them. However, GB MOSAIC was based upon national averages while the static demand classification was based only upon regional data. As we have seen, the pen portrait for Cluster 4 suggests high levels of affluence and thus a link with GB MOSAIC Group A. However, evidently this benchmark is not the same as the national standards and the level of geographical similarity is much lower than expected.

Table 3.2 The degree of similarity between the static demand classification and GB MOSAIC

SDC Cluster	n	GB MOSAIC	n	No. of zones in both	% of SDC	% of GB MOSAIC
1	165	C	206	103	62.42	50.00
2	45	E	22	16	35.56	72.73
3	15	H	30	12	80.00	40.00
4	201	A,B,G & J	181	82	40.80	45.30
5	20	H	30	8	40.00	26.67
6	112	D	145	93	83.04	64.14
7	122	K	92	60	49.18	65.22
8	103	F	93	64	62.14	68.82
-	-	I	14	-	-	-
Total	**783**	**-**	**783**	**438**	**55.94**	**55.94**

It is possible that the method of attributing postal sectors to MOSAIC Groups may also be problematic. Following the logic of The Modifiable Areal Unit Problem (Openshaw 1984) that the averaging out of small area attributes will occur as the typologies are made at the postal sector level, it is likely that some of the distribution of Groups will be lost. Labelling a postal sector on the basis of which MOSAIC Group has the most households in the zone may not be the best way to estimate a postal sector classification from one based on postcodes and may explain why only 56% of postal sectors are in the matching Cluster. However it is the most immediate method from the data available.

Despite this, many of the key characteristics of both systems are replicated in each other; for instance the illumination of areas of council house purchasing and the precise geographical location of high rise estates. It is apparent that the static demand classification retains the spirit of the selected proprietary classification system and is a suitable base from which to try to extend the analysis by adding indicators that measure other dimensions.

3.2 Assessing the performance of the classification system

A robust assessment of the classification system can be made by looking at the goodness of fit of cases (postal sectors) to their clusters. The *K-means* clustering algorithm uses the average value of each variable to plot the cluster centre in multivariate taxonomic space. The Euclidean distance of each case from its cluster centre can then be calculated using the square root of the sum of squares of the difference between that zone's attributes and the

cluster averages. By analysing these distance values it is possible to detect how well zones have been fitted to clusters. As discussed previously, one of the weaknesses of *K-means* is that the procedure can force a zone into a cluster because the algorithm is mutually exclusive, collectively exhaustive and is bound to satisfy the pre-determined value of *K*. A large number of high distance values will suggest that cases have not been optimally clustered and that consequently the cluster averages have been distorted, thus affecting the overall character of the cluster.

3.2.1 Distance from cluster centre

Figure 3.7 displays a histogram of distance values for the entire classification. The histogram shows a distinct positive skew, thus suggesting that the majority of values are in the lower distance categories. The mean distance value is reasonably low at 0.62 and the standard deviation just 0.23. However the range of values is quite high, the minimum is 0.27 and the maximum 1.73 (a range of 1.46) and there are some 52 zones (6.6%) with distance values in excess of 1. Nevertheless, the classification looks to be performing reasonably well; in general the distance values are low and there appear to be very few genuine 'outliers' in the taxonomic space. It will be interesting to see how the addition of extra variables in subsequent classification systems affects this distribution.

The map of zone distances from cluster centres (Figure 3.8) shows no real discernible pattern although there does appear to be an urban bias to the zones with the highest distance values (the underlined labels). The zones with the lowest values (plain text labels) appear more spread out although they are rarely very rural. The zone with the highest distance is BD1 3 (central Bradford – Cluster 8), the zone with the least distance is HD8 9 (rural fringe beyond Huddersfield – Cluster 4). The labelling system in Figure 3.8 is designed to display the cluster membership of each zone with a distance value less than 0.36 or greater than 1.46 and suggests that there is some cluster specific pattern to the distance values that we can examine in more detail in the next section.

Figure 3.7 Histogram of distances from cluster centre by case

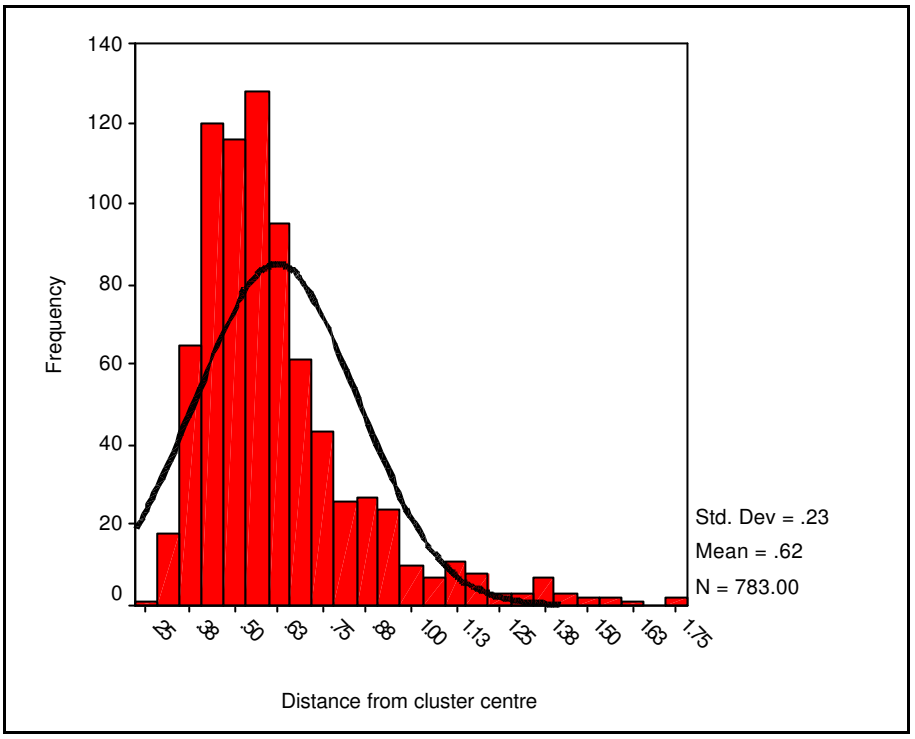
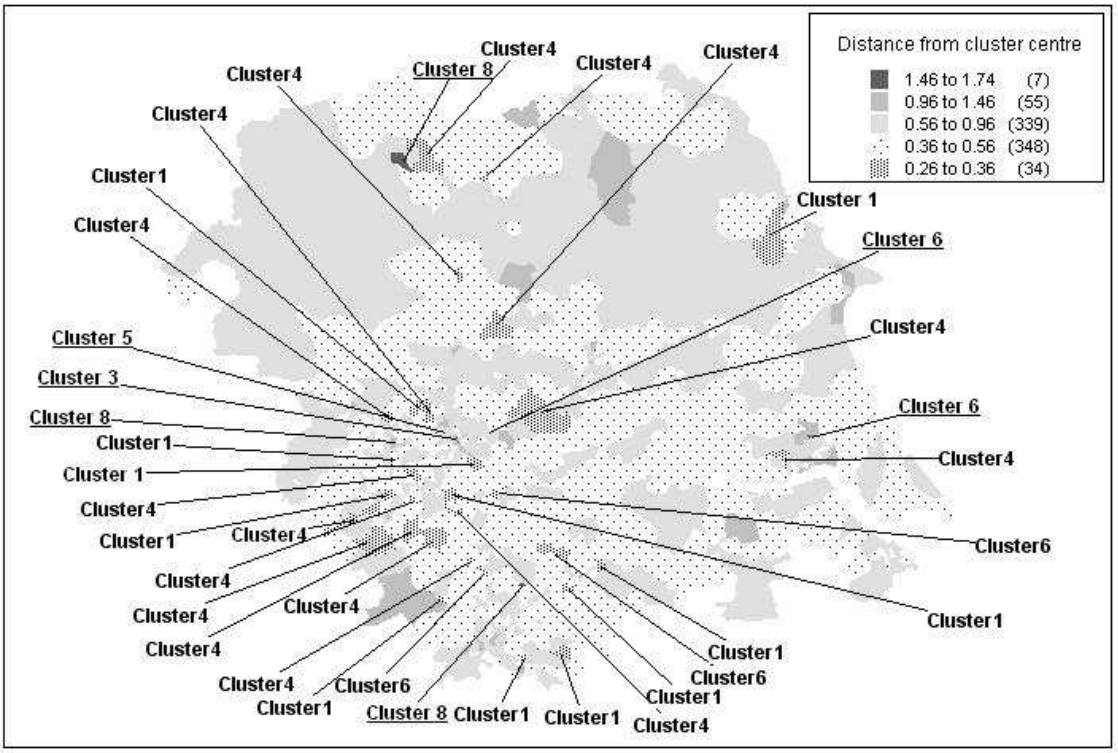


Figure 3.8 Mapping the distance from cluster centre by postal sector



3.2.2 Cluster specific distance values

Table 3.3 shows a set of basic statistics describing the distance values organised by cluster. Cluster 1 has the smallest range at just 0.592 and the lowest maximum value. The mean value of 0.504 is very low and the standard deviation of just 0.109 further suggests a particularly strong cluster. At the other extreme, Cluster 8 has by far the largest range (1.373) containing, as it does, the single furthest distance of a case from its cluster centre (1.733 – HD8 9). However, the range of values does not tell us the whole story. The highest average value belongs to Cluster 5 at 1.069, over 1 and a half times the average for the entire system. The standard deviation in this cluster is reasonably low, thus suggesting that Cluster 5 is made up of consistently larger distance values than the others. Furthermore it has the second lowest membership value with only 20 postal sectors. These zones each have high distance values suggesting that Cluster 5 may be made up of 'outlying' zones in the system and suggesting that it may be the weakest cluster.

Table 3.3 Descriptive statistics of distance values by cluster from the static demand classification (ranked by the mean)

Rank	Cluster	n	max	min	range	mean	st. dev	no. of "outliers"	no. of "extremes"
1	**1**	165	0.876	0.283	0.592	0.504	0.109	2	
2	**4**	201	1.172	0.267	0.905	0.555	0.172	9	1
3	**6**	112	1.536	0.322	1.214	0.594	0.199	3	3
4	**7**	122	1.398	0.364	1.034	0.638	0.186	4	-
5	**8**	103	1.733	0.360	1.373	0.718	0.266	5	1
6	**2**	45	1.427	0.422	1.005	0.840	0.299	-	-
7	**3**	15	1.731	0.469	1.261	0.869	0.292	-	1
8	**5**	20	1.470	0.701	0.769	1.069	0.236	-	-
Total		***783***	***1.733***	***0.267***	***1.466***	***0.620***	***0.229***	***23***	***6***

Such a suggestion is sensible if we consider the geographic location of the zones in Cluster 5. As they are all city centre locations it is likely that they will have quite small populations. With such a small denominator, the raw data is likely to be translated into some quite high percentages, and this will push these cases to the margins of the taxonomic space. The pen portraits of Cluster 5 suggest some confusion in this type of area. There is strong mix of household types and some very contrasting features, such as, on average, higher proportions of households in social class II but also high unemployment. However, this may not be so detrimental. In the pen portraits, these areas were highlighted as possibly being gentrified neighbourhoods. In such areas there is, by nature, an eclectic mix

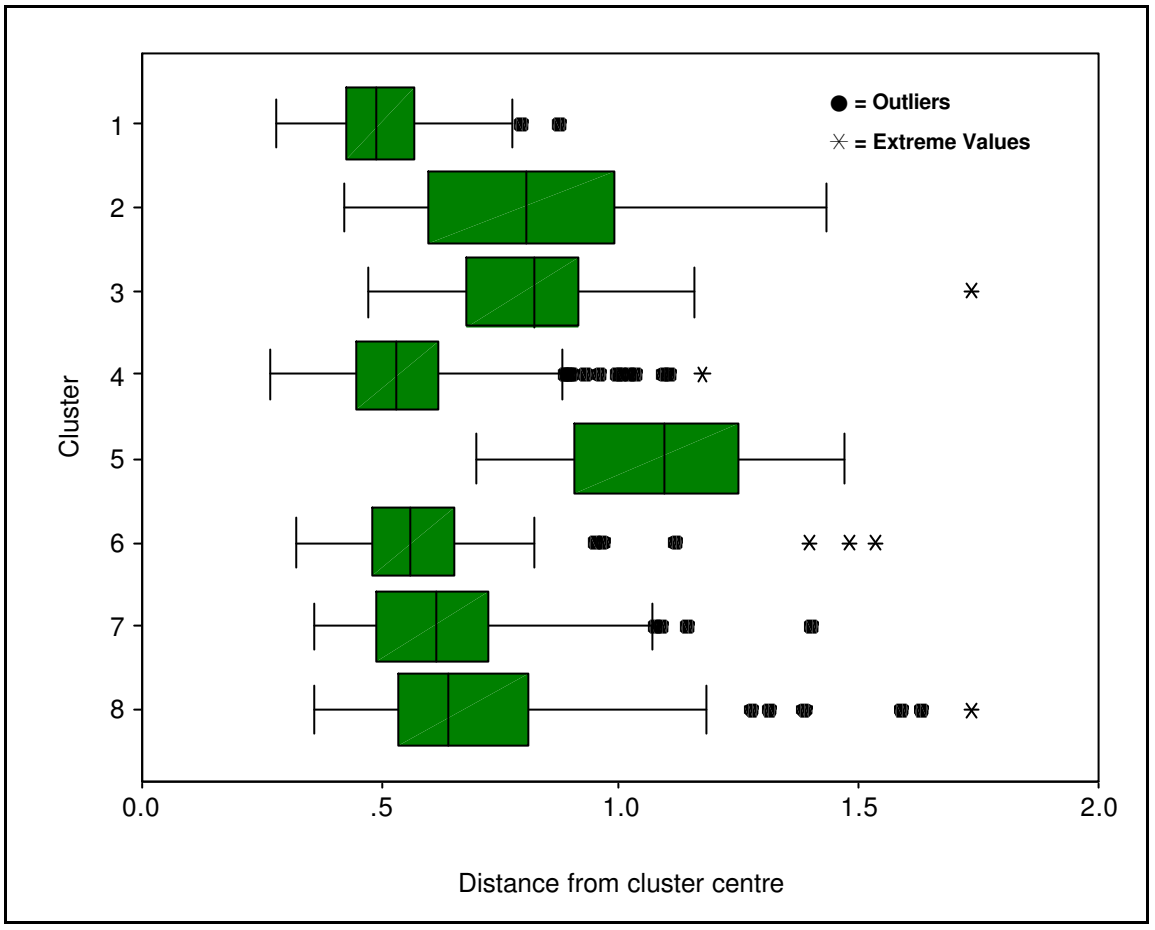
of populations as poorer areas are redeveloped into more up-market residential areas. Cluster 5 may be a 'weak' cluster for a good reason.

