WORKING PAPER 05/01

DERIVING SUPPLY-SIDE VARIABLES TO EXTEND GEODEMOGRAPHIC CLASSIFICATION

James Debenham,
Graham Clarke &
John Stillwell

PUBLISHED NOVEMBER 2001

For further copies contact the Working Paper Secretary,
School of Geography, University of Leeds, Leeds, LS2 9JT
Telephone 0113 233 3300

CONTENTS

ABSTRACT

The traditional proprietary geodemographic information systems that are on the market today use well-established methodologies. Demographic indicators are selected as a proxy for affluence and are then often linked to customer databases to derive a measure of the level of consumption expected from the different area typologies. However, these systems ignore fundamental relationships in the retail market by focusing upon demand characteristics in a 'vacuum' and ignore the supply side and consumer-supplier interaction.

This paper argues that there may be considerable advantages to including supply-side indicators within geodemographic systems. Whilst the term 'supply' in this context might imply the number of consumer services already in an area, equally important for understanding demand are variables such as the supply of jobs and houses. We suggest that profiling an area in terms of its labour market characteristics gives a better insight into the income chain while the supply of houses could be argued to be a crucial factor in household formation that in turn will impact upon demographic structure. Using the regional example of Yorkshire and Humberside in northern England, we indicate how a suite of supply-side variables relating to the labour market can be assembled and used alongside a suite of demand variables to generate a new area classification. Spatial interaction models are calibrated to derive some of the variables that take into account zonal self-containment and catchment size.

Keywords: geodemographics, supply-side, interaction

LIST OF TABLES

LIST OF FIGURES

DERIVING SUPPLY-SIDE VARIABLES TO EXTEND GEODEMOGRAPHIC CLASSIFICATION

1 INTRODUCTION

Geodemographics involves the *"classification of small areas according to their inhabitants"* (Rothman, 1989, p1) and geodemographic systems are built upon the principle that two people who live in the same neighbourhood are more likely to have similar characteristics (and consumption behaviour) than two people chosen at random. The widespread commercial use of geodemographics as a mechanism for analysing consumption patterns may be attributed to the fact that they produce what Beaumont and Inglis (1989) refer to as *"actionable information – information viewed strategically as an important resource or asset"* (p. 587). One of the defining features of geodemographic systems hitherto has been the demand-side nature of the variables included in the systems, i.e., demographic and social characteristics of the population that result in different propensities to consume different products or services. As the most reliable and comprehensive source of socio-demographic information on small areas in Britain, the Census of Population has been very important. Whilst all the major proprietary systems now make use of non-census variables (such as county court judgements, credit applications or the electoral registers), geodemographic typologies are still almost entirely based upon the demand in a given area that is determined by the characteristics of the resident population on census night.

Our contention is that the traditional systems pay no attention to the supply-side characteristics of the market that also vary spatially, and therefore we argue that existing systems might not be totally fulfilling the criteria of business need. We suggest that no indication is given of the

economic, social or environmental conditions that might influence the consumption of retail goods and services in an area, let alone how these conditions might change. Yet it is clear that areas with good employment opportunities, housing provision and environmental conditions are likely to be areas where demand for goods and services is buoyant. In contrast, areas lacking in jobs, with low levels of housing development and poor environmental quality are much less likely to be areas identified as having the potential for business exploitation, unless they are likely targets for gentrification or policy-focused regeneration. Thus, it seems appropriate to extend the traditional framework of geodemographic systems to include other variables that indicate the potential of an area. This means drawing upon information about the level of employment, the provision of housing and the condition of the environment in an area, as well as details of the existing infrastructure of retail and service facilities. Furthermore, there is a strong argument that the suite of variables used to classify areas should include the dimension of change over time. The process of population decentralisation, for example, could render a once attractive looking investment totally unworkable in 15 to 20 years. Alternatively, spending power might be drastically reduced if an area suffers the closure or downsizing of a major employer or industrial establishment. These examples indicate that, in addition to static or cross-sectional measures pertaining to demand and supply, it may be appropriate to take into account temporal dynamics in the spatial system through the inclusion of variables such as population and employment change.

In this paper, we focus on the demand side of the labour market and aim to construct a suite of variables that represent characteristics of the local economy in the areas used for classification.

Only static variables are defined in this instance. It is important to acknowledge that conventional classifications in business geodemographics, such as MOSAIC or Superprofiles, include census variables indicating the employment characteristics of the residential population in a small area. The variables are usually residence-based since the Census is a survey of households. However, the Annual Employment Survey (formerly the Annual Census of Employment) provides information about the jobs provided in a zone and therefore it becomes possible to define the characteristics of the zone according to the people who work there rather than those who live there. Workplace-based statistics indicate the nature and pattern of the jobs and the establishments that are available across the zonal system, regardless of where the employers or employees are living. Moreover, indices of specialization can be computed that provide summary measures of the employment structure of zones and give some indication of the extent to which a zone is dominated by one industrial sector.

The addition of workplace characteristics therefore adds a further dimension to the classification of small geographical areas based on residential characteristics. However, whilst these 'stock' variables provide evidence of how the levels of population and employment variables vary spatially, they do not give any indication of the relationship between residential areas and workplaces. Some zones may attract large numbers of workers from across a wide area; other zones may provide jobs in their workplaces for those living within the locality; other zones may have no dwellings or no workplaces at all. Consequently, we suggest that a further set of variables should be derived that measure the interaction that each zone has with other zones in the same system. These variables reflect concepts such as zonal self-containment and catchment

size, both of which require careful definition and modelling based on Special Workplace Statistics (SWS) from the Census of Population.

Thus we contend that supply-side stock variables, and variables that indicate the degree of interaction that zones experience with those that surround them, may prove to be measures that would enhance the segmentation that geodemographics systems deliver. The aim of this paper to build upon these ideas by proposing a series of variables that might be included along with those representing population characteristics when constructing a new zonal classification that is based on a postal geography, the spatial units preferred by businesses.

We begin with a brief synopsis of some of the key developments in the evolution of geodemographics over the last two decades and short review of problems associated with the packaged solutions that have been on offer (Section 2). These comments provide a context in which to focus on the derivation of a set of new variables for inclusion with the more traditional set of demand variables that are introduced in Section 3. The spatial units selected for the clustering algorithm are the set of postcode sectors in Yorkshire and the Humber, the largest of the English planning regions. Some additional variables measuring residential property transactions by postal sector also introduced in this section.

Section 4 of the paper focuses on employment, showing the types of data that are available for direct incorporation as well as a synthetic index that represents the extent of specialisation in the industrial structure of each zone. The use of postcode sectors as the spatial units presents some

problems when the interaction variables are computed since the Census journey-to-work data sets are only available for flows between and within Census wards. In Section 5 of the paper, we therefore explain why spatial interaction models are used and how they are calibrated. Destination-specific distance decay parameters are mapped to show their spatial variation and used to estimate flows between postcode sectors that are subsequently incorporated in the calculation of zone self-containment and catchment size.

In Section 6, we present a classification system that includes both demand variables and those variables that represent the labour market and the interactions between residence and workplace zones. The K-means method of clustering is explained and the clusters that are identified are mapped and interpreted. Finally, some conclusions are contained in Section 7.

2 CONTEMPORARY GEODEMOGRAPHIC SYSTEMS

2.1 Developments in geodemographics

The summer of 1993, when the results of the 1991 Census were disseminated, was a 'watershed' (Sleight, 1997) in the evolution of geodemographic systems, reviewed extensively by Batey and Brown (1995) and Birkin (1995). By this time *"the geodemographics industry had matured substantially and those who were involved ... had a much clearer view about what they needed"* (Batey and Brown, 1995, p. 89). The 1991 Census itself was seen as a great boost to the geodemographics industry because of the inclusion of questions relating to dwelling type and 'life stage' represented by household structure. In addition, the release of a directory linking census enumeration districts (EDs) to postcodes provided a much more accurate link between

census and postal geography than the 'proximal matching' approach that had been used previously in constructing the 1981-based systems (Raper *et al.*, 1992). Sleight (1997) reports that six companies registered with the Office of Population Censuses and Surveys (OPCS) to become 'census agencies', thus having the right to hold and handle the 1991 Census data: CACI Ltd, CCN Marketing, CDMS Ltd, Pinpoint Analysis, Equifax Europe (UK) Ltd and Infolink Decision Services Ltd. EuroDirect, a census licensee rather than an agency, joined the suppliers in the early 1990s and launched two new systems called 'Neighbours' and 'PROSPECTS'. Rationalisation in the late 1990s has seen the market move back towards its mid 1980s situation with four major general purpose geodemographic systems available: 'ACORN' from CACI, 'MOSAIC' from Experian, 'CAMEO UK' from EuroDirect and 'SuperProfiles' from Claritas UK.

The post-1991 product range also marks something of a watershed in the amount of available information about the variables included in these systems and the methods used to define the clusters. The development of MOSAIC, SuperProfiles and ACORN (to an extent) after the release of the 1981 Census data were reasonably well documented in the literature. However, after 1991, the 'secrets' of these proprietary systems become more guarded. Birkin (1995) reports that the 1991 version of the ACORN classification allocated the 147,000 or so census EDs to one of six neighbourhood 'Categories', each with its own shorthand label: *'thriving'*, *'expanding'*, *'rising'*, *'settling'*, *'aspiring'* and *'striving'* (CACI, 1993). The categories could then be dissagregated into 17 ACORN 'Groups' and then 54 ACORN 'Types'. This

segmentation is very different from the 1981-based 11 'Groups' from 38 clusters yet there is no literature in the public domain to explain these decisions.

There have been a number of notable theoretical or product developments during the 1990s. The first has been the inclusion of far more non-census variables, especially associated with data sources relating to lifestyles. Lifestyles systems involve the collection and classification of enormous databases derived from consumer surveys. In contrast to geodemographic systems that might use a handful of non-census variables to supplement the information in the census-based clusters, lifestyle systems are totally independent of census data. Consumer Surveys ltd., NDL and Experían are among the biggest collectors of lifestyle data in the UK, using postal surveys (Experían and Consumer Surveys ltd.) or product registration guarantees (NDL). Sleight (1997) reports that the 'critical mass' of lifestyles information has risen considerably in the latter half of the 1990s with the larger companies claiming to have anything between 8 and 15 million records in their databases.

The advantage of lifestyles systems is the immediacy of the data. Lifestyles companies like to boast that unlike the census, which is updated only decenially, no information in their databases is more than three years old (Sleight, 1997). Lifestyle classification systems therefore have a much more up to date feel about them. One area where lifestyles systems might have a significant advantage over geodemographic systems is with the question of income. Most lifestyles questionnaires ask for income information to be given (normally voluntarily) in the form of reasonably broad income earnings brackets. Webber (1992) argues that the use of a

single income variable from a lifestyles database may have similar discriminatory power to that of an entire geodemographics system based on a large number of census variables.

A second key development has been the growth of the regional or market-specific geodemographic classifications (offered by CACI and EuroDirect in particular). In total, there are 17 such systems on the market in the UK (Table 1).

Experian	EuroDirect	CACI
MOSAIC for Ireland	CAMEO*Northern Ireland*	Scottish*ACORN
Scotland MOSAIC	CAMEO*Republic of Ireland*	UK*ACORN
EuroMOSAIC	CAMEO*International*	
Financial Strategy Segments	CAMEO*Income*	Investor*ACORN
Grocery MOSAIC	CAMEO*Financial*	Financial*ACORN
	CAMEO*Investor*	
	CAMEO*Unemployment*	
	CAMEO*Property*	

Table 1: Regional and market-specific geodemographic systems available in the UK

The immense amount of consumer information that lifestyle databases contain means that the systems are more easily made into product specific systems, thus avoiding the complicated and rather subjective decisions about which census variables best fit together to make a financial market classification system such as *Financial**ACORN or CAMEO*Investor*. Furthermore, because the surveys are georeferenced to individual postcodes, the error prone transition from census geography to postal geography is avoided.

2.2 Shortcomings of geodemographic systems

The development of geodemographic systems has not been without considerable debate about the relative merits and robustness of the procedures. This section attempts to draw upon some of the critical literature surrounding geodemographics (for there is significantly more literature critiquing the classification process than explaining how it is done) to summarise some of the main criticisms and limitations. First, the quality of (Census) data has been called into question. The standard statistical blurring of Census data to ensure confidentiality can introduce sources of error into the enumeration and it is clear that the reliability of the SAS is dependent upon the accurate completion, processing and compilation of the data (Harris, 1999a). Furthermore, it could be said that the census is not totally comprehensive. It has been suggested that anything up to 1 million people are missing from the 1991 census (Simpson and Middleton 1999, Tye 1995). No SAS tables are generated for EDs with less than 50 residents or 16 households and Martin (1999) has suggested that this suppression of data, *"runs counter to many business applications in which an objective is to identify individuals or households with particular characteristics as precisely as possible"* (p.79).

As we have seen already, some geodemographics systems use non-census data to add to the classifications, particularly those more indicative of income such as the credit scoring data used by MOSAIC. Webber (1989), however, reveals that this in itself is not without problems. Non-census data are usually not comprehensive in coverage because they are case specific. For instance, the average value of County Court Judgements (CCJs) cannot be obtained in a large number of postcodes and there are many other with no CCJs to report. It is difficult to think of

easy solutions to these problems. The vendors of lifestyle data claim great rigour in the handling and screening of responses but it is unlikely that any data sets will ever be 100% accurate.

A second major concern has been the *modifiable areal unit problem* (MAUP) and the ecological fallacy: classic problems of geodemographics. The reliance of geodemographics on the relationship between areal characteristics and consumer behaviour makes it very susceptible to such problems. Openshaw (1984a) defines two related components of MAUP. The 'scale problem' is described as *"the variation in results that can be obtained when data for one set of areal units are progressively aggregated into fewer and larger units for analysis"* (p.8). Clustering algorithms group EDs that may or may not be spatially contiguous and the characteristics of an individual ED become lost in the cluster average. In current systems, an ED may only belong to one classification and there is a strong chance that its characteristics may be very different from the descriptive statistics that define the cluster of which it is part. A second component of MAUP is the 'aggregation problem' where, at any given scale, the areal units can be grouped together in different ways. Openshaw (1994b) also refers to this as a combinatorial problem. In geodemographics the 150,000 or so EDs in a national classification may be grouped into the relevant number of clusters in many different combinations and each may result in a different 'geodemographic geography'.

