A new migration classification for local authority districts in Britain

Working Paper 09/2

Adam Dennett

John Stillwell

School of Geography

University of Leeds

LS2 9JT

May 2009

v2.0 Revised September 2010

CIDER

UNIVERSITY OF LEEDS

This working paper is an online publication and may be revised

Full contact details of the authors are:

School of Geography

University of Leeds

Leeds

LS2 9JT

Tel: +44 (0 in UK) 113 343 3300

Fax: +44 (0 in UK) 113 343 3308

Adam Dennett

http://www.geog.leeds.ac.uk/people/adennett.html

a.r.dennett@leeds.ac.uk

John Stillwell

http://www.geog.leeds.ac.uk/people/jstillwell.html

j.c.h.stillwell@leeds.ac.uk

Adam Dennett is the contact author

Contents

Table of Figures

Table of Tables

Introduction

Work by Dennett and Stillwell (2008b, 2009, Forthcoming), Raymer and Giulietti (2009) and Stillwell and Hussain (2008) has demonstrated the utility of area classifications in the study of migration in Britain. The Vickers *et al.* (2003) classification adopted by Dennett, Stillwell and Hussain is a general purpose area classification which uses a suite of variables from the 2001 Census to define different groups of local authority districts in a three-tier hierarchy. It does not, however, include any migration or commuting variables. Whilst migrants and commuters will be included amongst the groups of individuals present in each area, their status as such is not explicitly recorded and consequently does not directly influence the clusters represented in the classification. This work has demonstrated that by examining migration flows between areas defined by their socio-demographic characteristics, associations can be made between the flows and the area types. For example, at an aggregate national level, outflows occur frequently from urban area types characterised by poorer health, higher unemployment and lower economic activity. Inflows, on the other hand, tend to be to rural area types characterised (amongst other things according to the classification used) by lower population densities and higher home ownership rates. By excluding migration variables from the classification, any associations between particular flows and area types are not confounded by the presence of these variables. Observing high in-migration to an area partially defined and characterised by a high proportion of young in-migrants adds no value to our understanding, whereas high in-migration to an area partially defined and characterised by low population densities might tell us something about the aspirations of migrants.

Of course, just because an object (or area) can be classified in one way, it does not mean it resides exclusively within that classification's typology. Objects can be classified very differently depending upon the purpose of the classification. For example, a tree might be classified by a biologist as a particular species within a certain genus; by an architect as a source of one type of building material with specific qualities for construction; or by a person sheltering from the rain as a more or less effective shelter than a nearby building. The tree's position in each classification fits a specific purpose. In much the same way, if an area is classified for one particular purpose, it could be classified entirely differently for a different purpose – the same area could be classified as a low crime area or as an area with a high proportion of elderly residents. In the context of this work, it may well be that in an effort to understand migration between and within defined geographical areas in Britain, the use of an area classification constructed without migration and commuting variables might not be most appropriate. Populations of migrants moving to and from

areas may well differ significantly from underlying populations; as such it may be sensible to classify areas according to these migrants and their associated socio-demographic characteristics.

This paper details the arguments for, and the process of creating, a new internal migration classification for local authority districts in Britain. The overarching research hypothesis is that where the processes of internal migration in Britain are complex, our spatial understanding can be enhanced through identification of the particular characteristics that migrants and migrant flows can contribute to defining different types of areas. Districts of Britain can be usefully classified by the types of migrant and the particular flows that they exhibit such that each classified area will have distinct profiles. In demonstrating the validity of this hypothesis, subsequent work (which is not within the scope of this paper) will seek to prove that a migration classification can enhance our understanding of the complex migration patterns occurring in Britain, and will enable more effective monitoring and analysis of patterns derived from data sources available since the last census in 2001.

1 The case for an interaction data classification

1.1 Why develop classifications?

So if an area classification based around interaction data variables is an alternative to other classifications, one question that might arise is ‘should such a classification be created?’ How might a classification be useful in helping develop our understanding of migration in Britain? Or perhaps taking the question even further – why create a classification at all?

1.1.1 Background to classifications

Taking the latter, it could be argued that human brains constantly classify our lived experiences in order for us to make sense of what is happening around us. Indeed, speaking from a biological perspective, Crowson (2006, p1) states that “classifying things is perhaps the most fundamental and characteristic activity of the human mind”. Through classification we are reducing the amount of data the brain has to deal with, thus aiding us to make sense of situations more readily. Taking an evolutionary perspective, it is understandably an advantage for any animal to mentally classify food and non-food items, or categorise other animals as dangerous or benign. But because classification is useful and indeed necessary for some fairly fundamental life processes, does this necessarily mean it is applicable outside of the sphere of everyday lived experiences?

This question can be answered through examining where else the process of classification has flourished. The introductions to a number of textbooks on the subject (the existence of which

already suggest a wider applicability) such as those by Aldenderfer and Blashfield (1984), Kaufman and Rousseeuw (2005), Everitt *et al.* (2001) and Gordon (1999) all mention the long history of the creation of classifications and taxonomies in fields such as biology, chemistry, physics and astronomy, as well as in the social sciences. When looking at the historical uses of classifications, it becomes clear that whilst they are frequently put to use in a variety of situations, the motivations behind classification creation can differ from those supporting the more common practice of every-day mental classification and the basic desire to sort objects to aid comprehension. Of course, aiding comprehension is one very useful end product, but once comprehension is improved, we are then more able to take what is known and apply it to alternative situations. For example in medicine, Everitt *et al.* (2001) describe the classification of diseases as both a useful aid to treatment, but also as a basis for research into the causes of disease.

To view the end result of the classification process – the taxonomic groupings – as the only benefit of creating a classification would be to ignore the value in the process which needs to be followed to arrive at this final product. The identification of particular data features which define the groups within the classification may prove even more useful than the classification itself, despite this perhaps not being the reason for embarking upon the classification development process in the first instance. Indeed, it was partially through the process of cataloguing and classification of new species that Charles Darwin began to develop the ideas which led to the publication of arguably one of the most influential works of all time: 'On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life' (Darwin, 1859), containing ideas on the theory of evolution by natural selection – ideas which changed the world as we see it and the received wisdom on the origin and evolution of our entire existence.

1.1.2 Classifications in Geography

Whilst perhaps not as fundamentally significant as the ideas which evolved from Darwin's classification exercise, there has been a long history and development of area-based classifications in geographically related disciplines which have sought to make sense of complex environments and the various population attributes characterising those environments. In all cases, these classifications have stimulated further research. Vickers *et al.* (2005) cite the work of Charles Booth in the nineteenth century as perhaps the earliest such example. Booth attempted to map and classify areas of London according to the socio-economic characteristics (specifically poverty, employment and religion) of the residents living in those areas – work which influenced both subsequent academic studies (the work of Orford *et al.* (2002) being one of the more recent

examples) and more applied political policy (Bales, 1999). Burgess (1925) in the early twentieth century, whilst focusing on the growth and expansion of Chicago also succeeded in classifying the 'types of areas differentiated in the process of expansion' (Burgess, 1925) – area types identified in part by their residents. This seminal work influenced later work on urban structure and classification by authors such as Hoyt (1939) and Harris and Ullman (1945). The work of both Booth and Burgess, whilst having contemporary influence, can also be seen as the forerunner of far more recent work which, whilst perhaps more detailed in its scope, complex in its methodology and arguably more accurate in its definition, actually seeks to do exactly the same thing – to classify areas according to set of particular key characteristics.

A case could be made for one of the main drivers behind much of the recent work on area classifications being commercial interest, which has led to the growth of an industry concerned with developing and applying area classifications for commercial gain. A commonly used term, for both the development and application of these (small) area classifications (as well as the industry stewarding it) is 'geodemographics'. There is a large literature documenting the development of geodemographics – a development which has largely occurred in parallel with improvements in computational and processing power, software and geographic information. Batey and Brown (1995) and Yano (2001) provide succinct historical overviews. The commercial imperative has helped spawn companies and organisations such as CACI Ltd, CCN marketing (now Experian) and EuroDirect, all producing their own geodemographic area classifications such as ACORN, MOSAIC and CAMEO respectively for commercial customers. The continuing growth of the industry (a brief visit to the press release section of any of CACI, EuroDirect or Experian's websites will present a selection of news stories documenting new updates of their classifications and expansions into different countries) might be evidence enough that there is real value in classifying areas according to certain key characteristics, tailored for specific needs and purposes. Other evidence, however, can be found in Harris *et al.* (2005) where an objective evaluation of whether geodemographic classifications 'work' is carried out though a case study of the application of the ACORN classification. ACORN is used to assess differences in product consumption patterns in a British town with a conclusion that, for this particular application, the classification did indeed work.

Further evidence as to the utility of geodemographic area classifications outside of the commercial sphere can be found in the renewed academic interest in the creation of area classifications and the ongoing development of area classifications by the Government. Longley (2005) postulates that this revival has been driven through a combination of a desire for evidence-based policy from local government, improvements in data and related infrastructures and a need

for setting service delivery targets at a local level. Examples of the former include some early work by Openshaw and Blake (1996) on a 'GB profiler' using 1991 Census data. More recently, a large number of geodemographic profiling projects have been undertaken by the Centre for Advanced Spatial Analysis (CASA) at University College London, including those focusing on health, ethnicity, education, and awareness and access to digital technologies (http://www.spatial-literacy.org/), each with applications in policy and resource targeting. Bespoke classifications have also been created within the School of Geography at the University of Leeds for specific local purposes. Work by Shepherd (2006) profiling neighbourhoods to aid community safety and Debenham *et al.* (2003) focusing on supply-side variables to extend commercial geodemographic classifications within Yorkshire and the Humber are such examples.

National classifications have been constructed for the major Census geographies by ONS and others: output areas, super output areas/data zones, wards, health authorities and local authorities, (http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/) as well as indices created from multiple variables for other geographical areas such as the Index of Multiple Deprivation (ODPM, 2003). Whilst the Index of Multiple Deprivation does not seek to cluster areas with similar characteristics at the outset as most geodemographic classifications do, by ranking each area by its index score and then dividing the ranked areas into proportions (quintiles or deciles) and allocating areas to these groups, an area classification comparable to those created through clustering methodologies can be created. Rees *et al.* (2002a) provide a comprehensive historical summary of census-based area classification typologies. In addition, some of the growing range of applications to which the ONS output area classification is being put are documented by the Output Area Classification User Group (http://www.areaclassification.org.uk/). These include higher education student profiling, analysis of crime and antisocial behaviour, analysis of transport need, various commercial applications and local authority housing market analysis – the breadth of applications no-doubt indicative of the open-source nature of both the classification and its construction methodologies.

Within geographical analysis there is another family of classification techniques concerned less with classifying single areas, but more with identifying the relationships between areas and classifying them by these relationships. These are known as 'functional regionalisation' techniques. Whilst standard geodemographic classifications can classify a number of areas based upon their similar attributes, there is no inference that because areas may fall into the same category they share any kind of connection or interaction. Openshaw (1989) picks up upon this drawback, and, using the example of interaction data in the form of credit card company

□□□□□□[a]□□ □ □ □□□□□▪ ▪ ▪ ▪ ▪ ▪ ation) and the residential address

highlights the potential these data have for identifying the catchment areas of shopping centres. By defining catchment areas an indication of the importance of key nodes to surrounding zones is presented.

In identifying the inability of standard geodemographic classifications to deal with interactions such as those which could define catchment areas, Openshaw (1989) highlights functional regionalisation methods as a potential solution. Coombes (2000, p1502) defines functional regionalisation as a *"form of area classification within which each class is normally a single group of contiguous areas"*. Brown and Holmes (1971) (quoted in Feldman *et al.*, 2006) suggest that functional regions are *"areas or locational entities which have more interaction or connection with each other than with outside areas"*. So it is the contiguous nature coupled with the attribute homogeneity of the smaller areas within each class area that sets functional region classifications apart from geodemographic area classifications. Assessing catchment areas for flows is a particularly geographical or spatial problem and functional region classifications deal with grouping common flows rather well.

The functional regionalisation approach is one that has been adopted widely in the analysis of commuting data, principally because it has long been recognised that the poor definition of geographical areas can lead to statistics giving a distorted view of the reality underlying them (Coombes, 2002). The need to create a set of geographical areas relatively consistent with the phenomena being examined in those areas has been recognised more recently by Martin (1998, 2000, 2002) in relation to the creation of relatively socially homogenous census 'Output Areas', but was identified by the UK Government in relation to locally specific unemployment rates and the allocation of financial assistance to those areas in greatest need (Coombes *et al.*, 1986) a number of decades ago. The result of this need was that Coombes and others, on behalf of the Department of Employment developed and successively refined a set of Travel-To-Work Areas (TTWAs). TTWAs were designed such that they reflected labour market areas within which the local supply and demand of labour interact (Coombes, 2002). With commuter flows inextricably linked to labour supply and demand, the origin-destination elements of commuting data can be used to identify labour demand nodes surrounded by labour supply areas such that the boundary of each TTWA surrounds an area that is relatively self contained in terms of its commuter flows. From 1981 census data and then with each successive wave of the census, Coombes and colleagues (Coombes, 2000, 2002; Coombes *et al.*, 1986) have developed new sets of TTWAs using variations on a functional regionalisation algorithm which essentially identifies employment nodes (or foci) through functions of job ratio and residence-based self containment, before amalgamating adjacent foci where they were strongly linked, and then iteratively

allocating residual non-foci areas to the foci with which there is the heaviest commuting association. A very similar methodology was used with migration data by Coombes *et al.* (2004) to create a set of 'Housing Market Areas' for housing policy developments. Here areas of relative in-migration self containment were defined.

1.1.2 An interaction data or a migration/commuting variable classification?

At this early stage, before a full discussion of the benefits of developing a classification is had, an important decision needs to be made: is this to be a true *interaction data* classification – a classification which concerns itself with both origins and destinations and the flows between them – or is this to be a classification based on *migration and/or commuting variables* – a classification which is concerned more with single areas? The answer to this question will guide the rest of the discussion and the rest of this paper. Choosing the former will mean that the work at this stage is likely to continue down the route of functional region creation, whereas the latter will mean that the work follow the route of geodemographic classification.

One of the benefits of geodemographic classifications is that although areas in the classification may not be in close geographic proximity, they might share similar characteristics meaning they can be classified similarly. One of the drawbacks, in this context, of using functional regionalisation techniques is that contiguity tends to be a constraint – areas with similar origin/destination flows will be in close geographic proximity. An area in Scotland would almost certainly not be grouped with an area in London, even if the types of migrant flowing into and out of these areas are similar.

Of interest at this stage of the research is whether areas in different places exhibit similar characteristics. As such the classification at this stage will head along the geodemographic route and will be interested in the profiles of single areas rather than the interaction between areas. This is not to say, however, that the classification of migrant flows will not be of interest at a later stage in the research process. Indeed data on flows created from a set of migration based functional regions could well be used in a standard geodemographic classification, with variables such as the proportion of migrants from inside or outside of an area's functional region being used as an alternative to standard distance measures. At this stage, however, functional regionalisation and interaction data will give way to geodemographics and migration/commuting variables.

1.2 Why create a migration/commuting data classification?

The preceding discussion has illustrated why the idea of classification has been appealing and has looked at how the creation of different types of classification (both geodemographic and functional region) for a variety of purposes can be beneficial, both as an initial aid to understanding by reducing the amount of information we need to process and understand in order to appreciate phenomena, but also as a foundation for subsequent analysis or research through the use of the taxonomy directly and also through the by-products of the data clustering process – the key variables which help define the groups and clusters within the classification. A classification, whilst useful in its own right is often only the starting point for additional exploration. Having answered the question as to why create classifications at all, attention must now be turned to the slightly more difficult question of 'why create a classification based upon migration/commuting data variables?' At least part of this question has been answered in the previous discussion, although there are facets of the question specific to migration/commuting data which need to be dealt with separately.

1.2.1 Classification to aid understanding

Dealing first with what has already to an extent been answered; creating a classification (or indeed *classifications*) based upon migration/commuting variables will aid the understanding of what are inherently more complex data than the standard counts of people residing in places displaying particular attributes (which comprise the bulk of the data in most social surveys). When examining standard census or other social survey data, the counts will relate to a defined geographical area. Migration/commuting data on the other hand, relate to both an origin and a destination or numerous origins and destinations. The two can be seen to be connected through the flow of individuals between these locations. Taking permanent migrants as an example, these individuals residing in an area will of course display many of the same attributes as the non-migrant population: they will be male or female; of a certain age, socio-economic category or ethnicity, etc. In addition, however, they will have a number of attributes associated with them which separate them from the non-migrant population: whether they have moved in, out or within the current area; whether they have moved short or long distances; if they have moved into the area, whether it is from an overseas origin; whether they have moved as part of a household or moving group or as an individual migrant. It is these unique and complex features of migrants that mean areas hosting them can be classified separately from existing classifications. It may be very useful from a policy perspective to know, for example, if an area is particularly prone to receiving relatively high numbers of elderly in-migrants or losing high numbers of skilled workers (human capital). Furthermore, it may also be that migrant populations either leaving or

moving into areas are not representative of the underlying resident populations, something which could certainly influence policy decisions – a poor inner-city area with a large transient student population may require different resource targeting to an inner city area with a sedentary young population. It is also the case that while some areas are very popular origins or destinations for migrants, there are also areas which are very isolated. There are some areas which will not send or receive very many migrants and consequently will have distinctive characteristics of their own.

1.2.2 Classification as part of the research process

Both the classification itself and the development of a classification can be seen as part of the wider research process. The nature of this research will of course vary, but some key themes can be identified: the concept of change is an important one to consider both as it will provide avenues for this research, as well as obstacles to overcome. It could be argued that due to the fluid nature of populations in most areas (an average of around 10% of the population across the UK lived at a different address in the year preceding the 2001 Census), the moment any data are recorded at a given time-point, the further away from that time point we move, the less likely it is that those data remain relevant. Of course this is precisely why there is a continuing programme of collection with most social surveys and considerable interest in longitudinal analysis across the waves presented by these surveys. This is also why organisations such as CACI and Experian are keen to publicise their ACORN and MOSAIC geodemographic classifications as 'latest' versions (http://uk.experian.com/business/products/data/113.html) and (http://www.caci.co.uk/acorn/whyupdated.asp) and why the ONS have released new sets of area classifications in tandem with census results since 1961 (Rees *et al.*, 2002a).

What a classification with its roots in one specific time period does do, however, is allow for the exploration of change over time. A classification based upon migration/commuting variables will be tied inextricably to the time period associated with the collection of those variables. Whilst this means that, potentially, the further away from that time period we move, the less relevant the classification will be; testing a hypothesis along those lines should reveal both information about change over time as well as the extent of the change. Such a hypothesis could potentially be tested by examining flow data from an alternative, more regular source such as the NHSCR (in the case of migration flows), over successive years for the different classification areas. Of course it may also be that the underlying structures which define migration in Britain – such as the interregional structures defined by Raymer and colleagues (Raymer *et al.*, 2007; Raymer *et al.*, 2006) which demonstrate a certain stability in the origins and destinations of migration flows,

might also be applicable at finer geographical scales. In this case, it is the absence of change that is telling.

A classification in one time period also allows for the opportunity for a similar classification to be developed at a later date using a similar methodology and data from later waves of the same sources; the differences and similarities between the two revealing the extent of any change. In addition the classification will reveal areas which are more or less susceptible to change, and in doing so will give clues as to the temporal validity of other area classifications. Research by Orford *et al.* (2002) already cited reveals that in the case of the early classification of London developed by Charles Booth in the nineteenth century, there were significant similarities between areas of poverty and affluence in Victorian London and areas of poverty and affluence today. Indeed, Orford *et al.* demonstrate that Victorian socio-economic conditions are a strong predictor of present day mortality.

Research stemming from the development of a migration/commuting data classification need not be limited to uses directly related to the classification itself. The process of analysing and clustering data to create a classification will produce a set of significant variables (Vickers *et al.*, 2005) which might be put to subsequent, alternative uses. These variables only become apparent through the classification building process: As Everett and Dunn (2001) relate, a classification will consist of a small number of homogenous groups or clusters. The Vickers *et al.* (2003) area classification already discussed consists of a number of districts grouped according to the similar characteristics of the individuals residing within, however, whilst the classification groups were created using 2001 Census data variables, all available variables from the Census were not used; rather a selection of as limited a number of variables as possible were chosen. These selected variables represent the main dimensions of the parent data source and were chosen to have as much variation as possible across the whole spatial system whilst also showing as little correlation with each other as possible (Vickers *et al.*, 2003). A series of analyses were needed to decide which variables should be chosen to summarise the whole dataset. Principal Components Analysis (PCA) was used to initially to identify which variables *drive* the dataset. That is, because principal components explain all of the variance in any particular data matrix (Kline, 1994), variables which comprise more of any one principal component can be seen to be more important. In addition, correlation matrices were used to exclude highly correlated variables and standard deviation statistics were employed to select variables which varied more across the range of districts in the spatial system. By using these techniques together, an initial large set of variables was reduced to a smaller set of significant variables which could be used to classify any particular area.

Through adopting a similar process of variable selection for a migration/commuting data classification, important variables which reveal the most about population flows will be exposed. It is likely that some variables will be distributed over space far more evenly than others, with others showing greater spatial concentration. For example, male and female migrants are likely to be distributed widely and relatively evenly, whereas migrants of certain ages are more likely to be concentrated in particular areas – university towns for those in their late teens and early twenties or coastal resorts for those at retirement age (Dennett and Stillwell, 2008a). Where particular variables exhibit much greater spatial variation, they could be used to explore what is determining, maintaining and/or changing population flow patterns. This could be explored by using these variables in models of the interaction data system – models which could eventually be used to predict future flows.

1.3 Considerations for a migration/commuting data classification

There are a number of questions that need to be answered before the process of developing a classification can begin. Perhaps the two most important questions, closely related, are: what should be the scale of analysis and which data should be included? The two are linked as data will be available for discrete geographical areas, the choice of scale affecting both the availability of variables as well as the application of the classification and vice versa.

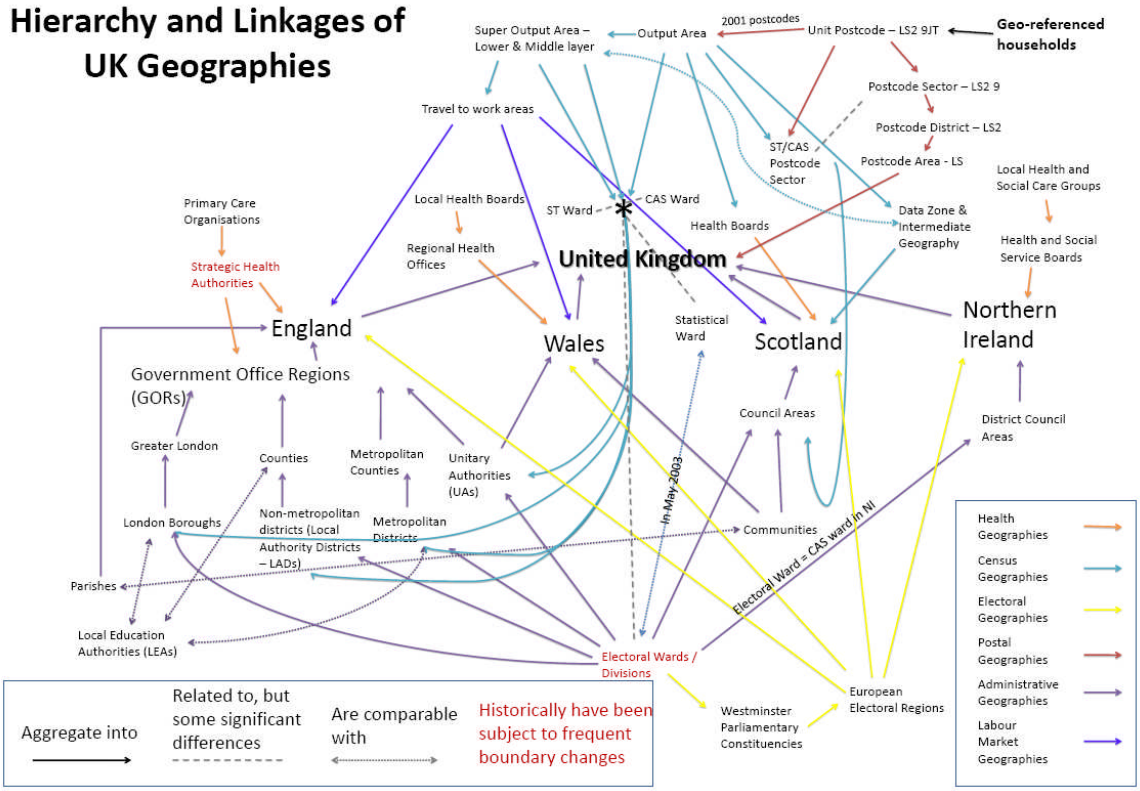
1.3.1 Scale and interaction data

All interaction data and associated migration/commuting variables available in the UK are available for discrete geographical units (Dennett *et al.*, 2007). Within the UK there is a complex hierarchy and linkage of geographies, depicted in Figure 1.1, in which all small area geographies eventually aggregate up to the coarsest country level, although as is also evident, not all lower level geographies are compatible. For example, electoral wards aggregate into districts and parliamentary constituencies, but these two geographies cannot be harmonised. This is an issue for the creation of any national classification where data from disparate sources are produced for different geographies.