Figure 3.9 shows an alternative visualisation of this information in the form of box plots. Box plots allow us to view the distribution of distance values within each cluster and give us an alternative range of values by determining *outliers* and *extreme values* in the data. Outliers in the data are deemed to be any point that falls further than one and a half box lengths from the third quartile; extreme values are those that lie more than three box lengths from the third quartile[1]. Therefore we can look at the minimum to maximum ranges of the clusters without these outlying points. With the extraneous data removed we can see that Cluster 1 begins to look stronger as it is even more firmly located in the lower distance values. Cluster 8 also appears stronger than before, having a much lower range of values and appearing positively skewed. Cluster 3, which has the second highest range, is shown to have a much better distribution of values once its extreme value is ignored. This value (LS2 3) is the second highest distance value and appears to be a long way from the 'maximum' value imposed by the box plots. This suggests it is something of an outlier and may be significantly different to the other zones in Cluster 3. Cluster 6 is shown to have 6 cases that have significantly distorted its distribution in real terms. With these zones removed we can see that it now has a very compact range of values that tend heavily towards the lower end. Its maximum value is half what it was before and is not much greater than Cluster 1. Cluster 4 also fares well in the box plots once the outliers are ignored because it has the most outliers of all the clusters. Again this is intuitive if we consider the nature of the cluster. The cluster map shows a wide variety of area types falling into this cluster, both very rural and semi-urban zones.

With the erroneous values removed Cluster 2 now appears to be one of the weaker clusters. While it does not hold the largest (artificial) maximum value, with no outliers at all its range of values has remained the same and is now the largest (1.005). However, Cluster 5 remains the 'weakest' cluster.

[1] The third quartile is the right hand side of the box in a box plot. In such a graph the box represents the inter-quartile range

Figure 3.9 Box plots of distances from cluster centre by cluster



Nevertheless, we can say that this classification performs quite well. There seem to be no significant problems. The two highest distances (BD1 3 and LS2 3) are somewhat removed from the other values (they lie at 1.73 while the next highest is 1.63) but this does not seem to unduly affect the classification. It will be interesting to see how these cluster specific distance values change as the new variables are added. BD1 3 and LS2 3 will need to be monitored. If they continue to be 'outliers' in taxonomic space then some consideration may be needed as to what to do with them.

3.3 Summary

In general terms, the classification performs reasonably well. Despite one zone having to be removed from the system a meaningful, sensible and comprehensive eight cluster classification has been created for the postal sectors of Yorkshire and the Humber that

shows a similar pattern of clustering to that of GB MOSAIC, a commercial geodemographic system. The analysis of the goodness of fit of cases to clusters also suggests that the classification is robust. Although there are some outliers in the data, the overall impression is that the zones have been clustered optimally. The distance analysis also indicates a satisfactory level of performance. There are a few outliers but this is to be expected. Geodemographic classifications do not necessarily provide an effective categorisation of every zone; some zones will always be difficult to fit into a cluster and as generalisations, the cluster descriptions cannot be expected to totally describe every element of their constituent zones in detail, only the major characteristics they have in common.

4 EVALUATING VARIABLES WITHIN THE CLASSIFICATION SYSTEM

Knowing how well a classification performs is only part of the story. It is important to understand which variables within the information system drive the classification. Understanding this will allow us to make informed decisions when choosing variables for subsequent extended classifications.

When defining the cluster characteristics, the Z-scores tell us which variables are, on average, higher or lower than the global mean. However, there is no indication of which variables are the most important in the cluster formation. Some will be 'important' in that they will have a strong bearing on the location of each case in multivariate taxonomic space; others may simply be replicating the information (and therefore position) already provided or may be drawing a zone away from a cluster it might otherwise be more suited to.

This section will attempt to analyse the regional dataset used to create this classification and try to discover which are the 'key stock variables' and which are the 'weakest links'. To undertake this analysis three statistical procedures of differing levels of complexity are employed. The first makes use of the ANOVA table produced as an output when a *K-means* clustering algorithm is implemented in SPSS. The second looks at the ecological relationships between the variables by studying a correlation matrix. The final method

takes the pairwise correlation analysis further by identifying the major *factors* in the data and performing a principal component analysis on the dataset.

4.1 Analysis of ANOVA table from *K-means* classification

The *K-means* clustering in SPSS produces a one way ANOVA table as part of the output. When ANOVA tables are used for regression the *'F'* statistic is used to test the overall fit of a regression model to a set of observed data and is based upon the ratio of the improvement due to the model and the difference between the model and the observed data using the mean sum of squares (Field, 2000). When used in conjunction with the *K-means* algorithm the *F*-statistic describes the ratio of the between-cluster mean square and the within-cluster mean square and therefore provides information about each variable's contribution to the segmentation of the dataset (Phipps et al. 2001). It can therefore be used to identify the driver variables in cluster analysis (Phipps et al. 2001, Röder 2000, Remmel 2000).

Table 4.1 shows the 10 variables with the highest and lowest *F*-statistics in the ANOVA table. According to this measure, the two car ownership variables are the most important in the segmentation process. It is satisfying to see that the top 10 variables all contain significant indicators of affluence and life stage. It is intuitive that they should be strong 'drivers' in the cluster formation process. More interesting, however, are the proportions of the population aged 75 to 84 and over 85 appearing to be the two least important variables with similar *F*-statistics. Four age variables appear in the list of 10 least 'effective' variables. The cluster descriptions in Section 2 show that the age variables often appear as defining features of the clusters. This underlines the earlier assertion that the Z-scores used to create cluster descriptions are not adequate for defining important variables in the segmentation.

The presence of the retired population variable in the set of least effective variables is also interesting as one would expect this, together with the numbers of pensioner migrants, to be a significant indicator of life stage. The presence of the other four variables, households with two or more families in private or council accommodation, number of bedsits and

number of households lacking bath or shower may be highlighted here due to there only being small percentages of each in any postal sector.

Table 4.1 Top 10 most and least 'effective' variables according to ANOVA table

Top 10		Bottom 10	
Variable	***F*-statistic**	**Variable**	***F*-statistic**
2+ cars	641.75	85 plus	5.24
No car	535.96	75-84	5.30
Detached	376.14	Pensioner Migrants	7.22
Local Authority rent	373.54	2+ families & Owner Occupied	11.32
Married Population	368.81	Bedsits	16.60
Single Population	356.45	2+ families & Local Authority rent	16.62
7+ rooms	337.61	Retired Pop	16.95
No family & Local Authority rent	265.43	65-74	17.01
Flats	259.07	25-44	18.14
Households in Social Class II	256.90	Households lacking bath or shower	19.91

As a tool for selecting the strongest variables the ANOVA table has proven useful. The variables used most frequently as proxy indicators for affluence and life stage have been revealed as strong drivers in the segmentation. However, as a selection process for the removal of variables we must be more cautious. According to the ANOVA table, we could justifiably remove the age variables from the classification. However, this would appear to be a rather rash step since, without the age variables, it would be almost impossible to identify the characteristics of certain clusters in a meaningful way. Although certain variables have little effect in cluster formation, they may be useful nevertheless in describing key characteristics.

4.2 Correlation analysis for static demand classification

Correlation analysis allows us to explore the ecological relationships between the 51 Census (sic) variables selected for this classification. We can assess the relative importance of certain variables by looking at their correlations with all the others. If the correlation between two variables is sufficiently high enough we could decide to focus on just one variable to describe an area. The variables that have the most predictive capabilities, i.e. those that are significantly correlated with the most variables could be described as the most important in the classification process because they will help explain the variance in other variables. However, Voas and Williamson (2001) suggest that using correlation analysis as a rationale to remove or retain variables presents two opposing tendencies.

> *"On the one hand attributes that are correlated with others will have predictive (and hence descriptive) power. We are inclined to retain them since they provide information about a range of qualities. On the other hand we may want to drop at least some of them, precisely because they are partially predictable on the basis of others. By contrast we will have no information about relatively uncorrelated variables unless we include them in [the] set"* (p. 64).

From Voas and Williamson's argument it is difficult to tell whether multicollinearity in the variable selection is undesirable or even avoidable. It is true that we do not want two variables describing the exact same phenomena. If two variables have an association of ± 1, then one should be removed because it would essentially be giving one aspect of an area's characteristics extra weighting. However, even though a pairwise correlation might be very high, as long as it is not perfect then there is an argument for retaining both since the outliers in each association will add nicely to the pattern of points in multivariate taxonomic space and might draw a point towards a cluster that it might otherwise not have been placed in if there was not that extra dimension.

Table 4.2 shows that '2+ cars' is the variable most highly associated with others. It has correlation coefficients greater than ± 0.5 with 23 of the 50 other variables. 'Renting from local authority', 'households in Social Class II' and 'Single population' are also significantly associated with large numbers of other variables. In total, 47 variables have one or more association greater than ± 0.5 and, as we can see from Table 4.2, 23 have such associations with more than 10. The largest correlation between any pair of variables is between the married and single populations with a value of -0.991. Voas and Williamson (2001) explain that when two variables share the same denominator, the aggregate is capped at zero (a situation known as *closure*) and therefore there is an automatic tendency towards negative correlation. *"In some cases the effect of closure may be difficult to disentangle from a genuine negative association"* (p. 64).

This is a good example of the 'opposing tendencies' described above. Where we find such a high correlation we might feel justified in removing one of the variables. In areas where a high proportion of the population is married we know by default that there will be a low number of singles there, and vice versa. This, at least, is Voas and Williamson's argument

for rationalising the number of variables. However, by removing one of the variables we are depriving ourselves of a dimension in taxonomic space and the chance of finding interesting outliers in the data that may characterise certain clusters. A number of the associations in the correlation matrix occur because of closure. The correlation between 'no car' and '2+ cars' is –0.834.

Table 4.2 Variables with correlation coefficients greater than ± 0.5 with 10 or more other variables

Variable	Number of correlations	Variable	Number of correlations
2+ Cars	23	No family household and council rent	16
Rented from HA, LA or New Town	22	Persons with LLTI	16
Households in Social Class II	20	Population aged 45-64	15
Single Population	19	Lone Parents	15
Detached	19	Self-employed	15
No Car	19	Unemployed	15
Married and cohab couple & OO/Rent	19	Mortgage owners	14
Married Population	18	Married and cohab couple & council rent	13
Home owners	18	Married and cohab couple w/ children & owner occupied/ privately rented	13
Households > 7 Rooms	18	Flats	12
Workers with other qualifications	18	Married and cohab couple w/ children & council rent	10

Some high correlations also occur because variables belong to the same question on the Census form. Thus we find a correlation of 0.801 between the married population and the married population living in owner occupied or privately rented premises, and a correlation of –7.86 between the latter and the proportion of households renting from a local authority.

The more interesting correlations are those that do not necessarily have a technical explanation. There is a correlation of –0.881 between the married population and households with no car. Such ecological relationships are not simple to explain. Again, we would be unwise to lose one of the variables on the criteria that it is predicted by another. By retaining both we stand the chance that the clustering algorithm will pick out areas because they are different for this characteristic, i.e. have high marital status and a high proportion of households with no cars.

One interesting finding is the high number of correlations experienced by the unemployment variable and the age variables. Although only one age variable is found in

Table 4.2, they all have at least one correlation with other variables. The unemployment and age structure variables are taken from different datasets and, crucially, at different times to the census data. However, this correlation analysis has shown that the patterns of unemployment in 1999 are still strongly correlated with the socio-demographic patterns recorded in the census eight years earlier. This helps to justify the use of data from different time periods. A further important observation is that the proportion of the population aged 64-74 is correlated with 17 other variables. This variable was highlighted in the analysis of the ANOVA table as being of less significance to the segmentation. Yet its ecological relationship with so many other variables corroborates the point made earlier that the ANOVA table is less useful for selecting variables to remove from the information system.

Only four variables do not have any correlations with an absolute value greater than 0.5; Proportion of bedsits, households lacking bath or shower, households containing 2 or more families that are rented from the local authority and households in Social Class III(N) (Skilled non-manual). It is impossible to tell how important these variables are to the classification process just by looking at the pairwise correlations. As Voas and Williamson suggest, we might want to retain these variables these variables because we have no information on their pattern from any of the other variables. However, some of these variables were also identified in the ANOVA table (Table 4.1); households in social class III(N) was the 11th least effective variable.

Thus, while some variables show significant associations with several others, some are not strongly related to any. This discussion has shown that, while we can highlight some examples of multicollinearity within the variable set, we cannot necessarily suggest which variables are acting strongly. A more comprehensive multivariate analysis is required than just pairwise correlations.

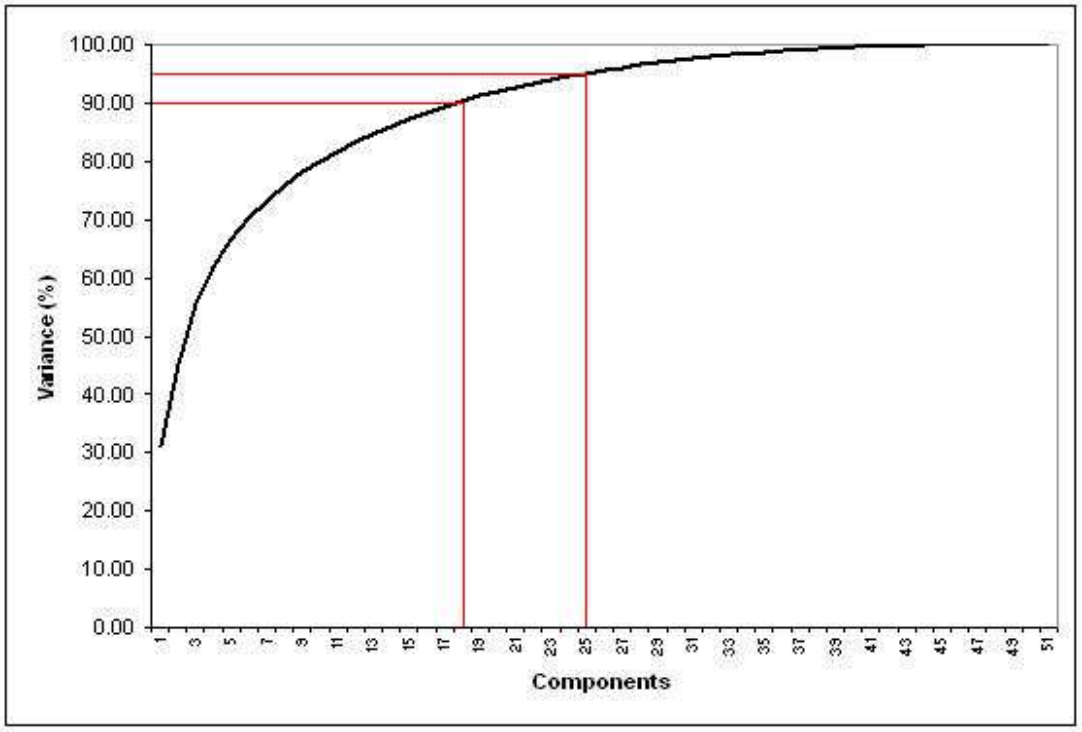
4.3 Assessing the importance of variables using Principal Component Analysis

We can arrange the pairwise correlations used in the previous section in an *R-matrix*, a table of correlation coefficients between variables. The concentration of large correlation

coefficients between subsets of variables suggests that those variables could be measuring aspects of the same underlying dimension; such dimensions are known as *factors*. Principal components analysis (PCA) allows us to reduce a data set into a smaller set of uncorrelated factors and explain the maximum amount of common variance in the correlation matrix using the smallest number of explanatory concepts (Field, 2000). Essentially, therefore, PCA will reveal which combination variables explain the most variation in the data set, thus informing of us of the most and least important variables in the cluster formation.