The ecological fallacy is well described by Martin and Longley (1995 p.17): *"put simply, in geographical analysis we cannot ascribe even the very dominant characteristics of areal data to individuals and to point locations within those areas"*. Sleight (1997) describes the ecological

fallacy as the *caveat emptor* of geodemographics and then as the stick used to beat it (1998). Harris' work testing geodemographic typologies in Bristol (1999, 1998a) suggests that the proximal matching technique used to link EDs to postcodes may often be to blame for areas that might find themselves misrepresented or wrongly labelled. *"At least one postcode in Bristol can be classified as an 'Affluent Achievers' neighbourhood or as 'Have Nots' depending upon where the population-weighted centroid of the postcode is taken to be and relating this to the SuperProfiles typology through the ED-to-postcode Directory"* (1999 p.60).

This leads on to a third issue. Birkin and Clarke (1998) point to the cluster titles themselves as a possible source of misrepresentation because of the possible misunderstanding of labels such as 'young married couples'. Such semantics may not always be an issue but there is a problem of prejudiced thinking towards certain cluster names. MOSAIC's use of terms such as *'Coalfield Legacy'* (D13), *'Smokestack Shiftwork'* (C12) and *'Rural Disadvantage'* (K38) all run the risk of promoting negative feelings towards an area (and this is before anyone tries to decipher what a *'Bohemian Melting Pot'* (F24) might be!). Harris (1998b) asserts that often the association between the propensity to buy products and a geodemographic cluster is ambiguous and the discriminatory power of the geodemographic typology vague. Nevertheless, *"it is within these ill-defined boundaries of representation and behavioural clustering that geodemographic systems are able to operate to the (low) demands of commercial users"* (p.21).

So what solutions have been suggested? First, Harris (1998b) and Sleight (1997) both suggest that it would be unwise and a little premature to discard geodemographics on the basis of this

misrepresentation. In many ways, these issues are similar to the age-old criticisms of generalisation that have always surrounded quantitative geography. Harris (1999) argues against Cathelat's (1990) assertion that the demographic, economic and physical criteria of the cluster must totally classify the social individual. While the geodemographics industry takes a positive stance on the association between the individual's consumer behaviour and the wider neighbourhood environment, it is in fact only an association that is asserted. However, as Harris (1999 p.61) points out, *"such an admission requires a rewording of the maxim of geodemographic advertising from, 'we know who you are because we know where you live' to, 'we've got a broad idea about your consumer choices and preferences inferred from the kind of neighbourhood we think you reside in'!"* Such an admission is unlikely although there are signs that it does happen. The ACORN classifications given on www.upmystreet.com come with a warning stating *"while in general ACORN profiles present a reasonably accurate picture, it won't be the case 100 per cent of the time in 100 per cent of the postcodes"*. However concerns with these issues has led to consideration of solutions, most notable from the work on fuzzy geodemographics (see Openshaw, 1994, Feng and Flowerdew 1999, and See and Openshaw, 2001).

A fourth and final issue surrounds the use of demand-side data only. To date, no systems have included supply-side variables to supplement the characteristics of areas only by persons and households. As noted earlier, supply-side variables may help when longer-term dynamics are considered. It is worth repeating why this may be advantageous to certain business users. Using conventional geodemographics, a business user may decide to locate in an area classified as high

income (perhaps Sainsbury's in the grocery market) or low income (perhaps Netto the discount retailer). That, of course, is a useful first step. However, it is likely that a business user would also be interested in long term stability – that is, to be reassured that the area will be similar in spending power terms in 5 or 10 years time. The spending power or income of a locality is as dependent on supply-events as it is on population ageing/migration; for example, population change side (caused by the building of a new housing estate) and/or changes in employment/unemployment (caused by a major disinvestment or investment in an area). To assess the implications of the latter it is important to build a labour market model that links residential areas to employment areas, and to understand the drivers of economic change at the small area level. For example, the dependency of a residential area on one employer can have drastic impacts when that employer no longer operates from that area. Similarly, the profits of a company can also be affected by other more direct supply-side changes: for example, the opening of a brand new shopping centre or discount warehouse. Thus, it is important to supplement those that measure the demand-side with new variables that consider the characteristics of the supply-side.

It is this issue that is the focus of this paper, the remainder of which explores the integration of demand and supply side variables shown in Figure 1 into a new geodemographic classification for Yorkshire and Humberside.

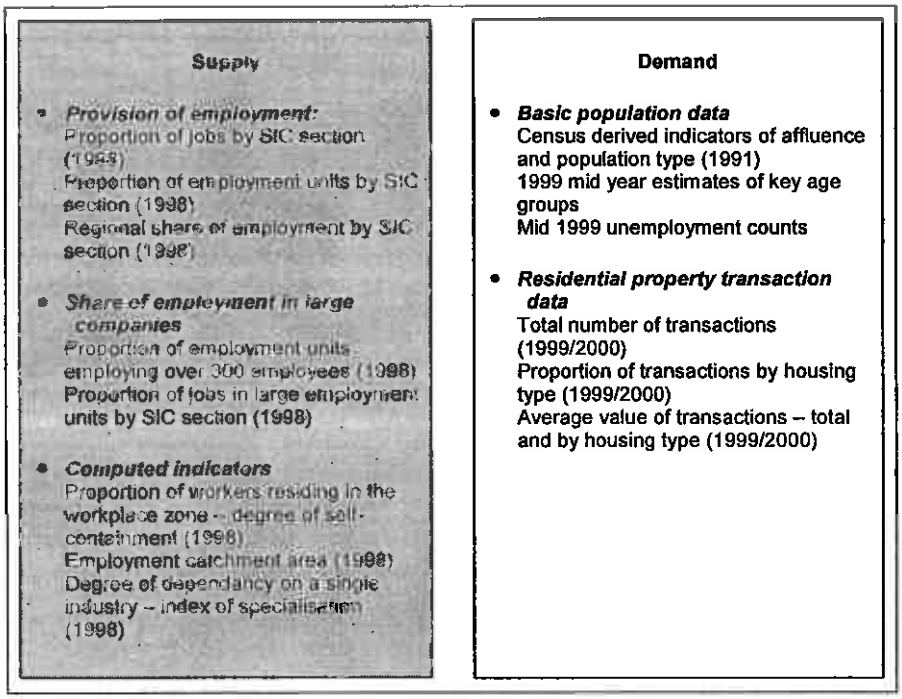


Figure 1: The proposed groups of supply and demand indicators

3 MEASURING DEMAND IN YORKSHIRE AND HUMBERSIDE

Because the main proprietary GDIS are in competition with one another, the companies who market them rarely release exact details of the census variables used in their clustering procedures. We have chosen to use 51 variables to measure demand characteristics. They are similar in nature to those used by the main geodemographics companies. We have been a little more parsimonious in the selection of variables than the geodemographic companies probably would be. For instance, no ethnicity data has been used, nor data from the tables in the 1991 Small Area Statistics that cross-reference variables, such ethnicity and housing tenure. Nevertheless, the variables incorporated here are good proxy indicators for affluence and life stage. The variables are considered following an outline of the geographical units.

3.1 Postal sectors

Businesses usually require geodemographic systems that relate to postal geographies and the variables proposed here are held in the database at the *postal sector* level. Postal geography is different to census geography. It is based upon the administrative structure of the Royal Mail postal service, as opposed to the arbitrary statistical boundaries used by the Office for National Statistics (ONS) for the dissemination of small area census data and by Central Government for the demarcation of local government boundaries. Postal geography is now more widely used as a basis for study in the UK, especially in the private sector (Martin, 1995, Raper *et al.*, 1992), although any analysis using a long time-series can be hampered by regular alterations of boundaries on the ground as development occurs. Postal sectors are areal units created by the aggregation of point postcodes, for which no geographical boundaries exist. The first five characters of a UK postcode (e.g., LS2 9JT) reveal the postal sector (LS2 9). Postal sectors aggregate into *postal districts* (LS2), and then to *postal areas* (LS).

Until recently, the main stumbling block to conducting small area research based on postal geographies was the paucity of available data. However, in the last three or four years a concerted effort has been made make more statistics available, especially for postal sectors. This is in part due to the UK Government's National Strategy for Neighbourhood Renewal, which was launched early last year and detailed the need for more neighbourhood statistics to be made available, especially on-line (ONS 2001a). Furthermore, the methodologies for converting data held at other administrative or statistical units into postal geographies are also becoming more

available, such as the look up tables derived from the All Fields Postcode Directory (AFPD) (Simpson and Yu, 2001).

Although the established geodemographic systems are national (or even European in the case of MOSAIC), the system proposed here is a prototype based on data for the region of Yorkshire and the Humber. This region provides an interesting example of the problems of mismatch between census and postal geographies. One key problem is that it is impossible to select postal sectors that fall neatly within the boundary of region of Yorkshire and the Humber that follows the boundaries of its constituent local authorities (Figure 2). The solution has been to adopt only those postal sector polygons whose centroid falls within the regional boundary. This gives a total of 784 postal sectors and, as Figure 2 shows, a reasonably good match at the boundaries of the region.

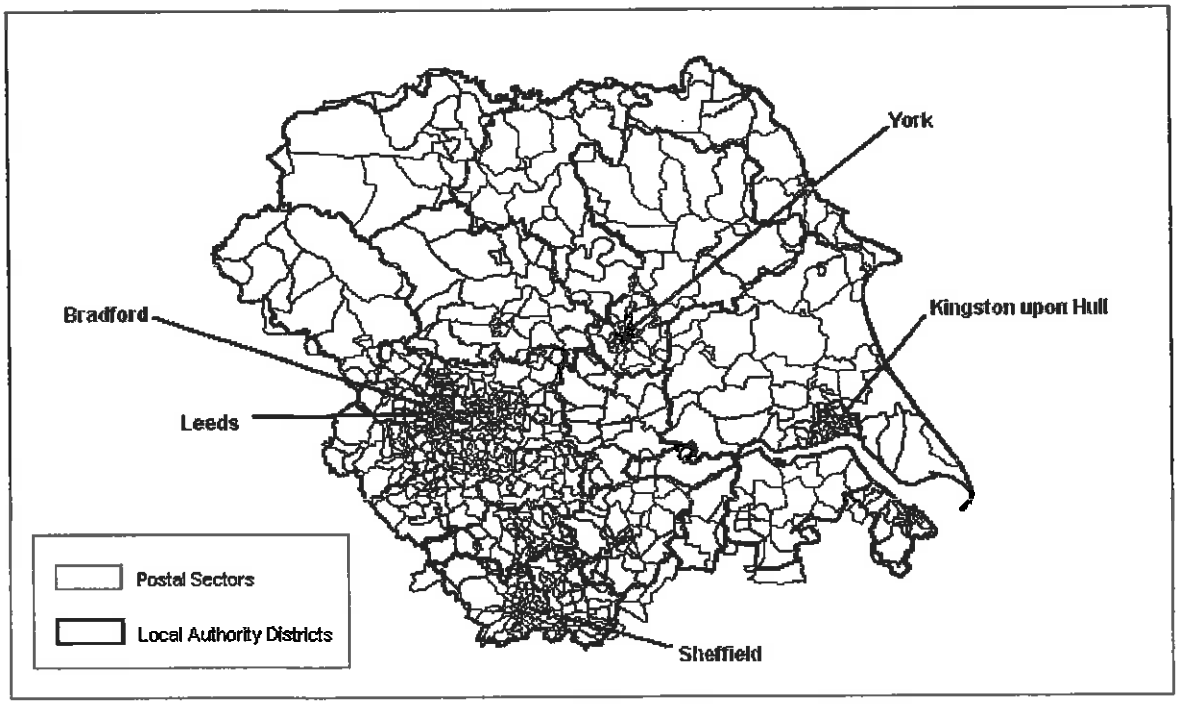


Figure 2: Postal sector and local authority boundaries in Yorkshire and the Humber

3.2 Demand variables

Table 2 indicates the set of demand variables that were used. With the exception of variables 1 to 8 and variable 51, all the data have been extracted from the 1991 SAS available from MIMAS. The key problem here is that census data is not held at the up-to-date postal sector level, but only for the postal sectors as they were on census night in April 1991 (MIMAS, 1999). The conversion of census data to modern postal geographies has long been recognised as a problem (Raper *et al.*, 1992). However, the situation has been made a little easier now thanks to the development of a series of look-up tables from the AFPD for moving between 25 different administrative and statistical geographies (Simpson and Yu, 2001). The SAS data required to compute these variables were obtained at the enumeration district level and converted using this

method. All the variables have been computed as proportions of the total population/households unless otherwise stated.

However, the age structure variables (1 to 8) were taken from the newly released Experían postal sector data that was made available earlier in 2001 through a joint ESRC/JISC agreement. The population estimates are produced using an innovative two-stage process. The first makes estimates at the postcode level using data from such sources as the Electoral Register and the Postal Address File. The second constructs an age profile from a 'weighted-combination' of results from the ageing in situ of the population in the previous year and the maintenance of a constant age profile over time (Experían, 2001). Throughout the whole process the estimates are constrained to JICPOP's (Joint Industry Council for Population Standards) targets for local areas (such as local authority districts), generated using the most recently available central government population estimates and projections (Ibid.).

The only other non-census variable is the unemployment rate (variable 51). This is taken from the Computerised Claimant Count for July 1999, made available through NOMIS. Unemployment data is available from the census tables but this is now 10 years old so the Claimant Count data provides a more up to date picture. The variable here is measured as a percentage of the Experían mid-year estimates of the working age population.