Different datasets have different levels of geographical detail. For example, as is summarised in the sixth column of Table 1.1, interaction data from survey data sources in the UK are generally not available below the level of Government Office Region (GOR). The only exceptions are the Labour Force Survey (LFS) where, with special permission, data can be accessed at the local authority (district) level, and the International Passenger Survey (IPS), where destinations can be obtained at the county level. Dennett *et al.* (2007) explain in detail all the interaction data sources available in the UK and the particular geographies applicable for each data set. To summarise,

other than survey data sources, census data are available between output area and GOR geographies (depending on the data set in question) with administrative data sources available for a range of different geographies ranging from output area (in the case of Hospital Episode Statistics data (HES) – although national availability of this data, whilst theoretically possible, is subject to various confidentiality and financial restrictions) to GOR levels.

Figure 1.1 A schematic representation of the hierarchy and connectivity of UK geographies, 2008.



Of relevance to the interaction data and associated migration and commuting variables for single areas, as has been discussed previously by Dennett and Stillwell (2008a), are issues of national geographical compatibility, particularly with census data. As seen later, the 2001 Census proves to be the most useful source of data for classification purposes. Where interaction data are concerned, however, geographical compatibility issues will occur at the national scale at particular data levels. For example, at level 1 (districts) origin-destination statistics are not available for comparable geographies across the whole UK. In Northern Ireland, data at level 1 are available for parliamentary constituencies rather than district council areas. This is an issue, for example, if comparisons were to be made with other national, district level classifications. Furthermore, looming over both the issue of scale and data are the more conceptual issues of the modifiable areal unit problem (MAUP) and the related problem of the ecological fallacy.

1.3 Considerations for a migration/commuting data classification

Table 1.1 Summary of survey data sources from which interaction data are available (Taken from Dennett *et al.*, 2007).

Survey	LFS	LFS NI	IPS	GHS	CHS	IHS	NTS
Start date	1973	1973	1993	1971	1983	2008	1965/66
Current sample size	53,000 Households, 126,000 individuals annually.	8500- 9000 individuals annually.	250,000 passengers annually.	8000-10,000 households, 15,000-20,000 individuals annually.	4,500 households (around 1% of Northern Ireland total).	204,000 households annually.	16,000 households annually.
Current timing	Calendar quarterly sampling and release.	Calendar quarterly sampling and release.	Continual sampling, quarterly compilation, annual release.	Annual release.	Annual release.	Rolling annual release of calendar quarterly datasets.	Data collected on sample 'travel week' for study sample over course of a year. Annual release.
Main variable types covered (Variables flows can be disaggregated by).	Age, gender, ethnicity, level of education, marital status, religion, number of dependent children, employment type, sick days, socio-economic classification.	Age, gender, ethnicity, level of education, marital status, religion, number of dependent children, employment type, sick days, socio-economic classification.	Age, gender, UK port or route, type of vehicle, type of fare, purpose of visit, intended length of stay, money spent on beer, wine, spirits and cigarettes, overseas origin or destination.	Household members, household and family information, household accommodation, housing tenure, consumer durables including vehicle ownership, employment, pensions, education, health and use of health services, income.	Household members, household and family information, household accommodation, housing tenure, consumer durables including vehicle ownership, employment, pensions, education, health and use of health services, income.	Covering everything included in the LFS and GHS, with additional information on all aspects of household income and expenditure, as well as much information that is normally included in the omnibus survey.	Accessibility of public transport, access to amenities, household vehicle access, household composition and household socio-economic information, age, gender and marital status, employment, occupation and industry details, income, place of work and travel to work details.
Interaction data	GOR to GOR and International country to GOR interaction matrices possible. Disaggregation by any variable of choice. UA/LAD to UA/LAD with special permission.	GOR to GOR and International country to GOR interaction matrices possible. Disaggregation by any variable of choice. . UA/LAD to UA/LAD with special permission.	International country of origin to UK county matrices possible. Disaggregation by any variable of choice.	Very little. GOR of destination is all that can be accurately measured. Origin is either current GOR or 'elsewhere.' There is no way of telling which.	Only for 1983. NI electoral ward or council area can be origin or destination. Immigration from GB or Eire also available for this year alone.	Expected to be the same as the LFS.	GOR to GOR commuting data is available readily for most recent years. This data should be available in theory for other years too, although in practice availability is variable.

1.3.2 The MAUP and the ecological fallacy

The modifiable areal unit problem can create difficulties when analysing aggregate data for discrete geographical areas. O'Sullivan and Unwin (2002, p30) describe the problem succinctly thus: *"aggregation units used are arbitrary with respect to the phenomena under investigation, yet the aggregation units used will affect statistics determined on the basis of data reported in this way."* The problem is duel faceted: the first relates to scale, the second to zoning. Taking the former, patterns identified in that data at one scale of aggregation may not present themselves at a different level of aggregation. For example, ten deaths related to cancer in a year for one output area might represent a high cancer death rate for that level of aggregation. However it may be that when examining other output areas in the same district there are no more cancer deaths. This would present a low cancer death rate for the district. Taking the latter, it might be that the boundary for the output area in this example is different to previous enumeration district boundaries, and in fact if previous boundaries were in place, the ten cancer deaths in the new output area would be distributed across three old enumeration districts, thus presenting lower, less significant rates. For a more detailed explanation of the problem and its effects on spatial data, see Openshaw and Taylor (1979).

The ecological fallacy has some commonality with the MAUP, although is a slightly different problem. Where the two are similar, as pointed out by O'Sullivan and Unwin (2002), is that both issues make it apparent that statistical relationships can change at different levels of aggregation. The ecological fallacy emerges from the practice of 'ecological inference,' described by King *et al.* (2004) in the preface to their book as the *"reconstructing of individual behaviour from group-level data."* That is to say that the ecological fallacy is the problem of inferring something at a lower level of aggregation, from something observed at a higher level. We make ecological inferences commonly in everyday life – perhaps when deciding upon a holiday destination, because a particular country has a reputation for good beaches, and inferring (rightly or wrongly) that because a particular resort lies within that country, it too will have good beaches. The ecological inferences also form the basis of many governmental decisions – the recent ban on smoking in public buildings was in part an effort to reduce the numbers of smokers, and this was largely due to compelling aggregate evidence that cases of heart disease and lung cancer are more prevalent among patients who smoke. Of course at an individual level, there are always exceptions to this general rule and one cannot say that *all* smokers will develop heart or lung problems. It is understandable, however, that ecological inferences are made in this context, as it would be impossible to tailor manageable policy to individual needs.

1.3.3 Additional issues of scale

Bringing the discussion back to the issue of scale and data choice, it is inevitable that any decision has the potential to create problems and these will need to be acknowledged. The question is will any particular scale of analysis create any more or less problems? Harris (2005) asserts that all users of geodemographic typologies will need to contend with issues of representation. Whereas classifications describe areas, some users will try to infer the characteristics of individuals from these areas and it is inevitable that general classifications will not be entirely representative of the whole population. This is perhaps more of an issue where geodemographic classifications are constructed from micro-data and apply to small areas – for example output areas or unit postcodes. The temptation is to assume that as the level of areal aggregation is reduced, the likelihood of generalisations being accurate increases – indeed Farr and Webber (2001) assert that analyses have shown data at the level of the person discriminate better than more aggregate data. Of course, even at the household level, generalisations can be inaccurate. The key to the utility of the classification lies in the purpose for which the typology was created and the use to which it is eventually put. If an area classification is created at the district level (for example, the ONS classification of local authorities) and is designed to summarise districts in terms of their key characteristics, then providing the methodology is sound, the classification should be fit for that purpose. Problems will only start to arise if assumptions are made about the population residing within any given classification area below (or indeed above) the level of that described by the classification. Just because the district features high proportions of elderly residents, it does not mean that all areas within the district will also feature high proportions of elderly residents. It sounds a very obvious point to make, but this type of assumption is what lies at the heart of the ecological fallacy. Returning to the original question, however, it is unlikely that any particular level of analysis will create any more or less problems for the user; issues associated with the MAUP and the ecological fallacy will propagate at any level of analysis – the extent to which they will matter will depend upon the final use of the classification.

Where scale may present some problems is in the design and creation of the classification. The larger the area being classified, the more chance there is that variation within the area will have a cancelling effect. Consider Figure 1.2. The large box representing a spatial system sub-divided into four smaller zones containing equal numbers of white and coloured circles (which could represent either groups or individuals). When looking at this whole system, concentrations in smaller areas are ignored. Furthermore, by having three categories of circle that are not white, concentrations of coloured circles, even at a finer spatial scale do not become apparent. For

example, taking the right-hand half of the whole area, the blue and the white circles are in equal concentration, exactly as the concentrations in the smaller bottom-left square, the bottom right square and the top right square. However, if the blue circle were classified as non-white along with the green and red circles, then suddenly concentrations that were masked at the aggregate level become more obvious.

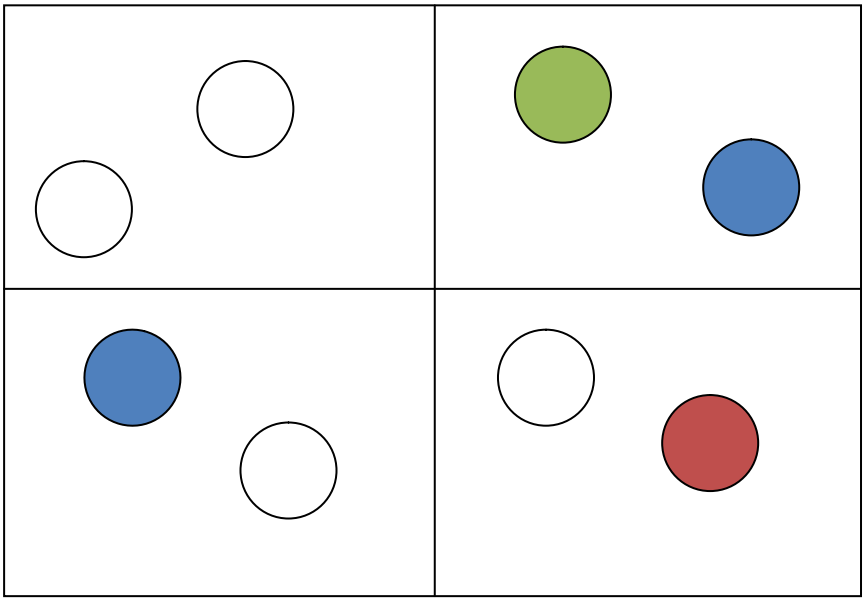


Figure 1.2 Cancellation of variation through area size effects.

When creating an area classification, it is likely that, in order to observe variations in larger areas, variables may need to be more generalised, especially where concentrations of one variable might be low. If the large square in Figure 1.2 represented a city district and the circles represented ethnic groups, then in order to discern concentrations of the less populous non-white groups they would need to be classified generally as 'non-white' rather than separate individual groups. It is difficult to predict, however, exactly how this effect will present itself before initial clustering experiments are run.

1.3.4 Data: Internal migration, international migration, commuting or a combination?

Where scale has a significant effect on the amount of data available and accuracy of the data, a decision about scale of analysis and choice of data source will be more influenced by these issues than by any inherent theoretical problems associated with using one scale over another. Thus far in this discussion, reference has been made to a 'migration/commuting data classification', indicating that there is potential at least for the inclusion of all types of interaction data derived variables. It follows, therefore, that an important decision which needs to be made is whether the

classification should indeed incorporate all origin/destination derived variables, or instead merely elements of internal migration, international migration or commuting data. This is a decision that is both crucial and one that cannot be made satisfactorily in advance of testing variables in experimental classifications for their spatial variability and discriminatory properties. It is, however, important to make a clear distinction between the types of data that are available for inclusion:

1.3.4.1 Internal migration data

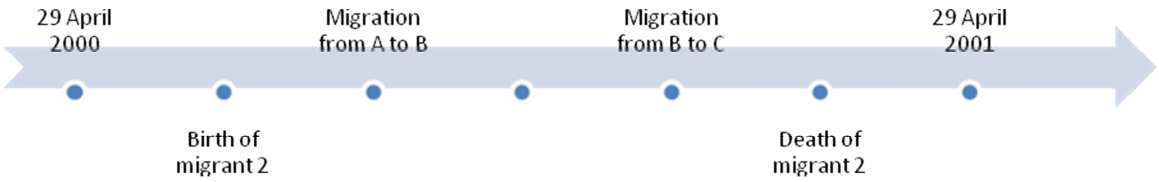
Internal migration data in the UK can be conceptualised as either 'movement' or 'transition' data (Rees, 1977). Movement data counts the number of migration moves that a migrant makes over a given time period – the age of the migrant recorded with each movement. Transition data, on the other hand, records a single migrant transition for a given time period regardless of whether the migrant has moved more than once, and the age of the migrant is recorded at the end of the period. In the UK, the two principal sources of data on internal migration, the NHSCR and the Census, are examples of movement data and transition data respectively. The two sources have a number of generic benefits and drawbacks, which have been described in detail by Stillwell *et al.* (1992), Champion *et al.* (1998) and Dennett *et al.* (2007). To summarise, there are some more obvious differences between the two datasets: the decennial census versus the quarterly/annual NHSCR; the huge range of variables in the Census versus the age/sex dimensions of the NHSCR; the inclusion of all moves in the Census including within areas versus only between area moves with the NHSCR. There are, however, some less obvious differences between the datasets related to their respective movement and transition compositions that were not covered in detail earlier.

Figure 1.3 Representation of two migrant histories over a 1 year period.

Migrant 1

Migrant 2

1.3 Considerations for a migration/commuting data classification



The exact nature of these differences is demonstrated by the two hypothetical one year migrant histories depicted in Figure 1.3. Consider migrant 1. In the one year period, the migrant moves twice, has a birthday and then moves a third time. In the NHSCR movement data, not only would all three migration events be recorded, but the age of the migrant would be recorded differently between the first two migration events and the third migration event. The census transition data, on the other hand, would not record the migrant at all. The third migration event sees the migrant move back to original location A. As the address at the beginning of the one year period is the same as the address at the end of the period a migration is not recorded. Even if the third migration event either did not happen, or the migrant moved to a fourth location 'D', there would be differences between the NHSCR and Census data. If the third migration did not happen, the census would record only one migrant transition, and this transition would record the age of the migrant at the end of the period, rather than the different age when the two migration events actually happened. If a fourth migration event happened, then whilst the age recorded would be accurate for that event, previous movements would not be recorded. Considering migrant 2, the migrant is not alive at the beginning of the period, but is then born, migrates twice and then dies before the end of the period. The census would not record this migrant. Even if either the birth or death events did not happen, the migrant would still not be recorded. The NHSCR, on the other hand, would record both migration events.

Using both NHSCR and census data concurrently in any classification could present problems because of these conceptual differences. The issues of harmonising movement and transition migration data are covered by Duke-Williams and Blake (2003). To summarise, consider Figure 1.4.

Figure 1.4 Migration data representation. Source: Duke-Williams and Blake (2003).

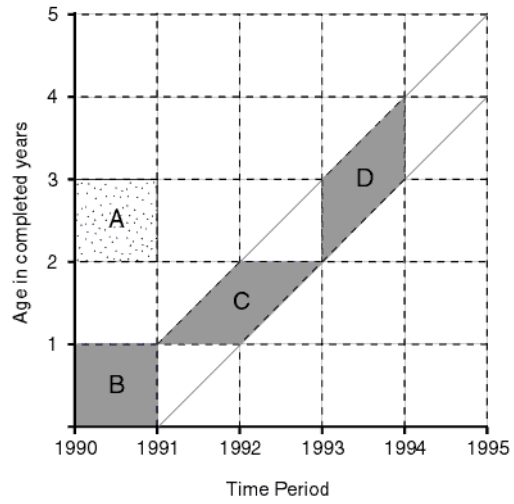
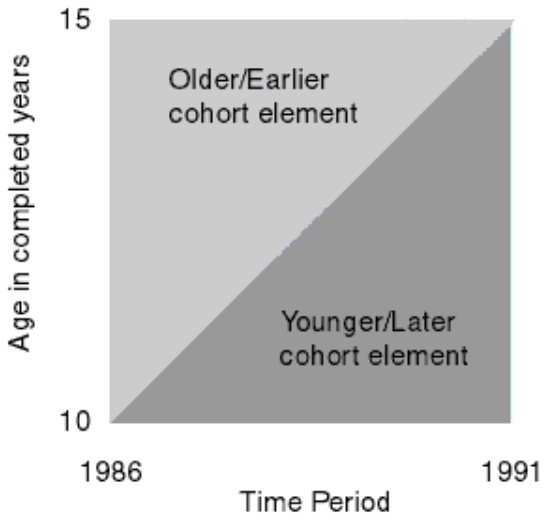


Figure 1.4 is an example of a Lexis diagram (Carstensen and Keiding, 2005; 2001; Vansershrick, 1992) which can be used to depict the age/time period/cohort elements of demographic data. The different shaded parts of the lexis diagram represent different ways in which migration data can be represented. Shapes A and B show age and period data. As is stated by Duke-Williams and Blake (2003), the dots in A each represent a single movement event of a migrant of a precise age at a precise time. B on the other hand, represents a similar age/period set of data but this time the data represents anyone who migrated during the 1990-91 period whose age at the time of migration was between 0 and 1. These representations are comparable to the NHSCR movement data. Shape C represents data classified by age and cohort, and could represent all migrants (who had moved at some point in the previous year) who turned two during the 1992-93 period (Duke-Williams and Blake, 2003). Shape D on the other hand, represents data classified by cohort and period and could show all migrants who moved during 1993-94, and were aged 3 at the end of the period. This representation is comparable to census transition data available in the UK. As is evidenced by the different shapes in the Lexis diagram, movement (B) and transition (D) data measure subtly different phenomena.

If movement and transition data are to be harmonised, Duke-Williams and Blake propose dividing the age/period square into two component triangles (Figure 1.5). With information about the precise age of the migrant and the exact date of the migration event, it is possible for migrants recorded in an age/period format to be re-assigned to either an older/earlier cohort element or a younger/later cohort element. These cohort elements can then be re-aggregated into either an age/cohort dataset, or perhaps more commonly into a cohort/period dataset comparable with census transition data.

Figure 1.5 Cohort elements of age/period data



So, because of these important conceptual differences between movement and transition data, any attempt to use one or the other, or a combination of both data types, in a migration classification, should be made with caution. The question that necessarily follows on from this discussion relates to how these movement and transition data are currently represented in the UK, and which internal migration datasets might feasibly be used in the construction of a migration classification. In answering this question, one is initially drawn to the UK Census of Population. As observed by Rees *et al.* (2002b), the census is the most comprehensive source of population data available in the UK, the last census being conducted in 2001. The range of variables, geographical coverage and the scales at which data are available are incomparable and furthermore, Census data are freely available. As such it is an obvious first destination when considering variables for inclusion. Dennett *et al.* (2007) offer a detailed account of the range of interaction data available from the census, including counts and proportions of migrants moving into, out of and within district areas included in Key Statistics table KS24, as well as other Standard Tables, data available in the samples of anonymised records (SAR), longitudinal study (LS) and specially commissioned tables. Of particular interest here, however, are the origin-destination statistics.

The 2001 Census Origin-Destination statistics provide counts of internal migrants, as well as international migrants and commuters through the Special Migration Statistics (SMS), Special Workplace Statistics (SWS) and the Special Travel Statistics (STS - in Scotland only, which also include travel to place of study as well as work). In 2001, data are available at three geographical levels: District (level 1), Ward (level 2) and Output Area (level 3) (Stillwell and Duke-Williams, 2007). The level of geographical detail has a direct influence on the level of attribute detail due to

concerns about maintaining respondent confidentiality. Consequently, data are disaggregated by more variables at district level than at ward or output area level. This relationship between attribute detail and geography is demonstrated in Table 1.2

Table 1.2 Summary of origin-destination statistics data, 2001.

Data Sets	Level 1 (District)	Level 2 (Ward)	Level 3 (Output Area)
2001 SMS	10 tables (996 variables)	5 tables (96 variables)	1 table (12 variables)
2001 SWS	7 tables (936 variables)	6 tables (354 variables)	1 table (36 counts)
2001 STS	7 tables (1176 variables)	6 tables (478 variables)	1 tables (50 variables)

In addition, the accuracy of the data becomes more compromised at finer spatial resolutions. The Small Cell Adjustment Method (SCAM) alters small cell counts of 1 or 2. With more of these small counts appearing at finer spatial resolutions, data for the coarsest geography are the most reliable. So, despite Farr and Webber's (2001) assertion that classifications discriminate better at finer scales of analysis, when one is left with only one table containing data and potentially over 90% of SMS and 70% of SWS data being adjusted at output area level (Stillwell and Duke-Williams, 2007), it might be preferable to choose a more aggregate geographical scale.

With this in mind, using the most aggregate, district level (level 1) data from the Census would appear to be the more sensible choice. Other considerations enhance this assertion. For example if additional data from outside of the Census are to be incorporated into the classification, then using a spatial scale below that of the district would make this extremely difficult. In their review of all interaction data sources available in the UK, Dennett *et al.* (2007) show that other interaction data sources are not readily available below the level of the Local Authority District. Only data on student migrations to higher education institutions from the Higher Education Statistics Agency (HESA) and data from the Pupil Level Annual Schools Census (PLASC) are available at a sub-district scale. Interaction data, even at the district level are scarce, with only NHSCR/Patient Register data available to study internal migration and National Insurance Number Statistics (NINo) available for international migration. Even if data from outside of the Census are not used, this argument is still applicable where the classification might be used as a framework for monitoring inter-censal migration from these other sources – it will be impossible to analyse district level migration data from another source using a classification made for output areas, wards or another geography. Furthermore, districts are the areas used for local governance in the UK and therefore are important for planning and policy decisions which may stem from the information presented in a classification. Adopting a district level classification does have

implications for full national coverage however: as previously mentioned, level 1 data are not available for district council areas in Northern Ireland. Therefore the classification would need to omit Northern Ireland if other district level data from outside of the census were included or if only census data were included and comparisons were to be with other district classification typologies.

1.3.4.2 Commuting data

One substantial issue that needs to be dealt with relates to whether commuting data should also be included in this classification. Commuting data are conceptually different from migration data, however practically, the nature of these differences are perhaps more down to the way in which commuting data are defined through being recorded, rather than a definitive conceptual boundary separating the two phenomena – this is especially the case where commutes are of a longer distance and involve spending significant amounts of time in another location as the result of work. As with migration data, the census provides the most comprehensive and detailed source of commuting data in the UK, with commuting data featuring as part of the Origin-Destination Statistics already described. In the census, commuting data are generated through a question relating to the address of the usual place of work. Whilst this will not capture all commuting movements, with many employees tending to work at one main location, a large proportion of commutes will be captured. The only other source of travel-to-work data in the UK is the National Travel Survey (NTS) (Dennett *et al.*, 2007), although this is only a sample of 16,000 households and can only offer GOR to GOR commuting flows.

Many would accept that whilst both migration and commuting involve moving between origins and destinations, the difference between the daily journeys to and from work and a permanent change of residential location is easily identifiable. What happens though when an individual lives in more than one location, or if during the week an individual works and lives in one location, but at the weekend returns to a place they more readily identify with as home? Here the difference between a commuter and a migrant becomes a little more difficult to define, both for those living a life where the boundaries between migration and commuting are more blurred, and for those wishing to record the phenomena. With a census form offering no opportunity to account for these less common movements, it becomes difficult to define exactly where some migration and commuting events cross over. This is perhaps exemplified when one looks at commuting data taken from the 2001 SWS (level 1) showing the main method of transport to work for commuters who stated they worked in London, but lived in other regions of the UK. (Table 1.3).

Table 1.3 – Commuters to London from UK regions by method of travel to work, 2001.