As with the *K-means* clustering algorithm, each variable functions as an axis and every postal sector may be represented as a point in multi-dimensional space. However, PCA allows us to investigate whether instead of needing 51 axes, a much smaller number will be adequate to locate the point in variate space (Voas & Williamson 2001). In PCA, each *component* represents a weighted combination of the original variables although Voas and Williamson (2001) argue that a potential drawback of this approach is that nothing obliges these components to be meaningful in substantive terms.

Figure 4.1 Percentage of variance accounted for by components in the static demand classification



By plotting the cumulative amount of variation explained by the different components we can see that there are diminishing returns from adding further components. Figure 4.1 shows that 90% of the variation in the data can be explained by the first 18 components while 25 are needed to explain 95%. The data for this classification was arranged in a matrix of 783 × 51 cells[2] (39,933). The graph in Figure 4.1 suggests that we can achieve 95% of the variation in the data set with just half of those cells. Furthermore, the first four components account for 62% of the variance in the data.

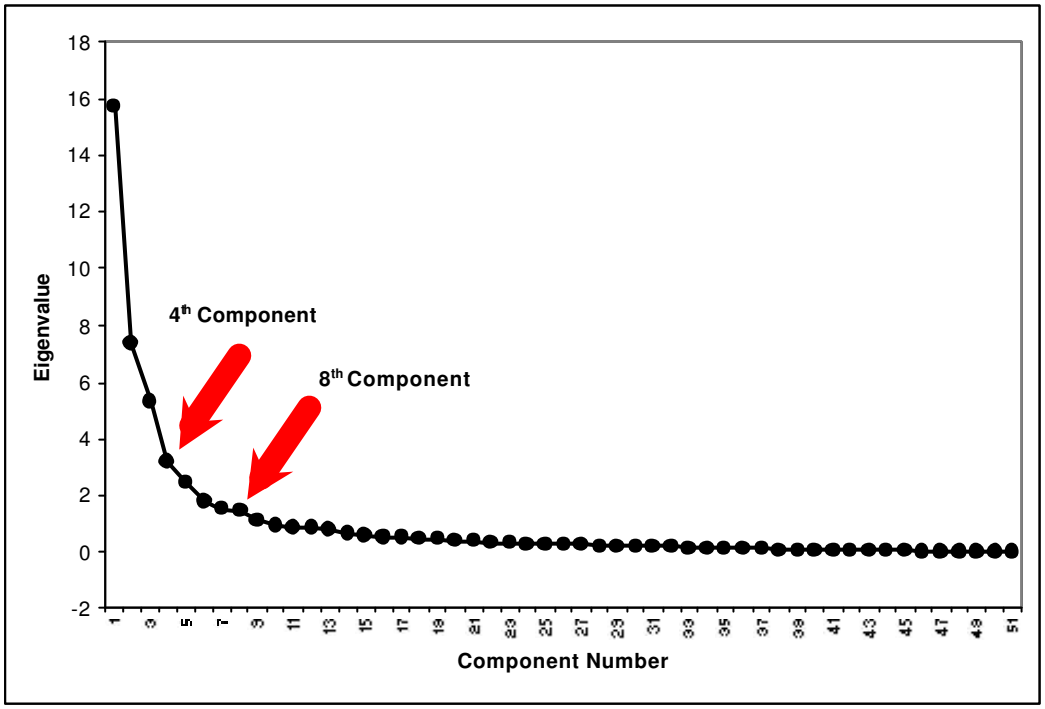
By determining the most important factors we can make assumptions about the importance of the variables within those components to the whole classification system. However, in the same way that there is no robust method of choosing the right number of clusters in a classification, there is no way of determining how many components are important. When PCA is used for data reduction it becomes necessary to *extract* the most important components. Although at this stage we do not want to remove variables, it is still a useful exercise as it will identify the components that contain the key *cluster formative* variables. The eigenvalues that are calculated indicate the substantive importance of each component. Typically there would be few factors with high eigenvalues and many factors with relatively low scores. It would make sense, therefore, to retain the factors with the highest eigenvalues. The relative importance of the factors can be seen by plotting a graph of each eigenvalue (y-axis) against the relevant component (x-axis). This is known as a *scree plot.* According to Field (2000), Cattell (1966) argues that the cut-off point for selecting factors should be at the point of inflection of the curve. The curve in Figure 4.2 is a little difficult to interpret because it begins to tail off after four components, but there is also a small increase at the eighth component before a more stable descent is achieved. It would be reasonable to consider both options.

Field (2001), however, suggests that factor selection should not just be based on this criterion. Kaiser (1960) recommends just looking at the eigenvalues and retaining all factors with a value greater than one. As we can see from Figure 4.2, these two approaches give different results. If we were to select factors with eigenvalues greater than one we

[2] 783 because we have removed LS1 8 from the classification because of its extreme values due to a paucity of population.

would want to take the first nine components, not the fourth or eighth. Furthermore, Jolliffe (1972) argues that Kaiser's criterion is too strict and therefore suggests retaining the factors with eigenvalues greater than 0.7. Adopting this approach would leave us with the first 13 components; cumulatively this would explain 85% of variance.

Figure 4.2 Scree plot of eigenvalues for components in the static demand classification system



However, we do not have to limit our criterion for selection to eigenvalues. In order to find out which variables make up the components we need to look at the *factor loading* of each variable to its factor. If we again visualise our points in multivariate space and accept that each axis represents a factor, then the constituent variables can be plotted according to the extent to which they relate to a factor. The factor loading is therefore the Pearson correlation between a factor and a variable. From this factor loading we can find the most important variables in each component. Voas and Williamson (2001) selected variables that had a factor loading of greater than ± 0.45. We can also use this criterion to help us find the most important components. If ± 0.45 is the cut-off point for significance to a factor, then we can say that a factor with fewer variables with these loadings is less likely to be important to the classification as a whole. Figure 4.3 shows the number of variables greater than ± 0.45 in the first nine components. We can see a very sharp decline from 27

in the first component to seven in the third. The fourth component has just three variables with factor loadings greater than ± 0.45, the eighth and ninth have none. Therefore we can argue that we can remove the fifth and subsequent components after from our analysis.

Figure 4.3 Number of 'significant' variables (> ± 0.45) in the first nine components of the static demand classification

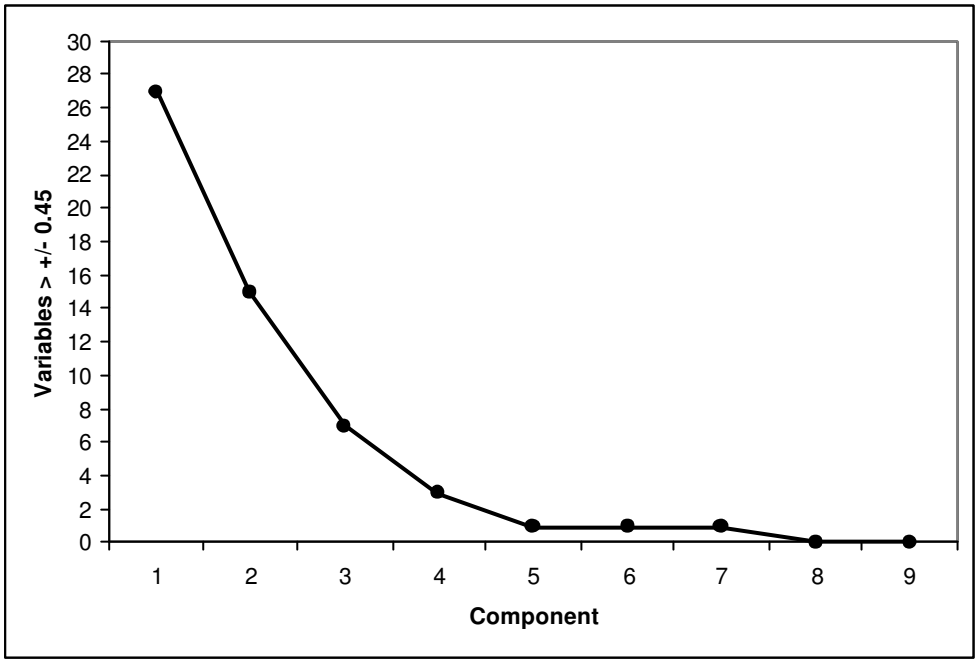


Table 4.3 shows the four principal components selected according to the above criteria. PCA has been criticised earlier for not requiring the components to be meaningful in substantive terms. However, here we can see that the first three components in Table 4.3 seem to be relatively easy to interpret. The first component appears to measure an underlying dimension of socio-economic disadvantage contrasted with prosperity; hence the importance of such variables as unemployment on the *positive* side and households with 2 or more cars on the *negative*-side. The second component seems to compare students and young professionals with more established working class households. The third measures younger families and older populations. The fourth is a little less straightforward as there appears to be less of a distinct characteristic to the variables that are most significant to that factor and there are no variables with factor loadings under –0.45.

Table 4.3 Principal components for static demand classification (variables with factor loadings > ± 0.45)

Positive		Negative	
Component 1			
No Car	0.886	2+ Cars	-0.920
Rented for HA, LA or New Town	0.858	Detached	-0.829
Single Population	0.823	Married & cohab couple & OO/Rent	-0.823
No family household & council rent	0.777	H'holds in Social Class II	-0.810
Unemployment	0.754	Total Married population	-0.800
LLTI	0.725	H'holds > 7 rooms	-0.790
Lone Parents	0.686	Home owners	-0.774
married and cohab couple & council rent	0.676	Self-employed	-0.729
H'holds in Social Class V	0.647	Mortgage owners	-0.716
H'holds in Social Class IV	0.628	Workers with other qualifications	-0.700
married & cohab couple w/ children & council rent	0.611	married and cohab couple w/children & OO/rent	-0.673
Flats	0.565	45-64 est	-0.647
Terraced	*0.496*	H'holds in Social Class II	-0.575
		Economically active residents aged 16+	*-0.456*
Component 2			
Privately rented	0.772	Households with dependents	-0.697
Movers last year	0.703	Semi-detached	-0.660
No family household & OO/Rent	0.674	H'holds in Social Class III(M)	-0.635
Students	0.624	5-14 est	-0.635
Flats	0.572	*Total married population*	*-0.469*
workers with higher degrees	0.563	*0-4 est*	*-0.456*
15-24 est	0.547		
workers with other qualifications	0.507		
Single population	*0.470*		
Component 3			
25-44 est	0.608	65-74 est	-0.841
0-4 est	0.521	75-84 est	-0.750
Economically active residents aged 16+	*0.491*	Retired population	-0.649
		85+ est	-0.588
Component 4			
Terraced	*0.476*		

Note: Variables in *Italics* had coefficients $< \pm 0.5$

The principal component analysis suggests that the affluence and socio-economic disadvantage variables in the data set (component 1) are the most important in the classification, explaining 31% of the variance in the data. Within this component the two most important variables are those relating to car ownership. 'No car' has a factor loading of 0.886 while '2+ cars' has a score of –0.92. This is intuitive as car ownership is strongly related to affluence and deprivation.

Field (2000) argues that, in an ideal world, a variable should have a large factor loading for one axis and low scores for any other factors, indicating that the variable was only related to one factor. For the main part we can see this situation occurring here although there are a handful of variables that are duplicated across the three components. These include; total married population, flats, workers with other qualifications, residents aged 0-4 and economically active residents aged 16+. This does not mean that these variables are any less important to the classification, but simply that they can explain more than one underlying dimension in the data. However, it could be argued that those variables that are not included in the four principal components at all might be less significant.

A total of 44 variables appear in the first four components, leaving seven that are not included. Table 4.4 lists these latter variables and details the components where they do have significant factor loadings. Using the criteria of greater than ±0.45 we can see that three variables are not considered to be significant to any of the components in the PCA; households lacking bath and shower, households with more than 1.5 persons per room and households with two or more families and are owner occupied or privately rented (variables VAR26, VAR29 and VAR37, respectively). Once again it is interesting to note that VAR26 and VAR37 were highlighted in both the ANOVA and pairwise correlation analyses as potentially weak variables.

One of the reasons for this may be a 'small number problem'. The number of households falling into these three categories is likely to be small. Table 4.5 shows some descriptive statistics for these variables and suggests that these variables are poorly represented across the region. The maximum number of households lacking a bath or shower is just 2.3%, the average only 0.2%. Similarly, the average number of households with more than 1.5 persons per room in any postal sector is just 0.4% and although there is one postal sector (BD1 3, central Bradford) with 14.4% of households in VAR37, the average across the region is just 0.7%. It is clear, therefore, why these variables do not account for any variation in the data set.

There is a case for removing these variables from the classification system. Their ability to monitor socio-economic disadvantage across postal sectors is limited. When mapped at the enumeration district level these variables might indeed reveal pockets of poor living

standards or affluence. However, these extremes will be lost when aggregated up to the postal sector level.