No.	Variable
1.	Persons aged 0-4 (1999 Experian estimates)
2.	Persons aged 5-14 (1999 Experian estimates)
3.	Persons aged 15-24 (1999 Experian estimates)
4.	Persons aged 25-44 (1999 Experian estimates)
5.	Persons aged 45-64 (1999 Experian estimates)
6.	Persons aged 65-74 (1999 Experian estimates)
7.	Persons aged 75-84 (1999 Experian estimates)
8.	Persons aged 85+ (1999 Experian estimates)
9.	Total married population
10.	Single population
11.	Retired (pensioners)
12.	Lone parents
13.	Students (16+) in term-time addresses
14.	Movers last year
15.	Pensioner migrants
16.	Home Owners
17.	Mortgage owners
18.	Privately rented
19.	Rented from Housing Association, Local Authority or New Town
20.	Detached
21.	Semi-Detached
22.	Terraced
23.	Flats
24.	Bedsits
25.	No central heating
26.	Lacking bath & shower
27.	No car
28.	2+ cars
29.	Households > 1.5 persons per room
30.	Households with > 7 rooms
31.	No family household and owner occupied or privately rented
32.	No family household and council rented
33.	Married + cohabiting couple, no children and owner occupied or privately rented
34.	Married + cohabiting couple, no children and council rented
35.	Married + cohabiting couple, dependent children and owner occupied or privately rented
36.	Married + cohabiting couple, dependent children and council rented
37.	Households with two or more families and owner occupied or privately rented
38.	Households with two or more families and council rented
39.	Economically active residents aged 16+
40.	Households with dependants
41.	Self-employed
42.	Households in Social Class I (Professional)
43.	Households in Social Class II (Managerial & Technical)
44.	Households in Social Class III (N) (Skilled Non-manual)
45.	Households in Social Class III (M) (Skilled Manual)
46.	Households in Social Class IV (Partly Skilled)
47.	Households in Social Class V (Unskilled)
48.	Workers with higher degrees
49.	Workers with other qualifications
50.	Persons with Long Term Limiting Illness
51.	Unemployment (claimant count – July 1999) as proportion of working age population

Table 2: The suite of demand variables

3.3 Residential property transactions

Whilst the Census and Experían data sets provide useful information about the population and household characteristics of postcode areas, data on residential property transactions are also available from the Experían database that indicate the numbers of houses sold per quarter and the average price at the postal sector level as documented by HM Land Registry. Data are available from the beginning of 1995 to Quarter 2 of 2000. These data on housing turnover and value (Table 3) 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 in the set of demand variables

No.	Variable
52.	Total number of transactions (Quarter 3 1999 to Quarter 2 2000)
53.	Number of transactions: Detached (Q31999 - Q2 2000) (proportion of total)
54.	Number of transactions: Semi-detached (Q31999 - Q2 2000) (proportion of total)
55.	Number of transactions: Slats/maisonettes (Q31999-Q22000) (proportion of total)
56.	Number of transactions: Terraced (Q3 1999 – Q2 2000) (proportion of total)
57.	Average value of transactions: All (Q2 1999 to Q2 2000)
58.	Average value of transactions: Detached (Q2 1999 to Q2 2000)
59.	Average value of transactions: Semi-detached (Q2 1999 to Q2 2000)
60.	Average value of transactions: Flats/maisonettes (Q2 1999 to Q2 2000)
61.	Average value of transactions: Terraced (Q2 1999 to Q2 2000)

Table 3: Residential property transaction data

It surprising that this data is not used more in mainstream GDIS, although spatial variability in the data may preclude this. The data are only the average values of the houses sold in an area, not an accurate survey of the full housing stock. Some areas will see many transactions while others may have very few and this will affect the averages. The data is dissaggregated by four housing types, detached, semi-detached, flats or maisonettes and terraced houses so variables 52

to 56 not only detail the total number of sales in the 12 months running from mid 1999 to mid 2000 but also the breakdown by housing type. Variables 57 to 61 do the same for the average value of those transactions.

4 SUPPLY SIDE INDICATORS

It is clear that there are many factors that drive retail consumption in addition to the proxy indicators of affluence and population type that geodemographic systems try to create. The level of supply of retail goods and services themselves are important measures but, in addition to this, it could be said that provision of housing supply, infrastructure and employment are equally important factors that can determine the level of potential consumption or stability of development in an area. Furthermore, there may also be a case for the importance of local environment factors and indications of the quality of life. All of these elements can indicate the level of 'buoyancy' in an area and therefore the key components of an area's potential for retail consumption are likely to be more complex than the purely demand based approaches that have been used in the past. In this paper, we concentrate our attentions on the supply side of the labour market.

4.1 The supply side of the labour market

One of the key determinants of the potential of an area for investment is the structure of its labour market. Traditional geodemographic systems might include the number of employees in certain industries that live in a particular place but they do not take into account the actual provision of jobs in that area. It can be argued that employment is one of the key factors

underpinning retail consumption levels because it provides the income that creates the opportunities to spend. The conventional view of the labour market is based upon the notion that the workforce sells their *supply* of labour to employers who *demand* this factor of production to make goods and services to sell for a profit. However, it may be possible to view this situation another way. Since income from employment provides the means to consume retail goods, then it follows that the population will *demand* the jobs that the workplaces *supply*.

A large supply of jobs in an area may suggest a buoyant local economy while the specific industry in which these jobs are provided might indicate likely income levels. However, the dependence of an area on one industry might indicate a vulnerability to decline if national or even international economic conditions promote a collapse of such industries. The Office for National Statistics recently published figures showing that UK manufacturing output in the second quarter of 2001 fell by 2% from the previous quarter (ONS 2001b), prompting fears that the industry was heading into a recession. If such a situation were to arise and jobs were threatened in the same way as they were in the early 1990s then it would be reasonable to expect the investment potential of areas that are heavily dependant upon manufacturing to be depleted. The supply side variables proposed here are designed to test for such dependencies, while also building up a picture of the employment characteristics of an area. Table 3 details the suite of supply-side labour market variables used in this classification.

Basic labour market data is provided using the Annual Employment Survey (AES) that has been downloaded from the National Online Manpower Information System (NOMIS). Since 1995,

the AES has been the principal source of data on employee jobs and is the result of an annual survey of a maximum of 125,000 business enterprises. This sample data is used to estimate employment in non-responding or non-sampled businesses (NomisWeb Reference Centre, 2001). Before 1995, the AES was known as the Census of Employment and was undertaken biannually.

AES data for 1998 was obtained at the 'Section' level of the UK Standard Industrial Classification (SIC92). There are 17 SIC Sections in total but only 15 were included as follows:

- **Section A** (Agriculture, Hunting and Forestry)
- **Section B** (Fishing)
- **Section C** (Mining and Quarrying)
- **Section D** (Manufacturing)
- **Section E** (Electricity, Gas & Water Supply)
- **Section F** (Construction)
- **Section G** (Wholesale/ Retail Trade: Repair etc.)
- **Section H** (Hotels & Restaurants)
- **Section I** (Transport, Storage & Communication)
- **Section J** (Financial Intermediation)
- **Section K** (Real Estate, Renting Business Activities)
- **Section L** (Public Administration/ Defence; Social Security)
- **Section M** (Education)
- **Section N** (Health & Social Work)
- **Section O** (Other Community/ Personal Service)

because Section P (Private Households with Employed Persons) and Section Q (Extra-Territorial Organisations and Bodies) have no employees in the region.

Two sets of 15 variables are derived from this data set: the proportion of jobs by SIC section in each postal sector and the proportion of employment units by SIC and postal sector. These variables are designed to characterise the provision of labour in each postal sector. The number of employment units is used in conjunction with the number of jobs to try to highlight areas dominated by small or large employers. Areas with a large number of employees yet only a small number of employment units in a given SIC section might indicate a dependency on a large employer. However, it should be noted that this comparison is not perfect because the term 'employment units' does not refer to an individual business *per se* but the individual site from where a business operates. Some confusion is therefore possible in areas of high 'job density' such as industrial estates or business parks.

In addition to these two basic indicators, the share of the regional workforce in each section is also calculated for each zone along with the total proportion of employment. This may help identify particularly large employment zones.

4.2 Share of employment in large companies

In addition to the employee analysis, the AES also provides a workplace analysis that gives the number of data units and employees in each postal sector broken down by the size of the unit and by industry (SIC). Therefore a variable was included that measured the proportion of units employing over 300 employees, as were 13 further variables that quantify proportions of jobs in large employment units (over 300 employees) in each of the 15 SIC sections - with the exception of Agriculture and Fishing (sections A and B) which have no units employing over 300

employees in the region. These variables are therefore used to inform of the presence of particularly large employment units in an area, providing a more definitive indication than the relationship between the number of jobs and the number of units described above as it is possible that these might get separated in the segmentation process.

Essentially this indicator will show where areas have a high dependency upon large employers in particular industries and will serve as a warning should that industry or company be in decline. Past experiences in the region have shown the devastating effects of the loss of a major employer to the local economy such as the decline of the steel industries in and around Sheffield, the closure of the coal mines in the Yorkshire coalfields and the loss of textile jobs in Leeds and Bradford. The downsizing of Vickers in Leeds is one example of employment change exerting a major influence on the local community. Such dependency upon a certain industry can be quantified using a pair of indicators that have also been built into the classification system.

4.3 Index of specialisation

Stillwell and Palmer (1986) describe the index of specialisation as a variant of the index of dissimilarity, which is often used in population geography to measure such things as the residential segregation of ethnic populations within cities (Rees and Birkin, 1983). The index of specialisation for zone j can be defined as one of two indices, $SP_j^{(A)}$ and $SP_j^{(B)}$ that represent the characteristics of employment in the workplace. The first indicates the extent to which the

structure of employment by sector k in zone j, E_j^k, differs from the employment structure of the entire system (excluding zone j) and is defined as:

$$SP_j^{(A)} = 0.5\sum_k \left| \frac{E_j^k}{\sum_k E_j^k} - \frac{\sum_j E_j^k}{\sum_j \sum_k E_j^k} \right| \qquad (1)$$

The second index compares each zone with all the other zones (excluding zone j) and is derived from the first index as:

$$SP_j^{(B)} = SP_j^{(A)} \Big/ 1 - \frac{\sum_k E_j^k}{\sum_j \sum_k E_j^k} \qquad (2)$$

These two variables are designed to pick out any zones that are heavily dependent upon one industry. To a certain extent, these two variables are just a compression of the basic labour market variables. However, it is possible that, with all the other variables, the clustering algorithm might be pulled away from the structure of employment provision in each zone. It is therefore hoped that these variables might draw the segmentation process back to areas with a high dependency on particular industries. This may serve to give warning of a dependence upon a particular industry and thus warn of the likely impacts of closures or downsizing. However, the impact of closure depends to a large extent on where the workforce lives.

5 MODEL-BASED INTERACTION VARIABLES

The employment variables may tell us a lot about the extent of the provision of employment in a given postal sector but there is very little indication of the level of interaction between the jobs and the population who might demand them. The impact of changes in the labour market on the surrounding areas can only be 'guestimated' using such indicators of employment provision because they are calculated with little reference to other zones and only use data on the geographical variation in the level of supply (Clarke and Wilson, 1994a; 1994b). Bertuglia and Rabino (1994) maintain that a model-based approach can be employed to reduce this reliability on 'one-dimensional' indicators. By focusing on the performance of zones in a system rather than individually, the level of *interaction* and *interdependence* can be measured. This can be assessed by operationalising a number of 'performance indicators' developed by Clarke and Wilson (1994) and Birkin *et al.* (1994) that are specifically designed to analyse the journey to work flows in an interaction matrix.

5.1 Modelling the journey to work

In order to compute these indicators, we need to know the volume of movement between zones in the system. No journey to work data for postal sectors exist so workflows in the region have to be simulated using a spatial interaction model. This model is calibrated using ward to ward flows obtained from the 1991 Census Special Workplace Statistics and the estimated beta values derived from this calibration used to generate postal sector to postal sector flows.

Since we know the number of jobs provided in each zone, a simple residential location model can be used to allocate individuals who work in workplace zone *j* to residences in residence zone *i*. This residential location model takes the form of the attraction constrained spatial interaction model originally proposed by Wilson (1971). An attraction (destination) constrained model is used because the number of jobs in each zone serves as observed data on the total in-flow into each workplace. We want every job to be accounted for and this is ensured using the balancing factor, B_j^k. The model takes the form:

$$T_{ij}^k = B_j^k W_i D_j^k d_{ij}^{-\beta_j^k} \qquad (3)$$

where:

T_{ij}^k = flow from residence *i* to jobs in workplace zone *j* in industry *k*,
W_i = origin attractiveness factor (population at working age),
D_j^k = number of jobs in workplace zone *j* in industry *k*,
B_j^k = destination balancing factor,
d_{ij} = distance from zone *i* to zone *j*, and
β_j^k = destination-specific distance decay value for industry *k*.

The power distance function on the far right hand side of the equation measures the distance-decay effect of the attraction between two zones. In this case the physical distance between two zones was measured by calculating the Euclidean distance between the centroids of the two zones using the equation:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (4)$$

where x is the easting co-ordinate and y is the northing. Using this method, intra-zonal distances are calculated as zero, i.e. when the origin and destination zone are the same (the diagonal of the matrix). Duley (1989) computes the intra-zonal distance as the radius of a circle with the same area as the zone in question. A similar method was used here although intra-zonal distance was taken as being half the radius of the circle. Many urban postal sectors, by their very nature as aggregations of postcodes, are small and obscurely shaped. Using a whole radius suggests that people are always travelling half way across the zone in question to reach their destination. While this may be true in some instances, the intention of these interaction indicators is to pick up the local flows that might suggest a self-contained workforce. Thus, using half the radius of the circle replicates a smaller local travel to work distance.