Origin	Destination	Total	Works or studies mainly at or from home	Underground etc	Train	Bus etc.	Taxi	Car – driver	Car – passenger	Motorcycle etc.	Bicycle	On foot	Other
North East	London	3140	0	303	340	293	9	1472	135	33	24	221	310
North West	London	8676	0	673	1162	520	48	4780	259	72	94	506	562
Yorkshire and The Humber	London	6473	0	485	1117	524	36	3345	252	36	54	429	195
East Midlands	London	13767	0	559	5078	475	36	6404	391	93	140	456	135
West Midlands	London	10435	0	618	2043	709	27	5906	325	54	93	534	126
East of England	London	283750	0	18785	123721	5262	361	121605	6842	4420	746	1473	535
South East	London	374829	0	6736	144693	11847	424	190233	8930	6321	2063	2657	925
South West	London	16243	0	1218	4607	667	30	7966	344	180	210	811	210
Wales	London	3687	0	302	615	158	6	2060	144	30	30	253	89
Scotland	London	0	0	0	0	0	0	0	0	0	0	0	0
Northern Ireland	London	674	0	0	76	54	15	240	24	6	13	81	129

As is shown by Table 1.3, some of the statistics relating to the method of travel to work do not appear to make empirical sense. For example there are 506 people who walk and 94 people who cycle as their main method of transport to work while living in the North West and working in London! It is quite obvious that these people will not be making the return journey of several hundred miles by foot and on bike on a daily basis, so it raises the question why do these 'improbable journeys' exist? The likelihood is that many of these people will have what they class as their permanent home address in one part of the country, but will spend a large amount of time living and working in London and travelling to work by these methods. If this is the case then it is debatable whether these individuals should be recorded as migrants or commuters. Unfortunately the census does not currently record (and does not look in the future like recording) these more complicated interaction flows.

So there is an argument for including commuting data in an interaction classification as a certain proportion of commuting moves could feasibly be reclassified as migration moves. The problem is, however, that it is impossible to accurately differentiate these ambiguous flows. A further argument for the inclusion of commuting data would be that the relationship between employment related migration and commuting is a close one. Work by Eliasson *et al.* (2003) states that there is clear evidence that the accessibility of job opportunities has an impact on whether individuals choose to migrate or commute to those jobs; increased accessibility having a more noticeable negative impact on migration than a positive one on commuting. With such a close relationship between some employment-related migration events and commuting, it might be that including commuting data in an interaction classification would prove useful in discriminating different areas with different interaction profiles.

1.3.4.3 International migration

The final type of interaction data that are available are international migration data. As with internal migration and commuting data, the census provides some measure of international

migration. Census Key Statistics table KS24 records the proportion and count of people who move into each district in the UK from outside of the UK. SAR, LS and Origin-Destination data also feature similar 'outside the UK' variables for differing geographies. Specially commissioned data does feature some tables with more detailed international origin locations, with specially commissioned table C0711 disaggregating the foreign origin to British district by 56 countries.

One other major source of international migration data at the district scale are the national insurance number statistics (NINo) produced by the Department for Work and Pensions (Boden and Rees, 2009; Dennett *et al.*, 2007). Produced since 2001-2002, the NINo statistics provide a record of any foreign worker entering the UK to work legitimately, regardless of their length of stay. In this way they are conceptually different from the census (which records international migrants as permanent residents who lived abroad in the year preceding the census – regardless of their employment status). The other principal source of international migration data is the International Passenger Survey (IPS) (which records international migrants as individuals who stay in the UK for at least a year following arrival) which is the main source of data used to produce the Total International Migration (TIM) estimates by the ONS (ONS, 2008).

There is a case for including international migration variables as there is some evidence of linkage between immigration/emigration patterns and more localised in-migration/out-migration. Stillwell and Duke-Williams (2005) show that in the case of at least some UK districts, a relatively high rate of immigration is coupled with a relatively high rate of net out-migration. This is not the case everywhere however, and Stillwell and Duke-Williams do concede that a closer look at the types of districts where this is not the case is needed. With this being the case, the inclusion of international migration variables in a classification could well help discriminate areas with similar internal migration profiles.

The quality of international migration data, however – especially that which comes from the IPS which has a very small sample – has been debated, especially where data are estimated below the regional level. Recognising the limitations of migration statistics in the UK, a programme known as the Improving Migration and Population Statistics (IMPS) (Inter-departmental Migration Task Force, 2006), led by the ONS has been set up, part of which addresses the limitations of the current IPS data. Additionally, a broader Migration Statistics Improvement Programme (MSIP) has been ongoing since 2008 which is also examining limitations of immigration and emigration data. As the work of both IMPS and MSIP is still ongoing, the quality of any data from TIM will be restricted. As such, if international migration data are to be used in the classification, it may be wise to include only data from the census – even if this means excluding emigration.

1.3.6 The unique problems associated with creating a classification based on migration/commuting data

In some respects, creating a classification from standard area-based data is relatively straightforward. The data refer to particular attributes related to the population in that area at a given time. The classification is classifying an area by a static population. Each attribute therefore is related to only one state of the population. With interaction data on the other hand, flows of individuals create additional complexity which makes classification more challenging.

For example, any given area can have variables related to counts or rates of in-migration, out-migration or within-area migration, and these can be combined to create even more variables, such as net migration, migration effectiveness, turnover or churn (Dennett and Stillwell, 2008b). There can be similar measures for commuters, and all of which can be further categorised with measures of distance. There are also international migrants to consider, as well as migrants where the origin is unknown.

So any variable can be disaggregated by a whole range of interactions. This means that when creating a classification based upon interaction data, a decision may need to be made about which types of movement to include: whether to include all types of movement in one classification, some types of movement or to create separate classifications for each type of movement.

1.4 Initial decisions

The preceding discussion has set out the case for a migration/commuting classification and has highlighted some of the issues that will need to be addressed as the research progresses. It will be possible to address some of these issues, as the classification is developed, however before continuing some key, the inter-related issues of what scale, which data and whether to classify flows or individuals in the first instance needs to be taken. Work leading up to this paper by Dennett and Stillwell (Dennett and Stillwell, 2008b, 2009, Forthcoming) was at the district scale, so it would seem appropriate that this work continues at this scale – certainly in the first instance. Whilst this in itself may not be a solid reason, other considerations do enhance the suggestion. For example, data availability and accuracy is improved at this scale, and it would seem sensible at the beginning of an exercise such as this to begin with as much data as possible. This does, however, mean that the classification will be limited to Britain. The best source of data in terms of coverage, accuracy and detail is the census, so the sensible first choice would be to use census data.

2 Developing a migration classification

"If the process of clustering is likened to an animal then it is a very peculiar beast! It has the front legs of automation but the back legs of user intervention; eyes for data-led classification but the ears of a priori expectation; it feeds on a variety of data sources but generally prefers a census; displays a patchwork coat mixing the qualitative and the quantitative, the objective and the subjective; and is born of a cross-breed between art and science!" (Harris et al., 2005).

2.1 An initial district level area classification based upon migration variables

As is noted by Založnik (2006, p10), geodemographic classifications 'invariably produce plausible results'. That is to say whatever data are input into a clustering procedure, the resulting outputs can often be interpreted in a way that can make intuitive sense. As Založnik points out, this is both a great strength and great weakness of the process. How then can one be sure that the classification output from a clustering procedure is 'optimum' – i.e. most accurately reflects the key patterns in the underlying data? The answer is probably 'never', as with any generalisation, detail will be lost that some may argue is important. In practice though, it is possible to create a more robust classification though careful decision making at each step of the process. This presupposes that the process can be theorised as a series of steps, which indeed it can. A number of authors have considered the process of designing and creating a classification and a general framework for this process which has been suggested several times (Everitt *et al.*, 2001; Shepherd, 2006; Vickers, 2006) is that proposed by Milligan and Cooper (1987) and Milligan (1996). It consists of seven sequential steps which organise the clustering process from start to finish. Creating a classification from the beginning, it would seem appropriate, therefore, to adopt Milligan's approach. The steps outlined below from 2.1.1 to 2.1.7 are those suggested by Milligan. As with any piece of work, it is unlikely that the first 'draft' will be the same as the final piece. An initial 'draft' of the classification will be created here and reviewed. Where improvements to the initial methodology and decisions can be made, these will be discussed and implemented in section 2.3.

2.1.1 Objects to cluster

Local authorities have been chosen as the areas to cluster within the whole spatial system. Here the whole system is Britain rather than the UK for the reasons already mentioned, so the districts to cluster will be the 408 districts of England, Wales and Scotland.

2.1.2 Variables to be used

Earlier in this paper, reference was made to the difficult decision concerning the variables to include in an interaction data classification. As discussed, sources of data other than the census are less attribute rich and sample far fewer individuals. Consequently the classification taxonomy in this initial classification will be developed solely from 2001 Census data. A summary of the data tables available from the census is given below in Tables 2.1:

Table 2.1.1 - 2001 SMS tables

Table Reference	Table Name	Cells/variables within table
Table 1	Age by sex	75
Table 2	Family status of migrant	54
Table 3	Ethnic group by sex (GB destinations)	24
Table 3n	Ethnic group by sex (Northern Ireland destinations)	9
Table 4	Whether suffering limiting long term illness by whether in household by sex by age	84
Table 5	Economic activity by sex	42
Table 6	Moving groups	16
Table 7	Moving groups by tenure	32
Table 8	Moving groups by economic activity by sex	336
Table 9	Moving groups by NS-SEC of group reference person	288
Table 10	Migrants in Scotland/Wales/Northern Ireland with some knowledge of Gaelic/Welsh/Irish	36

Table 2.1.2 - 2001 SWS tables

Table Reference	Table Name	Cells/variables within table
Table 1	Age by sex	108
Table 2	Living arrangements by employment status by sex	216
Table 3	Method of travel to work	156
Table 4	NS-SEC by sex	144
Table 5	Industry by sex	156
Table 6	Ethnic group by sex (England and Wales residences)	96
Table 6n	Ethnic group by sex (Northern Ireland residences)	36
Table 7	Employment status by sex	36

Table 2.1.3 - 2001 STS tables:

Table Reference	Table Name	Cells/variables within table
Table 1	Age by sex	199

Table 2	Family status by sex	258
Table 3	Method of travel to place of work or study by sex	184
Table 4	NS-SEC by sex	174
Table 5	Industry by sex	201
Table 6	Ethnic group by sex	120
Table 7	Employment status by sex	51

Even with the source of data decided upon, a decision also needs to be made about which variables to include from this source. An important choice is whether to include all families of interaction variable: internal migration, international migration and commuting; or to select only one or two. Internal and international migration variables are available from the same tables in the SMS and consequently are disaggregated identically. Commuting data available from the SWS are disaggregated slightly differently, even when tables are given the same title. For example the 'age/sex' table in the SMS is disaggregated by small age groups (between one and five years) whereas the comparable table in the SWS is disaggregated by age groups which vary between two and ten years. This, however, this is not a major issue as age groups from each table could be re-aggregated to form comparable groups. Whilst migration and commuting phenomena can be related, (as discussed earlier), at this early stage it will be more manageable to deal solely with migration variables. There may be scope for incorporating commuting variables into another classification (if indeed it is felt commuting variables will add something to a classification) at a later stage in the process.

Throughout the literature warnings abound that choosing appropriate variables is very important if not key to the success of the final classification produced. Whilst the use of statistical techniques can certainly help whittle down the choice of variables systematically, as will be shown later on, Openshaw and Wymer (1995 p244) suggest that *'[t]here is no statistical technique that is a good substitute for thinking about choice of variable, yet!'* Certainly in the case of a migration classification, careful thought should be put into whether particular variables are likely to influence migration events or patterns. With this in mind it is useful to assess groups of variables as to their suitability for inclusion. The tables listed in Table 2.2 usefully sort variables into sets for consideration.

Age and sex

As has been shown in chapter 4 and in a number of other pieces of work (Bates and Bracken, 1982; Dennett and Stillwell, 2009; Raymer *et al.*, 2007; Raymer *et al.*, 2006; Rogers and Castro, 1981; Rogers *et al.*, 2002), age has a significant influence on the propensity to migrate, as well as the direction and volume of migration, with very large numbers of migrants in their late teens and

early twenties gravitating towards larger conurbations; migrants in the family rearing ages moving out of cities into rural areas; and post-retirement migrants moving to coastal areas (Uren and Goldring, 2008). Therefore the inclusion of age variables is of great importance to any migration-based classification.

The case for the inclusion of sex variables is less clear-cut. Evidence from past analysis (Champion, 2005) has tended to indicate that there is little difference between the migration patterns of males and females. However, as has been demonstrated elsewhere, (Dennett and Stillwell, 2010) there are some variations by sex, especially at different ages. The propensity for females to migrate at the age of peak migration (late teens and early twenties) may warrant the inclusion of sex variables in a Migration Classification.

Family status

Cooke (2008) provides a comprehensive review of research which has been carried out on the many complex family-based influences which can affect migration flows, from marriage to family formation to divorce. The influence of family status can influence both the motivations for moving and the moves themselves (Geist and McManus, 2008), and can interact with other influencing factors. For example, work by Boyle *et al.* (1999) and Cooke and Bailey (1999) has made the link between the differing employment status of female migrants who move either alone or as a part of a family. Certainly, therefore, a case can be made for the importance of including family status variables in a migration-based classification as origins and destination particularly favoured by migrants moving in families may have labour market implications. Furthermore, as Castro and Rogers (1981 p vii) note '*many internal migrations are undertaken by individuals whose moves are dependent on those of others'*. It may well be that the origins and destinations of group or family movers are markedly different from those who move independently of others.

Ethnic group

The particular patterns of migrants of different ethnicities within the UK have been the focus of a number of recent pieces of work (Faggian *et al.*, 2006; Finney and Simpson, 2008, 2009; Hussain and Stillwell, 2008; Owen, 1997; Raymer and Giulietti, 2009; Raymer *et al.*, 2008; Simpson and Finney, 2009; Stillwell and Hussain, 2008; Stillwell *et al.*, 2008). All of this work suggests there are differences in the migration propensities between ethnic groups. It may be that in some cases the patterns are confounded by other variables such as age and socio-economic status, although despite this, with concentration of non-white groups predominantly in urban areas, particularly cities, the identification of areas where ethnic minority migrants are more commonly moving in or out will be useful in developing a clearer migration picture for Britain.

Limiting long-term illness

Research carried out by Norman *et al.* (2005) has focused on the health of migrants and the implications for the origins and destinations associated with healthy or less healthy migrants. Norman *et al.* (2005) discovered that whilst (amongst the young) migrants are generally healthier than non-migrants, in less deprived areas migrants are healthier than non migrants, but in more deprived areas migrants are less healthy than non-migrants. They also found that healthier migrants move away from deprived areas, increasing the rates of ill health and mortality in these areas, and interestingly a significant number of unhealthy migrants move into more deprived areas, exacerbating this increase in ill health and mortality rates still further. With this in mind, the inclusion of variables related to limiting long-term illness may certainly highlight areas, which, if characterised by flows of unhealthy migrants could flag important changes in the concentrations of ill health.

Economic Activity

Much has been written on the influence of economic activity on direction and volume of migrant flows. From the work of Ravenstein (1885; 1889) well over one hundred years ago which observed the pull of urban areas for rural workers, to more recent work by Fielding (1992) and Findlay et al. (2009) which characterises the South East of Britain as an 'escalator region' for economic migrants, the influence of employment availability on migration flows has been well documented. Whilst the economic condition of origins or destinations may influence the flows of migrants, the economic condition of migrants themselves may also be influential. Work by Bohara and Krieg (1998) in the United States provides evidence of a linkage between levels of income and the propensity to migrate. Dixon (2003) recounts a similar story in the UK, showing through time-series analysis that those in the highest socio-economic groups are far more likely to migrate between regions than those who are less educated and employed in less skilled jobs, with Böheim and Taylor (2002) making an alternative observation of a strong link between unemployment and migration propensity. If economic reasons are the influencing factor for a very large number of migration events, then while examining whether migrants are employed may not tell us a great deal as there are large differences between the earnings of those employed at the bottom of the socio-economic scale and those at the top, examining those who are not employed may tell us more. Furthermore, additional categories of economic activity such as 'retired' or 'student' are likely to present distinct flows for certain areas.

Housing Tenure

Links between the housing market/housing tenure and migration events have been observed before. Boyle (1998) notes that whilst those living in council housing are more likely to move than owner occupiers, these moves are likely to be over shorter distances – longer distance moves

being constrained by administrative barriers. Other work has shown that housing availability influences flows of owner-occupant migrants (Cameron *et al.*, 2005; Murphy *et al.*, 2006), with new private housing influencing in-flows (Boyle *et al.*, 1998), and Clark and Huang (2004) linking the distance of migration moves with tenure in the UK. With housing tenure also being a proxy for affluence as well as an indication of the potential ease at which individuals can move, the inclusion of tenure related variables in a Migration Classification can be justified.

Socio-economic status

As outlined by Champion *et al.* (2007), migration, historically, has been a selective process with (in the case of counterurbanisation) predominantly wealthy people moving out from the cities to the suburbs, or from cities to rural areas. Whilst international migration has frequently involved some more disadvantaged individuals, recent major internal migration flows in Britain have tended to involve those slightly older migrants who have been able to afford to move (the counterurbanisers) and those younger skilled migrants who have been attracted to urban agglomerations perhaps by higher education opportunities and who have then remained, or who have been tempted to move between larger urban areas (particularly in London) in search of tertiary sector jobs with higher salaries.

Champion *et al.* (2007) suggest that higher skilled migrants are tending to migrate over much longer distances than their lower skilled counterparts, redressing labour supply and demand imbalances in different locations – findings which echo the early work of Sjaastad (1962). With socio-economic status influencing the direction, volume and distance of migration it is evident that the inclusion of such variables in a Migration Classification will be important.

So it is clear from previous research that a case can be made for the inclusion of at least some variables from all of the main tables. Therefore, selected for initial inclusion in the classification were data from tables 1, 2, 3, 4, 5, 7 and 9. Table 3n was not included as it only applies to Northern Ireland and is thus irrelevant for this classification. Table 6 was not included as the information in this table was also contained in other tables. Table 10 was not included as it does not apply to the whole of Britain. The eight tables selected, therefore, cover as far as possible the dimensions of the whole dataset; an approach advocated in the methodology adopted for other area classifications (Bailey *et al.*, 2000; Vickers *et al.*, 2003). From these selected tables, a suite of variables to be considered for inclusion in the classification was created. The tables contain a total of 599 count variables, however unlike a standard area classification where the variable relates to a static individual residing in that place, for every district in a migration-based classification, each variable needs to be further defined by its movement component.

For example for any one area in a standard classification there could be a count of individuals of perhaps of a particular age and sex. These individuals could be divided by a total population in that area and be represented as a proportion. When we add an interaction component, the area could be either a destination for in-migrants, or an origin for out-migrants, or indeed both for within-area migrants. Straightaway, just by identifying migrants as in, out or within district migrants we have created three times as many variables from our original count. Where these counts can be divided by populations to create in, out and within-area migration rates this number is doubled again.

In addition to standard rates and counts associated with internal migration, there are a number of additional counts and rates that can be attached to most variables for all areas. A selection of these are outlined below in Table 2.2:

Table 2.2 – Migration components of SMS variables

Variable sub-classification	Interaction components	Description
1	internal in-migration count	count of in-migrants to area i
2	internal out-migration count	count of out-migrants from area i
3	within-area migration count	count of migrants moving within area i
4	in migrants from an international origin count	count of in migrants to area i from an international origin
5	in migrants from no usual address count	count of in migrants to area i who had no usual address 1 year ago
6	internal in-migration rate	in-migrants to area i / the population at risk in area i
7	internal out-migration rate	out-migrants from area i / the population at risk in area i
8	within-area migration rate	migrants moving within area i / the population at risk in area i
9	international in-migration rate	international in migrants to area i / the population at risk in area i
10	no usual address in migration rate	migrants to area i of no usual addess 1 year ago / population at risk in area i
11	net migration rate	(in-migration to area i - out-migration from area i) / population at risk in area i
12	population turnover rate	(in-migration to area i + out-migration from area i) / population at risk in area i
13	population churn rate	(in-migration to area i + out-migration from area i + within area i migration) / population at risk in area i
14	migration efficiency rate	(in-migration to area i - out-migration from area i) / (in-migration to area i + out-migration from area i)
15	in-migration/out-migration ratio	in-migration to area i / out-migration from area i
16	average internal in-migration distance	Sum(in-migration count to area j from origin i * distance between population weighted centroids of areas j and i) / in-migration count to destination j
17	average internal out-migration distance	Sum(out-migration count from origin i to destination j * distance between population weighted centroids of areas i and j) / out-migration count from origin i

Where populations at risk are available for a variable, it is possible to calculate the rates shown in the variable sub-classifications 6-13. Where populations at risk are not available, only counts, distance measures, in/out ratios or efficiencies are possible. Interaction components 1, 3, 4 and 5 in the selection have associated PAR data, the rest do not. Therefore taking each variable and sub-classifying it by the categories in the table above, the final count of variables which could be included in the classification was 5,559!

A suite of over five and a half thousand potential variables poses a number of practical problems. It is highly unlikely that all of these variables would be relevant for the classification. Examining the numbers of variables included in district level classifications created by Vickers *et al.* (2003)

and ONS (2004) it becomes apparent that considerably less than five thousand variables are needed to develop a useful area typology. The Vickers *et al.* classification uses 56 variables, whereas the ONS classification uses only 42, therefore it is imperative that a systematic process of reducing these five thousand variables is employed. Indeed, Aldenderfer and Blashfield (1984) warn that there is a temptation to include as many variables as possible in an analysis and hope that cluster analysis techniques will produce meaningful output when doing so could cause problems. They state that *'the importance of using theory to guide the choices of variables should not be underestimated'* (Aldenderfer and Blashfield, 1984, p20) as cluster analysis is beset with unsolved problems and is ultimately still a heuristic technique. Whilst there is debate within the literature about whether in the classification process the use of more variables is better with Harris *et al.* (2005) advocating a general approach of including as many variables as possible, there does appear to be consensus from a number of writers (Everitt *et al.*, 2001; Shepherd, 2006; Vickers, 2006) based on the ideas of Milligan (1996) that variables should only be included if there is a good reason to think they will define the clusters and that *'irrelevant or masking variables should be excluded if possible'* (Everitt *et al.*, 2001, p179).

Candidates for irrelevant or masking variables would be those which are highly correlated. For example, two highly correlated variables could be 'age group 30-44' and 'economically active' as the majority of 30-44 year olds will also be economically active. The inclusion of highly correlated variables serves to effectively positively weight the underlying common factor. Aldenderfer and Blashfield (1984, p21) point out that *'if three highly correlated variables are used, the effect is the same as using only one variable that has a weight three times greater than any other variable'*. Such weighting would skew the results of any analysis, therefore it is desirable, where possible, to only include variables that are *not* highly correlated (Shepherd, 2006; Vickers, 2006).

Whilst it is desirable to drop one variable from a pair of highly correlated variables to reduce the size of the dataset, doing so is not necessarily straightforward, especially when presented with a symmetrical correlation matrix with over five thousand variables on each axis. One approach to data reduction could be 'top-down'. That is, to start with all variables and successively reduce the numbers through a systematic and logical process, perhaps starting where easily identifiable divisions in the data occur. For example, with this approach, a logical first step could be to discard all non-rate data (as differing area size is likely to bias the classification towards larger areas with more flows) and then to discard either the male, female or total individual elements of a variable depending on the correlation coefficient. In many cases, males and females are likely to be highly correlated, so where they are, only the combined male and female counts should be

used. Once the data has been cropped substantially in this way, further data reduction through the elimination of other highly correlated variables can take place.