Table 4.4 Variables not included in the first four principal components of the static demand classification

Variable	Components where variable loading is > ± 0.45
Pensioner Migrants	10
Bedsits	11
Lacking bath and shower	None
Households > 1.5 persons per room	None
Households with 2 or more families & owner occupied/private rental	None
Households with 2 or more families & council rental	11, 13
Households in Social Class III(N)	5

Table 4.5 Descriptive statistics of households lacking a bath or shower, households with > 1.5 ppr and households with 2 or more families and owner occupied or privately rented

	Households lacking bath or shower	Households > 1.5 persons per room	Households with 2 or more families & owner occupied or privately rented
Variable	VAR26	VAR29	VAR37
Max	2.31481	7.83133	14.3713
Min	0	0	0
Avge	0.215305	0.392433	0.680604
St Dev	0.272538	0.783588	1.023978

There may also be some doubts as to how well these variables actually measure deprivation or affluence. There is no explicit link between the number of households with 2 or more families in an owner occupied or privately rented house and socio-economic (dis)advantage and no studies to prove its worth. However, as it is part of a set of variables relating to household structure and tenure it may be wise to retain it for the sake of completion. Census variables relating to households lacking basic amenities and overcrowding are more commonly used in deprivation studies. For instance, the Jarman underprivileged area index, the Townsend deprivation index and the Carstairs index of deprivation all use a measure of overcrowding although in both cases 'overcrowding' is defined as there being more than one person per room, not 1.5 as used here (MIMAS 1999). Indeed, as overcrowding is used in so many studies of deprivation there is a strong case for its

retention here, although it may be wise to change the variable so that it reflects households with more than one person per room.

Census data describing the lack of basic amenities such as a bath or shower are less frequently used in deprivation studies. There is no variable of this sort in the Carstairs, Jarman or Townsend indices. However, this is not to say that it has never been used; the DETR 1998 Index of Local Deprivation uses this indicator (DETR 1998) and it has been used in studies of housing stress (Simpson 1993). However, it is clear that the success of an index is dependent upon the relative importance of the variables to deprivation and these may change over time (Martin, Senior & Williams). Whereas households lacking a bath or shower may have been a useful measure of deprivation 20 or 30 years ago, it is unlikely to have much resonance now as very few households will fall into this category. The provision of basic amenities is one of the main focuses of housing regeneration schemes and it is likely that, by 1991, there would have been very few households with such a paucity of provision. Indeed, Simpson (1993) suggests that, when mapped at the smaller spatial scales, this variable tends to identify dwellings made up of flats or bedsits that share bathroom or toilet facilities and thus would hardly describe high levels of deprivation.

4.4 Summary

The evaluation of the variables used to make this classification has highlighted a number of interesting issues. The stronger variables that are suggested to be most important in cluster formation are recognised in all three analyses as car ownership, house size and type and family status. This suggests that the static demand classification is capable of recognising particular socio-economic patterns in the region and has incorporated them into the cluster types.

Furthermore, despite the suggestion in the ANOVA table that certain age variables do not segment well, the same variables that perform poorly are also consistently identified. In the principal components analysis three variables were finally identified as being the least effective. As discussed before, the variable recording households with 2 or more families in private housing should be retained for completion as the other variables in this family are all important to the segmentation. The proportion of households with more than 1.5 persons per room should probably be changed to more than 1 person per room to bring it in

line with data commonly used in accepted deprivation indices. The only variable that could be dropped is the households lacking bath or shower. However, it will be interesting to see if this variable continues to perform badly when others variables are added in subsequent classification systems. Therefore the variable will be retained although a close monitoring of the ANOVA, correlation and PCA analyses as the classifications are developed will reveal whether, in the end, the variable can be dropped from the final system.

Overall we can say that the static demand classification is a robust and sensible classification that will act as a useful springboard from which to extend the concept of geodemographics (see Debenham et al. 2001a, 2001b). The following sections will show how the classification is affected by the addition of a suite of variables relating to the housing market.

5 ADDING PROPERTY TRANSACTION VARIABLES TO CREATE A NEW CLASSIFICATION

Much has been written about the development of geodemographic classification using certain non-census variables such as county court judgements (CCJs), credit ratings and the electoral register (Experían 2001, Sleight 1997, Birkin 1995) however there is very little literature describing how this data actually affects the basic census data classification. As discussed before, geodemographic companies are rarely willing to release details of their systems, let alone admit whether or not the variables that they are adding have any positive impact upon the taxonomies they create. The purpose of the subsequent sections is to use the evaluation analysis performed on the static demand classification upon a classification created with a new suite of variables.

The variables chosen for use here are derived from property transaction data available in the Experían Ltd. Postal Sector Data (Experían 2001). Postal sector level data is available from the beginning of 1995 to Quarter 2 of 2000 on the number of houses sold per quarter and the average price of such transactions, as documented by HM Land Registry. The data is disaggregated by housing type and Table 6.1 shows the 10 variables that have been derived from this data. These variables may be considered as representing the housing market since they reflect both demand and supply. House prices are a clear indication of the

buoyancy of an area yet they can also provide a reasonably good measure of affluence, hence their inclusion as a set of demand variables.

Table 5.1 Suite of property market variables

No.	Variable
var52	Total number of transactions (Quarter 3 1999 to Quarter 2 2000)
var53	Proportion detached sales (Q3 1999 - Q2 2000)
var54	Proportion semi sales (Q3 1999 - Q2 2000)
var55	Proportion Flats sales (Q3 1999 - Q2 2000)
var56	Proportion terraced sales (Q3 1999 - Q2 2000)
var57	Average value all transactions (Q399 to Q200)
var58	Average value detached (Q399 to Q200)
var59	Average value semi (Q399 to Q200)
var60	Average value flats (Q399 to Q200)
var61	Average value terraced (Q399 to Q200)

It is surprising that this data is not used more in mainstream GDIS, although spatial variability in the data may preclude this. Some areas will see many transactions while others may have very few, thus affecting the averages. Furthermore there is no data for Scotland.

Table 5.2 Cluster memberships for the 7,8 and 9-cluster solutions

9-Clusters Number of Cases in each Cluster			8-Clusters Number of Cases in each Cluster			7-Clusters Number of Cases in each Cluster		
Cluster	*1*	116	**Cluster**	*1*	150	**Cluster**	*1*	112
	2	9		*2*	43		*2*	27
	3	91		*3*	98		*3*	203
	4	128		*4*	1		*4*	30
	5	106		*5*	43		*5*	100
	6	172		*6*	270		*6*	108
	7	22		*7*	161		*7*	203
	8	42		*8*	17		-	-
	9	97		-	-		-	-
Total		783	**Total**		783	**Total**		783

The 10 property market variables were added to the Census variables. The only change to the original dataset was that, as recommended in Section 4.4, the overcrowding variable was changed from households with more than 1.5 persons per room to households with more than one ppr. The variables were classified using the same *K-means* algorithm as before and all values of *K* were tested from 100 to 2. Once again LS1 8 was removed because it was found to be a particularly difficult case to classify and was invariably placed

on its own until forced into a cluster by a low *K* value. However, as Table 6.2 shows, *K*=8 still resulted in a poor segmentation with a single zone cluster while the 7-cluster solution had too high an average distance component and a rather homogenous segmentation. Therefore the optimum number of clusters was taken to be nine.

5.1 Cluster pen portraits for a static demand and property data classification

Figure 5.1 displays the nine clusters created by the static demand and property data classification and the cluster pen portraits are described below.

5.1.1 Cluster 1 *116 Zones*

Zones in cluster one are largely suburban or ex-urban and are predominantly found in West Yorkshire. The population in these areas appears to be made up of married or cohabiting couples in privately rented or owner occupied housing. Although there is no predominant housing type there is a larger than average number of households with more than 7 rooms and there tends to be fewer flats. The population appears reasonably affluent with a higher number of households in social class II and with two or more cars. Self-employment and mortgage ownership levels are also higher here. There is a mix of age groups in these zones although the younger age groups are less well represented. This suggests more middle-aged households with older children. There tends to be less unemployment, LLTI and local authority renting in these zones.

The housing market is very buoyant with an average of 106 sales per zone. Because of the mixed housing stock, each type of dwelling is well represented in the market with higher average prices. The average price for all transactions is £72,000. Detached houses tend to sell for over £100,000 and Semi-detached houses for £66,000. The average price of a terraced house in this cluster is 150% of the regional average with some zones seeing values in excess of 200%.

5.1.2 Cluster 2 *9 Zones*

This cluster has the smallest membership and is only found in the very centre of the major urban areas (Leeds, Sheffield, Hull and York). The housing stock is dominated by flats and

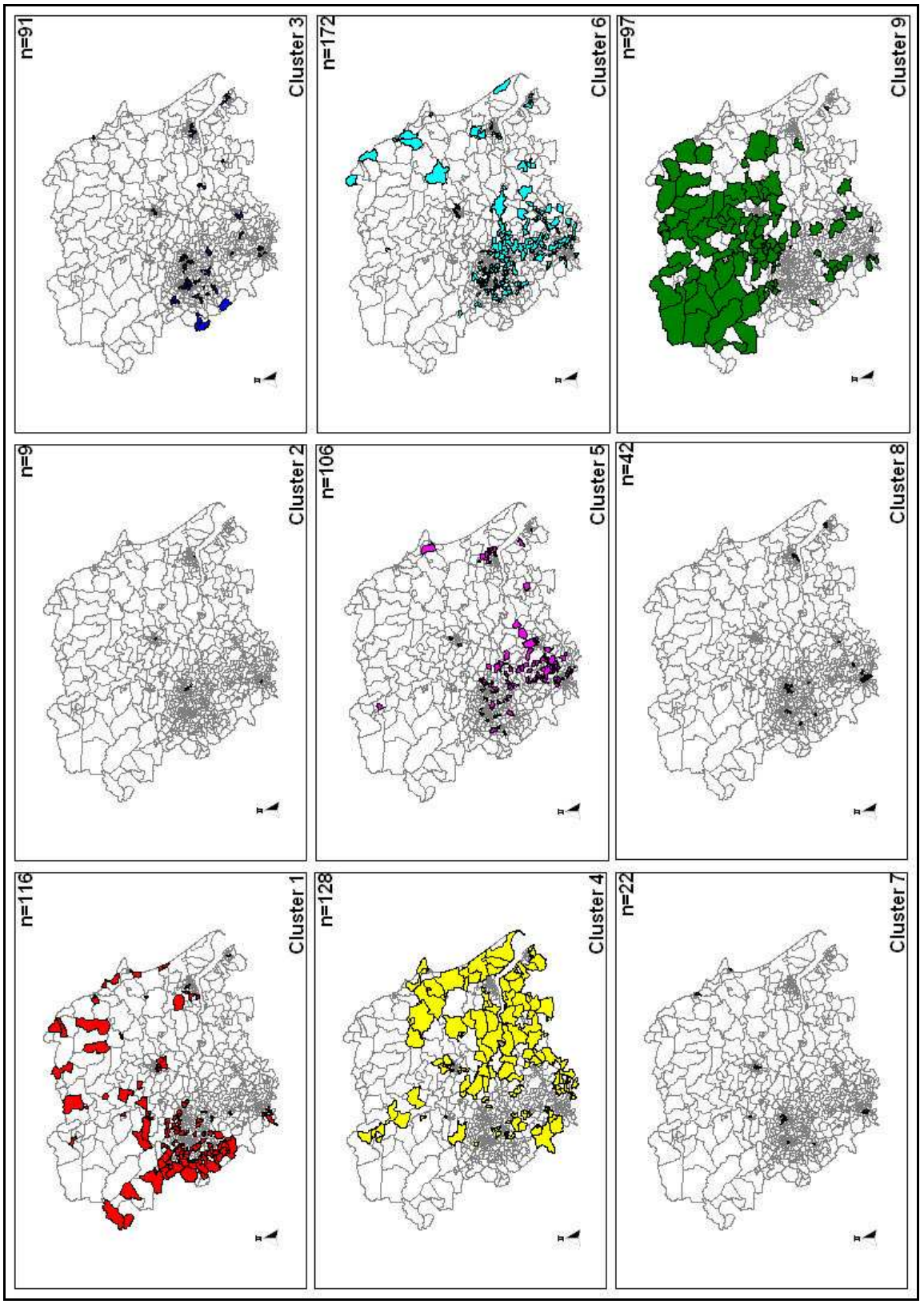
the housing market reflects this pattern. On average 88% of all transactions are flats and in seven of the nine zones they account for 100%. The population here tends to be a mixture of younger single adults and those over 75. Mobility is high with a very high number of movers last year and pensioner migrants. The total number of sales is low but it is more likely that people live in rented accommodation here as home and mortgage ownership is noticeably low. There tends to be a mix of private and council rented property, possibly suggesting some level of gentrification. Car ownership is low but as these are central locations it may not be a sign of deprivation. The average transaction price of a flat in these zones is over £80,000, emphasising the likelihood that these gentrifiers are young professionals.

5.1.3 Cluster 3 *91 Zones*

Postal sectors in this cluster are predominantly found in the inner city of the major urban areas and the sectors that make up the smaller towns in the region such as Barnsley, Mexborough and Doncaster in the south, Goole, Scunthorpe and Grimsby in the east and Kieghley, Todmorden and Hebden Bridge to the west. A large proportion of the housing stock is terraced and often overcrowding can be a problem (especially in the metropolitan inner city postal sectors). Many of these houses lack basic amenities, especially central heating. A high proportion of households are in social class IV or V and unemployment is high. LLTI is not much more than the national average and this may be due to the predominance of younger families amongst the population; there tends to be less elderly. There are higher proportions of lone parents in these postal sectors and car ownership is low. Mortgage ownership is low but houses tend to be privately rented rather than council rented. Mobility tends to be a little higher than average due to the number of younger, less established families.

Although the number of sales per postal sector is equivalent to the regional average, the housing market appears quite depressed. There are very few detached or semi-detached houses and therefore few sales of such types; where sales are made the average value is low (around £40,000 for a semi-detached house). The average value of all transactions is just £33,000 (nearly half the regional average) and terraced housing can sell for under 30,000.

Figure 5.1 9- Cluster solution of the static demand and property data classification



5.1.4 Cluster 4 *128 Zones*

The postal sectors that fall into Cluster 4 tend to be found in rural areas with some suburban zones around Lees and Sheffield. The majority are found in the south or east of the region, with the exception of a belt of zones running north along the route of the A1 through Thirsk, Northallerton, Catterick and Richmond.

The population in these zones is made up of late middle-aged couples, more than likely with grown up children as the younger age groups are less well represented. Car ownership is high and unemployment low. The population is still economically active with self-employment common and a low proportion of retired persons. LLTI, lone parents and local authority renting are low. Households tend to be in Social Class II although group III(M) is also common, suggesting mixed levels of income. Houses tend to be large detached or semi-detached dwellings that are either mortgaged or owned outright.