Calibration was achieved using routines adapted from the Inter-area Migration analysis and Projection package (IMP) developed by Stillwell (1984, 1991). An iterative Newton Raphson automatic search routine is used to find a best-fit beta value on the basis of the difference between the average trip distance in an observed matrix and that predicted by an attraction constrained spatial interaction model. In this case the observed matrix was the ward-level journey to work flows.

As we have shown, postal and census geographies are not coincident. However, postal sectors and wards are roughly similar in size, both demographically and physically. The main difference

comes in the urban areas where postal sectors are much smaller than wards due to the density of postcodes. As these postcodes are mostly business addresses, these postal sectors have very small populations. Nevertheless, we made the assumption that the beta values calculated for a system of wards will be appropriate to use for predicting flows for a system of postal sectors.

The commuting data in the 1991 SWS are dissaggregated into 11 industrial groups that are not exactly equivalent to the 15 SIC sections used so far in the creation of the basic labour market indicators. Therefore the SWS data and the AES data had to be amalgamated to create nine *'interaction groups'* that adequately represented the characteristics of both datasets without combining dissimilar industrial groups. Figure 3 outlines the two disaggregated datasets and shows the newly created interaction groups.

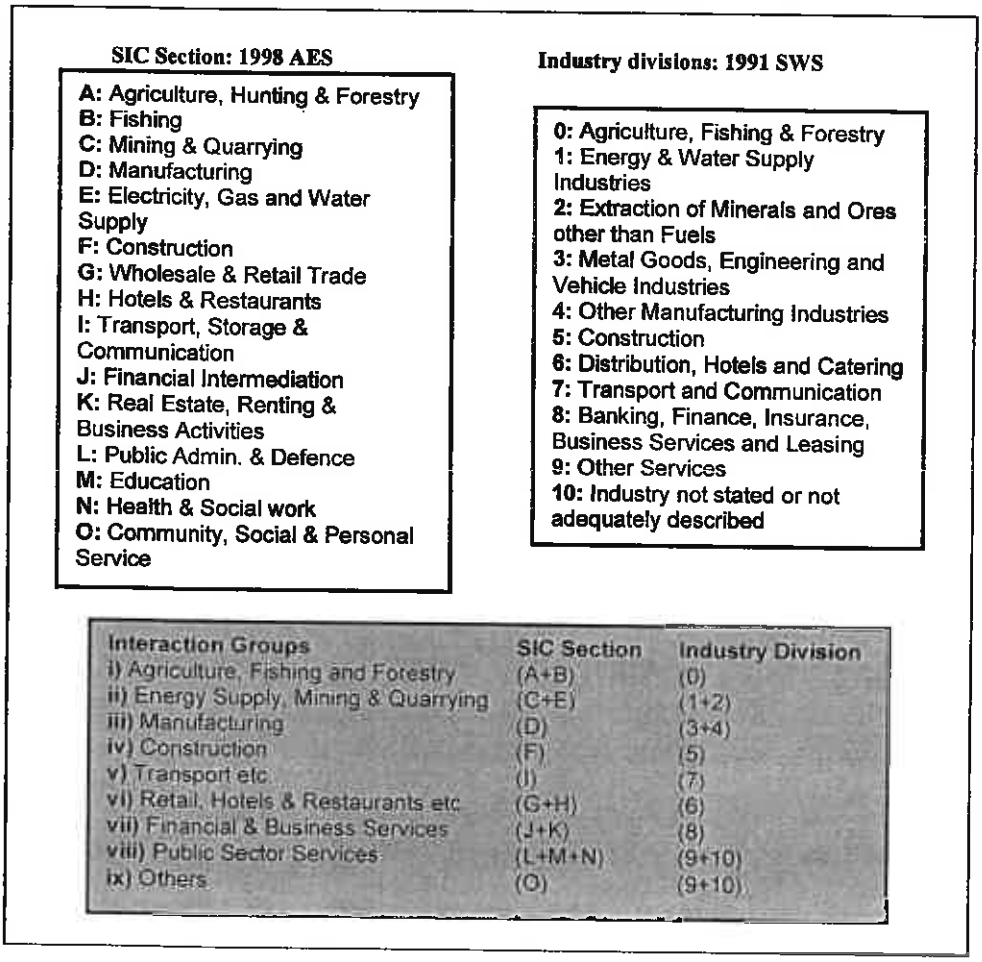


Figure 3: New interaction groups created from amalgamations of AES and SWS disaggregations

The procedures derived from IMP allow the creation of destination-specific beta values for each of the 626 wards in Yorkshire and the Humber. This was done for each of the nine interaction groups. The spatial distribution of the decay parameters for interaction group iii (Manufacturing) is illustrated in Figure 4 and indicates how the frictional effect of distance on travel to work for behaviour is more significant in rural areas for manufacturing jobs.

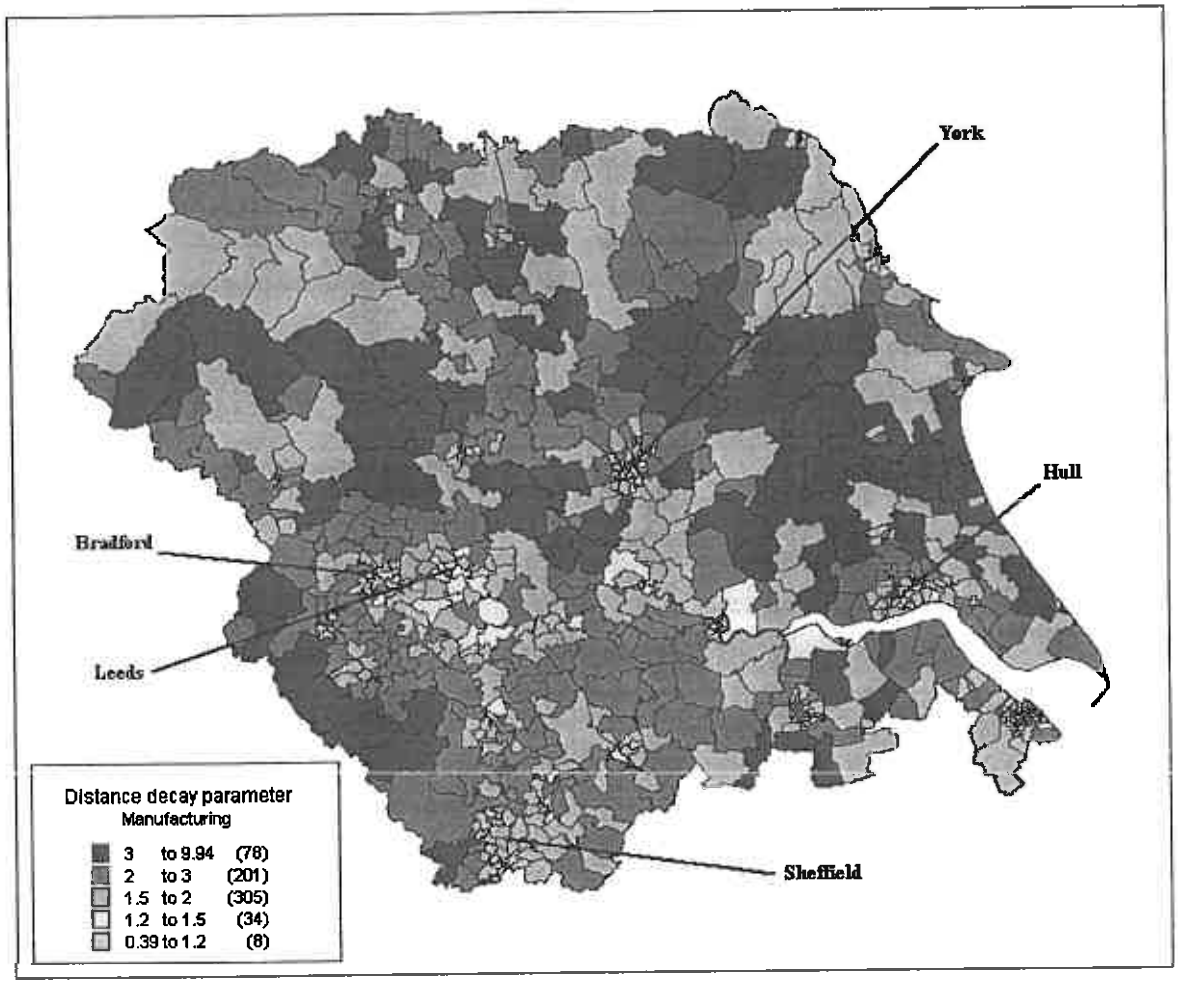


Figure 4: Destination-specific decay parameters for Manufacturing

The ward-level beta values were assigned to the postal sectors using a point-in-polygon search in a GIS to determine which postal sector centroids fall within each ward. With these beta values assigned, the residential location model could be run to create the large array T_{ij}^{k}. The origin zone attractiveness factor (W_i) was created using the working ages of the 1999 mid year population estimates from the Experían postal sector data set. The number of jobs in each interaction group k in each workplace zone j, D_j^k, was obtained by aggregating the AES as

detailed in Figure 3. The predicted flows between postal sectors can be used to compute indicators of self-containment and catchment size.

5.2 Self-containment and catchment size variables

The degree of self-containment measures the extent to which jobs in zone j in industrial sector k are taken up by residents from within that zone. It is formally described as:

$$SC_j^k = \frac{T_{jj}^k}{E_j^k} \qquad (5)$$

where E_j^k is the number of jobs in industry k in zone j (equivalent to D_j^k in equation 3).

Nine variables were created when the degree of self-containment was calculated for each of the interaction groups in each workplace zone. This indicator is specifically designed to look for postal sectors that are heavily dependent upon the supply of jobs in one industry. A high degree of self-containment will suggest a local workforce that will be very vulnerable to changes in that industry and may serve as a warning that such an area might not be so stable for investment under certain economic conditions.

Very few zones will be totally self-contained. Most destination zones will see workers arrive from a number of different locations. Therefore, in order to predict the impact of changes in the labour market upon the surrounding area, it is important to know how far such effects will be

felt. Estimating the average distance travelled by persons working in the zone will allow us to do this and give an idea of catchment size. Nine more variables were therefore calculated using the equation:

$$CA_j^k = \frac{\sum_i T_{ij}^k d_{ij}}{\sum_i T_{ij}^k} \qquad (6)$$

6 PRELIMINARY RESULTS: A NEW GEODEMOGRAPHIC CLASSIFICATION

The last two sections have proposed a number of variables to be classified using geodemographic clustering techniques in an attempt to build a more informative picture of small area populations and their labour markets. A number of key small area data sources have been utilised, from the older, more established sources such as the SWS and the Census of Employment/AES, to the newly released Experían postal sector data. This section shows the preliminary results of a clustering exercise that makes use of the variables proposed in the previous section. A brief synopsis of the basic principles and methodology behind clustering is presented and some of the issues surrounding the choice of the optimum number of clusters are discussed. The clustering results in a new supply and demand-based area taxonomy, which is mapped and interpreted.

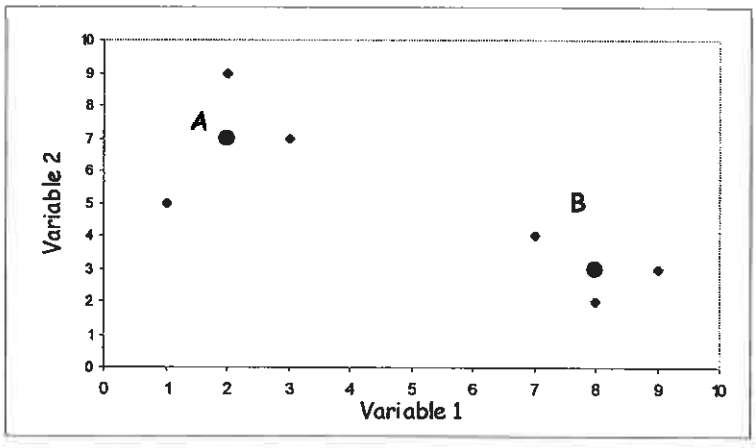
6.1 Basic principles of geodemographic clustering

The overriding principle of geodemographic clustering is that systems should be *mutually exclusive* and *collectively exhaustive*; a zone can only belong to one cluster grouping and every

zone must be accounted for. The clustering process was undertaken using the procedures available in SPSS for Windows to classify an $m \times n$ data matrix (where m is the number of variables and n is the number of cases). An iterative relocation algorithm, otherwise known as *'K-means'*, was used as it is regarded as being less computationally intensive than the stepwise (hierarchical) approach also available.

The number of clusters (K) is specified before the clustering process starts. The data is randomly split into K clusters and the distances between the cluster centres and the observation values in m-dimensional space are measured using the Pythagorean equation for Euclidean distances. Each postal sector is then allocated to the cluster it is nearest to. A new cluster centre is calculated by averaging the observations that fall in each cluster and the process is repeated until there are no more changes in the location of the clusters or some minimum average distance criteria are satisfied.

Figure 5 shows a diagrammatic example of the calculation of cluster centres through the average of the observations in that cluster. Six hypothetical observations are grouped into two clusters in two-dimensional (bi-variate) space. It can be seen how the value for each of the variables becomes the 'co-ordinate' for locating a point in the Cartesian plain. It is impossible to visualise this situation in anything more than three dimensions, let alone the 140 dimensional cluster space demanded by this classification process.



Observation	Variable 1	Variable 2	Cluster
1	2	9	A
2	1	5	A
3	3	7	A
4	7	4	B
5	8	2	B
6	9	3	B
Cluster Center A	2	7	
Cluster Center B	8	3	

Figure 5: *K-means* clustering

The algorithm has two principal aims: to minimise the distance from the cluster centre for all observations belonging to a cluster and to maximise the distance between clusters. However, the clustering process depends upon a pre-emptive decision about the value of *K* before the classification is undertaken. There is no optimum value for *K*; it depends entirely on the data being classified and also the user's personal impression of how many typologies the segmentation process should create. It is therefore instructive to repeat the process with different values for *K* to find the best results.