Although the top-down approach to data reduction is logical, it requires the analysis of very large correlation matrices and can be extremely time consuming. An alternative approach is to start from the 'bottom-up'; starting with no variables, successively adding those viewed as most likely to produce a useful classification before changing and adapting the variables selected as necessary. This approach fits with the recommendations of Aldenderfer and Blashfield (1984) to use theory to guide the choice of variables. It is also a much faster process than the top-down approach, dealing with only a few variables from the beginning. Table 2.3 shows the initial collection of 88 variables chosen for inclusion in the classification:

Table 2.3 – Initial variables chosen for inclusion in the classification

	Variable
1	Internal in-migration rate of persons aged 0 to 15
2	Internal in-migration rate of persons aged 16 to 29
3	Internal in-migration rate of persons aged 30 to 44
4	Internal in-migration rate of persons aged 45 to 59
5	Internal in-migration rate of persons aged over 60
6	Internal out-migration rate of persons aged 0 to 15
7	Internal out-migration rate of persons aged 16 to 29
8	Internal out-migration rate of persons aged 30 to 44
9	Internal out-migration rate of persons aged 45 to 59
10	Internal out-migration rate of persons aged over 60
11	Internal within-area migration rate of persons aged 0 to 15
12	Internal within-area migration rate of persons aged 16 to 29
13	Internal within-area migration rate of persons aged 30 to 44
14	Internal within-area migration rate of persons aged 45 to 59
15	Internal within-area migration rate of persons aged over 60
16	International immigration rate of persons aged 0 to 15
17	International immigration rate of persons aged 16 to 29
18	International immigration rate of persons aged 30 to 44
19	International immigration rate of persons aged 45 to 59
20	International immigration rate of persons aged over 60
21	In-migration rate from no previous address of persons aged 0 to 15
22	In-migration rate from no previous address of persons aged 16 to 29
23	In-migration rate from no previous address of persons aged 30 to 44
24	In-migration rate from no previous address of persons aged 45 to 49
25	In-migration rate from no previous address of persons aged over 60
26	Internal in-migration rate of whites
27	Internal in-migration rate of non-whites
28	Internal out-migration rate of non-whites
29	Internal out-migration rate of whites
30	Internal within-area migration rate of non-whites
31	Internal within-area migration rate of whites
32	International immigration rate of non-whites
33	International immigration rate of whites
34	In-migration rate from no previous address of non-whites
35	In-migration rate from no previous address of whites
36	In-migration rate of economically active individuals
37	In-migration rate of economically inactive individuals
38	Out-migration rate of economically active individuals

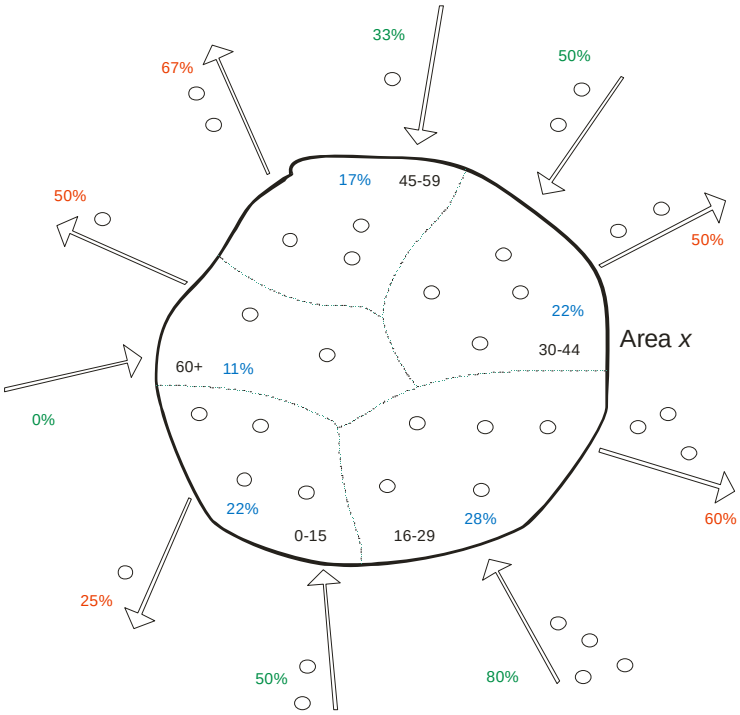
39	Out-migration rate of economically inactive individuals
40	Within-area migration rate of economically active individuals
41	Within-area migration rate of economically inactive individuals
42	International immigration rate of economically active individuals
43	International immigration rate of economically inactive individuals
44	In-migration rate from no previous address of economically active individuals
45	In-migration rate from no previous address of economically inactive individuals
46	In-migration rate of individuals with a limiting long term illness
47	In-migration rate of individuals with no limiting long term illness
48	Out-migration rate of individuals with a limiting long term illness
49	Out-migration rate of individuals with no limiting long term illness
50	Within-area migration rate of individuals with a limiting long term illness
51	Within-area migration rate of individuals with no limiting long term illness
52	International immigration rate of individuals with a limiting long term illness
53	International immigration rate of individuals with no limiting long term illness
54	In-migration rate from no previous address of individuals with a limiting long term illness
55	In-migration rate from no previous address of individuals with no limiting long term illness
56	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.1
57	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.1
58	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.2
59	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.2
60	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 2
61	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 2
62	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 3
63	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 3
64	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 4
65	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 4
66	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 5
67	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 5
68	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 6
69	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 6
70	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 7
71	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 7
72	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 8
73	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 8
74	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Full Time Student
75	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Full Time Student
76	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Not Classified
77	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Not Classified
78	Migration efficiency of wholly moving households moving into owner occupied accommodation
79	Migration efficiency of other moving groups moving into owner occupied accommodation
80	Migration efficiency of wholly moving households moving into socially rented accommodation
81	Migration efficiency of other moving groups moving into socially rented accommodation
82	Migration efficiency of wholly moving households moving into privately rented accommodation
83	Migration efficiency of other moving groups moving into privately rented accommodation
84	Migration efficiency of individuals living alone
85	Migration efficiency of individuals not living in a family but with others in a household
86	Migration efficiency of individuals who are part of a couple family
87	Migration efficiency of individuals who are part of a lone parent family
88	Migration efficiency of individuals living in a communal establishment

Vickers *et al.* (2003) recommend that for an area classification to be comprehensive, all domains within the dataset need to be included. Here the initial set of 88 variables cover all domains available in the SMS, with age (1-25), ethnicity (26-35), economic activity (36-45), health (46-55), socio-economic status (56-77), housing tenure (78-83) and family status (84-88) all being

accounted for. For each variable, net rates rather than absolute numbers have been chosen to avoid area size creating bias. Where it has been impossible to calculate rates using related PAR data (variables 56-88), rates of migration efficiency have been used. Distance variables were not included at this stage as preliminary experiments with clustering variables including average distance moved, tended to create concentric rings around London – an undesirable result which may affect the cluster solutions when the first clustering trials proper are conducted.

Before a first cluster run can be carried out on the data, however, the list of variables needs to be reduced further, specifically to reduce instances of correlation. One solution, which has been used to reduce the number of cross-correlated variables in classifications based upon standard area counts, is to remove one variable from a related family of variables. Vickers (2006) suggests that where there are *n* groups within a variable, the optimum number of groups from that variable to include in a classification is *n*-1. Where a classification is being constructed from count data this makes sense. Figure 2.1 represents an area *x* with a count of 18 individuals residing within. These 18 individuals can be grouped according to their age. There are five different age groups from 0-15 to 60+. If there is information about the proportion of total individuals that each age group contains, then in order to obtain information about the number of individuals in all age groups *n*, only information about *n*-1 is required. The sum of the proportions in the youngest four age groups means that the proportion in the eldest (60+) age group has to be 11%. By only including *n*-1 variables all the information is still included.

Figure 2.1 – Why *n*-1 groups within a variable is not optimal for flow related data



However, where flow data for an area are being used rather than count data, knowing *n*-1 does not mean it is possible to deduce *n*. As is shown in the example in Figure 2.1, knowing the inflow or outflow rates for the four youngest age groups reveals nothing about the inflow or outflow rates for the oldest age group. Therefore, in this instance, *n*-1 is not automatically the optimum number of groups to use from a parent variable. It may well be that if correlations with other variables are low, then the inclusion of *n* variables in a family of variables is permissible.

A matrix of Pearson's correlation coefficients has been calculated for every variable with every other variable in the list of 88. From this matrix, pairs of highly correlated variables can be identified in order that one from the pair might be dropped. The question here, however, is what constitutes a 'high' correlation coefficient? A coefficient of +1 or -1 signifies a perfect correlation, whereas 0 signifies a complete lack of correlation. What though is a suitable cut-off? Is anything over 0.5 or under -0.5 a high correlation, or should this figure be higher? The decision that is made will obviously affect subsequent analysis, but is also highly subjective. As a guide, a correlation coefficient of 0.7071 is equivalent to around 50% of one variable being associated with the other (Vickers *et al.*, 2003). A higher coefficient means that even more of a variable's information can be gained from looking at the other variable in the pair. It seems appropriate that where more than 50% of a variable's information can be gained from elsewhere, then this variable would be a candidate for omission from the classification. For each variable in the initial list of 88, a count of the correlation coefficients over 0.7071 was created to flag those variables that it might be useful to omit from the classification. Particularly numerous instances of high correlation were found with variables relating to White ethnicity and an economically active status. This is unsurprising as the majority of individuals in Britain are both White and economically active. As a result, these variables were dropped from the list. Similarly, some variables related to no limiting long-term illness showed higher instances of high correlation, so were also dropped from the list, leaving 73 remaining variables.

Examining correlation, however, should not be the only technique used to choose variables for inclusion in a classification. Shepherd (2006, p112) notes that in some instances a high correlation 'may not be a good justification for removal'. For example, the age variables 0-15 and 30-44 (for all interaction types) have a consistently relatively high correlation with each other, as well as with other variables, although the age 30-44 variable shows a slightly higher correlation with other variables. The inclination, based purely on correlations, would be to drop the latter variable; however, with the majority of migrations of young people happening only as a result of parental migration, it is likely that age 30-44 is empirically a more important variable to keep. Another technique, therefore, is required to help make a more effective decision on the

inclusion/exclusion of some variables. Principal Components Analysis (PCA) is a technique advocated by a number of authors (Everitt *et al.*, 2001; Harris *et al.*, 2005; Shepherd, 2006; Vickers, 2006; Vickers *et al.*, 2003) in the variable selection stage of classification creation, and can be used in conjunction with correlation analysis to choose variables where correlation does not help, as in this example.

PCA is a method which takes a matrix of correlated variables and reduces it down to a new set of un-correlated components derived from the original variables. Whilst this new set of components can be used as surrogates in the analysis, where the interpretation of the clusters created is desirable this is not necessarily a good option. Harris *et al.* (2005) warn precisely of this problem. They also note that commercial geodemographic companies such as Experian have avoided PCA claiming that the distinctions between cluster types become blurred when these surrogate variables are used.

Where PCA can be used in the variable selection process, however, is as an exploratory analysis tool whereby variables which have higher amounts of their variance accounted for within a particular component can been seen as more important within the dataset. Everitt *et al.* (2001 p26) refer to this as a measure of *'interestingness'* – more interesting variables are more desirable to include in a cluster analysis. A number of outputs from a PCA can be used to assess the 'interestingness' of the component variables. The amount of variance explained by each factor is explained by its eigenvalue. The larger the eigenvalue the more variance is explained (Kline, 1994), with initial components having larger eigenvalues than latter components. As well as the list of eigenvalues and the related variance explained by each component, PCA also produces a component loadings matrix whereby the proportion of a variable associated with a component is displayed. Variables with large proportions in the early components are important in the context of the entire dataset. Care should be taken, however, to 'rotate' a component matrix before it is interpreted (Kline, 1994). The purpose of rotation is to pick the most simple principal component solution. Whilst there are a number of ways to rotate a component matrix, the Varimax solution produces for each component variable loadings which are either high or near zero – a feature of a simple solution (Kline, 1994).

PCA was run on the 73 remaining variables, producing 12 components with eigenvalues greater than 1, accounting for around 78% of the total variance in the dataset. From this, a list of variables with low component loadings for the first 6 rotated components (accounting for around 70% of the data) was created (Table 3.6). Featuring consistently low component loadings, these variables could now be considered for omission from the group used in the initial classification.

Table 2.4 Variables exhibiting low component loadings in the first 6 rotated components produced by PCA.

Variable
Migration efficiency of other moving groups whose household reference person is in NS-SEC category 4
Migration efficiency of other moving groups whose household reference person is in NS-SEC category 5
Migration efficiency of other moving groups whose household reference person is in NS-SEC category 7
Migration efficiency of other moving groups whose household reference person is in NS-SEC category 8
Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 8
Migration efficiency rate of other moving groups moving into socially rented accommodation
Migration efficiency rate of wholly moving households moving into socially rented accommodation
In-migration rate of individuals with a limiting long-term illness
Within-area migration rate of individuals with a limiting long-term illness
In-migration rate from no previous address of individuals with a limiting long-term illness
Internal out-migration rate of non-whites
International immigration rate of non-whites

Returning to the initial problem relating to the choice of age groups, PCA reveals that in the first component, the component loading scores for age group 30-44 are generally higher than they are for age group 0-15, suggesting that it may be more useful to include 30-44 age group variables in the classification rather than 0-15 age group variables.

Before a final decision is made, however, consideration should also be given to the variation of the variables across the areas comprising the spatial system for the classification. Vickers *et al.* (2003) suggest that by examining the standard deviation of each variable, an appreciation of the extent to which they vary across space can be gained. Shepherd (2006) warns that variables with particularly low standard deviations will probably add little to cluster definitions, whereas those with high standard deviations may feature undesirable outliers. Examining the standard deviation statistics for age groups 0-15 and 30-44 (Table 2.5), it is evident that whilst all standard deviations are low, age group 30-44, with a higher average standard deviation, is likely to prove a more discriminatory variable across the spatial system than age group 0-15.

Table 2.5 Standard deviation of problematic age variables

Variable	Standard deviation
out_mig_rate_Age_30_44	0.0185
in_mig_rate_Age_30_44	0.0169
within_mig_rate_Age_0_15	0.0161
within_mig_rate_Age_30_44	0.0132
in_mig_rate_Age_0_15	0.0126
out_mig_rate_Age_0_15	0.0124
international_in_mig_rate_Age_30_44	0.0081
international_in_mig_rate_Age_0_15	0.0060
no_addr_in_mig_rate_Age_30_44	0.0029
no_addr_in_mig_rate_Age_0_15	0.0023
All interaction categories average 30-44 age group	0.0119
All interaction categories average 0-15 age group	0.0099

So, with the additional information gained from PCA and the examination of standard deviation statistics, it was decided that, for this initial trial classification, the collection of variables that would be used would not include age group 0-15 and would not include those variables with consistently low component loadings from the PCA. These exclusions are in addition to the variables already excluded for having high correlations with greater numbers of other variables. Consequently, the list of variables to be included in the initial classification was reduced to 56. These are shown in Table 2.6:

Table 2.6 – Variables used in the initial migration classification.

	Variable
1	Internal in-migration rate of persons aged 16 to 29
2	Internal in-migration rate of persons aged 30 to 44
3	Internal in-migration rate of persons aged 45 to 59
4	Internal in-migration rate of persons aged over 60
5	Internal out-migration rate of persons aged 16 to 29
6	Internal out-migration rate of persons aged 30 to 44
7	Internal out-migration rate of persons aged 45 to 59
8	Internal out-migration rate of persons aged over 60
9	Internal within-area migration rate of persons aged 16 to 29
10	Internal within-area migration rate of persons aged 30 to 44
11	Internal within-area migration rate of persons aged 45 to 59
12	Internal within-area migration rate of persons aged over 60
13	International immigration rate of persons aged 16 to 29
14	International immigration rate of persons aged 30 to 44
15	International immigration rate of persons aged 45 to 59
16	International immigration rate of persons aged over 60
17	In-migration rate from no previous address of persons aged 16 to 29
18	In-migration rate from no previous address of persons aged 30 to 44
19	In-migration rate from no previous address of persons aged 45 to 49
20	In-migration rate from no previous address of persons aged over 60
21	Internal in-migration rate of non-whites
22	Internal within-area migration rate of non-whites
23	In-migration rate from no previous address of non-whites
24	In-migration rate of economically inactive individuals
25	Out-migration rate of economically inactive individuals
26	Within-area migration rate of economically inactive individuals
27	International immigration rate of economically inactive individuals
28	In-migration rate from no previous address of economically inactive individuals
29	Out-migration rate of individuals with a limiting long term illness
30	International immigration rate of individuals with a limiting long term illness
31	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.1
32	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.1
33	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.2
34	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.2
35	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 2
36	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 2
37	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 3
38	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 3
39	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 4
40	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 5
41	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 6
42	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 6

43	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 7
44	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Full Time Student
45	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Full Time Student
46	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Not Classified
47	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Not Classified
48	Migration efficiency of wholly moving households moving into or from owner occupied accommodation
49	Migration efficiency of other moving groups moving into or from owner occupied accommodation
50	Migration efficiency of wholly moving households moving into or from privately rented accommodation
51	Migration efficiency of other moving groups moving into or from privately rented accommodation
52	Migration efficiency of individuals living alone
53	Migration efficiency of individuals not living in a family but with others in a household
54	Migration efficiency of individuals who are part of a couple family
55	Migration efficiency of individuals who are part of a lone parent family
56	Migration efficiency of individuals living in a communal establishment

Finally, in some instances it may be desirable to weight particular variables depending on their perceived importance. To a certain extent, a weighting exercise has already been undertaken through choosing the variables. All excluded variables have effectively been given a weight of 0, included variables 1. Shepherd (2006) outlines a range of mathematical techniques for weighting variables including the HINoV method; other approaches such as those used by Experian in their Mosaic classification, described by Harris *et al.* (2005), are more down to the judgement of the researcher. A common approach described by Everitt *et al.* (2001) is to weight variables according to their variability; a technique more often referred to as 'standardisation'. No weighting has been applied to the 56 variables in this initial classification.

2.1.3 Variable Standardisation

After the variables to be used have been selected and left un-weighted, it is necessary to standardise them over the same range. This is particularly important when the units used to measure the variables differ. For example, in this initial classification most variables are gross rates measuring the overall magnitude of flow. Efficiencies, however, measure the magnitude of flow in a particular direction rather than the volume of measured by the other rates and as a result will feature some negative flows where out-migration is greater than in-migration. Furthermore, whilst in this initial classification only gross rates are used, it is entirely feasible that, if appropriate, alternative variables measured over different scales could be included in the future. Whenever data measured across different ranges are used, it is appropriate to standardise across the range so that any individual variable will not bias the classification. Whilst Aldenderfer and Blashfield (1984, p21) debate the necessity of variable standardisation in all situations, they concede that, where units of measurement vary between variables, researchers classifying these data will *"undoubtedly want to standardise them."*

2.1 An initial district level area classification based upon migration variables

Once the decision has been made to standardise the data, the method of standardisation needs to be chosen. As with all other elements of the clustering process, the literature does not provide consensus on the most appropriate methodology to use. A number of researchers (Everitt *et al.*, 2001; Milligan and Cooper, 1987; Shepherd, 2006) cite work by Milligan and Cooper (1988) which suggests that the most effective way of standardising data is to standardise over the range of data for that variable. That is:

(1)

$$Z_i = \frac{[X_i - Min(X_i)]}{[Max(X_i) - Min(X_i)]}$$

where:

Z*i* = the standardised variable value for area *i*,

Xi = the value of variable *X* for area *i*,

Min(*Xi*) = the minimum value of variable *X* for all areas *i*, and

Max(*Xi*) = the maximum value of variable *X* for all areas *i*

However, work by Schaffer and Green (1996) contradicts the findings of Milligan and Cooper (1988). As the result of research carried out on empirical datasets (Milligan and Cooper did not use real data), they find that when six different types of standardisation are compared, all but one perform well, leading them to conclude that *"column variable standardization does not seem to affect clustering results nearly as much as other aspects [such as the] choice of clustering algorithm and the presence of noise variables"* (Schaffer and Green, 1996, p162).

One of the more common ways of standardising data is through the calculation of *z*-scores. *Z*-scores standardise variable data for each unit (in this case the local authority district) by its standard deviation from the mean for the entire variable. That is:

(2)

$$Z_i = \frac{X_i - \bar{X}}{\sigma_X}$$

where:

$\bar{X}$ = national mean for variable *X*

σ_X = Standard deviation for variable *X*

with:

(3)

$$\sigma_X = \frac{\sqrt{(X_i - \bar{X})^2}}{N}$$

Whilst there are also other methods that can be used to standardise data, in the light of the research by Schaffer and Green (1996) showing little difference in the clustering outcomes when different methods of standardisation were used, *Z*-scores were chosen as the method of standardisation for this initial classification. This is in line with the method chosen to create the Vickers *et al.* (2003) district classification (although not Vickers' OA classification).

2.1.4 Proximity Measure

The decision over which proximity measure to choose to judge the distance between the cluster centroids is another important decision which will affect the outcome of any clustering process. Generally different measures of proximity are suited to different types of data (nominal/binary, categorical or continuous). Where data are continuous (as they are here) Everitt *et al.* (2001) list six commonly used measures of proximity, the most common of all being Euclidean distance. Whilst euclidean distance may or may not be the most suitable measure to use, since the objective of this current exercise is to create a trial classification which will be refined at a later stage, for n, P. (2001). "The diversity of diversity: a critique of geodemographic classification." Area 33(1): 63-76.

Ward, J. (1963). "Hierarchical g.1.5 Clustering method

2.1.5 Clustering method

Any researcher browsing through the literature on clustering will be presented with a plethora of different clustering techniques which can be applied to find groups within data. Choosing an appropriate clustering technique, therefore, can be a challenge, especially when any one particular clustering algorithm will almost certainly produce a different output from another. Aldenderfer and Blashfield (1984) note this problem and suggest that, in such a situation, the wise solution would be to run more than one clustering algorithm and compare the different results from each.

Even if a researcher chooses to use more than one clustering method to analyse a data set it still may be the case that one method may be more logical than another to start with. A review of the

literature reveals that there are two main families of clustering method: hierarchical and partitioning. Summarised by Aldenderfer and Blashfield (1984), partitioning methods take *n* observations and classify them into *k* clusters or groups which satisfy the requirements of a partition (i.e. that each group must contain at least one data object and each object must belong to one group). When a partitioning method is used, the researcher must decide the value of *k* before the process begins. This, in itself, can be problematic when the optimum number of groups is unknown, therefore one approach to tackle this could be to go through the process several times with different values of *k* where the value of *k* can theoretically be as large as the number of observations. However, running the clustering algorithm for a number of values *k* could be a lengthy process. As an alternative to partitioning methods, hierarchical methods deal with all values of *k* at the same time and produce output from $k = 1$ to $k = n$. From this output with all possible solutions for *k*, the researcher is then able to select the solution with the most appropriate number of clusters *k*.

The logical way forward for an initial clustering run, therefore, would be to use a hierarchical clustering algorithm in order to ascertain what might be the most appropriate number of clusters, before optimising the solution at a later stage through the use of another clustering algorithm. Indeed, Everitt *et al.* (2001) suggest an initial partition may be created through a hierarchical technique before an optimisation algorithm such as *k*-means is used to re-arrange the original solution of *k* groups into a new solution of *k* groups – keeping the new partition only if an improvement is made; a methodology adopted in the past by the ONS in their local authority classifications (Bailey *et al.*, 2000).

Within hierarchical methods of classification there are a number of different algorithms to choose from, some of which are agglomerative (i.e. which start with *n* single member clusters before amalgamating these clusters into successively larger and fewer clusters until a final solution is reached with one single cluster featuring all the data), and some which are divisive (starting with one large cluster before successively splitting the cluster until there are *k* clusters each containing a single data item). One of the more common hierarchical methods employed in the creation of area classifications – used by both ONS (Bailey *et al.*, 2000) and Vickers *et al.* (2003) – is Ward's method (Ward, 1963). Simple methods of agglomerative clustering join cases to clusters if the case is similar to at least one case already in the cluster (single linkage/nearest neighbour); to all members of the cluster (complete linkage/furthest neighbour); or to the average for all members of the cluster (average linkage). Ward's method, on the other hand, optimises the minimum variance between cases within clusters. Cases are joined to clusters where the addition results in the minimum increase in the error sum of squares (Aldenderfer and Blashfield, 1984).

Whilst all methods have their benefits and limitations and a number of studies have found conflicting performance (Everitt and Dunn, 2001), in this initial clustering run Ward's method will be used as it minimises the loss of information associated with each cluster as it is created (Vickers *et al.*, 2003).

2.1.6 Number of clusters

Deciding upon the number of clusters is also a difficult challenge, although as previously pointed out, through using a hierarchical clustering method a solution is produced with all clusters enabling the choice of the appropriate number of clusters to be made subsequent to the clustering process. Dendrogram output (Figure 2.2) can give clues as to the best clustering solutions. As Everitt (2001) describes, the best solutions are likely to be where clusters below a selected distance (from the cluster centre) on the dendrogram are distant from each other by the least amount. In other words, large changes in the distances indicate solutions where the optimum number of clusters are presented. The example in Figure 2.2, suggests that, in this case, it might be sensible to include all cases in one group as the largest distance between clusters includes all cases. When dendrograms are produced for a large number of cases, however, interrogation of the tree to find the optimum number of clusters becomes more difficult. An alternative method therefore is to examine the numerical distance coefficients. Where large jumps occur in the coefficients between clusters, the points where the jumps occur signify the optimum cluster solutions. Figure 2.3 is a graphical representation of the gaps in the coefficients. In this case, the number of clusters just after a steep decline in the graph represent the optimum cluster solutions.

Figure 2.2 – Sample dendrogram output.