However, despite this apparent affluence the housing market is relatively flat. The number of sales in the year is lower than the regional level and the mean price of all transactions, at £65,000 is not much more than the regional average. Detached houses sell for just under £8,000 more than the regional average. The average price of a semi-detached house is £51,500, only £1,500 more than the regional level.

5.1.5 Cluster 5 *106 Zones*

The zones in Cluster 5 are predominantly found in inner city locations or in the smaller urban areas of South Yorkshire. Most households here rent from the local authority; mortgage or home ownership are well below the regional average. Lone parents, LLTI and unemployment are all much higher here than the regional average. Car ownership is much lower. The majority of households are in Social Class III(M) and there are higher than average numbers of households in classes IV and V. Very few households fall into classes I or II. The age structure and percentage of households with dependants suggests the presence of young families. There is a higher than average retired population but few very elderly.

The population is not particularly mobile and because of the proliferation of local authority renting the number of sales are low. Most dwellings are terraced or semi-detached and this is reflected in the transactions that are made. The average transaction price is low at just

£41,000 (just over two thirds of the regional average). Semi-detached houses sell for a little over £50,000 but the average transaction price of terraced housing is just £29,000.

5.1.6 Cluster 6 *172 Zones*

Cluster 6 zones are in similar locations to those in Cluster 5. The characteristics are quite similar too although this group of postal sectors appear to be more affluent than Cluster 5. Most households are in Social Class III(M) but there are also a higher number of households in III(N) and groups IV and V are much less prevalent. Furthermore, there is much less local authority renting and more mortgage ownership. This mix of tenure indicates that some households in these postal sectors might be former council tenants who have purchased their houses. There is less terraced housing in Cluster 6, most of the housing stock in semi-detached, suggesting large inner city housing estates. Economic activity is higher here than in Cluster 5 and unemployment is lower. Car ownership is a little higher, there are less households with no car than the regional average but there are also very few with two or more cars. Families tend to be a little more established with a smaller single population and slightly fewer infants.

The housing market is a little more vibrant. The population shows less of a propensity to move but the number of sales is more than the regional average; this could be seen as further evidence to support the suggestion of council house purchasing. The average transaction value of a semi-detached house is the same as the rest of the region (£49,000) although the average value of all sales is a little less.

5.1.7 Cluster 7 *22 Zones*

Postal sectors in Cluster 7 are only found towards the centre of the main urban areas of the region and are dominated by students. The population is characterised by young single adults in privately rented accommodation. Most households are "no family" households which suggests students sharing rented housing. There is also a high proportion of the population with higher degrees, suggesting that recent graduates also live in these areas. The higher number of households in Social Class I might suggest that these are young professionals. The population is prone to moving as students tend to change houses at the end of each academic year.

Most housing is either terraced or flats. Many dwellings lack amenities, particularly central heating, and overcrowding is common. Despite the predominance of privately rented dwellings, the number of sales in these areas is also high and the housing market buoyant. This may be due to properties changing hands among landlords, families moving out as the students take over and the increasing propensity for students or their parents to buy their own houses (Slade, 2001; McGhie, 1999). Because of this boom in demand for student housing, prices for terraced houses and flats are higher here than elsewhere in the region. Flats generally sell for around £50,000 while terraced housing prices are, on average, some £67,000.

It is likely that the non-student population in these areas, apart from the young professionals, will be less affluent households who have not been able to move elsewhere to avoid the student in-flux. Quite often the population might be quite elderly.

5.1.8 Cluster 8 *42 Zones*

Cluster 8 postal sectors are also located in central urban locations but would appear to be particularly deprived. A large proportion of the dwellings are flats and the majority are rented from the local authority. This, and the overall location of these zones, points to inner city high or mid-rise council estates. Unemployment is very high here, typically over 2.5%, and car ownership is low. The population is a mixture of young singles and the elderly. LLTI is also very high here, likely because of the elderly population. There appears to be a high turnover of population in these zones, probably as households become more established and are able to move into better surroundings. There is also a higher proportion of pensioner migrants as the poorer elderly move into smaller housing units. These zones have some of the highest proportions of households in Social Classes IV and V.

The housing market is very flat with very few sales over the 12-month period. This is understandable with such a high rate of rented property. Only 17% of households own a mortgage (compared to 41% in the region as a whole). Most sales are of terraced housing, often for under £30,000. Despite the predominance of flats as the most common type of

housing, there appear to be very few sales. What transactions are completed appear to be for about the same amount as the terraced housing.

5.1.9 Cluster 9 *97 Zones*

The geographical pattern of Cluster 9 appears to include the rural and ex-urban zones in the region that are not included in Cluster 4, i.e. those in the north and west of the region. The population here appears very affluent with large detached houses making up most of the housing stock. Car ownership is very high with many households having two or more cars. Most households are in Social Class II or I and many dwellings are owned outright. Mortgage ownership is also very high. The population is late middle aged or in early retirement. However there do not appear to be many pensioner migrants suggesting that the population is quite established and had aged *in situ*. There are very few singles or young families. Unemployment and LLTI are both very low in these areas, as are households in the lower social classes.

The average value of property sold in these zones in very high at well over £110,000. On average detached houses make up nearly 60% of sales and the average transaction price for these is nearly £145,000. These zones contain some of the most expensive housing in the region. However, although the market is very valuable there are actually fewer sales per year. The average number of sales in the 12-month period is just 57, again suggesting a more established rural population.

6 EVALUATING THE PROPERTY DATA CLASSIFICATION

The cluster pen portraits in Section 5 suggest that the property data has indeed added to the taxonomy of the original static demand classification. Key characteristics of the original clusters, such as affluence levels, type of housing stock or areas of deprivation, have apparently been enhanced by the addition of the property data. Clear distinctions can be made between areas where there are few sales because of the predominance of council renting or because the population is older, more established and less inclined to move. However, it is important to understand how, if at all, the cluster geography has been altered by the new variables.

6.1 Comparing this classification with the static demand classification

By comparing the 8-Cluster solution of the original static demand classification (SDC) with the 9-Cluster static demand and property classification (SDPC) we can see that there are a number of distinct similarities. 'Eye-balling' the pen portraits in sections 2.2 and 5.1 and cluster geographies in figures 2.4 and 5.1 suggests that the two classifications are very similar and, in most cases, reveals that the new property variables have added to the original taxonomies.

An immediate similarity can be seen between SDC Cluster 1 and SDPC Cluster 6. Not only are they very similar geographically but we can see that the principal characteristic, the purchase of council houses by their former tenants has been retained in both. This element has been enhanced by the addition of the property transaction data. SDC Cluster 8 is very similar to Cluster 3 in the SDPC as both have a predominance of privately rented terraced housing and are again similarly located. Both classifications have a cluster with a very distinct student population. However, the advantage that SDPC Cluster 7 has over SDC cluster 3 is that it draws upon the positive impact that students have on the housing market as it shows that terraced property prices are higher and the market more buoyant. A key omission, however, is that Figure 5.1 shows that the student areas of Hull have been lost from the property data classification. SDC Cluster 2 and SDPC Cluster 8 both display the same proliferation of mid or high rise council estates with the low property prices and flat housing market characterised in the latter classification highlighting the fact that these may be particularly deprived areas. SDPC Cluster 2 and SDC Cluster 5 both suggest the same sort of area with a mix of affluence and age types hinting towards gentrification in the major urban centres. As Section 5.1.2 shows, this suggestion of gentrification is corroborated by the high average value of flats in these areas.

The final set of matching clusters is a little less specific. The static demand classification distinguishes between two largely ex-urban or rural clusters. SDC Cluster 7 is more rural and there is a suggestion that the population is a little older, more established and potentially more affluent than the slightly younger families found in SDPC Cluster 4. While both clusters have fewer migrants, Cluster 4 has more pensioner migrants suggesting

retirement migration. Furthermore there is a clear geographic distinction between the two with Cluster 4 being found mostly in the south of the region and Cluster 7 in the north

The addition of the property data has further added to this picture by creating a third, predominantly ex-urban cluster. It is clear that the segmentation process has made a further distinction between these clusters. SDPC Cluster 1 appears to represent the postal sectors where the housing market is more buoyant. While Cluster 9 and Cluster 4 both have higher transaction prices, their markets appear a little flat with fewer sales in the 12-month period. Cluster 1 has more sales although possibly slightly lower average values than its rural counterparts. The creation of this cluster has also added to the geographical distinctions between these clusters. Cluster 4 and Cluster 9 appear to retain the north-south split evident in SDC Clusters 4 and 7, however the new Cluster 1 is concentrated heavily in the west of the region, creating more of a north-west-southeast distinction between the non-urban clusters. The geographical position of the Cluster 1 zones is intuitive; they occupy the wealthy commuter belt to the north of Leeds and the north and west of Bradford as well as areas surrounding Harrogate and York. More significantly, they are also found to the very west of the region, the postal sectors that include such towns as Hebden Bridge, Ripponden and Sowerby Bridge; areas that have vibrant housing markets because of their proximity to both the West Yorkshire and Greater Manchester conurbations. Table 6.1 shows that Cluster 1 has been predominantly created from zones that were previously in SDC Cluster 4 with 90 of its 116 zones having come from there. Four of these postal sectors were displayed as outliers to Cluster 4 in Figure 3.9. It is therefore fair to say that, in terms of matching clusters across the classifications, SDC Cluster 7 is most similar to SDPC Cluster 9 while SDC Cluster 4 is matched to SDPC Clusters 1 and 4.

Table 6.1 SDC Cluster membership of postal sectors in SDPC Cluster 1

SDC Cluster	No. in SDPC Cluster 1
1	7
4	90
7	13
8	6
Total	116

Table 6.2 monitors the extent of the geographical similarities between the matching clusters. In general terms, 666 (85%) of postal sectors fall into the correct pair of clusters. This suggests a high degree of similarity between the two classifications. The table is designed to show the number of matching zones as a percentage of the number of zones in both the SDC and SDPC cluster in question. The SDC cluster with the highest percentage of postal sectors in the corresponding SDPC cluster is Cluster 4. However, it is hardly surprising that this percentage is so high when its corresponding SDPC cluster is actually an amalgamation of two clusters. Beyond this, the most similar cluster is SDC Cluster 3 which sees 14 of its 15 zones placed into the corresponding SDPC cluster. The only zone not to have been retained is HU5 2; as discussed before it is interesting, and perhaps disappointing, that Hull is the only major urban area to be without a postal sector in this group when sectors of Scarborough and Bridlington (YO11 2 and YO15 2, respectively) – which can hardly be described as having a large student population – have been added instead. The SDC cluster with the least zones in the equivalent SDPC group is Cluster 5. Matched with SDPC Cluster 2 because they both seem to represent areas of high and mid-rise council estates, only 8 of the 20 postal sectors (40%) are successfully matched. In many ways this is a small number problem as both clusters have very small membership values but it is still noteworthy that less than half of the original cluster is placed into the (apparently) appropriate cluster when the property data has been added. However, in Section 3 it was suggested that Cluster 5 was the weakest of the 8 clusters in the original classification with the largest mean distance from cluster centre. It is therefore possible to argue that, in this case, the new variables have added to the segmentation process and allowed some of the outlying zones that were placed in SDC Cluster 5 to be allocated to other clusters in the new classification.

As a proportion of the static demand and property data classification, the most geographically contiguous pair of clusters is SDPC Cluster 9 and SDC Cluster 7. 95 of the 97 (98%) zones in Cluster 9 are also in the parallel cluster in the previous classification. The two extraneous zones (LS22 6 and LS25 3) were originally in SDC Cluster 4. Cluster 3 is also shown to be very closely linked with SDC Cluster 8 because of the high number of privately rented, inner city terraced houses. 82 of the 91 (90%) of postal sectors placed in this cluster were also in the equivalent cluster.

Table 6.2 Monitoring the degree of comparison between clusters in SDC and SDPC

SDPC Cluster	n	SDC Cluster	n	No. of zones in both	% of SDC	% of SDPC
1 & 4	244	**4**	201	192	95.5	78.7
2	9	**5**	20	8	40.0	88.9
3	91	**8**	103	82	79.6	90.1
5	106	**6**	112	95	84.8	89.6
6	172	**1**	165	145	87.9	84.3
7	22	**3**	15	14	93.3	63.6
8	42	**2**	45	35	77.8	83.3
9	97	**7**	122	95	77.9	97.9
Total	**783**		**783**	**666**	**85.1**	**85.1**

The lowest level of geographical association is found in parallel clusters SDPC 7 and SDC 3 where just 64% (14 out of 22) of postal sectors in the new classification are present in the original 'Student' cluster. As we have seen before, this is likely to be a small number problem because both clusters have such low memberships.

It can be said that, at face value, the addition of the property market variables has enhanced the original static demand classification. Despite there being a different number of clusters in the two segmentations, the cluster characteristics and geography have been shown to be very similar. The extra cluster created in the static demand and property data classification has added a further dimension and made an important distinction between postal sectors in the region that was apparently lacking in the original classification.

However, this analysis only suggests that the addition of property data has been successful on the basis of some very basic visual and geographical analysis. More robust statistical analysis will be needed to see how the classification performs and what effect the new variables have had upon the segmentation process.

6.2 Assessing the performance of the new classification system

6.2.1 Distance from cluster centre

Figure 6.1 shows the histogram of distance values for this new classification. The range of values is greater than in the static demand classification (Figure 3.7). The mean value in this new classification is 0.75 which is 0.13 higher than the static demand classification although the standard deviations are similar. However, despite this the histogram in Figure

6.1 suggests that the new classification 'performs' as well as the original demand classification. The histogram is again positively skewed, suggesting that most cases have lower distance values. Furthermore it appears that this new graph maybe slightly more skewed towards the lower values than its predecessor because the modal category is further to the right of the mean value than in Figure 3.7.