Therefore the process was run for values of $K = 2$ to $K = 100$. However, it can be argued that this only serves to add further subjectivity to the process because once all these cluster solutions have been created, the best one must be chosen.

6.2 Selecting an appropriate number of clusters

Some of the uncertainty in selecting a value for K can be removed 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 5 shows that the progression of change across the different schedules rises reasonably steadily until the around the 85^{th} procedure (15 clusters) when it starts to increment a little more sharply. However, an analysis of the graph would suggest that the 8-cluster solution (circled in red in figure 6) might be more appropriate as it is the final solution before a very rapid rise towards higher average distance values and, furthermore, represents a fall in value from the previous (9-cluster) solution.

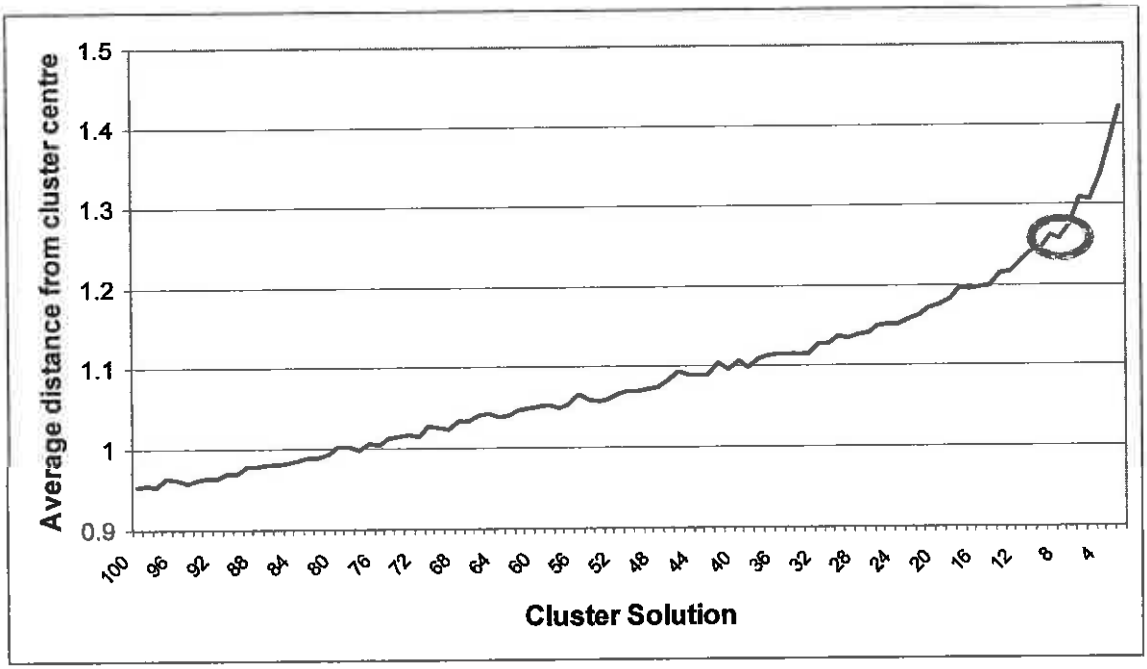


Figure 6: Monitoring the average distance from cluster centre with different values of k

Another way in which the optimum number of clusters can be ascertained is by monitoring the cluster membership. While we would neither expect nor want an equal number of cases in each cluster, the better segmentations are ones that avoid having the majority of cases in one or two clusters and then a number of sparsely populated groups.

Cluster memberships were checked for between 10 and 6 cluster solutions. The results of the 10 and 8 cluster solutions are presented in Table 3. We can see that both the solutions look reasonable, as there is a fairly good spread of membership values. However, both solutions contain clusters that have very small membership values. It happens twice in the 10-cluster solution and once in the 8-cluster solution. The problem is not avoided until the 5-cluster run, but by this time any valuable segmentation of the data set has been lost.

10 Cluster Solution			8 Cluster Solution		
Number of Cases in each Cluster			Number of Cases in each Cluster		
Cluster	1	1	Cluster	1	185
	2	36		2	119
	3	19		3	1
	4	80		4	105
	5	103		5	150
	6	85		6	152
	7	3		7	24
	8	139		8	48
	9	179	Valid		784
	10	139	Missing		0
Valid		784			
Missing		0			

Table 4: Cluster membership returns from the 10 and 8 cluster solutions

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, in the 10 and 8 cluster solutions, 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 5 rises so sharply.

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 nearly redundant 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.

10 Cluster Solution				8 Cluster Solution			
Cluster	Postal Sector	Population	Number of Jobs	Cluster	Postal Sector	Population	Number of Jobs
7	DL9 3	5,623	2,610				
7	DL9 4	6,691	1,026	3	LS1 8	11	2,652
7	YO30 2	1,459	784				
1	LS1 8	11	2,652				

Table 5: Basic demographic and employment statistics for the poorly clustered postal sectors in the 10 cluster and 8 cluster solutions

Table 4 shows the basic demographic and employment statistics of the zones that are poorly clustered in the 8 and 10 cluster solutions. It is clear why LS1 8 presents a problem in both segmentations. With a population of just 11 it has very small demographic statistics. LS1 8 is a particularly small postal sector in the very centre of Leeds, just 1.2 hectares in area. It is a reasonably important zone for employment with over half of its 2,652 jobs being in SIC Section J (Financial Intermediation). Nevertheless, if the intention of this new clustering system is to give an impression of how the population is served by the labour market and to monitor likely population dynamics to assess the stability of investments there, it would be sensible to remove LS1 8 because there is no significant population. However, the removal of the other postal sectors is less easy to justify. They all have significant populations and significant numbers of

jobs, the removal of such a zone from the information system would hinder the comprehensiveness of the final results.

One way to decide which postal sectors to remove is through trial and error. The intention was to remove the erroneous postal sectors one at a time and to re-run the clustering process, monitoring the segmentation through the cluster membership. Table 3 shows the results of removing LS1 8 from the data set before clustering.

10 Cluster Solution			8 Cluster Solution		
Number of Cases in each Cluster			**Number of Cases in each Cluster**		
Cluster	1	79	Cluster	1	186
	2	39		2	142
	3	9		3	47
	4	8		4	96
	5	95		5	134
	6	28		6	5
	7	129		7	22
	8	86		8	151
	9	139	Valid		783
	10	171	Missing		1
Valid		783			
Missing		1			

Table 6: The impact upon cluster membership of removing postal sector LS1 8

Table 5 suggests that the removal of LS1 8 has not been a total success. Both the 10-cluster and 8-cluster solutions have still failed to segment properly; both still have clusters with very low membership. In the 10-cluster solution there are two clusters (cluster 3 and cluster 4) with 9 and 8 cases, respectively. The 8-cluster solution also has one cluster with very low membership; group 6 has just five cases in it. Good classification systems do not have to have a high number

of cases in each cluster. As we have said, it is less desirable to have homogeneity in cluster membership because it suggests that the cases have not been properly grouped together. In many ways, therefore it helps the classification if there are clusters that contain minority of extreme cases. However, in the same way, very low cluster membership is also undesirable because often there is not enough information to make accurate comments about the characteristics of that cluster. In fact, because it is likely that such low membership clusters are made up of extreme outliers in taxonomic space, it is possible that they may not have any similar characteristics at all. In such situations it may be plausible to label them as "pseudo-clusters", rather than clusters.

It is for this reason that it is probably wise to reject the 10-cluster solution finally and accept the 8-cluster. With two potentially meaningless clusters with low membership, it can be argued that segmentation here is worse than in the 8-cluster solution, which only has one. Furthermore, if the 8-cluster solution has segregated the five "outlying" zones from the classification then this may mean that the rest of the cases may be more optimally clustered as the cluster centres are not being dragged away towards extreme cases. Therefore we can take this value of K forward and use it in our analysis of the cluster typologies.

6.3 Evaluating the clusters

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. Conventionally an 'index table' is produced that provides a convenient and simple means of comparing cluster diagnostics (Batey and Brown, 1995). Index tables compare the cluster averages for a given variable against the global average (standardised to 100) across all the clusters.

However, this can be a rather time consuming task; with 140 variables and 8 clusters there would be 1,120 means to evaluate. Furthermore, the mean value of the data can be unduly affected by; the non-normal distribution of census data; the size of the cluster (the number of zones it contains) and the use of percentages for most variables, which means that extreme values of 0 or 100% are unlikely to be reached. Cluster evaluation for this system was therefore performed using Z-scores derived from calculating the standard deviations that occur above and below the global mean as follows:

$$Z_{Km} = (A_{Km} - B_m) / S_m \qquad (7)$$

where:

A_{Km} = cluster mean for variable m,
B_m = global mean for variable m,
S_m = 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. This method produces the same number of values as the index tables, yet by ranking the Z-scores for

each variable in each cluster it is possible to pick out the major characteristics a little more easily.

6.4 Analysing and mapping the new geodemographic clusters

Figure 7 shows the geographical distribution of the 8 clusters created by the *K*-means algorithm. An analysis of the Z-score tables and the raw data itself allows some basic 'pen-portraits' to be created that synthesise the sort of areas that have been grouped together in the segmentation process. These follow Figure 7 and are given for each cluster. It should be recognised that not every case that falls into a given cluster will *exactly* match those characteristics, but that the descriptions instead represent the average conditions experienced in that cluster.

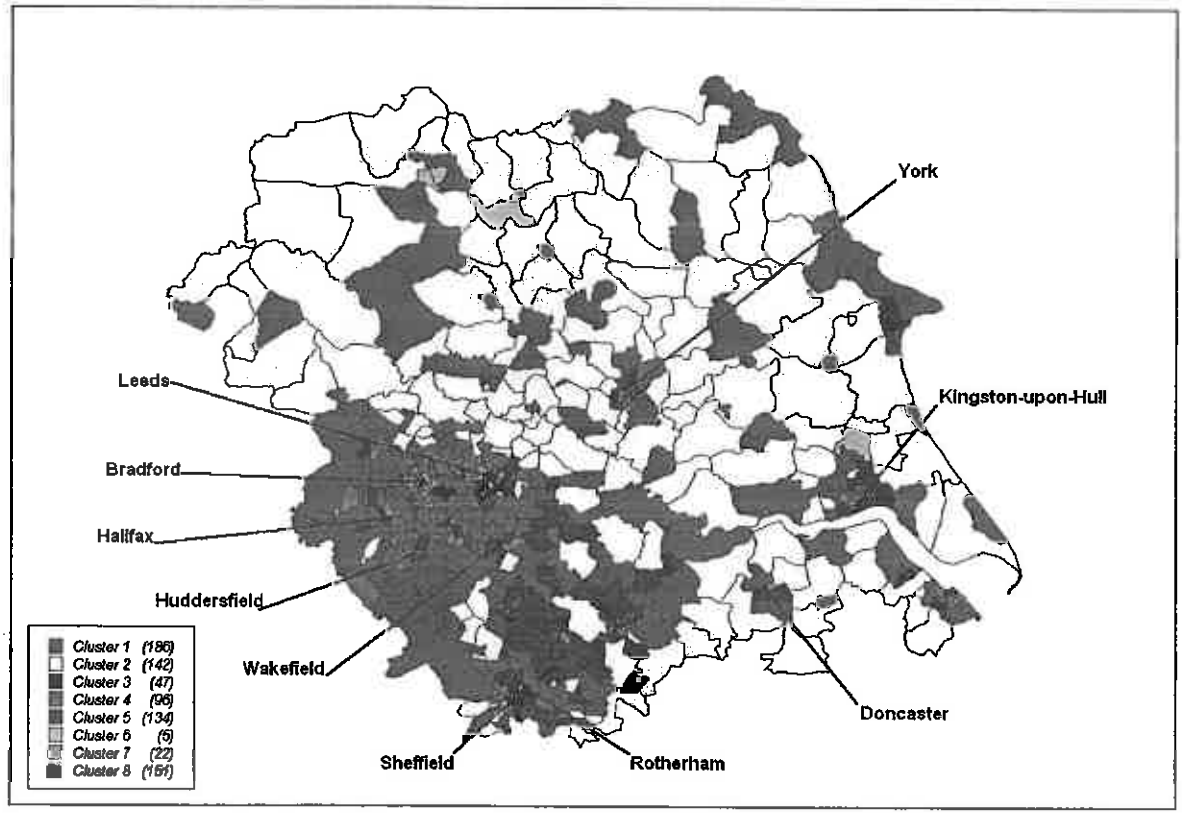
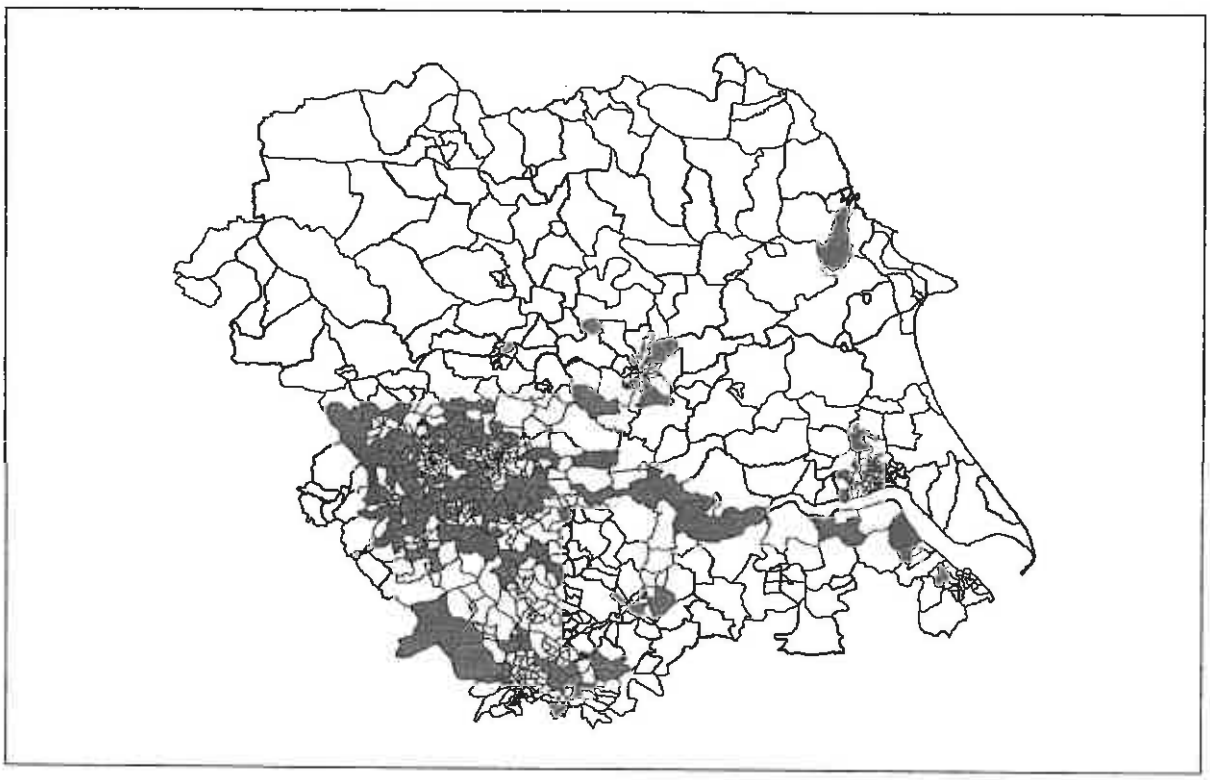