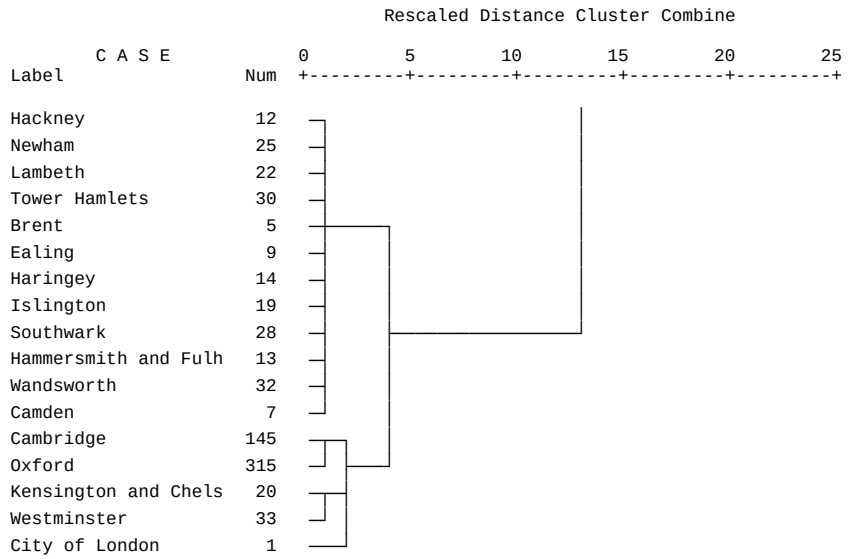
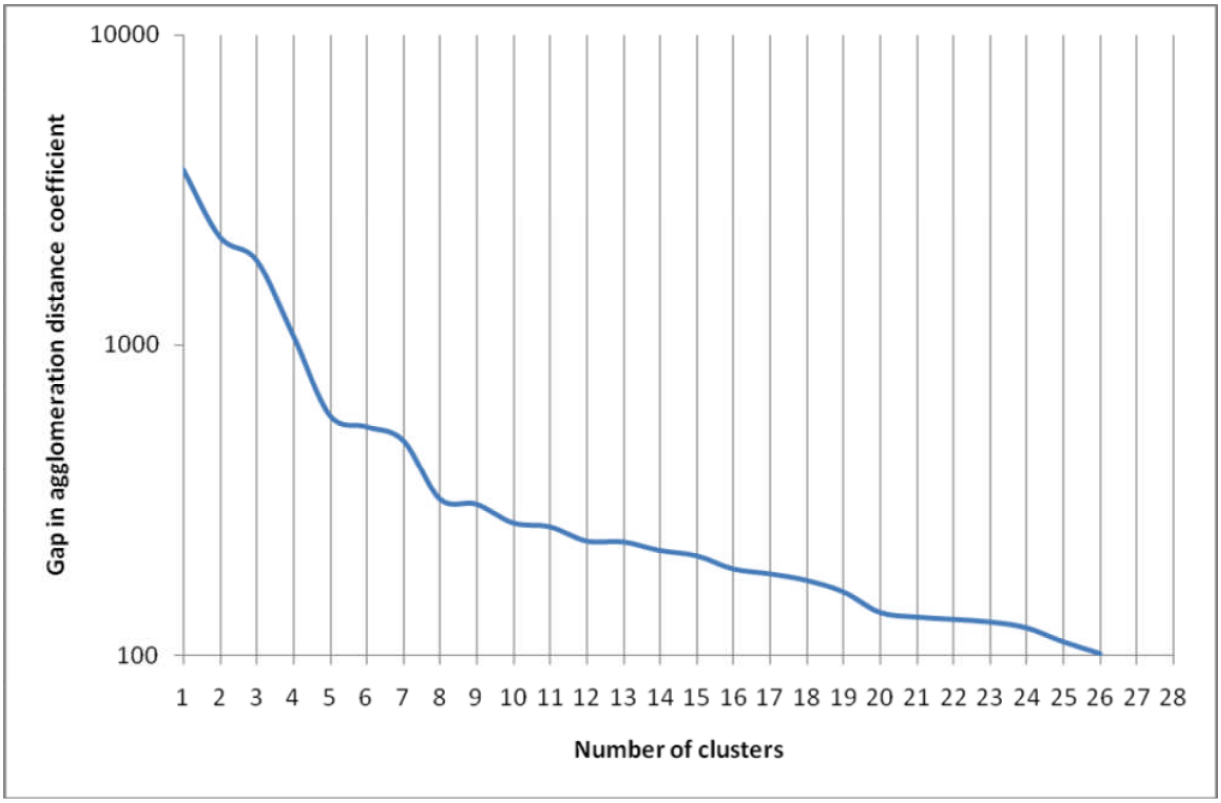


Figure 2.3 – Agglomeration schedule representing the distance between the most dissimilar areas within cluster groups



It is clear to see in Figure 2.3 that the steepest declines in the graph occur just before 5 and 8 clusters, signifying that solutions containing either 5 or 8 clusters would be the best for this set of data. Reading the graph from right to left, what this shows is that if there is a steep rise as the number of clusters declines – for example between 8 and 7 clusters – the data within the 7 clusters is more dissimilar than the data within the 8 clusters. Since the objective of the clustering exercise is to find clusters with similar characteristics, it is sensible to select 8 rather than 7 clusters.

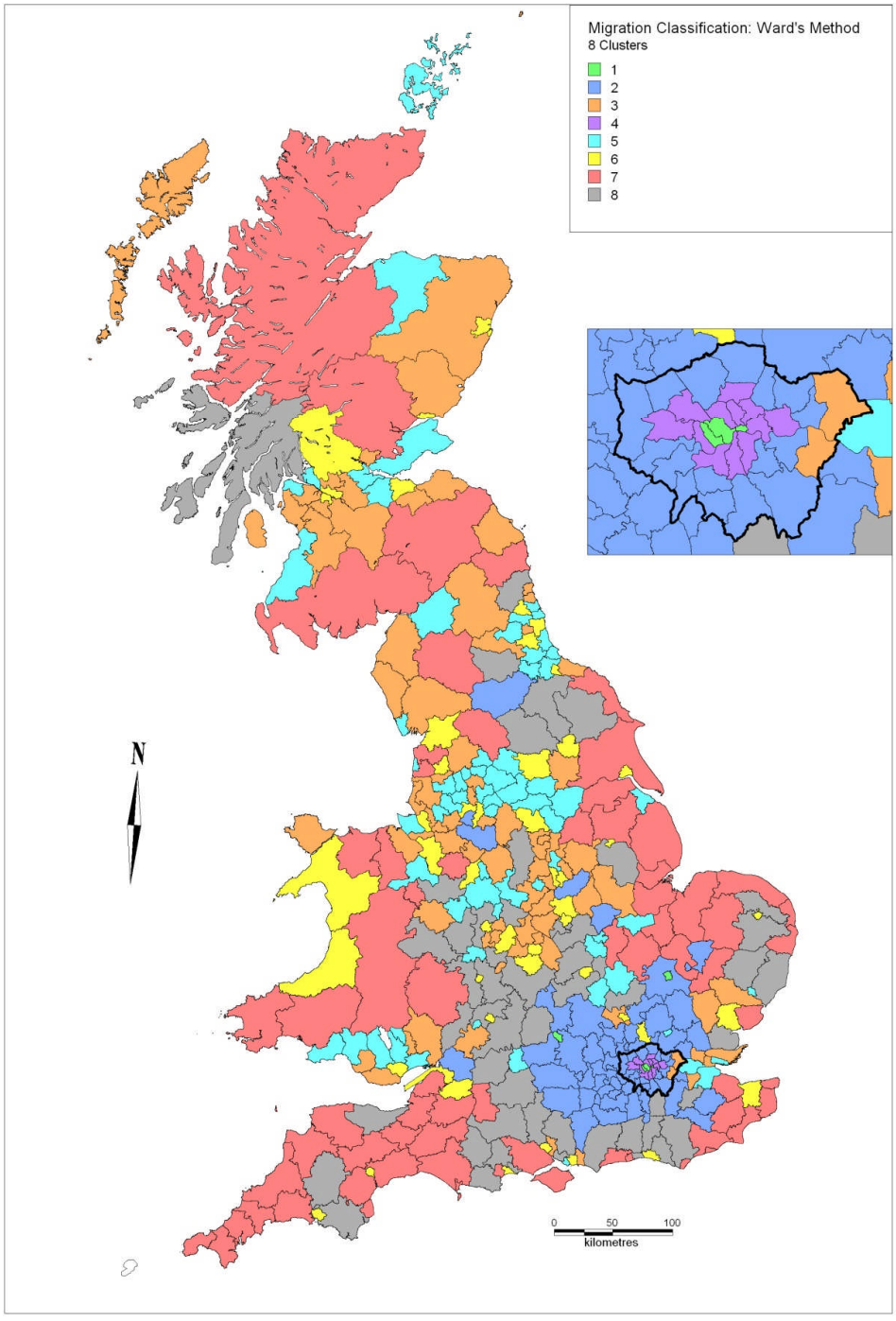
2.1.7 Replication testing and interpretation

The final stage in Milligan and Cooper's framework for carrying out a cluster analysis involves the cross-validation and test of any cluster output that is produced. Validation is a lengthy process and necessarily happens after an initial partition has been made. Consequently discussion related to this stage of the classification process will be carried out later.

2.2 Initial classification results

Following the information given in the agglomeration schedule in Section 2.1.6 an initial, draft classification partition was created for 8 clusters. The spatial distribution of these clusters is shown in Figure 2.4.

Figure 2.4 – Spatial distribution of 8 migration data clusters created using Ward's method



Even before detailed analysis of the variable composition of each cluster, a few remarks can be made in relation to the spatial patterning of the cluster groups. Cluster 1 contains the fewest districts – three of which are located in the most central London boroughs; City of London, Westminster, and Kensington and Chelsea; the other two being Oxford and Cambridge. Cluster 2 forms a Greater London hinterland encompassing almost all boroughs bordering the London region as well as a swathe of districts in the Home Counties and a few sprinkled beyond. Cluster 3 is more spatially diverse but perhaps most concentrated in the Midlands, Northern England, South Wales and Scotland. Cluster 4 is a selection of districts found solely in inner London buffering cluster 1 from cluster 2. Cluster 5 is most concentrated around the Northern ex-industrial areas, South Wales and the North East. Cluster 6 is spatially diverse, but features districts mainly associated with thriving cities characterised in many cases by the presence of higher education institutions. Cluster 7 features districts mainly in the rural periphery of Britain, including the South West, Eastern England, central Wales and remoter parts of Scotland. The final cluster, 8, can be found principally in areas to the south, west and north-west of the main body of cluster 2. Other areas are found in the North, south-west and east. Two districts were omitted from the final classification by the software used to run Ward's algorithm; these were Merthyr Tydfill and the Isles of Scilly.

Examining the spatial patterning of clusters reveals some coherent patterns, which whilst interesting, tell us little without knowledge of the variables which drive each cluster group. Indeed, summarising these clusters through labels and 'pen portraits' is a key part of the classification process (Harris *et al.*, 2005; Shepherd, 2006; Vickers *et al.*, 2003). This next section will therefore summarise the clusters in terms of their key characteristics.

2.2.1 Initial draft classification portraits

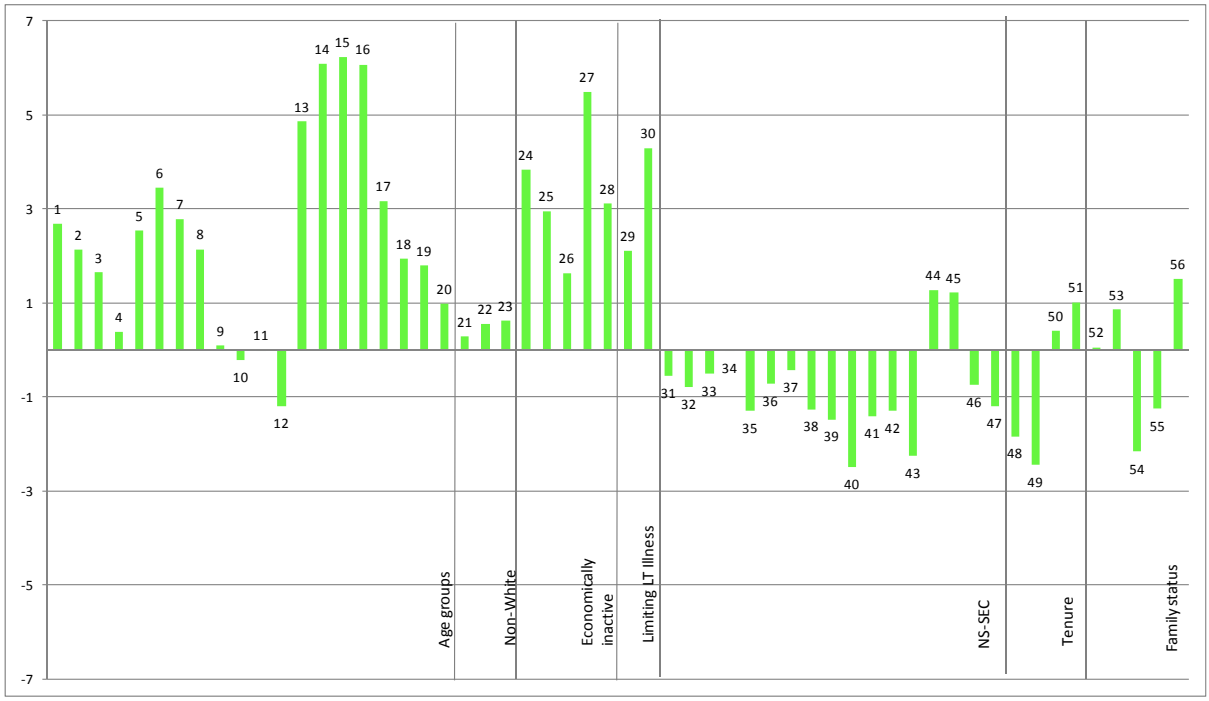
By taking the average *z*-score value for each variable across the districts comprising each cluster it is possible to ascertain which variables are more and less important within the cluster. High positive *z*-scores indicate more importance, high negative *z*-scores indicate less importance for variables 1-30, with 0 being the average *z*-score across all districts. Interpretation of the graphs differs slightly with variables 31-56. Here, because the variables are net efficiency rates (which have positive and negative values), the positive and negative *z*-score values are maintained. In this latter half of the graph, low importance is indicated by being close to 0. Of note also is the dominance of the London based clusters, with *z*-scores across a much larger range in clusters 1 and 4 than in any other cluster.

2.2.1.1 Cluster 1

Cluster 1 contains 5 districts:

City of London
Kensington and Chelsea
Westminster
Cambridge
Oxford

Figure 2.5 – *Z*-scores defining cluster 1



Numbers refer to variables in Table 2.6

This cluster is defined principally by international immigration. 13-16, 27 and 30 all relate to these migrants. There is also an above average turnover of the population in younger age groups (1, 2, 3, 5, 6, 7, 17) Other important variables are 24, 25, 26 and 28 which relate to the in, out and within area migration of economically inactive individuals. Districts in this cluster also feature slightly above average inflows of students (44 and 45) and individuals moving into communal establishments and privately rented accommodation (50, 51 and 56). Non-white migrants are also slightly above average (21,22 and 23).

2.2.1.2 Cluster 2

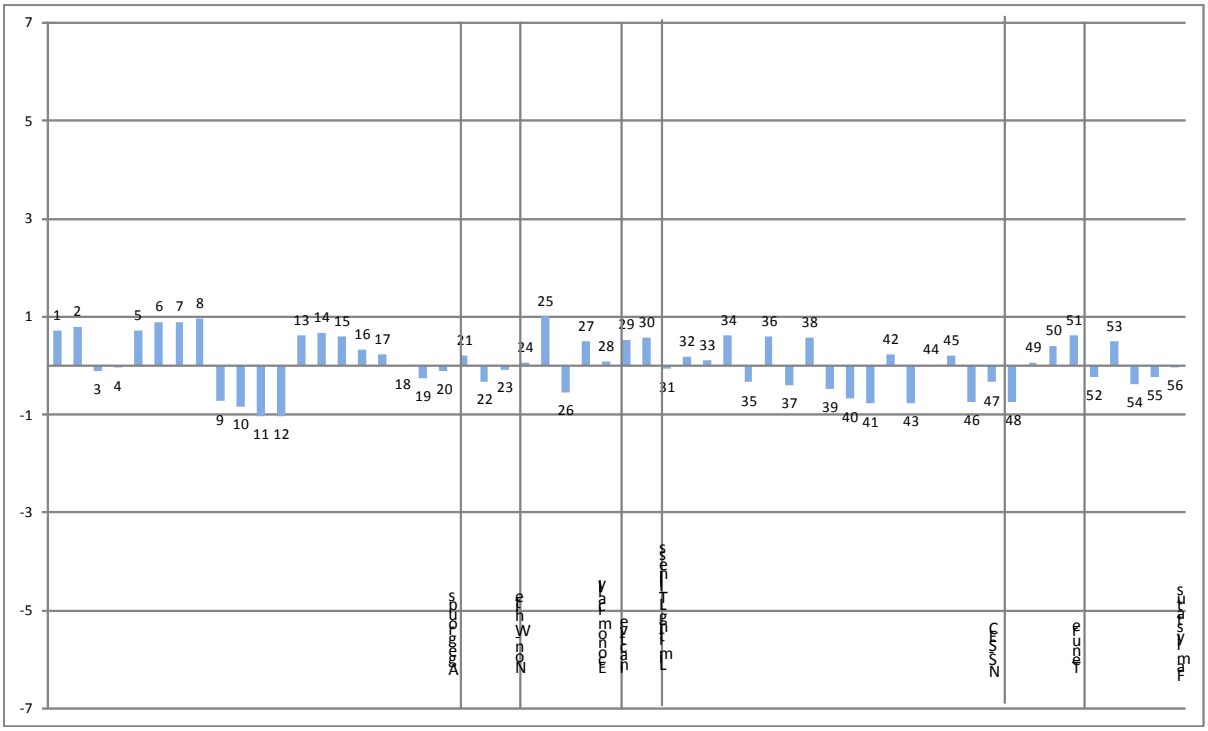
Cluster 2 contains 69 districts:

Barking and Dagenham
Barnet
Bromley
Croydon
Enfield
Greenwich
Harrow
Hillingdon
Hounslow
Kingston upon Thames
Lewisham
Merton
Redbridge
Richmond upon Thames
Sutton
Waltham Forest
Trafford
Rutland
South Gloucestershire
Bracknell Forest
West Berkshire
Reading
Slough

2.2 Initial classification results

Windsor and Maidenhead	Uttlesford	Three Rivers	Elmbridge
Wokingham	Basingstoke and Deane	Watford	Epsom and Ewell
Aylesbury Vale	East Hampshire	Dartford	Guildford
Chiltern	Hart	Maidstone	Mole Valley
South Bucks	Rushmoor	Sevenoaks	Reigate and Banstead
Wycombe	Broxbourne	Tunbridge Wells	Runnymede
South Cambridgeshire	Dacorum	Richmondshire	Spelthorne
Macclesfield	East Hertfordshire	Rushcliffe	Surrey Heath
Brentwood	Hertsmere	Cherwell	Waverley
Chelmsford	North Hertfordshire	South Oxfordshire	Woking
Epping Forest	St. Albans	Vale of White Horse	Crawley
	Stevenage	Forest Heath	

Figure 2.6 – *Z*-scores defining cluster 2



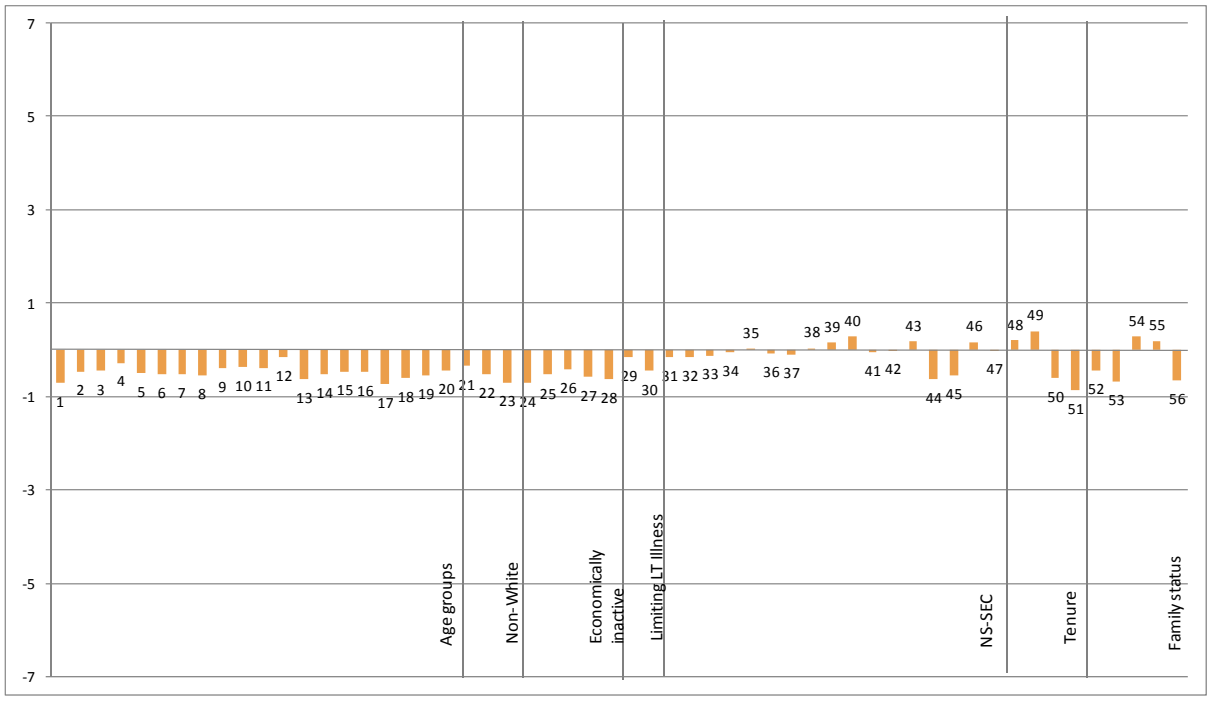
This cluster is defined principally by the in-migration and out-migration of young migrants (1, 2, 5 and 6) signifying higher levels of population turnover for this group. This higher turning over of the younger population is joined by the out-migration of older and economically inactive migrants (7, 8 and 25). International immigration of younger individuals (13, 14, 15) is also of heightened importance, as too are movements of those of higher socio-economic status who are not moving in household groups (34, 36, 38) and those individuals moving into privately rented accommodation (51). Out-migration and immigration of individuals with a limiting long term illness is of an increased importance as well (29 and 30). Within district moves and moves of those in lower socio-economic categories are of much less importance in this cluster (9-12 and 39, 40, 41 and 43).

2.2.1.3 Cluster 3

Cluster 3 contains 81 districts:

Bexley	Ellesmere Port and Nes	Gravesham	Tamworth
Havering	Vale Royal	Ribble Valley	Babergh
Bury	Allerdale	Rossendale	North Warwickshire
Stockport	Copeland	South Ribble	Nuneaton and Bedworth
Knowsley	South Lakeland	West Lancashire	Rugby
Sefton	Amber Valley	Hinckley and Bosworth	Redditch
Dudley	Bolsover	Melton	Aberdeenshire
Sandwell	Chesterfield	North West Leicestershire	Angus
Solihull	Erewash	South Kesteven	Clackmannanshire
Walsall	High Peak	Berwick-upon-Tweed	East Ayrshire
Redcar and Cleveland	North East Derbyshire	Blyth Valley	East Dunbartonshire
Halton	South Derbyshire	Tynedale	East Lothian
Warrington	Chester-le-Street	Wansbeck	East Renfrewshire
Isle of Anglesey	Wear Valley	Selby	Midlothian
Flintshire	Basildon	Gedling	North Ayrshire
Bridgend	Braintree	Mansfield	North Lanarkshire
The Vale of Glamorgan	Castle Point	Newark and Sherwood	Renfrewshire
Torfaen	Rochford	Shrewsbury and Atcham	Shetland Islands
Monmouthshire	Stroud	Staffordshire Moorland	South Lanarkshire
South Bedfordshire	Eastleigh		Eilean Siar
Congleton	Havant		

Figure 2.7 – *Z*-scores defining cluster 3



As is shown by Figure 2.7, cluster 3 can be defined more by the variables that are below average than the variables that are above. Starting with those that are above, migrants moving into or from owner occupied accommodation are of above average importance in this group (48 and 49), as too are migrants in the middle to lower socio-economic groups (38, 39, 40 and 43). Migrants who are part of couple families and single parent families are also of increased importance (54 and

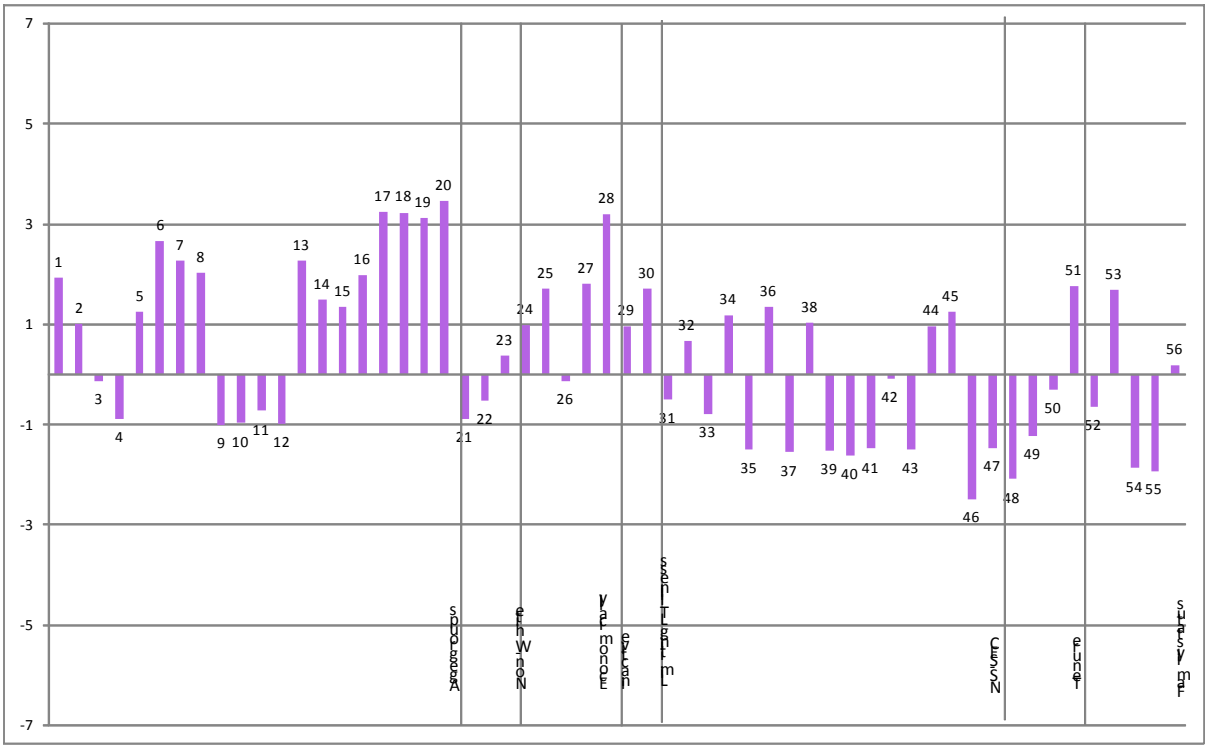
55). Of below average importance in this group are almost every other variable signifying generally low levels of in and out migration of all ages, low levels of international migration, student migration and fewer migrants in higher socio-economic groups.