Figure 6.1 Histogram of cluster distances in the static demand and property data classification

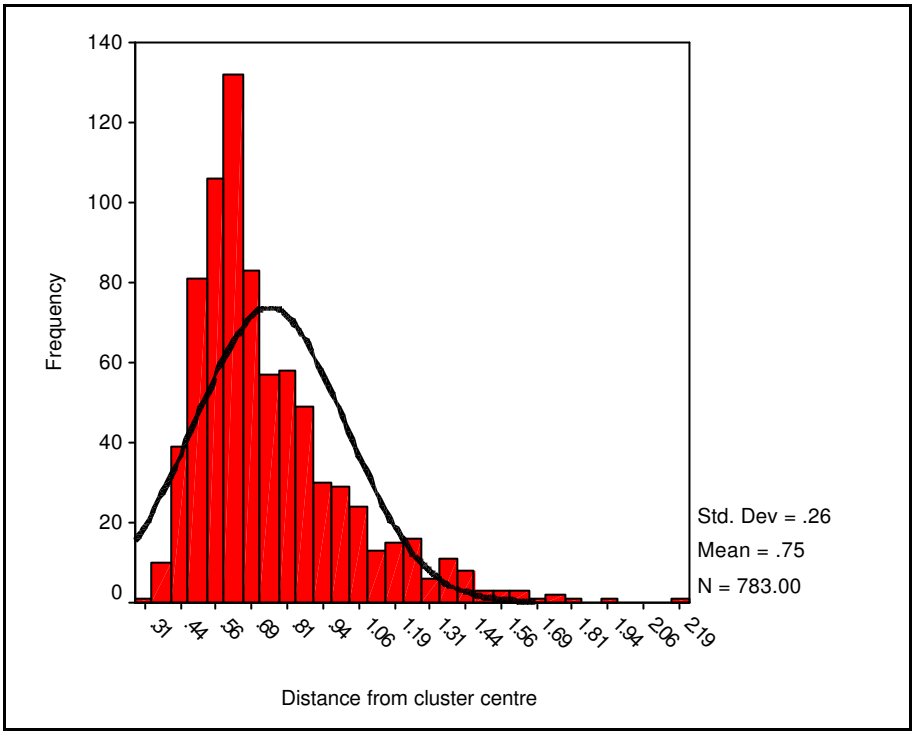
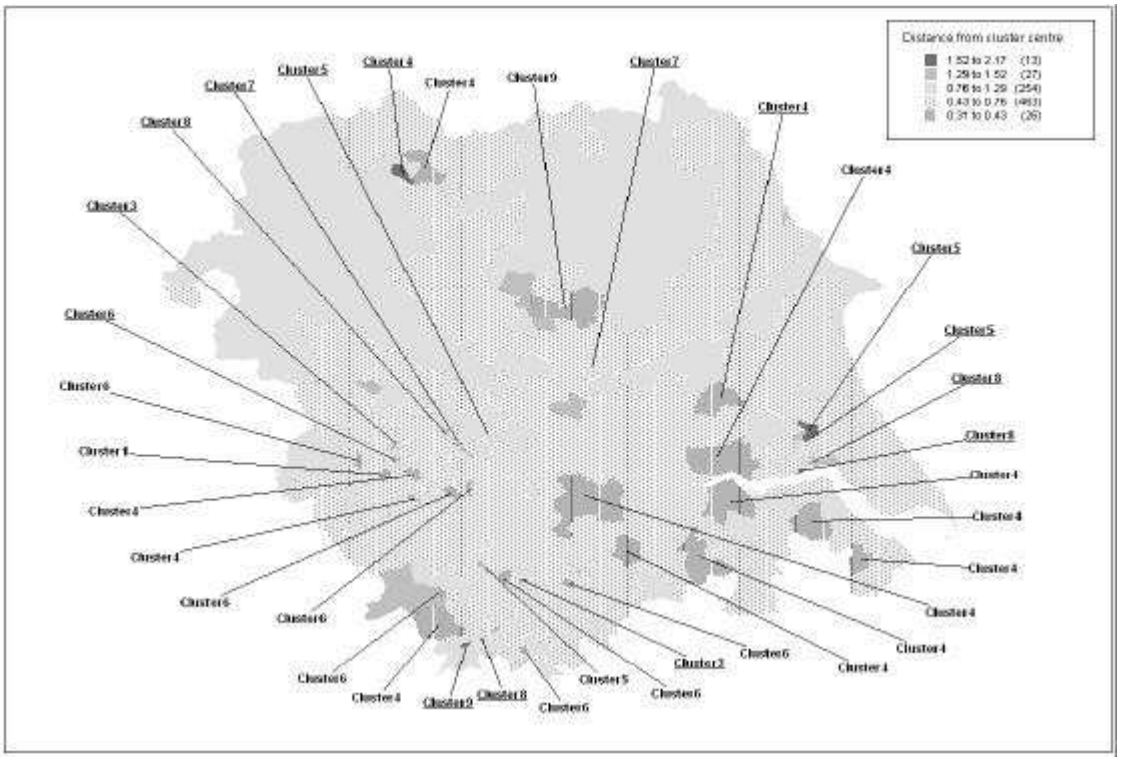


Figure 6.2 shows the geographical distribution of these cluster distances. Once more the labelling scheme shows the cluster membership of the zones that fall into the lowest category (0.31 to 0.43) in plain text and the highest category (1.52 to 2.17) as underlined. Many of the zones that fall into the lowest category are in Cluster 4 although the highest category is not so easily demarcated. In Figure 3.8 there was a clear urban-rural distinction between the zones in the highest and lowest distance categories and it would appear that this pattern has been replicated in Figure 6.2; the zones in the lowest category are mostly rural with a few urban (although rarely central urban) postal sectors, while the highest distance values tend only to be found in the centre of the urban areas. The exception to this is HU7 3, HU7 5 and DL9 3. DL9 3 contains the Armed Forces camp at Catterick and it is likely that the characteristically young population living in and around an Army base will

result in some extreme values that may draw the case away from the others in taxonomic space. However, it is less clear why HU7 3 and HU7 5 should have such extreme values.

Figure 6.2 Mapping the distance from cluster centre in the new property data classification



6.2.2 Cluster specific distance values

As with the static demand classification there would appear to be a cluster specific pattern to the distribution of distance values. Table 6.3 shows the descriptive statistics of the distance values within the different clusters. The immediate observation is that the range of values within each cluster is higher on average than in the static demand classification. Previously the lowest range (SDC Cluster 1) was 0.592, yet in this classification the lowest range is 0.823 (Cluster 6). According to this table, Cluster 6 is the strongest cluster as it has the lowest range, the lowest average value and the lowest standard deviation. Cluster 7 appears to be the weakest cluster as it has the highest range, mean and standard deviation. Once again we can see that the weakest cluster has a very low membership value. In the previous analysis the 'weakest' cluster (SDC Cluster 5) had just 20 zones, Cluster 7 has 22.

Table 6.3 Descriptive statistics of distance values by cluster centre in the new classification (ranked by the mean)

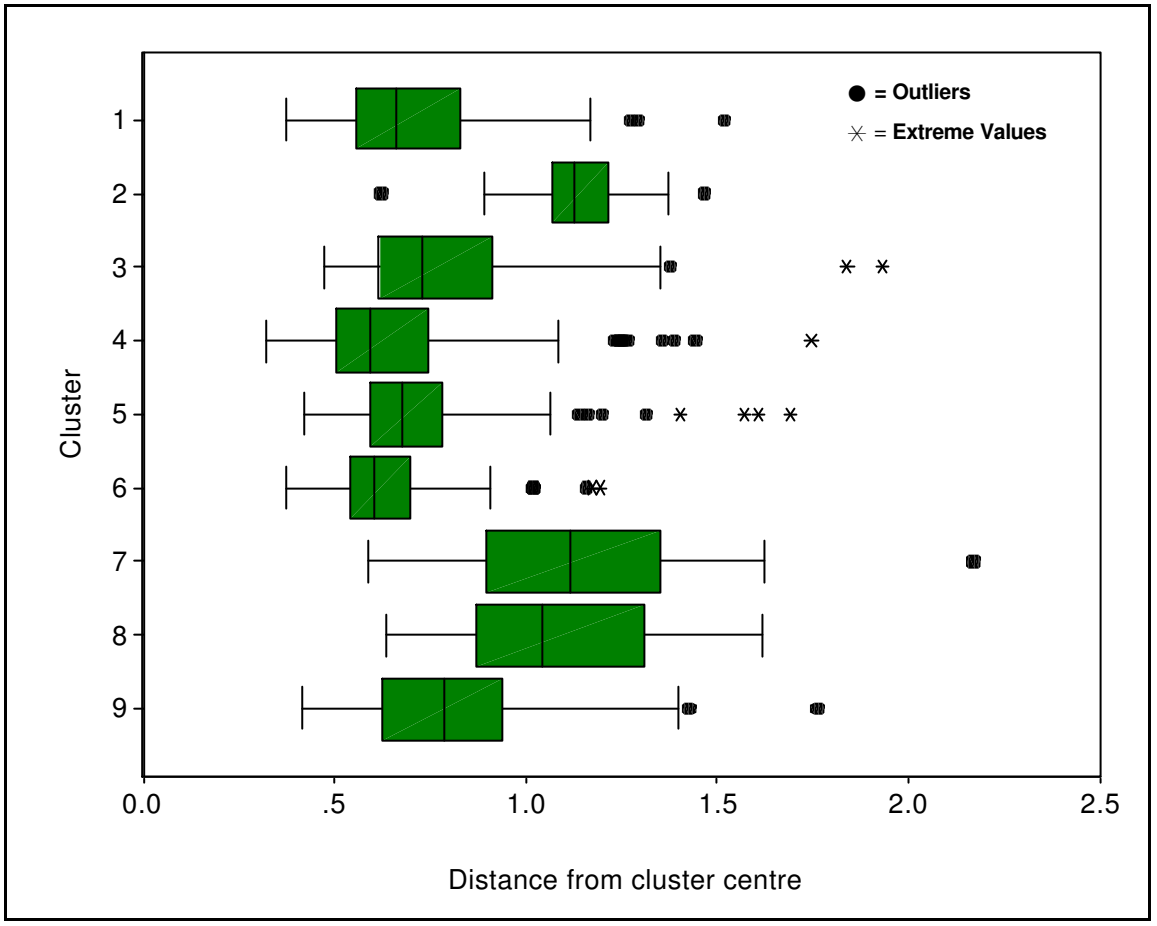
Cluster	Rank	n	max	min	range	mean	st. dev	no. of "outliers"	no. of "extremes"	equivalent SDC Cluster	Rank of SDC Cluster
6	*1*	172	1.196	0.373	0.823	0.630	0.141	2	2	**1**	*1*
4	*2*	128	1.747	0.322	1.425	0.672	0.261	9	1	**4**	*2*
1	*3*	116	1.520	0.376	1.143	0.714	0.207	3	-	**4**	*2*
5	*4*	106	1.688	0.417	1.271	0.736	0.240	4	4	**6**	*3*
3	*5*	91	1.935	0.476	1.459	0.803	0.266	1	2	**8**	*5*
9	*6*	97	1.762	0.416	1.347	0.810	0.242	2	-	**7**	*4*
8	*7*	42	1.618	0.632	0.986	1.094	0.264	-	-	**2**	*6*
2	*8*	9	1.467	0.625	0.842	1.113	0.249	2	-	**5**	*8*
7	*9*	22	2.167	0.591	1.576	1.158	0.354	1	-	**3**	*7*
Total		***783***	***2.167***	***0.322***	***1.845***	***0.752***	***0.265***	***24***	***9***		

Figure 6.3 shows the box plots of these cluster specific distance values and indicates that the positive skew identified in Figure 6.2 has been replicated in the individual clusters. Looking at the data without the *outliers* and *extreme values* confirms that Cluster 6 is the strongest cluster as it is even more heavily concentrated in the lower range of values. Furthermore, Clusters 4 and 5 are stronger than the descriptive statistics would imply. Cluster 4 has the largest number of *outliers* (nine) and Cluster 5 the most *extreme values* (four) and without these points the distance values in the two clusters begin to look a lot more favourable. Cluster 7 is confirmed as the weakest because, even with its outlier removed, it still has the highest (artificial) maximum value and the largest range. Cluster 2 presents an interesting case here as both its maximum and minimum values are displayed as *outliers*. Although it is now left with the smallest range of values, removing these erroneous cases only serves to make the cluster look weaker because it is left with such a high (artificial) minimum value.

It is apparent that the relative strengths of the clusters in the static demand classification have been continued in their equivalent clusters in the property data classification. Table 6.3 displays the SDC cluster and it's rank in Table 3.2. The link between the two classifications is clear, with most clusters holding the same or similar position in the table in the new classification. In Section 3.2.2, two postal sectors were identified as having the largest distance values, BD1 3 and LS2 3. The new property data has not changed their positions in taxonomic space as they still have the largest distance values and they have

been retained in the equivalent cluster that they were attributed to in the original classification.

Figure 6.3 Box plots of distances from cluster centre in the new classification



6.2.3 Summary

The new property transaction data does not seem to have affected the classification unduly. 85% of the postal sectors have been placed into the same group that they were segmented into in the static demand classification. The main characteristics of the classification have been retained and it would appear that the property data has assisted in further segmenting the dataset. The creation of an extra cluster is testament to this as it appears to distinguish between types of areas that were originally amalgamated into one cluster. The average distance from cluster centre component has increased but in many ways we should expect this. The new property data has added 10 more dimensions to the multivariate taxonomic space and as such it is conceivable that the distance from cluster centre should get larger.

However, from this analysis it can only be speculated that the new property data has added to the segmentation. The following section will investigate the contribution of the new variables in this new classification.

7 EVALUATING THE CONTRIBUTION OF THE NEW PROPERTY DATA VARIABLES

7.1 Analysis of the revised ANOVA table

Table 7.1 Top 10 most and least 'effective' variables according to the revised ANOVA table

Top 10		Bottom 10	
Variable	***F*-statistic**	**Variable**	***F*-statistic**
No Car	477.57	Population aged 75-84	3.89
Detached	469.95	Pensioner Migrants	6.95
2+ Cars	452.89	Population aged 65-74	8.79
Total Married Population	332.61	Population aged 85 plus	9.79
Single Population	315.78	Bedsits	11.30
Rented from HA, LA or New Town	302.81	2+ Families & Owner Occupied / rent	11.52
Households > 7 Rooms	271.22	2+ Families & Local Authority rent	12.54
No family household & council rent	236.68	H'holds in Social Class III(N)	13.94
Flats	236.09	Population aged 25-44	14.41
Average value all transactions	218.99	Retired Population	15.62

Table 7.1 shows the 10 highest and lowest *F*-statistics from the revised ANOVA table. The variables displayed here are not substantially different from the original ANOVA table although the *F*-statistics themselves may have changed somewhat. 'Households with no car' is now the most important variable, followed by the proportion of detached houses. The *F*-statistics are generally a little lower than in table 4.1 although the relative importance of the detached housing variable has increased. The same age ranges can be found amongst the least 'effective' variables and the bottom of the ANOVA table is again very similar with only one new variable, households in Social Class III(N), which replaces households lacking a bath or shower.