Figure 7: The distribution of the 8 clusters of the new classification



Cluster 1 **186 Zones**

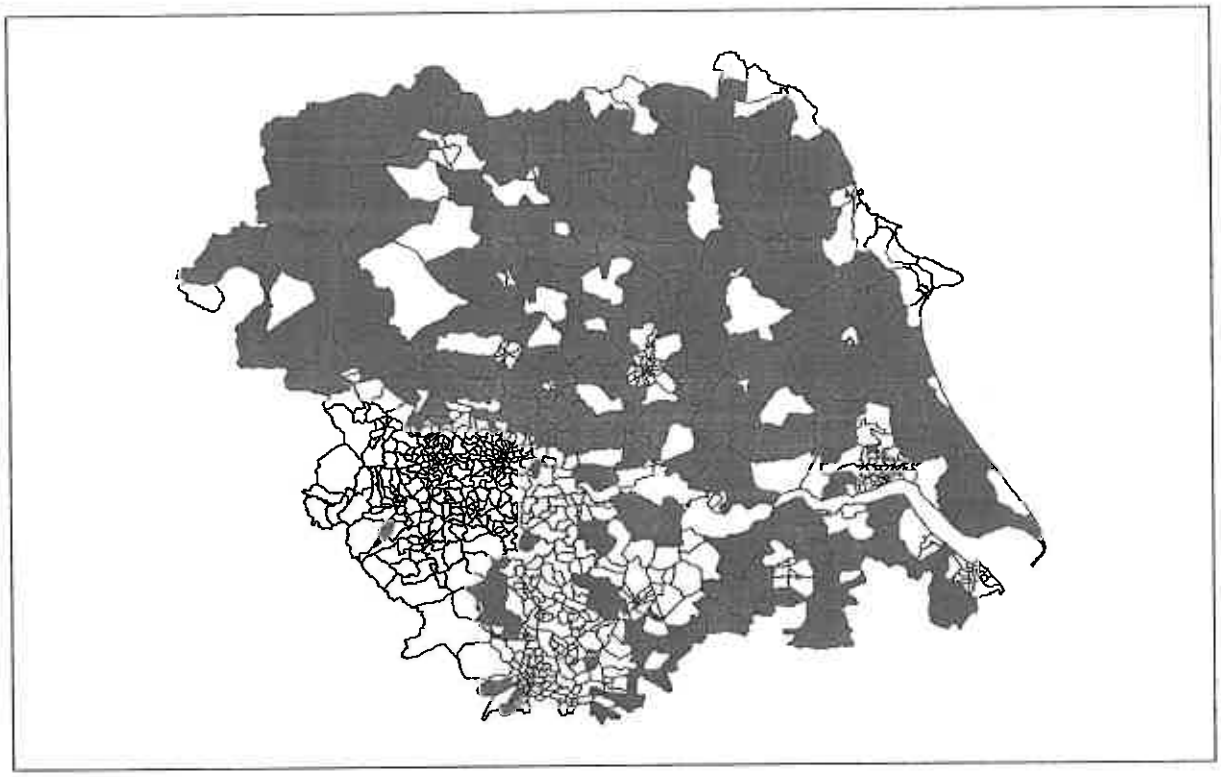
The zones in cluster 1 are situated in the areas surrounding the major urban centres. They are rarely located in the inner cities and are therefore likely to be suburban. The housing stock is mixed between semi-detached and detached with more mortgages and home ownership than average. The population appears to be reasonably transient as there are more sales recorded in the transaction data, especially of semi-detached and detached houses. The average price of a semi in this area is some £57,000, while the average price of all housing stock is a little higher at £61,000. The high proportions of households in census Social Class III(N) – non manual – and the high mortgage rate might suggest that this cluster includes the areas where it is common for people to have purchased their council houses as council renting rates are mixed across the cluster. Furthermore the numbers of households in Social Class II and the high proportion of households with 2 or more cars suggests a strong mix of affluence types. However there are few households in the lowest Social Class groups.

Generally the population is of mixed ages with a tendency towards established families.

Employment in these areas is mixed although there are more jobs in manufacturing and construction than anything else. There are also a larger number of construction units and a paucity of large employers, suggesting that most construction enterprises are small businesses.

Self-containment is generally low but this is to be expected with a reasonably mobile population. Catchment areas for most industry groups, especially manufacturing are higher than average but not by much, suggesting shorter cross-border journey to work trips.

Figure 8: Zones in Cluster 1



Cluster 2 142 Zones

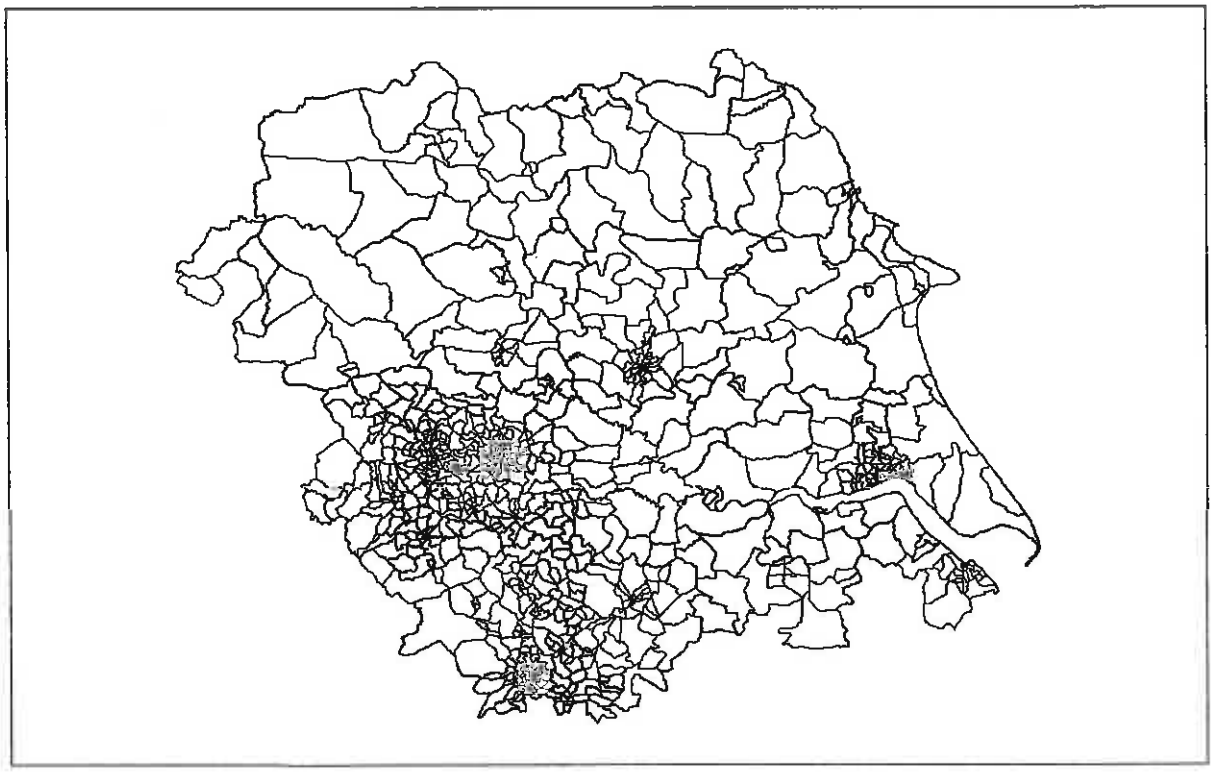
As the map shows, these zones are entirely rural in location. The population here tends to be slightly older with more in the 45 to 64 age group than on average. Meanwhile the 0 to 4 and 24 to 44 ages are less well represented, hinting at older families, possibly with grown up children. The married population is very large and the single population quite small, lone parents are rare.

The housing market appears to affluent. Most sales are in detached houses and the average value of this is around £120,000. The average value of all transactions is near to £100,000 and even semi-detached houses are more expensive then on average at some £62,000. Most households fall into Social Class groups I and II, there are very few in anything lower. However, despite the high property values, the actual number of sales is lower, suggesting a more established, less transient population

Unemployment is low, just 0.2% on average and car ownership is high, suggesting higher levels of affluence.

Employment in these zones is mixed but dominated by agriculture. Self-containment in agriculture is high although the catchment areas are also high due to the longer travel distances in rural areas. For this reason catchment areas are high in all industries. The total regional share of employment is low, especially in manufacturing and wholesale/ retail trade.

Figure 9: Zones in Cluster 2



Cluster 3	47 Zones

The zones that fall into cluster 3 give the impression of being particularly deprived, falling almost exclusively into inner city areas. Unemployment is very high, on average 2.5% but often as high as 3 or 4%. Car ownership is low. Limiting Long Term Illness rates are high and a majority of households fall into Social Class IV or V. Most households rent from the council or local authority and there is a larger proportion of flats in the housing stock, possibly suggesting inner city high rise estates.

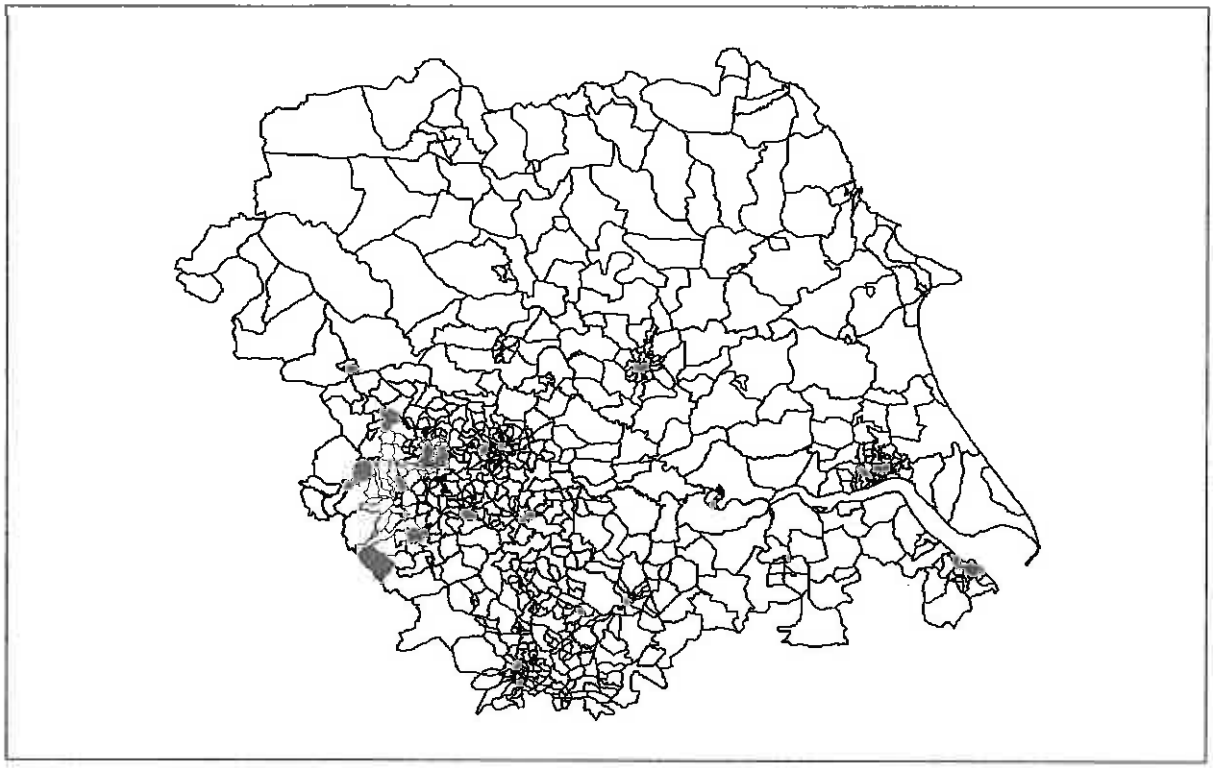
The proportion of the population that moved in the last year is high, particularly pensioner migrants, thus suggesting a quick turn-a-round in occupancy. However the housing market appears to be depressed with few sales, thus suggesting the dominance of the council rented accommodation.

The transactions that do take place are low in value. The average value of all stock is just £25,000 with terraced housing going for an average of just £20,000. Despite the high numbers of flats in the housing stock, they still make up very few of the transactions.

Employment in the area, however, is high. This is consistent with the zones' inner city locations and suggests the industrial areas of the urban centres. The regional share of jobs in construction and manufacturing is high here and so is the regional share of total employment.