2.2.1.4 Cluster 4

Cluster 4 contains 12 districts:

Brent	Hammersmith and Fulham	Newham
Camden	Haringey	Southwark
Ealing	Islington	Tower Hamlets
Hackney	Lambeth	Wandsworth

Figure 2.8 – *Z*-scores defining cluster 4



Variables of the most significance in this cluster relate to the in-migration of individuals from no previous address (17-20 and 28). Also of importance is the immigration of foreign migrants (13-16) and the general out-migration of individuals aged over 30 (6-8). Migration of the economically inactive also helps define this cluster (24 ,25, 27, 28) as too does the in-migration of those with a limiting long-term illness (29, 30). Interestingly, migrants in higher socio-economic groups are prevalent (32, 34, 36 and 38), however these moves tend to be by individuals not moving as households, but as other moving groups. Student moves are prevalent (44 and 45) as too are moves into or from privately rented accommodation (51) and into non-family households (53). Of less importance in this cluster are moves of individuals in the lower

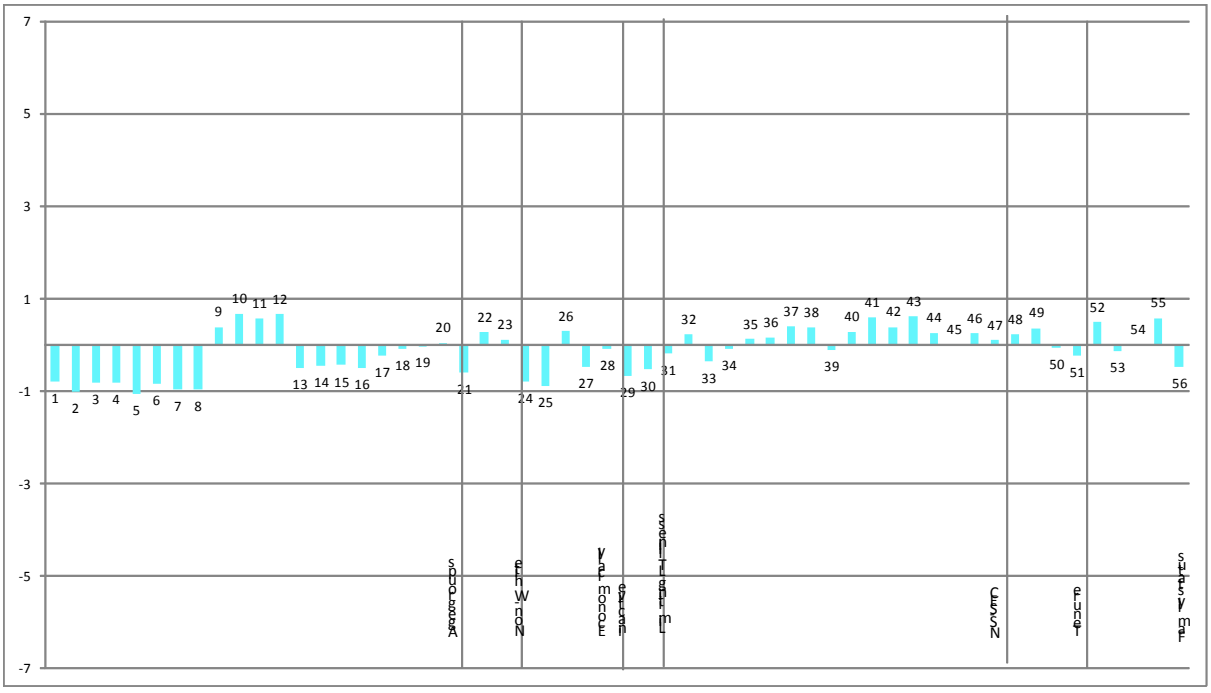
socio-economic categories (39-43) and moves into owner occupied or privately rented accommodation. Within-district moves are also of less relevance in this cluster (9-12).

2.2.1.5 Cluster 5

Cluster 5 contains 68 districts:

Bolton	Stockton-on-Tees	Neath Port Talbot	Corby
Oldham	Darlington	Rhondda Cynon Taff	Kettering
Rochdale	Blackburn with	Caerphilly	Ashfield
Tameside	Darwen	Blaenau Gwent	Cannock Chase
Wigan	Blackpool	Newport	East Staffordshire
St. Helens	North East	Bedford	Stafford
Wirral	Lincolnshire	Barrow-in-Furness	Ipswich
Barnsley	Derby	Carlisle	Wyre Forest
Doncaster	Telford and Wrekin	Derwentside	West Dunbartonshire
Rotherham	Stoke-on-Trent	Easington	Falkirk
Gateshead	Swindon	Sedgefield	Fife
North Tyneside	Peterborough	Harlow	Inverclyde
South Tyneside	Southend-on-Sea	Gloucester	Moray
Bradford	Thurrock	Gosport	Orkney Islands
Calderdale	Medway Towns	Burnley	South Ayrshire
Kirklees	Milton Keynes	Chorley	West Lothian
Wakefield	Wrexham	Hyndburn	
Hartlepool	Swansea	Pendle	

Figure 2.9 – *Z*-scores defining cluster 5



Cluster 5 features a higher prevalence of within district moves, represented by 9-12 and 22. Moves of individuals in lower socio-economic groups are significant (40-43), although lower managerial and intermediate occupations are also of importance (35-38). Migrants who are living alone or part of a lone parent family are of above average significance (52 and 55). Migrants

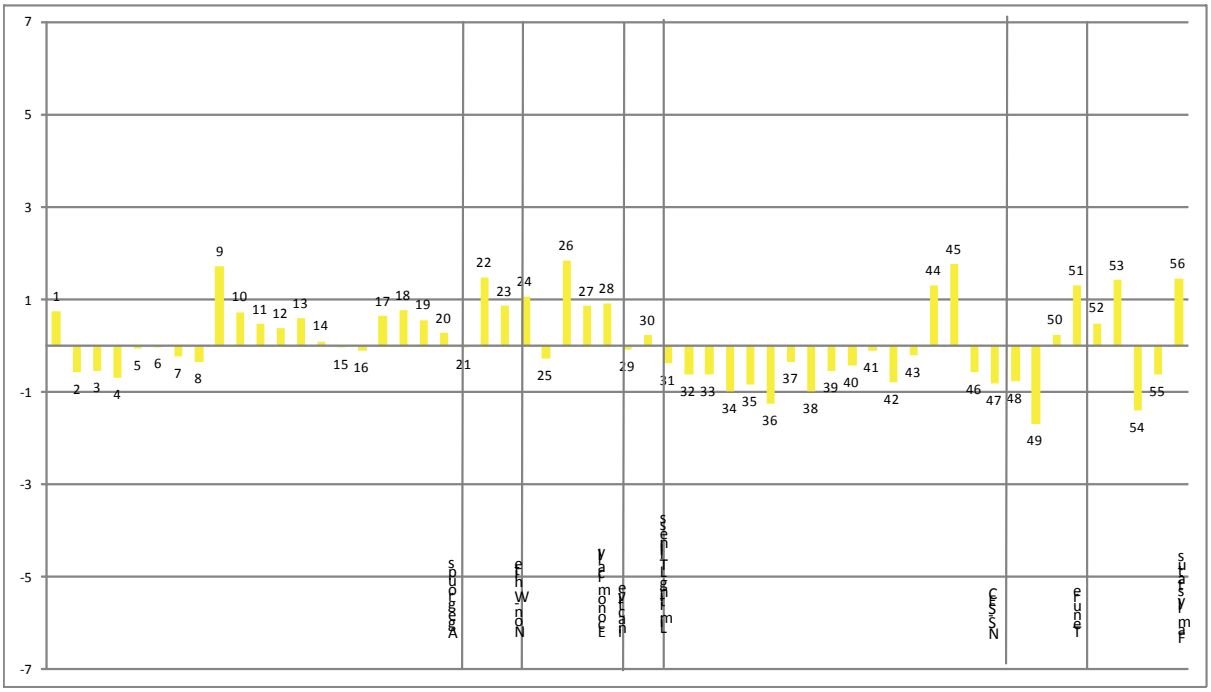
moving into owner occupied accommodation (48 and 49) have an above average prevalence, whilst moves into or from privately rented accommodation are below average. Population turnover is low with below average instances of in-migration and out-migration for all age groups (1-8) Immigration to districts in this cluster is also below average (13-16).

2.2.1.6 Cluster 6

Cluster 6 contains 68 districts:

Manchester	Leicester	Cardiff	Northampton
Salford	Nottingham	Chester	Broxtowe
Liverpool	Bath and North East Somerset	Exeter	Newcastle-under-Lyme
Sheffield	Bristol	Durham	Warwick
Newcastle upon Tyne	Plymouth	Colchester	Worcester
Sunderland	Bournemouth	Cheltenham	Aberdeen City
Birmingham	Luton	Welwyn Hatfield	Dundee City
Coventry	Brighton and Hove	Canterbury	Edinburgh
Wolverhampton	Portsmouth	Lancaster	Glasgow City
Leeds	Southampton	Preston	Stirling
Middlesbrough	Gwynedd	Charnwood	
Kingston upon Hull	Ceredigion	Lincoln	
York		Norwich	

Figure 2.10 – *Z*-scores defining cluster 6



Cluster 6 is defined principally by two main types of migrant. Firstly there are the young (often student) in-migrants and within district migrants (1, 9, 44 and 45). Then there are the slightly older, non-white and economically inactive within-district migrants (10-14, 22 and 26). Migration of individuals in other moving groups and non-family households is prevalent (51 and 53), as is migration into and out of communal establishments (56). International immigration of all ages is

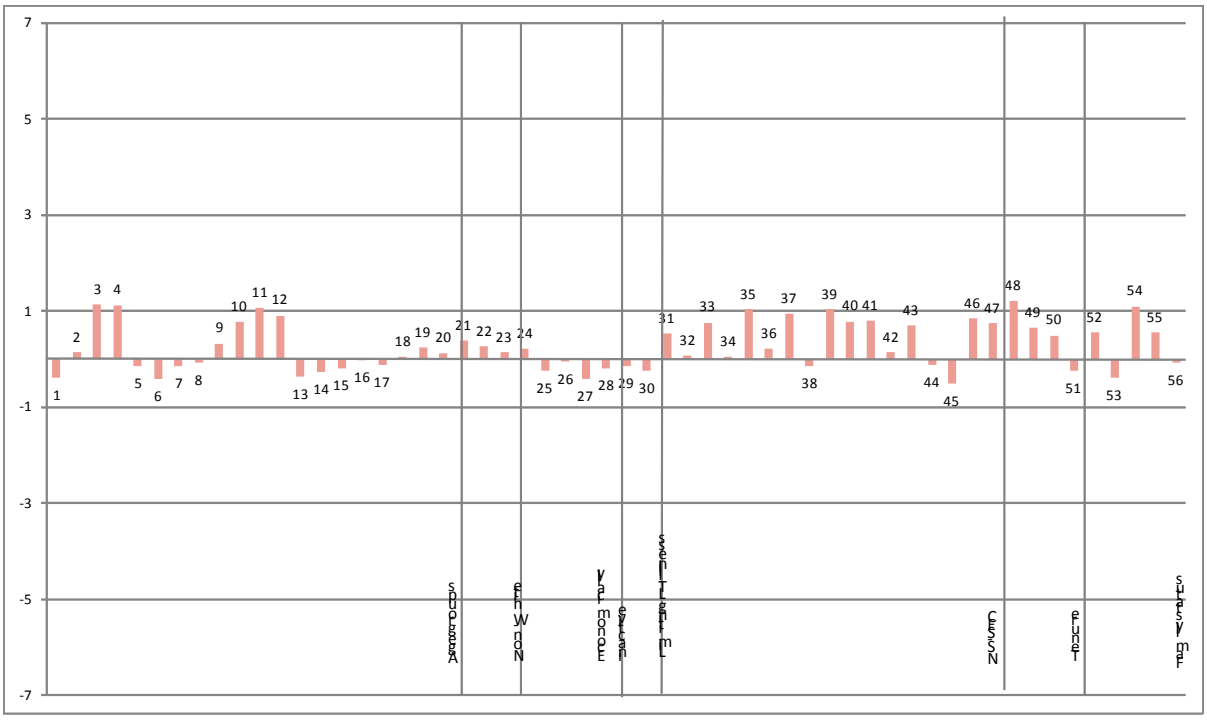
above average in this cluster (17-20 and 30) as is the migration of economically inactive migrants in general. Lone migrants and migrants moving into of from privately rented accommodation are also above average.

2.2.1.7 Cluster 7

Cluster 7 contains 69 districts:

East Riding of Yorkshire	Carrick	New Forest	Alnwick
North Lincolnshire	Kerrier	Ashford	Craven
Herefordshire	North Cornwall	Dover	Scarborough
North Somerset	Penwith	Shepway	Bassetlaw
Torbay	Restormel	Swale	Oswestry
Poole	Eden	Thanet	Mendip
Isle of Wight	East Devon	Fylde	Sedgemoor
Conwy	Mid Devon	Wyre	South Somerset
Denbighshire	North Devon	Boston	Taunton Deane
Powys	Teignbridge	East Lindsey	St. Edmundsbury
Pembrokeshire	Torridge	South Holland	Waveney
Carmarthenshire	West Dorset	West Lindsey	Arun
East Cambridgeshire	Weymouth and Portland	Breckland	Worthing
Fenland	Eastbourne	Great Yarmouth	West Wiltshire
Huntingdonshire	Hastings	King's Lynn and West Norfolk	Scottish Borders
Crewe and Nantwich	Rother	North Norfolk	Dumfries & Galloway
Caradon	Tendring	East Northamptonshire	Highland
			Perth & Kinross

Figure 2.11 – *Z*-scores defining cluster 7



Cluster 7 is characterised by in-migration of the post 30 age groups, particularly those 45 and over. Migrants tend to move in wholly moving households rather than other moving groups as shown by 31, 33, 35, 37, 39, 46, 48 and 50. These groups are across the socio-economic spectrum

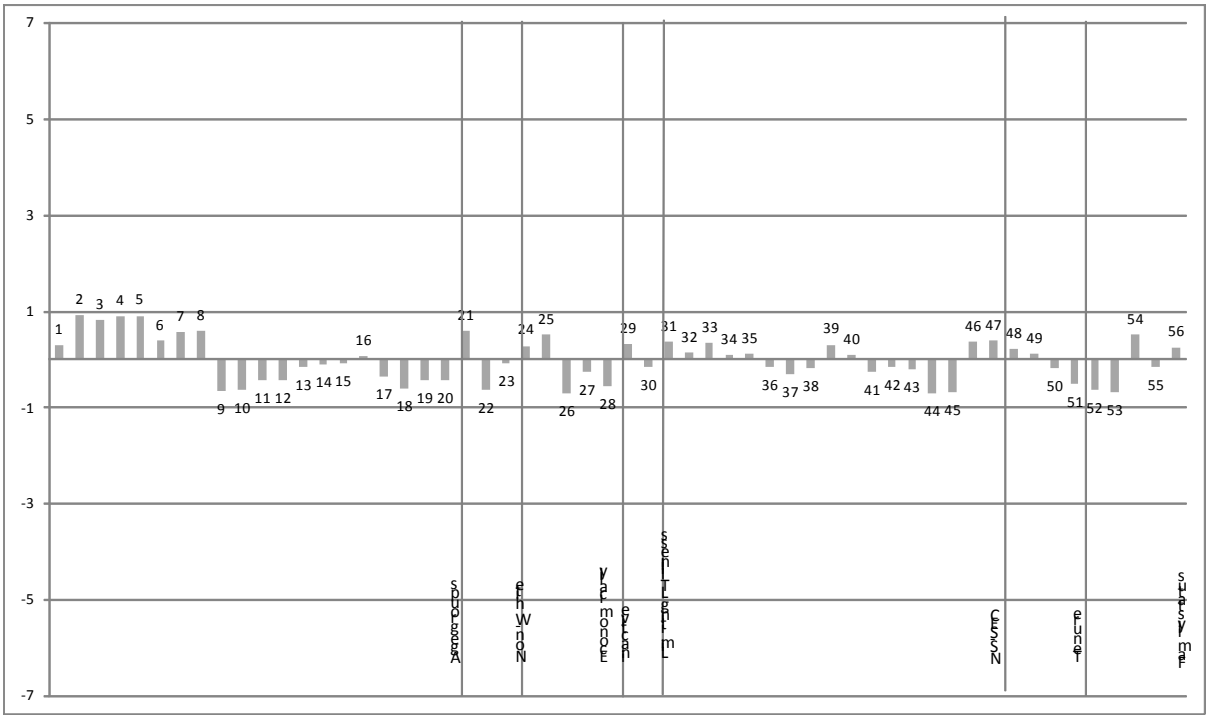
although NS-SEC categories 2, 3 and 4 predominate. Within district migration of all age groups is above average, whereas out-migration is below average. Migrants moving into or from owner occupied dwellings are of increased significance (48 and 49). Couple families are also of more significance than other habitation arrangements. Non-white migration is of slightly increased significance in this cluster (21-23), however economically inactive and ill migrants are of below average importance (25-30), as are students (44 and 45).

2.2.1.8 Cluster 8

Cluster 8 contains 54 districts:

Mid Bedfordshire	Tewkesbury	Wellingborough	Tandridge
Derbyshire Dales	Fareham	Castle Morpeth	Stratford-on-Avon
South Hams	Test Valley	Hambleton	Adur
West Devon	Winchester	Harrogate	Chichester
Christchurch	Tonbridge and Malling	Ryedale	Horsham
East Dorset	Blaby	West Oxfordshire	Mid Sussex
North Dorset	Harborough	Bridgnorth	Kennet
Purbeck	Oadby and Wigston	North Shropshire	North Wiltshire
Teesdale	North Kesteven	South Shropshire	Salisbury
Lewes	Broadland	West Somerset	Bromsgrove
Wealden	South Norfolk	Lichfield	Malvern Hills
Maldon	Daventry	South Staffordshire	Wychavon
Cotswold	South	Mid Suffolk	Argyll & Bute
Forest of Dean	Northamptonshire	Suffolk Coastal	

Figure 2.12 – *Z*-scores defining cluster 8



Cluster 8 is defined principally by in-migration and out-migration of all age groups (1-8) signifying relatively high levels of population turnover. Within-district moves, on the other hand, are of significantly below average importance (9-12). International immigration is also of less importance in this cluster. Non-white in-migration (21) is prevalent in this cluster as too are couple family migrants (54). Other variables which help define this cluster relate to the above average in and out-migration of economically inactive individuals (24 and 25); the out-migration of individuals with limiting long-term illnesses (29); and the prevalence of migrants in the higher socio-economic categories (31-35). Students, migrants living alone or in non-family households are particularly under-represented.

2.3 Refining the initial classification.

So it is clear from section 2.2 that through following a series of methodological steps, it is entirely possible to produce a plausible classification of districts based on migration variables. As was noted earlier, however, geodemographic classification results are often plausible. Clustering algorithms will always produce results regardless of the data input. Like any piece of work, however, the first draft is very rarely the final product. Once a draft has been produced, then it is reviewed and analysed, with improvements made where necessary to ensure the final result is the one which is based on the best decisions, and therefore produces the most robust end product. This next section, therefore, will review some of the key choices and additional considerations, and where necessary implement changes so that a final, definitive migration classification can be achieved.

2.3.1 Variable transformation

Something which was not confronted during the initial classification design was the issue of variable transformation. Milligan (1996) chose not to address this issue in his seven steps, yet throughout the classification literature there is much discussion about the need (or not) to transform variables which do not meet normal distribution assumptions.

Within the more general statistical literature (Field, 2005, provides a particularly accessible overview) the importance of a normal, Gaussian, distribution of frequency observations in data is frequently expressed, especially where parametric tests (which for their accuracy rely on such distributions) are employed. Often, frequency distributions do not follow a normal, symmetrical curve – observations may cluster at one end of the scale exhibiting either a positive skew (more frequent observations are at the lower end of the scale with fewer at the higher end) or a negative skew (more frequent observations at the higher end of the scale with fewer at the lower end). Where data are not distributed normally, typically, statisticians have ‘corrected’ the data to a more normal distribution in order that further analysis techniques can be used without the results of these analyses being unreliable. Field (2005) outlines the main ways in which skewed data can be corrected – the most common solutions being either to remove outliers (which may be affecting the frequency distribution), or to transform the data using either logarithmic, square-root or reciprocal transformations (Field, 2005, p80).

Whilst there is a need to transform skewed data for some parametric tests such as ordinary least-squares regression to be reliable, the necessity for such transformations in cluster

analysis is less clear-cut. It has been argued that skewed variables will bias cluster membership. In a migration classification, this might be apparent where, for example, inner London boroughs exhibit very high counts of international immigrants compared to all other districts in Britain. These cases may be clustered because of these very skewed variables, meaning that other interesting characteristics that these boroughs may exhibit for other more normally distributed variables will be ignored. The high values for these immigration variables will mean that it is by these variables that they are defined.

Both Vickers (2006) and Založnik (2006) advocate the transformation of skewed variables for the reasons mentioned above. There are others, however, who are less convinced of the case for transformation. Openshaw and Wymer (1995 p245) remark that: *"Some thought may sometimes be given to the possibilities of applying a data transformation. After all, this sounds like the correct statistical thing to do. Well, think carefully about it and, then, perhaps don't do it! It can be argued that there is little to be gained from data transformations, bearing in mind the exploratory nature of classification and the difficulties it might cause during interpretation."*

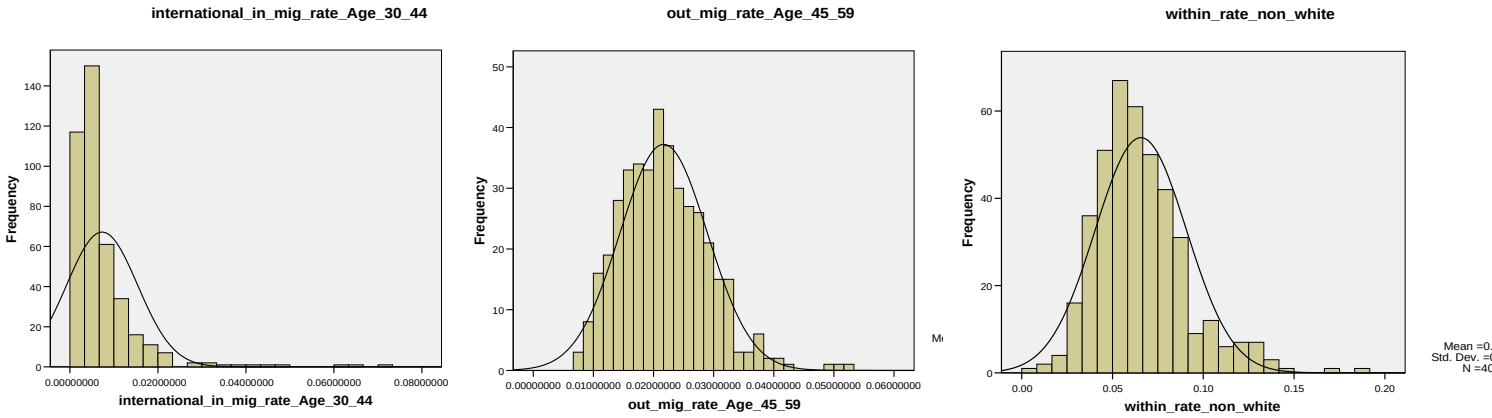
Grayson (2004) also warns of possible implications of transforming data. By transforming data, whilst relative differences remain the same – i.e. London is still a more popular location for immigrants than Leeds, and Leeds is more popular than Cornwall; exactly how much more popular is lost in the transformation. Whilst both London and Leeds are more popular than Cornwall, it may appear that Cornwall is less popular than Leeds by the same amount than Leeds is less popular than London, when in reality Cornwall may be vastly less popular than both.