The presence of the average value of all transactions in top 10 *F*-statistics and the fact that none of the new variables appear on the right hand side of the table is encouraging as it suggests that they have indeed played a significant part in the new segmentation. The average value of all transactions is the most effective new variable yet Table 7.2 shows that

four other variables have noticeably high *F*-statistics; the proportion of sales of detached houses, flats and terraces and the average value of detached houses. However, there does appear to be a range of *F*-statistic values amongst the new variables. The relative importance of the total sales of property appears quite low, as do the average transaction value variables for semi-detached, flats and terraces. The low *F*-statistic for total sales of property is surprising as one would expect this variable to be an important indication of the level of activity in the housing market. The data is sufficiently spread out to allow extreme cases to be identified, the mean number of sales is 80 per postal sector and the standard deviation is 53, with a maximum value of 267 (S10 1 – a suburban area on the western side of Sheffield). Nevertheless the ANOVA table suggests that it is not necessarily such an important variable.

Table 7.2 *F*-statistics for the new property transaction variables

Variable		*F*-statistic
VAR52	Total Sales of property	27.00
VAR53	Proportion detached sales	193.33
VAR54	Proportion semi sales	78.94
VAR55	Proportion Flats sales	191.47
VAR56	Proportion terraced sales	100.88
VAR57	Average value all transactions Q399 to Q200	218.99
VAR58	Average value detached Q399 to Q200	108.62
VAR59	Average value semi Q399 to Q200	50.63
VAR60	Average value flats Q399 to Q200	40.38
VAR61	Average value terraced Q399 to Q200	35.73

7.2 Correlation analysis

The ecological relationship of the 10 property transaction variables to the original 51 Census (sic) variables can be investigated by creating an *r*-matrix in the same way as before. Table 7.3 shows the variables that now have correlation coefficients greater than ±0.5 with 10 or more other variables. The number of correlations has increased for many of the variables; at the top of the table, '2+ cars' now has three more variables with correlations greater than ±0.5 while the single population and detached housing variables both have four more. However, seven variables have not seen an increase in the number of significant correlations and only one, 'lone parents', is not related to housing tenure. It is particularly surprising that the local authority renting and mortgage owning variables do not have any correlations over ±0.5 with the property transaction variables as one could

reasonably expect some level of ecological relationship. The *r*-matrix does reveal some level of association; the proportion of households renting from a Local Authority has correlations of –0.486 and –0.466 with the average value of all transactions and the average value of all detached housing transactions, respectively. Furthermore, it also has correlations greater than –0.4 with the proportion of detached sales and the average value of semi-detached houses. However the mortgage owners variable is even less well related to the property data as only one variable (average value of detached houses) has a correlation over 0.4. Nevertheless, there are clear associations between the original census variables and the property transaction data. Three of the new variables appear in Table 7.3 while all bar four (total sales of property and the average value of semi-detached, terraced and flats) have correlations greater than ±0.5 with at least one of the original 51 variables.

In Section 7.1 the total number of transactions was highlighted as having the lowest *F*-statistic of the new suite of variables and once again the analysis draws upon it here. Furthermore, not only does it not have a significant correlation with any of the original Census variables, it does not have any from within its own suite. Once again this is disappointing as one would expect this variable to be more important as an indicator of housing market buoyancy.

Table 7.3 Updated list of variables that have correlation coefficients greater than ±0.5 with 10 or more other variables

Variable	Number of correlations	Variable	Number of correlations
2+ Cars	26	Population aged 45-64	17
Single Population	23	Persons with LLTI	17
Detached Housing	23	Unemployment	17
H'holds in Social Class II	23	*No family household & council rent*	*16*
Total Married Population	22	*Lone Parents*	*15*
Rented from HA, LA or New Town	*22*	**Average value detached**	**15**
No Car	22	*Mortgage Owners*	*14*
married and cohab couple & OO/Rent	22	**Proportion detached sales**	**14**
Households > 7 Rooms	21	Flats	13
Home Owners	20	*Married and cohab couple & council rent*	*13*
Workers with other qualifications	20	*Married and cohab couple w/ children & OO/Rent*	*13*
Average value all transactions	**19**	*Married and cohab couple w/ children & council rent*	*10*
Self-employed	18		

Note: Variables in **bold** are new the property transaction variables. Variables in *italics* have seen no change in the number of significant correlations since Table 4.2

One interesting element of the *r*-matrix is the correlation between the movers last year from the 1991 Census and the total sales of property from the 1999/2000 transaction data. Not only is the correlation very low but it is also negative, -0.173. The reasons for this are unclear, although it may suggest that, for some areas, the focus of migration in the region has shifted since the 1991 Census. However, this does not mean that this variable is inaccurate or provides misleading information. Its correlation with other variables is more intuitive, particularly −2.36 with the Local Authority renting variable and 0.399 with mortgage owners. Furthermore, the relative positions of the migration and housing transaction variables in the Z-score tables reveals a lot about an area and helps to characterise the nature of the processes that are occurring there. For instance, a high number of sales but lower number of movers last year might indicate a level of development since 1991 or new areas of housing buoyancy (Cluster 1) or it may not involve the physical movement of people at all, simply a change in tenure such as the purchase of council houses (Cluster 6). A low number of sales but higher number of movers can indicate areas where the population is transient and therefore rents their accommodation; these can either be more affluent, possibly gentrified areas (Cluster 2), or less wealthy, often high rise estates (Cluster 8).

The average value of all transactions, and the proportion and average value of detached sales have the highest number of correlations greater than ±0.5. It was demonstrated in Table 7.2 that these three variables also have the highest *F*-statistics of the new suite. It is therefore apparent that the housing transaction variables that have the most effect on the classification are the ones that will enhance the picture of affluence in an area. However, we cannot necessarily prove this fact by just looking at the bivariate correlations. Once again, therefore, we must expand this pairwise analysis to a more multivariate approach.

7.3 Assessing the importance of the new variables using PCA

Using the same criterion for the selection of principal components as was used in Section 4.3 (components with one or more variable with a factor loading of ±0.45) there are now five groups of variables to consider (Table 7.4). The housing market variables are evident in all but one of the new components, immediately indicating that they have indeed had a role to play in the creation of this new classification.

The first three components describe the same underlying dimensions that were identified in Table 4.3. Once again the first component describes the contrast between socio-economic disadvantage and prosperity. The factor loadings and order in which the variables appear are reasonably similar in both tables with car ownership once again being at the top of both the negative and positive sides of the component. However, the 'polarity' of the variables has been reversed. In Table 4.3 the variables that described disadvantage all had positive factor loadings while the variables that would describe affluence were negative. In Table 7.4 this situation is inverted, suggesting that this component now specifically relates to affluence rather than deprivation. This first component explains 30% of the variance in the data which is a little less than the 31% explained by the first component in Table 4.3. However, as there are more dimensions in the multivariate analysis it is not surprising that the individual levels of variance might be diminished.

Four of the variables from the new suite have been added to this first component; the average value of all transactions, the average value and proportion of detached sales and the average value of semi-detached transactions. This confirms the suggestion made after the correlation analysis that the new property market variables are contributing to the picture of affluence in the region and have indeed enhanced the segmentation. A further notable inclusion in the first component is the overcrowding variable (percentage of households with more than one person per room). This variable has been now identified as a key indicator of deprivation, thus justifying the earlier decision to change it.

In the second component, the property market variables have again been combined with the original Census variables to enhance the distinction between the more transient student and young professional population, with a high value and proportion of sales of flats, and settled working class families living in semi-detached housing which, because the average value variable has not been included, may not be that valuable.

Table 7.4 Principal components in the static demand and house price classification

Positive		Negative	
Component 1			
2+ Cars	0.9295	No Car	-0.8813
Detached	0.8367	Rented from HA, LA or New Town	-0.8262
H'holds in Social Class II	0.8214	Single Population	-0.8233
Total Married Population	0.8179	Unemployment	-0.7538
married and cohab couple & OO/Rent	0.8131	No family household & council rent	-0.7526
Households > 7 Rooms	0.8039	Persons with LLTI	-0.7135
Average value all transactions Q399 to Q200	**0.7726**	Lone Parents	-0.6731
Home Owners	0.7650	married and cohab couple & council rent	-0.6454
Self-employed	0.7327	H'holds in Social Class IV	-0.6387
Average value detached Q399 to Q200	**0.7172**	H'holds in Social Class V	-0.6370
Workers with other qualifications	0.7113	married and cohab couple w/ children & council rent	-0.5808
Proportion detached sales	**0.6997**	Flats	-0.5505
Mortgage Owners	0.6899	Terraced	-0.5274
45-64est	0.6704	*Households > 1 ppr*	*-0.4879*
married and cohab couple w/ children & OO/Rent	0.6482	*No Central Heating*	*-0.4554*
H'holds in Social Class I	0.5936		
Average value semi Q399 to Q200	**0.5748**		
Component 2			
Privately Rented	0.7302	Households with dependents	-0.7480
Movers last year	0.6959	Semi- Detached	-0.6646
No family household & OO/Rent	0.6641	5-14est	-0.6487
Flats	0.6044	H'holds in Social Class III(M)	-0.6375
Students	0.5769	**Proportion semi sales**	**-0.5671**
Workers with higher degrees	0.5575	*Total Married Population*	*-0.4764*
Proportion Flats sales	**0.5340**	*0-4est*	*-0.4671*
Workers with other qualifications	0.5108		
15-24est	0.5044		
Average value flats Q399 to Q200	***0.4946***		
Single Population	*0.4682*		
Component 3			
65-74est	0.8203	25-44est	-0.5386
75-84est	0.7219	*Economically active residents aged 16+*	*-0.4925*
Retired Population	0.6957		
85+est	*0.4987*		
Persons with LLTI	*0.4876*		
45-64est	*0.4618*		
Component 4			
Home Owners	0.5112		
Proportion terraced sales	**0.5084**		
Terraced	0.5084		
75-84est	*0.4634*		
Component 5			
Total Sales of property	**0.6916**		
H'holds in Social Class III(N)	0.5603		
Average value flats Q399 to Q200	***0.4608***		

Note: Variables in *Italics* had coefficients $< \pm 0.5$. Variables in **bold** denote new variables

It would appear that the fourth and fifth components have proved Voas and Williamson's (2001) criticism that the components extracted by PCA are not always easy to interpret. Neither have any negative factor loadings and while component four appears to describe an underlying dimension of pensioners living in their own terraced homes, the fifth is even less obvious. Nevertheless, the two components together contain three of the housing market variables, thus further adding to the impression that the new variables have indeed added to the segmentation. The inclusion of the total sales of property in the fifth component is encouraging as this has twice been highlighted as potentially a weaker variable. However, as this component has little or no substantive meaning this is something of a double-edged sword.

In total, nine of the new variables appear in the first five components; the average value of flats appears twice. Only the average value of terraced sales is excluded. A full analysis of the component table reveals that this variable has no component factor loading greater than ±0.45 and its highest is just 0.39 (Component 1) or −0.25 (Component 17). Table 7.2 showed that this variable has the second lowest *F*-statistic of the property market variables and the correlation analysis identified that it did not have any significant correlations with the original variables. It is the only variable to be highlighted in all three analyses and there may, therefore, be a case for its removal. However, as with the retention of the weaker cross tabulated family structure and housing variable in Section 4.3, it would be unwise to remove this variable because it completes the disaggregation of the property value variables. Furthermore, as the cluster pen portrait of Cluster 7 (Section 5.1.7) reveals, if not a cluster formative variable, it is still useful for cluster description.

Table 7.5 Variables not included in the first five components of the property data classification

Variable	Components where variable loading is > ± 0.45
Pensioner Migrants	11
Bedsits	12
Lacking Bath & shower	None
households with 2 or more families & OO/Rent	None
households with 2 or more families & council rent	13
Average value terraced Q399 to Q200	None

It would appear that the extra dimensions in the dataset have had a positive influence upon the relative importance of some the original Census variables. 55 of the 61 variables are included in the first five components. As we have seen, nine of these 55 variables are from the property data, leaving 46 of the original variables; in the previous PCA there were 44. As discussed before, one of the new variables is the amended overcrowding variable; the other addition is households in Social Class III(N). Nevertheless, we can see from Table 7.5 that none of the other variables omitted from the original PCA (Table 4.5) have necessarily improved their position. Indeed, the three variables that do have a factor loading greater than ±0.45 are all found in a lower component than before[3]. It should be noted, however, that the lacking bath and shower variable only narrowly misses out on being included in the fourth component, its factor loading being 0.44 and this could be said to represent something of an improvement. It was suggested in Section 4.4 that this variable should be closely monitored as other variables are added. It may therefore be wise to continue to include this variable in subsequent classifications but there may already be a strong case for its removal from the final classification.

7.4 Summary

It is clear that the property transaction variables have indeed been a valuable addition to the original classification. The variables have mostly proven to be important drivers in the classification process. The average value of terraced housing is the only new variable that does not contribute in some way although this does not necessarily mean we should remove it. However the analysis has revealed that if anything the property market variables tend to accentuate the picture affluence rather than deprivation. This is probably because the level of transaction is dependent upon the level of affluence. The poorer areas of the region tend to have less property data for them and therefore the data becomes a little less meaningful. Furthermore, the census variables designed to indicate areas of poverty all have high values in such areas (for instance higher unemployment, more households with out a car). By definition, however, the property transaction variables will all be lower in such areas, particularly those relating to average value. This explains why we do not see such variables

[3] Households with two or more families and Local Authority renting was already in component 13 but it was also placed in component 11.

directly related to deprivation in the PCA. Nevertheless, as the cluster pen portraits show, they are useful for further characterising areas of lesser prosperity.

The new variables have only increased the overall importance of one of the original variables but they do not appear to have unduly affected the others. Changing the overcrowding variable to include households with more than one person per room (rather than 1.5) appears to have been a sensible decision. Although its *F*-statistic is still not particularly large (41.3) it is now included as one of the most important variables in the first component of the PCA. There might still be a case for removing the variable describing households lacking bath or shower but for the moment it will be retained. It will once again be interesting to see how this variable performs when further variables are added. In particular the addition of supply side and dynamic variables to this extended demand based classification (Debenham et al. 2001a; 2001b).

8 CONCLUSIONS

A static demand classification has been created using 51 Census (sic) variables that are similar to those likely to be used in commercial geodemographic systems. The result is a meaningful, sensible and robust 8-cluster taxonomy that shows a similar pattern of clustering to that of GB MOSAIC, a commercial geodemographic system. The choice of variables appears to have been reasonably sound with only one major change needed in the amendment of the number of persons per room that signified "over crowding".