Self-containment levels are low but catchment areas are also reasonably low which is to be expected if these are central employment zones as people might only be travelling short distances from neighbouring zones.

Figure 10: Zones in Cluster 3



Cluster 4 96 Zones

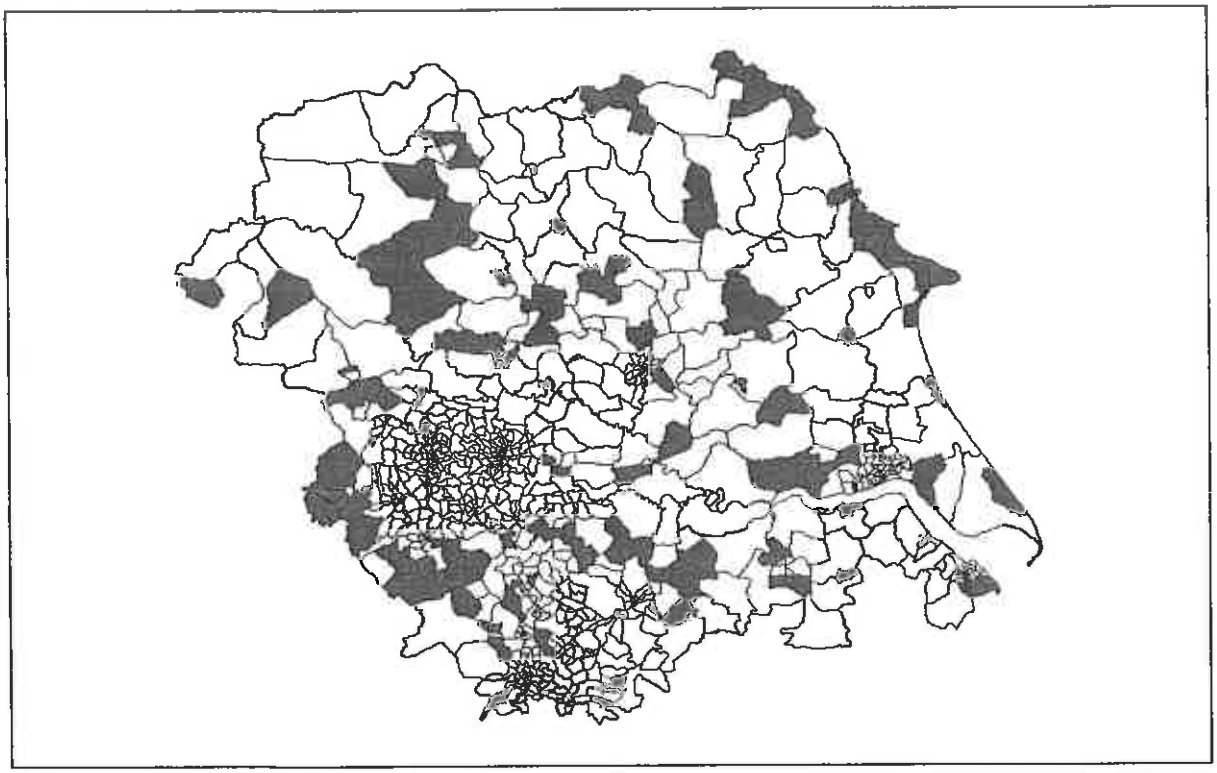
Again, the zones in cluster 4 appear to be concentrated in inner city areas, on the periphery of the city centre. There is a very high proportion of terraced housing here and sales in such housing are correspondingly high with the average value being around £35,000. The average price of all other housing is not much higher (at about £40,000) but there appears to be paucity of other housing stock, especially large detached houses. After terraced housing the next most common housing type is flats.

Although sales are high, homeownership is lower and private renting arrangements are more common. The high proportions of students and younger population (15 to 24) in these areas largely explain this. A much higher proportion of the population is single than on average and movement in the last year is common. There is also evidence of a married population and the proportion in the 0-4 age group would suggest the presence of young families, possibly in starter homes. Households in this area often lack basic amenities and there is a degree of overcrowding. Car ownership is low and unemployment is high at 1.5% on average.

Employment in these zones is largely concentrated in manufacturing and wholesale and retail trading, the regional share for both is commonly high. Employment in construction is characteristically low as are jobs in the 'public services' (SIC sections L, M & N).

Self-containment for the retail industry jobs is high in these zones but this is not the case for the other industry groups. However, catchment areas are not that high, particularly in manufacturing, which again suggests short cross-border flows from nearby residential zones.

Figure 11: Zones in Cluster 4



Cluster 5 **134 Zones**

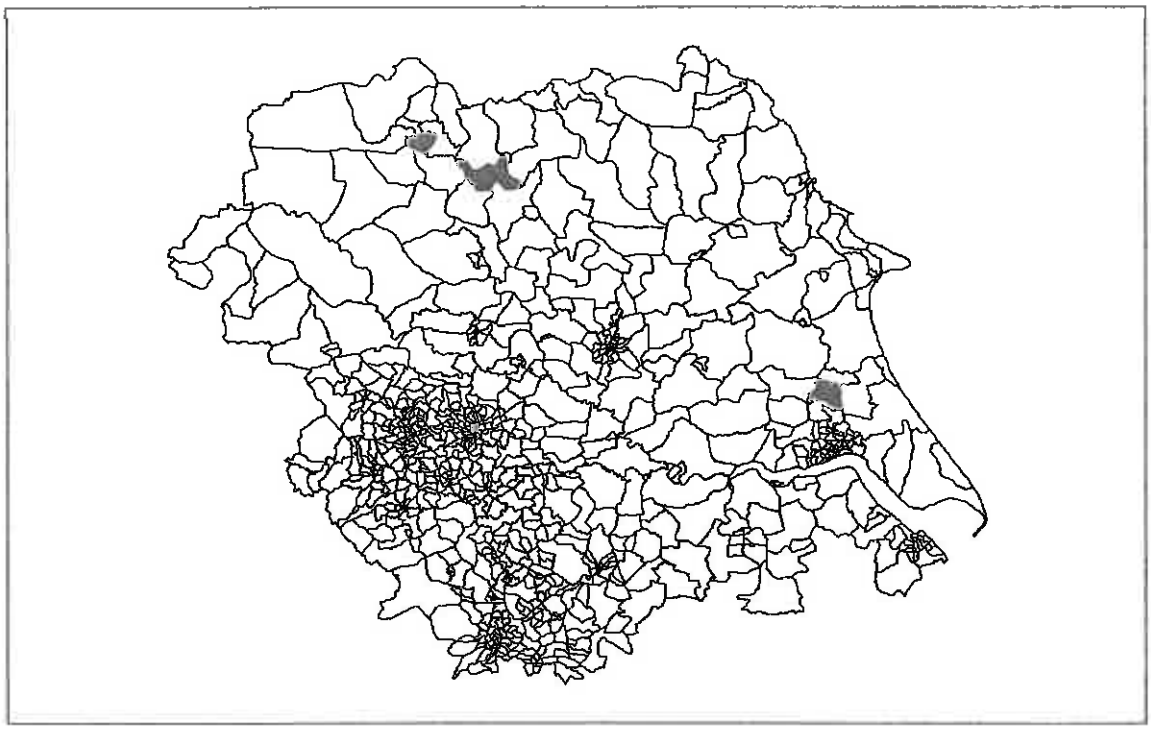
The postal sectors in cluster 5 are predominately found on the periphery of the urban centres or in rural locations.

The most striking aspect of these zones is that self-containment is very high in all industries, particularly public services, retail, construction and manufacturing. There are few large units (300+ employees) and catchment areas are correspondingly small due to the higher self-containment values.

The population is made up of a high proportion of homeowners and married population. Sales are high, particularly of detached and semi-detached houses and the average value of detached properties sold is some £100,000. Self-employment levels are high, unemployment is low and there is a large proportion of households in Social Classes II and I. As these zones are frequently found in rural locations, the regional share of agricultural jobs is high.

There are very few households renting from the local authority, few non-family households and the single population is poorly represented. The elder age groups, are more prevalent here, particularly the 45 to 64 group, possibly suggesting later middle age or retirement migration. The proportion of the 15-24 group hints at older, more established families. There seem to be few young families.

Figure 12: Zones in Cluster 5



Cluster 6 5 Zones

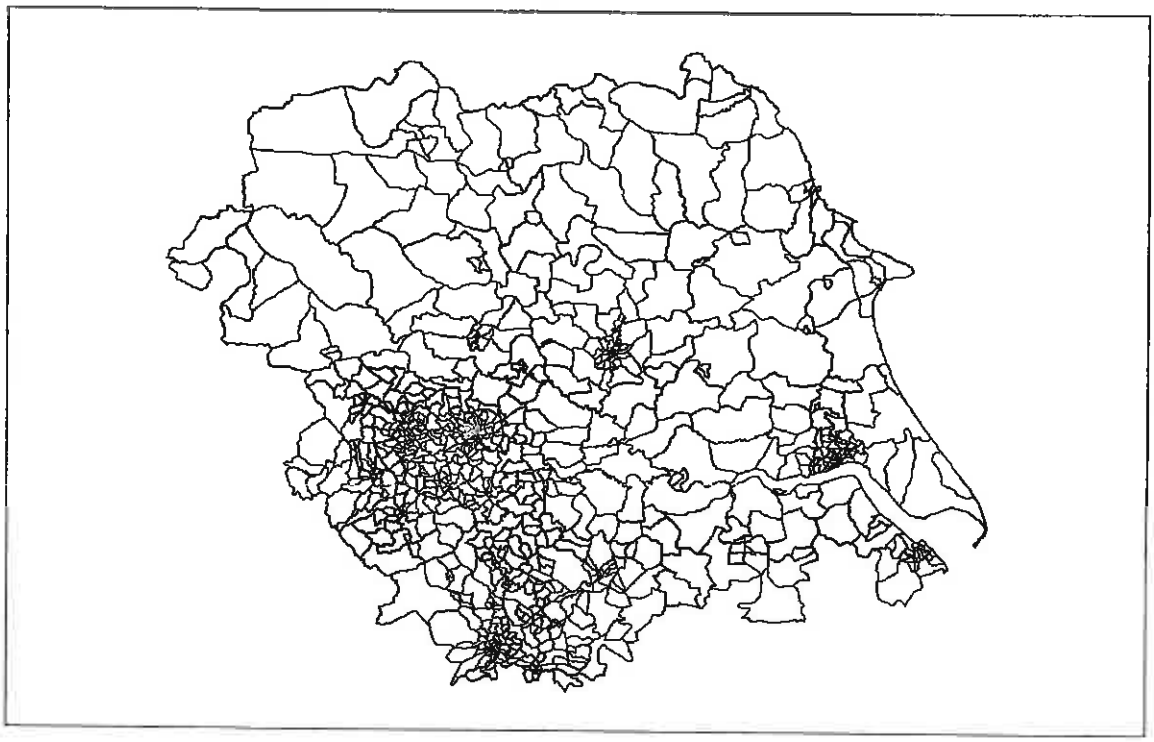
In the discussion above, cluster 6 was highlighted as a problem cluster because of its small membership. It was speculated that it may contain outliers in the data and where this is the case there is always a danger that the cluster characteristics might not be meaningful.

However, the cluster characteristics are not as random as one would expect from a series of outliers. The principal components of this cluster are very high numbers of jobs in SIC Section L - Public Administration, Defence and Social Security. The key aspect here is defence as each of the five postal sectors is linked to a major Armed Forces base. DL9 3 and DL9 4 contain Catterick Army camp. DL7 8 and DL7 9 contain RAF Leeming while HU17 9 includes Normandy Barracks, home of the Defence School of Transport.

Unsurprisingly, therefore, this aspect of employment dominates the characteristics of this cluster. On average 6% of the regional workforce in SIC Section L is found in these postal sectors. In HU17 9 and DL9 3 over 60% of employment is in this section. Self-containment is understandably high. The postal sectors also have very high numbers of jobs in the services that might support such Armed Forces camps, particularly retail and entertainment (Sections G & H). This is particularly the case in DL9 4 (Catterick town) and DL7 8 (Northallerton), the towns that are immediately adjacent to two of these bases.

The population characteristics are consistent with such installations. The population is dominated by younger adults with very few in the older age groups. Limiting Long Term Illness is very low, as is unemployment. Few households lack basic amenities. There are few sales of property but plenty of movement, the cluster average sees 19% of the population having moved in the previous year but in DL9 3 (Catterick Camp) this is 44%.

Figure 13: Zones in Cluster 6



Cluster 7 22 Zones

Cluster 7 would appear to be another geographically specific cluster. The postal sectors in this cluster are, without exception, those found in the very centre of the major urban areas, specifically, Hull, Leeds, Sheffield and York. Bradford is excluded from this but the centre of Harrogate is included.