So whilst there are both reasons and advocates for and against the transformation of 'poorly behaved' variables, any decision about whether or not a transformation is necessary for this classification will need to be based purely on the variables being used here, rather than what has or has not been common practice for other classifications. A useful start point may be to assess the impact on the classification of transforming the variables, if indeed poor distributions suggest they need transforming.

2.3.1.1 Examining variable skewness

One of the main difficulties when deciding whether or not to transform data is assessing 'how skewed is too skewed?' Consider Figure 2.13 below:

Figure 2.13 Variable distributions and skewness statistics for three example variables



Skew = 4.14 Skew = 0.73 Skew = 1.01

Standard error of skewness = 0.12

Figure 2.13 shows frequency histograms for three variables included in the initial migration classification. One common (but perhaps not strictly scientific) way of assessing skewness is to examine the frequency histogram for a variable. To anyone with even a rudimentary training in examining histograms for skew, it would be obvious that the histogram on the left representing the counts of international immigrants aged 30-44, is displaying a significant positive skew, with the majority of observations found to the left of the x axis. With the other two graphs, however, the presence of skew is much less obvious. Indeed the distributions of both graphs look relatively normally distributed. Should these variables be classed as such – is it acceptable to include them in the classification?

Rather than 'eyeballing' the data in a histogram, perhaps a less qualitative assessment would be to look at the skewness statistic for each variable. Most standard pieces of statistical software (such as SPSS) will provide these statistics as standard descriptive output. SPSS (2006) states that where a distribution is normal, the skewness statistic will be zero, and as a general rule when this statistic is more than twice its standard error, then the distribution is skewed. As can be seen in Figure 2.13, in all cases the skewness statistic is more than twice the skewness standard error, suggesting all of the variables are skewed. Indeed, analysis of all 56 variables included in the original classification reveals that only 7 have skewness which is less than twice their skewness standard error, despite many variables appearing to display relatively normal distributions in their histograms.

2.3.1.2 Transforming skewed variables

If we are to accept that the skewness statistics indicate all but 7 variables need transforming, what are results of such a transformation, both on the variables themselves, and on the classification? As previously mentioned, two of the most common transformations used to correct data are logarithmic transformations and square root transformations. Both will be applied to the data and assessed. Before any transformation is applied, however, constant values need to be added to the data. It is pointed out by Vickers (2006), that as the logarithm of zero returns no result, a constant should be added to the data before transformation, if indeed zeros exist in the data. Within the 56 variable dataset being used here, a number of zeros occur, so a constant of 1 was added. All variables were transformed using both methods and the results are displayed below in Table 2.7.

The results of the transformations are mixed. In some cases the transformations have reduced the skewness statistics, whereas in others, the original data remains more normally distributed. In fact, for the majority of variables, the original data shows the most normal distribution. It is not an option to apply different transformations to different variables – if any variables are to be transformed, all others must be transformed in the same way. With the original data being more normally distributed than the transformed data, then this suggests that the data should be left un-transformed. Before moving on, it will be interesting to explore the effect would any transformation has on a cluster solution. If the effect is small, then regardless of whether a transformation improves variable distribution or not, it may not be worthwhile pursuing it anyway. Put another way, will poorly behaved variables present in the dataset have an undue effect on the clusters produced? One way to explore this would be to compare the cluster solutions produced with the original data and with transformed data.

2.3 Refining the initial classification.

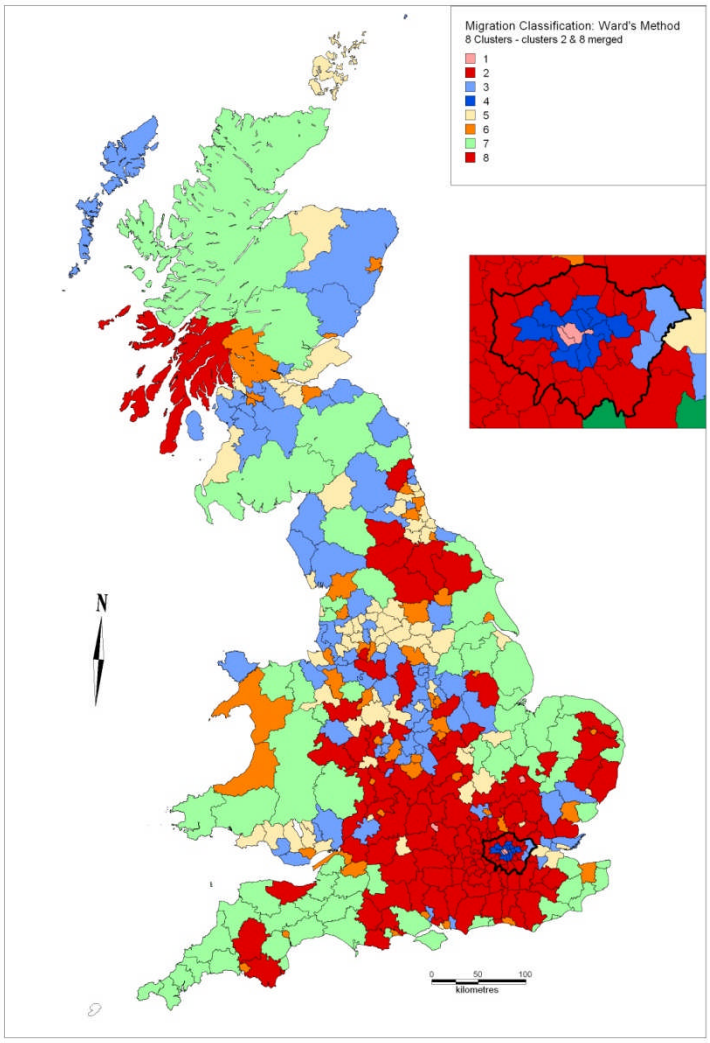
Table 2.7 Results of log and square root transformations on skewness statistics

Variable	Skewness	obsolute difference between skewness and skewness standard error	skewness after log transformation	skewness after square root transformation	data with best skewness statistic (original data - orig, log transformation - lg, square root transformation - sqrt)
in_mig_rate_Age_16_29	1.21	1.09	0.01	0.59	lg
in_mig_rate_Age_30_44	0.60	0.48	-0.49	-0.02	sqrt
in_mig_rate_Age_45_59	0.94	0.81	-0.12	0.38	lg
in_mig_rate_Age_60_plus	0.48	0.36	-0.40	0.04	sqrt
out_mig_rate_Age_16_29	0.71	0.59	-0.19	0.23	lg
out_mig_rate_Age_30_44	1.59	1.47	0.20	0.83	lg
out_mig_rate_Age_45_59	0.73	0.61	-0.30	0.19	sqrt
out_mig_rate_Age_60_plus	0.55	0.43	-0.34	0.09	sqrt
within_mig_rate_Age_16_29	0.98	0.86	-0.25	0.40	lg
within_mig_rate_Age_30_44	0.03	0.09	*	*	orig
within_mig_rate_Age_45_59	0.28	0.16	-0.28	0.00	sqrt
within_mig_rate_Age_60_plus	0.19	0.07	*	*	sqrt
international_in_mig_rate_Age_16_29	2.75	2.63	0.24	1.37	lg
international_in_mig_rate_Age_30_44	4.14	4.02	0.39	1.97	orig
international_in_mig_rate_Age_45_59	5.85	5.72	0.53	2.34	orig
international_in_mig_rate_Age_60_plus	5.15	5.03	-0.12	1.49	lg
no_addr_in_mig_rate_Age_16_29	2.16	2.04	0.87	1.47	orig
no_addr_in_mig_rate_Age_30_44	1.63	1.51	0.67	1.14	orig
no_addr_in_mig_rate_Age_45_59	1.55	1.43	0.44	0.50	orig
no_addr_in_mig_rate_Age_60_plus	1.94	1.82	0.34	0.44	orig
in_mig_rate_non_white	1.51	1.39	-0.72	0.27	orig
within_rate_non_white	1.01	0.89	-0.48	0.02	sqrt
no_addr_in_rate_non_white	0.82	0.70	-0.71	-0.51	orig
in_mig_rate_Economically_Inactive__Total	1.71	1.59	0.52	0.81	orig
out_mig_rate_Economically_Inactive__Total	1.04	0.92	-0.20	0.38	lg
within_rate_Economically_Inactive__Total	2.08	1.96	0.51	1.28	orig
international_in_rate_Economically_Inactive__Total	3.16	3.04	0.27	1.49	orig
no_addr_in_rate_Economically_Inactive__Total	1.78	1.66	0.51	1.23	orig
out_mig_rate_LLTI_in_HH_and_CE_Total	8.24	8.12	0.71	2.21	orig
international_in_rate_LLTI_in_HH_and_CE_Total	3.23	3.11	0.16		lg
efficiency_NS_SEC_11_Wh_move_hh_All_groups	0.30	0.18	0.18	0.24	lg
efficiency_NS_SEC_11_Oth_mvg_grp_All_groups	-0.40	0.52	-0.55	-0.47	orig
efficiency_NS_SEC_12_Wh_move_hh_All_groups	-0.26	0.38	-0.32	-0.29	orig
efficiency_NS_SEC_12_Oth_mvg_grp_All_groups	-0.62	0.74	-0.71	-0.67	orig
efficiency_NS_SEC_2_Wh_move_hh_All_groups	-0.23	0.36	*	*	orig
efficiency_NS_SEC_2_Oth_mvg_grp_All_groups	-0.71	0.83	-0.76	-0.73	orig
efficiency_NS_SEC_3_Wh_move_hh_All_groups	0.11	0.01	*	*	orig
efficiency_NS_SEC_3_Oth_mvg_grp_All_groups	-1.35	1.47	-1.47	-1.41	orig
efficiency_NS_SEC_4_Wh_move_hh_All_groups	-0.13	0.25	*	*	orig
efficiency_NS_SEC_5_Wh_move_hh_All_groups	-0.30	0.42	-0.36	-0.33	orig
efficiency_NS_SEC_6_Wh_move_hh_All_groups	0.30	0.17	0.20	0.25	lg
efficiency_NS_SEC_6_Oth_mvg_grp_All_groups	-0.06	0.18	*	*	orig
efficiency_NS_SEC_7_Wh_move_hh_All_groups	-0.48	0.60	-0.57	-0.52	orig
efficiency_NS_SEC_FT_student_Wh_move_hh_All_groups	0.48	0.36	0.39	0.43	orig
efficiency_NS_SEC_FT_student_Oth_mvg_grp_All_groups	0.90	0.78	0.83	0.87	orig
efficiency_NS_SEC_Not_class_oth_reason_Wh_move_hh_All_groups	-0.40	0.52	*	*	orig
efficiency_NS_SEC_Not_class_oth_reason_Oth_mvg_grp_All_groups	0.74	0.62	*	*	orig
efficiency_Owner_occupied_Wh_mvg_hh_All_groups	-0.47	0.59	-0.53		orig
efficiency_Owner_occupied_Oth_mvg_grp__All_groups	-1.38	1.50	-1.42		orig
efficiency_Private_rented_Wh_mvg_hh_All_groups	-0.60	0.72	-0.69		orig
efficiency_Private_rented_Oth_mvg_grp__All_groups	0.39	0.27	0.34		orig
efficiency_Alone_total	0.02	0.10	*	*	orig
efficiency_Non_Family_Household_Total	0.66	0.54	0.61	0.63	orig
efficiency_In_couple_family_total	-0.60	0.72	-0.63	-0.61	orig
efficiency_In_lone_parent_family_total	-1.64	1.76	-1.85	-1.74	orig
efficiency_Living_in_a_communal_establishment_total	0.50	0.38	0.43	0.47	orig

* = no requirement to transform original data due to normal distribution.

Figure 2.14 Comparison of areas created by Ward's clustering algorithm:

a) Original data



b) Log transformed data

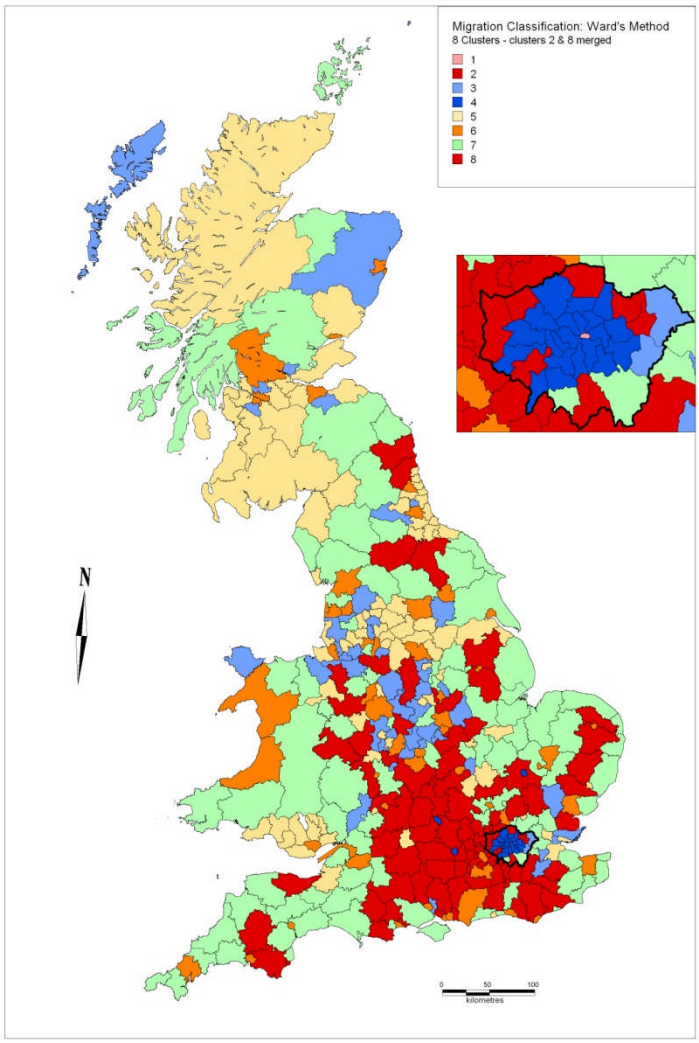


Table 2.8 Effect of a log transformation on cluster membership

	Count of districts		
Cluster	original data	logged data	difference
1	5	1	-4
2&8	123	93	-30
3	79	49	-30
4	12	24	12
5	67	84	17
6	47	52	5
7	69	99	30

In this example, clusters 2 and 8 have been combined to aid exemplification. The two clusters are principally defined by similar variables (as shown by the cluster profiles in section 2.2) so combining them here is not an issue. The first point of note is that transforming the data does not drastically alter the overall spatial patterns of the clusters. Both maps show comparable patterns, with the majority of districts remaining in the same cluster after transformation. Table 2.8 quantifies this by showing the number of districts comprising each cluster. At most the clusters change size by 30 districts, and across the whole classification, core areas appear to remain stable – university towns for cluster 6; ex-industrial areas for cluster 5; coastal and rural areas for cluster 7 etc. These results would suggest that transforming variables (especially where the transformation does not significantly improve skewness) does have a huge impact upon a final cluster solution. Taking all of this evidence into consideration, the conclusion would have to be to leave the data in its original state, and to not apply a transformation.

2.3.2 Dropping the most skewed variables?

The previous section has shown that transformations make little difference to the skewness of the 56 variables used in the initial classification. But while the decision has been made not to transform variables, there still remain a number of more highly skewed which may affect the final cluster solution. Skewness statistics showed that most variables were skewed, however some variables were significantly more skewed than others. These are shown in Table 2.7 but also can be identified easily by studying the associated histograms (Figure 2.15).

A number of variables are very obviously more skewed than others. The order of the histograms in Figure 2.15 is the same as Table 2.6, therefore histograms labelled output 12-15 represent international immigrant age variables, 16-19 represent the no usual address in-migrants, 25, 26 and 27 represent economically inactive within area migrants, international immigrants and no usual address migrants and 28 and 29 represent migrants with a limiting long-term illness. It has been suggested (Harris *et al.*, 2005; Vickers, 2006; Založnik, 2006) that these very skewed variables (which also contain a number of outliers) will bias some of the clusters within the classification. A sensible course of action, therefore, would be to examine whether indeed this is the case. Will the most skewed variables have a detrimental effect on a classification produced? Selectively dropping the most skewed variables and re-running the cluster analysis to examine the effect would allow for an assessment of this type to be made. However, even where the inclusion of very skewed variables creates biased

clusters in a classification, there are solutions to the problem. Harris *et al.* (2005) suggest that if such clusters are created, one approach would be to allow them to form, but then to run a separate cluster analysis on the clusters created by the skewed variables, linking them back to the rest of the classification. The results below describe experiments carried where some of the most skewed variables were dropped from a cluster analysis.

Figure 2.15 Frequency histograms for the 56 variables used in the initial classification

2.3 Refining the initial classification.

To establish a baseline for the experiment, the 56 variables from the initial classification were clustered using a *k*-means algorithm searching for 8 clusters. *k*-means was used instead of Ward's algorithm in this instance for reasons which will be explained fully in the next section. The algorithm ran through 1,000 iterations with different randomly selected initial cluster centroids in order to find the optimum cluster solution. The 8 clusters produced by this procedure are displayed below in Figure 2.16. In the second part of the experiment, exactly the same *k*-means procedure was applied to the data, although without most variables relating to no previous address (output 16-18 and 27 Figure 2.15), limiting long-term illness (output 28 and 29 Figure 2.15) and the two most skewed economically inactive variables (outputs 25 and 26 Figure 2.15) – a total of 8 variables were dropped leaving 48 variables left to be clustered. The skewed international immigration age variables were left in at this stage in order to ascertain their influence on the cluster solutions in the third part of the experiment where they are removed. In the third part of the experiment a final *k*-means cluster run was carried out, this time with the 4 international immigration age variables (outputs 12-15 Figure 2.15) additionally removed – leaving 44 variables. Figure 2.17 below reveals the districts in Britain which moved to a different cluster after the first group of variables were dropped and after the final international immigration variables were dropped.

As is shown clearly in Figure 2.17 a), dropping the first set of highly skewed variables has a negligible effect on the final cluster solution. Only 12 districts in Britain change cluster as a consequence. This is an unexpected result given the stated effect (Harris *et al.*, 2005; Vickers, 2006; Založnik, 2006) of skewed variables on cluster solutions. We might suspect, for example, that the cluster defined in a large part by no usual address variables (as these were the most numerous dropped) would be affected the most. Cluster 4 was the original cluster defined mostly by these variables, but there were not any districts in cluster 4 which moved group. Most changes, in fact, occurred in cluster 7. Even when the skewed international immigration variables were dropped (as shown in Figure 2.17 b), very few additional districts changed their cluster membership. An additional 9 clusters changed, principally from cluster 1 which forms the London periphery – a cluster not defined heavily by these international variables.

Figure 2.16 – Clusters produced from a *k*-means clustering run searching for 8 cluster solutions.

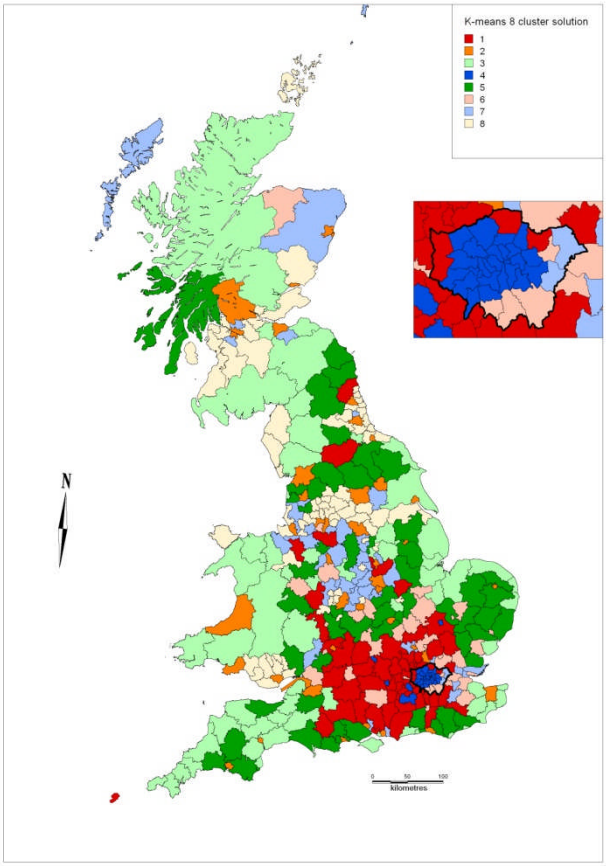
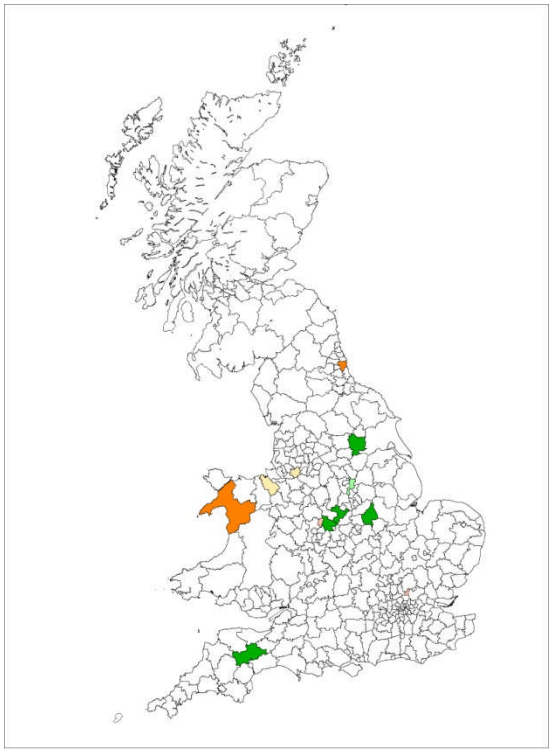
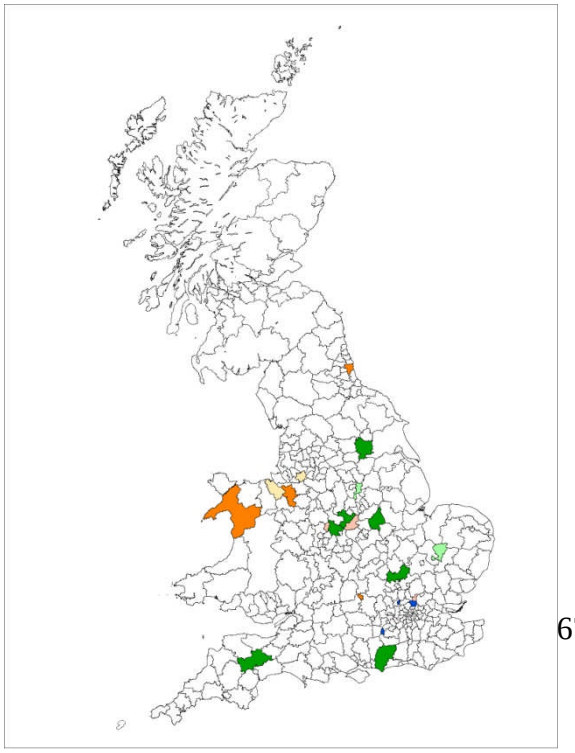


Figure 2.17 Districts changing cluster group after total number of variables reduced to.

a) 48 variables



b) 44 variables



What this experiment has shown is that the most skewed variables are not playing a very large role in the formation of clusters in this classification, despite the general consensus in the literature that skewed variables or variables with outliers will tend to create their own biased clusters. Whilst international migration, limiting long-term illness and no usual address variables can characterise some clusters when included, when they are removed, rather than districts being re-appropriated by other clusters, it appears that other variables are maintaining the cluster structures.