The main intention of this paper was to show how geodemographic classification can be extended using extra variables and it is fair to say that this has been achieved. It is clear that the addition of the ten property transaction variables has indeed enhanced the classification. The distinct characteristics of the 8 original clusters have been retained and in most cases (such as the purchase of council houses and the buoyancy of student housing areas) enhanced by the property market data. Furthermore, the creation of a new cluster that distinguishes between ex-urban areas with a more vibrant housing market than the more rural and more stable areas suggests that this new data has indeed assisted the segmentation.

There has been no undue degradation of the overall performance of the classification. Average cluster distances have increased but this is probably to be expected as there are now more dimensions in the multivariate taxonomic space. All of the new variables prove to be important drivers of clustering bar one, the average value of terraced housing. However, it would be unwise to remove this variable as it appears to be useful for cluster descriptive purposes. The new variables do not seem to have affected the performance of the original variables too much and the decision to change the overcrowding variable has proved to be a wise one. However, it would appear that the new variables only enhance the picture of affluence in the region. The three most important variables, average value of all transactions and the proportion and average value of detached house sales, are the three most important drivers of the new variables and would all indicate areas of affluence. Nevertheless as we have seen from the cluster pen portraits in Section 5 and the discussion in Section 7, the other variables still prove to be useful in describing the nature of the less prosperous postal sectors in Yorkshire and the Humber.

The extended system that makes use of property market data is still a purely demand-based classification. Further classifications will be developed that will see the addition of supply-side variables that report on the state of the labour market and the level of interaction between residential and workplace zones. Furthermore, retrospective and projective dynamic variables will be added on both the demand and supply side that will provide indicators on the level of change in each postal sector. It is believed that these indicators will add a crucial element to small area taxonomies that has hitherto been missing (Debenham et al. 2001a; 2001b). This paper has put forward a number of different analysis methods that have been shown to effectively measure the level of performance of a classification system and to highlight *key stock* variables in the cluster formative process and those that may be justifiably removed from further stages of the development of this new type of classification. The analysis of cluster distances, ANOVA tables, bivariate ecological relationships and principal component analysis have all proved their worth and can now be used to monitor the performance of subsequent supply-side and dynamic variables as they are added.

ACKNOWLEDGEMENTS

1. The author wishes to acknowledge the support of the Economic and Social Research Council (ESRC), CASE Award S00429937070, and CASE partner GMAP Ltd. for funding the PhD that produced the research in this report.
2. Digital maps of the postal boundaries in Yorkshire and the Humber (sic) are based on data from GEOPLAN 1999 postal boundary data made available under a Combined Higher Education Software Team (CHEST) agreement. GEOPLAN boundary data is Copyright of Yellow Marketing Limited, Postcodes are Copyright of the Post Office.
3. The 1991 Census statistics used in the research are Crown Copyright and made available by the Census Dissemination Unit through the Manchester Information and Associated Services (MIMAS) of Manchester Computing, University of Manchester. The 1991 Census data have been purchased for academic research purposes by ESRC and JISC.
4. The All-Fields Postcode Directory (AFPD) is produced by the Office for National Statistics (ONS) with information from ONS. GRO(S), NISRA, the Post Office and Department of Health. The Updated UK Area Masterfiles ESRC funded project (H507255164) has re-engineered the AFPD to link census geographies to other administrative geographies. Data in this lookup table is Crown Copyright, ESRC purchase.
5. Various labour market datasets are made available by ONS through the National On-line Manpower Information System (NOMIS) at the University of Durham. Employment data is taken from the Annual Employment Survey (AES) and is published by ONS. Unemployment data is taken from the Claimant Count which is also published by ONS and is based upon data from the Benefits Agency administrative system. All data is Crown Copyright.
6. The Experían Limited Postal Sector Data is made available through the MIMAS service and was purchased under a joint ESRC/JISC agreement. The HM Land Registry data contained in the Experían Limited Postal Sector Data is Crown Copyright.
7. The author would like to thank Dr. John Stillwell for his comments on a previous draft of this paper

REFERENCES

- Birkin, M. (1995) *Customer targeting, geodemographics and lifestyle approaches*, Ch 6 in Longley, P. & Clarke, G.P. (eds.) GIS for Business and Service Planning, GeoInformation, Cambridge
- Cattell, R.B. (1966) *The scree test for the number of factors*, Multivariate Behavioural Research, Vol. 1, pp245-276
- Debenham, J. Clarke, G.P. & Stillwell, J.C.H. (2001a) *Improving geodemographics for business: Adding supply side variables*, Paper presented to the 12th European Colloquium on Theoretical and Quantitative Geography. Saint Valery-en-Caux (France) 7th – 11th September 2001
- Debenham, J., Clarke, G.P. & Stillwell, J.C.H (2001b) *Deriving supply-side variables to extend geodemographic classification*, Working Paper no. 05/01, School of Geography, University of Leeds
- DETR (1998) Department of the Environment, Transport and the Regions Web Page: 1998 Index of Local Deprivation (Regeneration Research Summary No. 15), URL: http://www.regeneration.detr.gov.uk/98ild/indicate/htm (accessed 12/12/2001)
- Field, A. (2000) *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*, SAGE Publications, London
- Jolliffe, I.T. (1972) *Discarding variables in a principal component analysis, I: artificial data*, Applied Statistics, Vol. 21, pp160 - 173
- Kaiser, H.F. (1960) *The application of electronic computers to factor analysis*, Educational and Psychological Measurement, Vol. 20, pp141 – 151
- Martin, D., Senior, M.L. and Williams, H.C.W.L (1994) *On measures of deprivation and the spatial allocation of resources for primary health care*, Environment and Planning: A, Vol. 26, pp1911-1929
- McGhie, C. (1999) *"Digs that cut the rising cost of learning"*, Daily Telegraph 21/08/1999
- MIMAS (1999) Census Dissemination Unit Web Page: Deprivation scores based on 1991 area statistics, URL: http://census.ac.uk/cdu/Datasets/1991_Census_datasets/Area_Stats/Derived_data/Deprivation_scores/ (accessed 12/12/2001)

- Openshaw (1984) *The modifiable areal unit problem*, Concepts and Techniques in Modern Geography 38, Geo Books, Norwich
- Openshaw, S. (1994) *Developing intelligent geodemographic targeting systems*, Working Paper no. 94/13, School of Geography, University of Leeds
- Phipps, R., Shedd, J., Merisotis, J. and Carroll, D. (2001) *A classification system for 2-Year Postsecondary Institutions: National Centre for Education Statistics Methodology Report June 200*, U.S. Department of Education Office of Educational Research and Improvement
- Remmel, T. (2000) *Landscape structure dependence of various song birds in southern Ontario* Unpublished PhD Term Project, Department of Geography, University of Ontario. Available on-line URL: http://eratos.erin.utor onto.ca/remmelt/ggr1304/BirdProject.pdf (accessed 14/11/2001)
- Röder, K. (2001) *SPSS for environmental statistics*, In Ah Poe, A. (ed.) Statistics for environmental policy, Munich Centre for Economic Environmental and Social Statistics, Munich
- Simpson, S. (1993) *Areas of stress within Bradford District: a report from the 1991 Census and other sources*, report published by Research Section of the Strategic Management Unit, City of Bradford Metropolitan Council, Bradford
- Simpson, S. & Yu, A. (2001) *Updated UK Area Masterfiles: Full report of research activities and results*, University of Manchester, Manchester (available on-line URL: http://les1.man.ac.uk/ccsr/rschproj/lookup.htm#project)
- Slade, J. (2001) *"Bricks and mortarboards"*, Daily Telegraph 18/07/2001
- Sleight, P. (1997) *Targeting customers: How to use geodemographic and lifestyle data in your business*, 2nd edition, NTC Publications Ltd., Henley-on-Thames
- Voas, D. and Williamson, P. (2001) *The diversity of diversity: a critique of geodemographic classification*, Area, Vol. 33.1, pp63-76

Appendix – MOSAIC *'Group'* Pen Portraits

Group A – High Income Families

High Income Families are found in the more affluent and leafy suburbs, where professionals and wealthy business people can afford to live in highly priced, large, owner occupied housing. These are typically family neighbourhoods, where inter-war and post war houses tend to have four or more bedrooms and generous gardens.

First time buyers or pensioners are largely absent as *High Income Families* are dominated by two-income, two-upmarket car households, many with older children. Levels of educational and professional qualifications are particularly high and many people have accumulated substantial amounts of capital.

Group B – Suburban Semis

Suburban Semis represent the bedrock of middle class suburban taste. Within these neighbourhoods are found middle aged, middle income families, where parent often commute to work in middle management jobs in large service organisations. Living in satellite villages or in well established suburbs, these people live organised and agreeable lives and have time and income to pursue a wide variety of home based leisure interests.

Most are owner occupiers and have had children; many of the houses are inter-war semis with their own garages and reasonably sized gardens. Some neighbourhoods of *Suburban Semis* are becoming increasingly multi-cultural.

Group C – Blue Collar Owners

Blue Collar Owners comprises the less expensive neighbourhoods of owner occupier housing where skilled manual and junior white collar workers take pride in the exercise of practical skills in the home and garden. These are unpretentious rather than intellectual communities. Where sensible and self-reliant people have worked hard to achieve a comfortable and independent lifestyle.

Relatively few ethnic minorities or single people reside in *Blue Collar Owners.* Most occur in traditional industrial regions and many where council estates have been sold off to long standing tenants. Children tend to leave school early to get a job, whilst continuing to live at home. Family incomes are relatively high due to the large number of adults working and the absence of expensive mortgages.

Group D- Low Rise Council

Low Rise Council comprises neighbourhoods of Local Authority and Housing Association tenants who, for various reasons, have not exercised the right to buy their homes. Reasons may be that their wages are too low, they are retired or that they live in areas of the country, such as central Scotland, where the drive to own your own home is less pronounced than it is in England.

Most of these estates were developed to high standards of design in the 1930s and 1950s. They consist of two or three bedroom, two storey houses, typically built in closes and cul-de-sacs, and mostly at low residential densities. Many of the tenants are now middle aged or older; incomes are generally low, partly because there are many families where no one is at work. Despite high car ownership, many families are still dependent on public transport. Local shops are often expensive and supermarkets difficult to reach.

Group E – Council Flats

Council Flats are neighbourhoods that include high rise flats, large municipal overspill estates and smaller developments of Local Authority maisonettes and mid-rise dwellings. They have very low incomes and aspirations, where watching television is often the principal form of leisure activity and where consumers are often unable to afford more than basic brands and products.

Council Flats have a high demand for consumer credit and mail order is used for the purchase of durable products. Within these estates reside large numbers of pensioners, single parents, long term sick and unemployed. Few people have formal educaitonal qualifications and for many mobility is impaired by lack of a car. Much government money has been spent over the years in attempts to improve the social and physical environment in these neighbourhoods, which often suffer from high levels of crime and vandalism as well as financial poverty.

Group F – Victorian Low Status

Many *Victorian Low Status* neighbourhoods contain areas of genuine community feeling, where young families and the childless elderly live in owner occupied and privately rented terraces and tenements, often dating from the last centruy. These older established communities often lie close to the centre of large towns and offer less formal and pretentious environments than more recently built suburban areas.

Whilst offering high levels of local social contact, such neighbourhoods allow their residents to experiment with diverse lifestyles. Ownership of "lifestyle" products is less likely than elsewhere to impress the neighbours. Many *Victorian Low Status* are found in small towns which industrialised rapidly in the 19th century. They are also common in the older cores of large cities, many of which are becoming subject to gentrification.

Group G – Town Houses & Flats

Town Houses & Flats consists mostly of small properties providing middle income housing for junior administrative and service employees who don't have large families. Such neighbourhoods are found typically in small market towns and service centres, in the older areas of historic towns and in some inter-war suburbs of London. In some instances, larger houses in the better parts of smaller towns have been divided into small self-contained rented flats. Elsewhere big old houses have been demolished and replaced with privately owned flats.

Small market towns have much of this type of commu nity as do turn-of-the-century suburbs of high density terraced housing in London, designed originally for clerks and junior managers in service jobs. Today *Town Houses & Flats* comprises people who typically use inter-personal skills in service jobs rather than craft skills in industry, who are well-informed and sociable in their lifestyles, and whose aspirations centre primarily around material possessions.

Group H - Stylish Singles

Stylish Singles are people for whom self-expression, exploration, style and tolerance are important. Some are still students, others are highly paid young professionals in the service sector. Typically very well educated and very involved in their work, these people are highly aware of the behaviour of different social groups and enjoy living in a diverse, cosmopolitan and sometimes multi-cultural environment.

People are often so busy experimenting and experiencing life that they delay marriage and delay as long as possible the responsibility of looking after homes, gardens and children. *Stylish Singles* prefer the vitality of the large city to the tranquillity of outer suburbs and spend money freely on fashion, foreign travel, the arts, entertainment and eating out. Experiences are often valued more highly than material possessions.

Group I – Independent Elders

Independent Elders comprises neighbourhoods of owner occupied houses, bungalows and privately owned flats dominated by people over the age of 55. Many are found close to the sea, particularly along the South

Coast, as well as in some of the inter-war London suburbs. Here the people are conservative and self-reliant in outlook, but still fit enough to look after themselves. Income levels vary quite considerably within this group depending on age and the extent of private and company pensions.

Many *Independent Elders* depend on investment income – which despite being relatively low is sufficient as outgoings are modest since most houses are owned outright, children are grown up and homes are adequately supplied with consumer durables.

Group J – Mortgaged Families

Mortgaged Families mostly reside in areas of recently built private housing typically lived in by younger households often burdened by high levels of mortgage repayment. Whilst most of these neighbourhoods contain young families living on the outskirts of towns and cities, and increasing number of young people and childless couples are in this group, often living in new in-fill housing.

The furnishing and decoration of homes and gardens is a key focus of *Mortgaged Families*. Leisure activities and shopping trips are undertaken by the entire family to retail multiples in newly developed retail parks in out-of-town locations.

Group K – Country Dwellers

Country Dwellers consist of genuinely rural neighbourhoods, beyond the commuter belt of villages with their newly built estates, where houses have names rather than numbers and where agriculture and tourism are significant sources of local employment.

They vary considerably in their levels of affluence – from the gentrified villages of the New Forest and Sussex Weald, through the 'Ambridges' of the Midland Shires, to the impoverished upland farms of the Celtic fringes. All suffer poor access to shops, post offices, schools, medical services and entertainment and are heavily dependent on cars for work and leisure.

To most *Country Dwellers* small scale is still beautiful; people are expected to help their neighbours and many attempt to hold out against the depersonalising aspects of a mass consumption society.