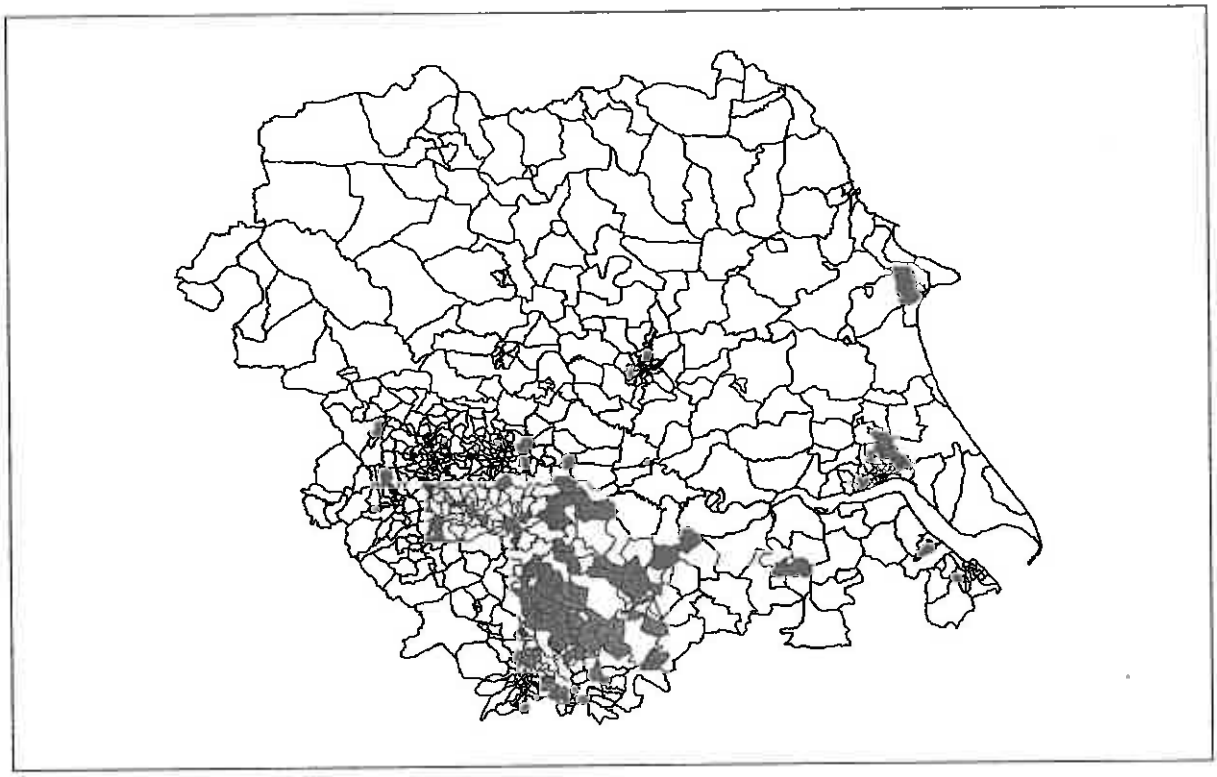
As these are city centre locations the regional share of employment is high, especially in renting and business services (section K), financial intermediation (section J) and hotels and restaurants (section H). The proportion of employment units in these sections is high and there are many large employment units in these sections too. Employment in manufacturing and construction is very low. Self-containment levels are very low and catchment areas are high. This is to be expected as the city centres draw in workers from a wide area. People tend to be travelling further distances into Leeds than anywhere else.

The population is dominated by the younger adults and movement is high, 25% on average. Most of the housing stock is flats, often of high value. There is a high proportion of the population with degrees and higher qualifications. Such information suggests that these areas could be characterised by young professionals in gentrified areas. Certainly this would be consistent with the location of these postal sectors.

There is an under-representation of the population in the younger age groups and few in the middle aged or elderly cohorts. The married population is also low.

Car ownership is low but this may not be a sign of deprivation in city centre areas. There are few lone parents and few households in Social Classes III(M) or below. Most are in Social Class III(N) or II.

Figure 14: Zones in Cluster 7



Cluster 8 151 Zones

These postal sectors are characterised by high proportions of the population in Social Classes III(M) and IV. There are very few in Social Class I or II. The proportion of the population in council rented accommodation is very high as is the proportion of the population aged 5 to 14. The high numbers of retired population and the percentage of households with dependants could suggest extended families. However lone parentage is also high. Car ownership in these areas is lower than the regional average, unemployment is high and Limiting Long Term Illness is also common. The housing stock would appear to be characterised by semi-detached houses although terraces are also common. However the average price of such housing is low at just £43,000 and the average price of terraced housing is as little as £30,000.

There is large amount of employment in manufacturing, retail and transport. Self- containment levels are high, particularly in manufacturing. Employment in education is common too. Self-employment is rare. The concentration of these zones in South Yorkshire ensures a noticeable amount of employment in mining and quarrying (section C). The levels of self-containment for this can therefore be high. Employment in real estate or business renting (section K) and financial intermediation (section J) is low.

Figure 15: Zones in Cluster 8

7 CONCLUSIONS

The dissemination of the 2001 Census will prove and exciting time for the geodemographics industry as the old systems are brought up to date. The tradition of geodemographics remains strong and although attempts to integrate lifestyles systems and GDIS have so far proved unsuccessful (Harris, 1999; Birkin 1995), it is likely that they will continue to thrive in business applications.

This paper has put forward a selection of variables that have enhanced the area taxonomies that are created using traditional demand variables. It has been argued that longer-term stability can be measured by including variables that measure the level of supply in addition to perceived demand as well as the interaction between demand and supply. Variables that relate to the provision of employment have therefore been proposed. In addition to this a further dimension has been added in the form of variables that are not solely reliant on percentage counts. Indices of specialisation have represented the structure of employment provision over and above the percentages and some model-based indicators have been used to measure the interaction between zones in the labour market.

Figure 7 and the subsequent cluster descriptions suggest that this first attempt at adding supply variables to more traditional demand data has indeed been successful. The small area characteristics now include measures of the level of employment provision in each small area and, importantly, the degree to which each zone is dependent upon particular industries and how

they are likely to change. The new supply variables have all proved to be important cluster formative variables and significantly add to the information provided by the demand data.

There is scope for a considerable amount of further work here. A significant starting point will be to fully evaluate the contributions of the new variables to the cluster segmentation in addition to that provided by the Z-score analysis. There is also a need to consider other supply-side variables; for instance those that relate to retail services, infrastructure and proposed housing development. With this achieved, the system needs to be extended to create a national classification.

Throughout this paper we have argued that this new system will be of considerable interest to the business sector. In Section 2 we suggested an application in the location of supermarkets. Traditional GDIS provide a suggestion of the level of retail consumption in an area by characterising demand but there is nothing to say how stable this is. By including the characteristics of supply it can be determined how prone an area is to significant changes in the local economy that might affect consumption. For this reason it is envisaged that a system such as this may find practical use in Spatial Decision Support Systems (SDSS). However, it may also be possible to see a social policy application here as well. There is at least one cluster in the segmentation presented here that shows some characteristics of deprivation and a dependency upon particular industries, especially manufacturing. This could be used to target government resources and investment and prepare for the impacts of local changes, such as the loss of a major employer, or a national recession.

ACKNOWLEDGEMENTS

1. The authors wish 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 administrative boundaries of Yorkshire and the Humber are based on data provided by the United Kingdom Boundary Outline and Reference Database for Education and Research Study (UKBORDERS) via Edinburgh University Data Library (EDINA) with the support of ESRC and the Joint Information Systems Committee of Higher Education Funding Councils (JISC) and boundary material which is Copyright of the Crown, the Post Office and the ED-LINE consortium.
3. 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.
4. 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.
5. 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.

6. 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.
7. 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.

REFERENCES

Batey, P. and Brown, P. (1995) From human ecology to customer targeting: the evolution of geodemographics, in Longley, P. and Clarke, G.P., (eds) *GIS for Business and Service Planning*, GeoInformation, Cambridge.

Beaumont, J.R. and Inglis, K. (1989) Geodemographics in practice: Developments in Britain and Europe, *Environment and Planning A*, 21: 587-604.

Bertuglia, C.S. and Rabino, G.A. (1994) Performance indicators and evaluation in contemporary urban modelling, in Bertuglia, C.S., Clarke, G.P. and Wilson, A.G. (eds.) *Modelling the City: Performance, Policy and Planning*, Routledge, London.

Birkin, M. (1995) Customer targeting, geodemographics and lifestyle approaches, Ch. 6 in Longley, P. and Clarke, G.P. (eds) *GIS for Business and Service Planning*, GeoInformation, Cambridge.

Birkin, M., Clarke, G.P., Clarke, M. and Wilson, A.G. (1994) Applications of performance indicators in urban modelling: subsystems framework, in Bertuglia, C.S., Clarke, G.P. and Wilson, A.G. (eds.) *Modelling the City: Performance, Policy and Planning*, Routledge, London.

Birkin, M. and Clarke, G.P. (1998) GIS, Geodemographics and spatial modelling in the UK financial services industry, *Journal of Housing Research*, 9(1): 87-111.

CACI (1993) ACORN Product brochure, CACI Information Services, London.

Cathelat, B. (1990) *Socio-Styles: The New Lifestyles Classification for Identifying and Targeting Consumers and Markets*, English edition, Kogan Page, London.

Clarke, G.P. and Wilson, A.G. (1994a) Performance indicators in urban planning: the historical context, in Bertuglia, C.S., Clarke, G.P. and Wilson, A.G. (eds.) *Modelling the City: Performance, Policy and Planning*, Routledge, London.

Clarke, G.P. and Wilson, A.G. (1994b) A new geography of performance indicators for urban planning, in Bertuglia, C.S., Clarke, G.P. and Wilson, A.G. (eds.) *Modelling the City: Performance, Policy and Planning*, Routledge, London.

Duley (1989) A model for updating census-based household and population information for inter-censal years, *Unpublished Ph.D. thesis*, School of Geography, University of Leeds.

Experían Ltd. (2001) Experían Ltd. Postal Sector Data documentation URL: http://www.mimas.ac.uk/docs/experian/ (accessed 5/2/2001).

Feng, Z. and Flowerdew, R. (1999) The application of fuzzy classification in geodemographics, Towards Digital Earth – Proceedings of the International Symposium on Digital Earth.

Harris, R. (1998a) Considering (mis-) representation in geodemographics and lifestyles, in Abrahart, R. (ed.) *Proceedings of the 3rd International Conference on GeoComputation* (CD-ROM). Also available on-line URL: http://rich-harris.freeyellow.com/geocomputation.htm Accessed 25/6/2000.

Harris, R. (1998b) Is there an 'I' in GDIS? The problem of representation in geodemographic and lifestyles systems, *Cybergeo* 63. Available on-line URL: http://www.cybergeo.presse.fr/revgeo/rostok/harris/harris.htm Accessed 17/7/2000.

Harris, R. (1999) Geodemographics and the analysis of urban lifestyles, *Unpublished Ph.D. thesis*, University of Bristol.

Martin, D.J. (1995) Censuses and the modelling of the population in GIS, in Longley, P. and Clarke, G. (eds.) *GIS for Business and Service Planning,* GeoInformation, Cambridge.

Martin, D.J. (1999) Spatial representation: the social scientists perspective, in,Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. (eds.) *Geographical Information Systems. Volume 1: Principles and Technical Issues* 2nd edition, John Wiley & Sons, Chichester.

Martin, D.J. and Longley, P.A. (1995) Data sources and their geographical integration, in Longley, P.A and Clarke, G.P. (eds.) *GIS for Business and Service Planning,* GeoInformation International, Cambridge.

MIMAS (1999) Census Dissemination Unit Web Page: SASPAC system files URL: http://census.ac.uk/cdu/Software/Naming_conventions/System_files.htm (accessed 8/2/2000).

NOMIS Reference Centre (2001) ONS Guide to Regional and Local Labour Market Statistics – Annual Employment Survey.
URL: http://www.nomisweb.co.uk/ref/guide/guide08_1.htm (accessed 20/8/2001).

ONS (2001a) Neighbourhood Statistics website – Introduction. URL: http://www.statistics.gov.uk/neighbourhood/general.asp (accessed 15/8/2001).

ONS (2001b) *First Release: Index of Production, June 2001,* ONS, London

Openshaw, S. (1984a) The modifiable areal unit problem, *Concepts and Techniques in Modern Geography*, 38, Geo Books, Norwich.

Openshaw, S. (1984b) Ecological fallacies and the analysis of areal census data, *Environment and Planning A*, 16: 17-31.

Openshaw, S. (1994) *Developing intelligent geodemographic targeting systems, Working Paper 94/13*, School of Geography, University of Leeds, Leeds.

Raper, J.F., Rhind, D.W. and Shepherd, J.W. (eds.) (1992) *Postcodes: The New Ggeography*, Longman, Harlow

Rees, P.H. and Birkin, M. (1983) Census-based information systems for ethnic groups: a study of Leeds and Bradford, Paper presented at the Regional Science Association British Section Meeting, University of Leeds, 7-9 September.

Rothman, J. (1989) Editorial, *Journal of the Market Research Society*, 31,(1): 1-7

Simpson, S and Middleton, E. (1999) Undercount of migration in the UK 1991 Census and its impact on counterurbanisation and population projections, *International Journal of Population Geography*, 5: 387-405.

See, L. and Openshaw, S. (2001) Fuzzy geodemographic targeting, in Clarke, G.P. and Madden, M. (eds.) *Regional Science in Business*, Springer-Verlag, Berlin.

Simpson, S. and 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).

Sleight, P. (1997) *Targeting Customers: How to Use Geodemographic and Lifestyle Ddata in Your Business*, 2nd edition, NTC Publications ltd., Henley-on-Thames.

Sleight, P. (1998) Opinion, *Database Marketing*, 1: 8.

Stillwell, J.C.H (1984) IMP: a program for Inter-area Migration analysis and Projection: user's manual (revised), *Computer Manual 12*, School of Geography, University of Leeds, Leeds.

Stillwell, J.C.H (1991) Spatial interaction models and the propensity to migrate over distance, in Stillwell, J.C.H and Congdon, P. (eds) *Migration Models: Macro and Micro Approaches*, Belhaven Press, London.

Stillwell, J.C.H. and Palmer, J. (1986) User's guide to PACE: A Program for Analysis of the Concentration of Employment, *Computer Manual 25*, School of Geography, University of Leeds, Leeds

Tye, R. (1995) The missing millions! (or What was the real population of small areas (wards and EDs) in Mid 1991?) MIMAS on-line newsletter March 1995 URL: http://mimas.ac.uk/newsletters/9503/95302.htm Accessed 7/4/2000.

Webber, R.J. (1989) Using multiple data sources to build an area classification system: Operational problems encountered by MOSAIC, *Journal of the Market Research Society*, 31(1): 103-109.

Webber, R.J. (1992) Streets ahead of the rest, *Precision Marketing*, 7/12/1992.

Wilson, A.G. (1971) A family of spatial interaction models and associated developments, *Environment and Planning* A, 3: 1-32.