This is a significant discovery. It suggests that international migration can be dropped altogether as a variable within the classification. The classification then becomes a classification purely for internal migration. Reducing the variables in this way is useful as one of the original aims of this classification was to provide a framework for monitoring internal migration. Removing international migration from the classification altogether means that it becomes an internal migration classification for monitoring internal migration, rather than a classification partially defined by immigration, but designed for studying internal migration. It may be of use in the future to construct an international migration (immigration and emigration) classification, however with data and aggregate estimates of international immigration being unreliable (Boden and Rees, 2009, Forthcoming; Rees *et al.*, 2009), details of individual immigrants more so, and estimates of emigration even poorer (Heasman, 2008) such a classification may not be feasible. The volatile nature of international immigration and uncertainties about the length of migrant stay – something which caused by changing economic circumstances at both origins and destinations, and is unlikely to become more stable as the current global economic crisis continues (Boden and Stillwell, 2006) – will affect the reliability and usefulness of any classification produced.

2.3.3 Cluster Optimisation: A Different Clustering Algorithm

The previous section of this paper has made it clear that it will be best to use non-transformed, internal migration variables in the final classification. Having dealt with that issue, the next stage in refining and arriving at a final classification concerns the clustering methodology itself. Ward's algorithm was used in the initial classification, principally because the method of creating a hierarchy of clusters allowed for a partition to be selected where the most suitable number of clusters was unknown. The main aim of partitioning data in a classification, as outlined by Gordon (1999), is to group objects that are similar to each other in one class, and dissimilar objects in another class. As mentioned during the

development of the initial classification, one of the issues with using a hierarchical algorithm is that it may not necessarily find the optimum solution where all cases/objects are allocated to the class to which they are most similar – i.e. closer to the cluster centroid of one cluster than to the centroid of any other. The nature of the agglomerative hierarchical algorithm means that once a case has been allocated to a particular cluster, it cannot then be removed and re-allocated to a more appropriate cluster. It is for this reason that Everitt *et al.* (2001) advocate the use of an optimisation algorithm, which, given a specific number of clusters to create, will iteratively allocate and reallocate points to different clusters until no improvement is made to a final solution.

One such algorithm is the *k*-means algorithm (Everitt *et al.*, 2001). For a given initial partition of *k* clusters with randomly selected cluster centroids, the algorithm takes each case in a dataset of *n* cases, allocates a case to a cluster, recalculates the centroid of that cluster and repeats the process until each case has been allocated to a cluster and the reallocation of any case does not improve the average distance of cases to the cluster centroids (for a more detail description of the *k*-means algorithm, see Založnik, 2006). Of course the distance to the cluster centroid can be measured in a number of ways, and this can affect the final cluster solution – this will be discussed in full later.

One of the main issues with using a *k*-means procedure is that the clusters created after the algorithm has been run once may represent a local optimum (i.e. cases are allocated to their closest cluster centroids, but these centroids may not be optimum, merely the artefact of their initial seed or partition), but may not represent a global optimum (where the centroids also represent the best possible solution) (Gordon, 1999). Frequently in *k*-means clustering, depending on the cases chosen as the initial cluster centroids, it can be that, even when all other elements of the clustering process remain constant, different final cluster solutions will be reach. This is exemplified clearly in Figure 2.18

Figure 2.18 *k*-means clustering of 408 cases, 56 variables in SPSS.

a) Cases sorted by district name b) Sorted by district code

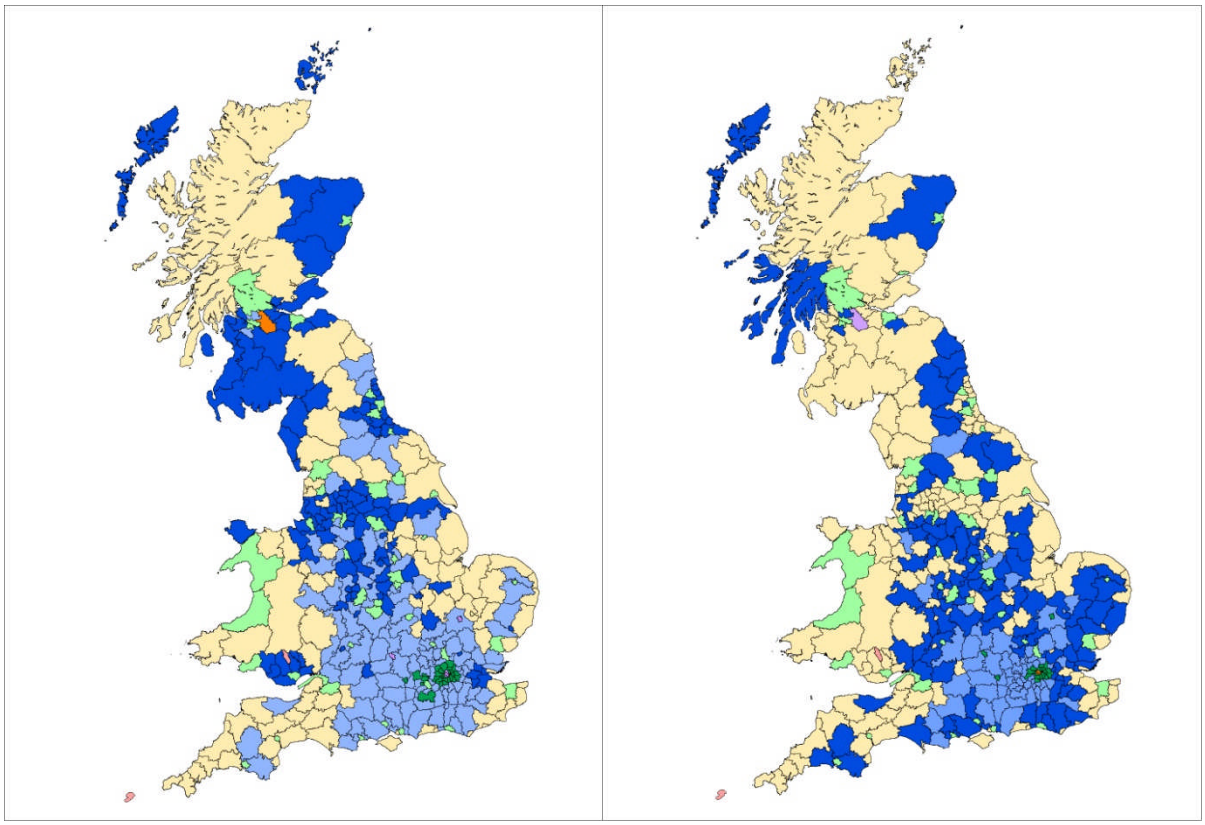


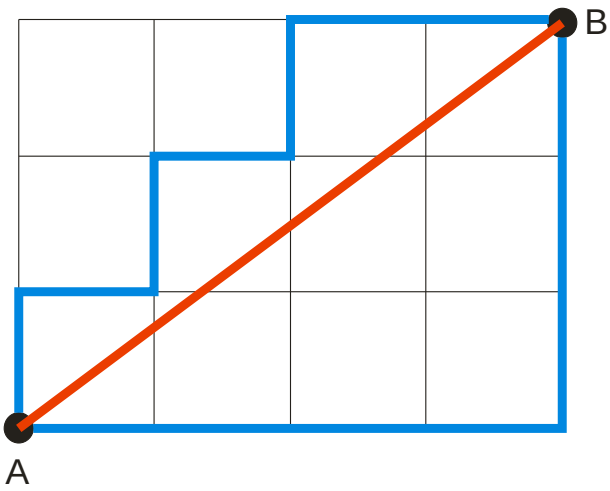
Figure 2.18 shows the cluster solutions produced when the 56 variables used in the initial classification were clustered into 8 clusters using the *k*-means algorithm available in the SPSS software package. In Figure 2.18 a), the 408 cases are sorted alphabetically by district name, in Figure 2.18 b) they are sorted by district code. Clearly the two maps are quite different. Sorting the districts in different ways has drastically altered the solutions produced by exactly the same algorithm using exactly the same data.

The issue with using the *k*-means algorithm as it is implemented in SPSS is that the user does not have control over the selection of initial cluster centroids. The software will allocate these centroids randomly (although the exact method for this is not made clear in the documentation) before iterating through until the end. SPSS (2005) acknowledges that the order of the cases will, in all likelihood, affect the final cluster solution, but only offers the (rather unsatisfactory) solution to sort the cases ‘in different random orders’ (SPSS, 2005, p489) to verify a cluster solution. This is impractical at best, especially where there are a large number of cases. At worst, it is likely that an optimum solution could never be reached intelligently. Work by Falkenauer and Marchand (2001 cited in Založnik, 2006) found that in

experiments with a dataset using 10,000 different initial partitions, 9,874 different cluster solutions were found.

An additional issue with using the *k*-means algorithm in SPSS is that the distance measure used to measure the distance of cases to a cluster centroids is Euclidean distance, and cannot be altered (SPSS, 2005). The solution SPSS offers to this problem is to use a hierarchical cluster analysis procedure. However, doing this would obviously be unsatisfactory as *k*-means is being used here to optimise the solution already produced using a hierarchical algorithm! Whilst Euclidean distance was used as the distance metric in the initial classification, it may well be that other distance measures produce better cluster solutions. Indeed research by Aggarwal *et al.* (2001) points to Manhattan (City) distance between points providing a better cluster solution than standard Euclidean distance where the data has many dimensions – many in the Aggarwal *et al.* example being more than 20 dimensions.

Figure 2.19 A representation of the difference between Euclidean and Manhattan (City Block) distances between two points.



To explain the difference between the two measures, consider Figure 2.19. Manhattan distance differs from Euclidean distance in that it is the sum of the absolute difference between two coordinates – put another way the distance between two points on a grid system where only the grid can be travelled along (analogous to the distance travelled by a taxi driver along the road grid network in New York City – hence Manhattan), whereas Euclidean distance is the straight line distance between two points in space. The side of each square in the grid represents one unit of space, and points A and B can be located on this two dimensional grid space. The red line represents the straight line Euclidean distance between them; the blue lines demonstrate two alternative ways the same Manhattan distance could be

calculated. In this case, the Euclidean distance is 5 units, whereas the Manhattan distance is 7 units. With the data being used in this research containing at least 44 variables (or *dimensions*) the research of Aggarwal *et al.* suggests strongly that using Manhattan distance to measure the distance to cluster centroids would result in a better definition of *k* clusters.

2.3.3.1 Using *k*-means in MATLAB.

So with a number of problems inherent in the way that the SPSS software implements the *k*-means clustering algorithm an alternative solution needs to be sought. One possibility would be to write a bespoke piece of software which implements the *k*-means algorithm with the option to run the algorithm through a user defined number of iterations, each starting with different initial cluster centroids, the final solution being the one fitting some defined 'best' criteria, and with the option to choose different measures of distance to the cluster centroids. In an ideal world, every researcher would have the computer programming skills to be able to do this. In reality few do – and even if they do could rarely afford the time to write such a piece of software – so rely on the skills and expertise of others to create tools for them to use which meet their requirements as far as possible.

Aside from SPSS, a number of other statistical analysis packages are available to researchers which have a version of the *k*-means algorithm pre-programmed. Packages such as Minitab, R and Stata will all run a *k*-means cluster analysis on a given dataset. The program, however, which met the needs of this analysis very well was the MATrix LABoratory (MATLAB) Statistics Toolbox (MathWorks, 2009). MATLAB incorporates a number of features in its *k*-means algorithm which makes it preferable to SPSS.

Firstly, MATLAB allows for a choice of five different distance measures. Both Euclidean and Manhattan distances are available as well as cosine, correlation and hamming distances. In addition to this, MATLAB also offers a solution to the problem of a local rather than global minima being reached at the end of the clustering iterations. An optional 'replicates' parameter can be included in the algorithm. This parameter will run the algorithm for the specified number of replicates (1-*n*) with each replicate starting the whole cluster run again with a new set of randomly selected initial cluster centroids. Once the specified number of replicates have been completed, the solution offered up by the program is the one with the lowest total sum of distances to which cluster centroids happen to have been chosen. (MathWorks, 2009). This solution is likely to be the global minimum, although obviously the more replicates used, the more confident one can be that this is indeed the case.

2.3.4 Choosing *k*.

One of the difficulties with using *k*-means over a hierarchial algorithm is that the number of clusters *k* needs to be defined at the outset. As with many elements of classification building the literature offers no definitive answer for deciding the most appropriate value of *k*. As stated by Everitt *et al.* (2001) the initial partition with an associated number of clusters might be chosen through prior knowledge or from result of a previous clustering method. Everitt *et al.* (2001) and Gordon (1999) review a number of other methods which could be used, ranging from the slightly more subjective, such as the assessment of 'large' differences in the distances between the most dissimilar areas in cluster groups in graphical representations (as used in the initial classification), to the more formal, such as those assessed by Milligan and Cooper (1985 – cited in Everitt et al., 2001 and Gordon, 1999), which in general use mathematical procedures to assess the within and between cluster differences – the best results generally being where within cluster distances are minimised and between cluster differences are maximised. Whilst both Everitt *et al.* (2001, p105) and Gordon (1999, p63) state (almost too similarly) that no one method assessed by Milligan and Cooper should be used in preference to another, rather researchers should *"synthesise the results of several"* techniques, it is impractical to exhaustively work through a large range of methods. A sensible option would be to select some of the more popular methods and use a combination of those.

One particular method invented by Rousseeuw (1987), recommended by Kaufman and Rousseeuw (2005), espoused by MathWorks (2009) and implemented as the principal method to decide the number of clusters in the classification developed by Shepherd (2006), is the interrogation of 'silhouette' plots and values. A silhouette plot is a graphical representation of the average dissimilarity between any object/case within a cluster and other objects within both its own cluster and those in other clusters. The plots are represented on an index of -1 to +1. A value close to +1 signifies that that object is nearer to its own cluster than any other. A value close to -1 suggests that the object might well be better placed in another cluster. Zero signifies that it is unclear whether that object is better placed in its current cluster or another.

A silhouette *s* value for any object (*i*) in a cluster can be defined thus:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \qquad (4)$$

where

$a(i)$ = the average dissimilarity of *i* to all other objects in the same cluster

$b(i)$ = lowest average dissimilarity of *i* to the objects in all other clusters in the whole solution

Silhouette plots are ranked (highest to lowest) silhouette values for the objects in each cluster. Better defined clusters will have fewer values close to or below zero compared to others.

In the construction of the ONS classification of Output Areas, Vickers (2006) chose not to use silhouette data to select the most appropriate value of *k,* but instead employed a selection of other methods – some more logically than others. For example, whilst Vickers used the average distance of points in a cluster from the cluster centroid as one of his methods, the utility of this method is unclear since the average distance to cluster centroids will always reduce as the number of clusters increases. Of more use is the assessment of cluster size. One key observation made by Vickers is that it is desirable to have clusters which are similarly matched in size – equal clusters being the optimum solution with very large and very small clusters being much less desirable as small clusters are more likely to contain outliers. This is logical, and Vickers uses the average distance from the mean number of cases in each cluster for a range of cluster solutions to help decide the most appropriate number of clusters for the OA classification.

Silhouette data as well as statistics for the size of clusters were produced for different values of *k* in order that both methods could be used in parallel to select the most suitable number of clusters for the final classification. These methods used in conjunction with different measures of cluster distance (Euclidean and Manhattan) along the replicates parameter in MATLAB will be discussed in the next section where the final district level migration classification is outlined.

2.4 Arriving at a final classification.

Now an initial trial classification has been produced, and both the variables selection and methodology reviewed with decisions made about how to improve both, the task still remains to build the final migration classification. For reasons discussed, variables relating to international immigrants, ill health and most relating to migration from no previous address,

were dropped from the initial set of 56 variables, reducing the final set of variables to 44. These are listed below in Table 2.9

Table 2.9 Final selection of internal migration variables used in the classification

	Variable
1	Internal in-migration rate of persons aged 16 to 29
2	Internal in-migration rate of persons aged 30 to 44
3	Internal in-migration rate of persons aged 45 to 59
4	Internal in-migration rate of persons aged over 60
5	Internal out-migration rate of persons aged 16 to 29
6	Internal out-migration rate of persons aged 30 to 44
7	Internal out-migration rate of persons aged 45 to 59
8	Internal out-migration rate of persons aged over 60
9	Internal within-area migration rate of persons aged 16 to 29
10	Internal within-area migration rate of persons aged 30 to 44
11	Internal within-area migration rate of persons aged 45 to 59
12	Internal within-area migration rate of persons aged over 60
13	In-migration rate from no previous address of persons aged over 60
14	Internal in-migration rate of non-whites
15	Internal within-area migration rate of non-whites
16	In-migration rate from no previous address of non-whites
17	In-migration rate of economically inactive individuals
18	Out-migration rate of economically inactive individuals
19	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.1
20	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.1
21	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 1.2
22	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 1.2
23	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 2
24	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 2
25	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 3
26	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 3
27	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 4
28	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 5
29	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 6
30	Migration efficiency of other moving groups whose household reference person is in NS-SEC category 6
31	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category 7
32	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Full Time Student
33	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Full Time Student
34	Migration efficiency of wholly moving households whose household reference person is in NS-SEC category Not Classified
35	Migration efficiency of other moving groups whose household reference person is in NS-SEC category Not Classified
36	Migration efficiency of wholly moving households moving into or from owner occupied accommodation
37	Migration efficiency of other moving groups moving into or from owner occupied accommodation
38	Migration efficiency of wholly moving households moving into or from privately rented accommodation
39	Migration efficiency of other moving groups moving into or from privately rented accommodation
40	Migration efficiency of individuals living alone
41	Migration efficiency of individuals not living in a family but with others in a household
42	Migration efficiency of individuals who are part of a couple family
43	Migration efficiency of individuals who are part of a lone parent family
44	Migration efficiency of individuals living in a communal establishment

2.4.1 A decision on *k*

This set of final variables then needed to be clustered using the *k*-means algorithm in MATLAB. Manhattan (City) distance was selected as the most appropriate distance measure (however Euclidean distance was also tested to compare the solutions produced). Where the most appropriate number of clusters was not known, *k*-means was run for a range of clusters from 2 to 14 each using 200 replicates to attain a global minimum. This range of clusters was

chosen as it was felt likely that the optimum solution would fall somewhere within this range. The initial classification had suggested 8 clusters were most appropriate, Vickers (2006) suggests that around 6 may be a useful place to start, whereas Shepherd (2006) tests a range between 5 and 25. A range somewhere around these numbers would in all probability produce the optimum solution. Silhouette and cluster size metrics were produced for each of the cluster solutions.

Figure 2.20 Average silhouette width values for solutions between 2 and 14 clusters

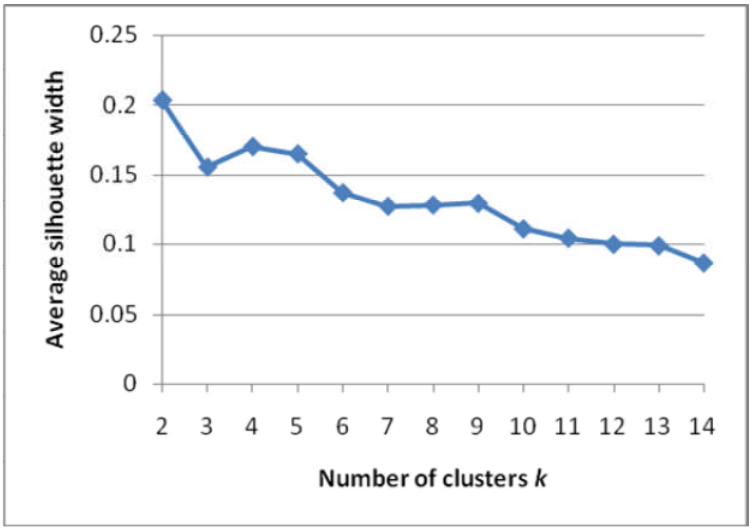
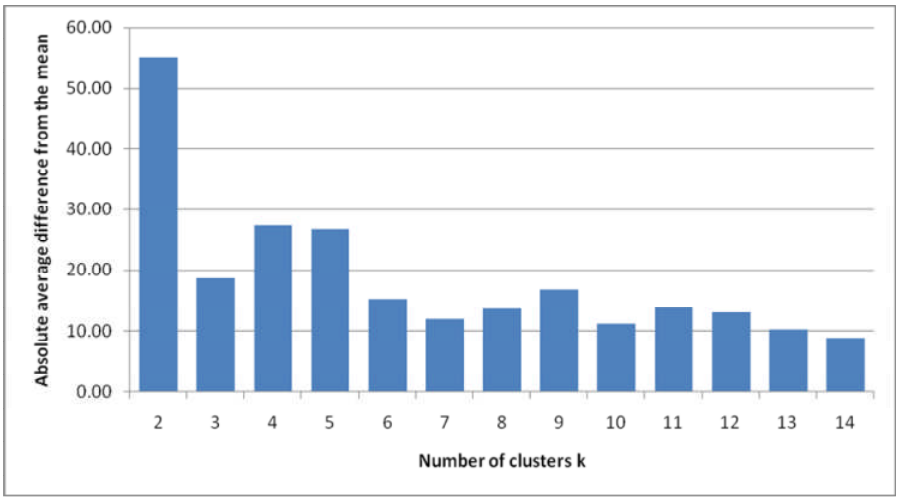


Figure 2.21 Absolute average difference from mean cluster size



Taking Figure 2.20 first, a better cluster solution will have an average value closer to 1 than 0. Whilst Kaufman and Rousseeuw (2005) state that an average silhouette width <0.25 represents a poor cluster, Shepherd (2006) successfully employs the technique to assess

cluster solutions with average values of around 0.1. It is evident that generally, as the number of clusters increases, the average silhouette value decreases, indicating at least for this metric, fewer clusters represent a more desirable solution. Taking Figure 2.21 also into consideration, however, a different conclusion might be reached. If we accept that more evenly sized groups are the most desirable outcome, then lower values in Figure 2.21 represent the better solution. Here clusters of 7, 8, 10, 13 and 14 groups could be candidates for selection.

Using both of these measures, it would appear that 3 clusters could be a good overall solution. A classification with only 3 clusters, however, is undesirable as fewer groups will represent much broader generalisations in the data. If somewhere between the 6 clusters suggested by Vickers (2006) and the 8 clusters suggested by the initial classification is aimed for, then Figure 2.21 suggests that solutions with either 7 or 8 clusters might be suitable as they both perform relatively well in both tests. With comparable scores in both metrics a decision between the two is a difficult one to make. The silhouette plots for both (Figure 2.22) enabling an assessment of the quality of each individual cluster are also very comparable, where perhaps if one solution featured a cluster with a large negative spike (representing a number of cases which could be very easily associated with another cluster) it would be a clear candidate for being dropped, here no such spikes are apparent, so additional data are required to assist the decision.

Figure 2.22 Silhouette widths for 7 and 8 cluster solutions – *k*-means, Manhattan distance, 200 replicates.

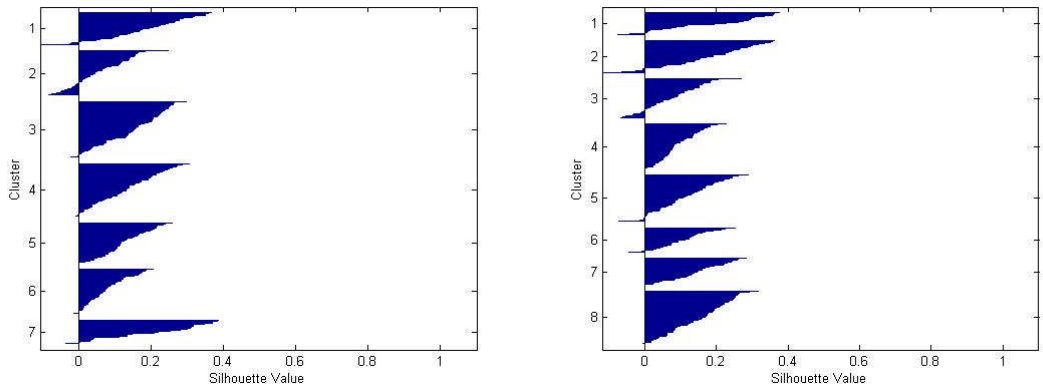


Table 2.10 – Summary of silhouette data for $k = 7$ and $k = 8$ cluster solutions

Cluster	Cases in cluster	Count Silhouette <0	Sum Silhouette <0	Avg Silhouette
1	73	1	-0.004	0.133
2	78	3	-0.028	0.146
3	45	4	-0.153	0.188
4	55	0	0.000	0.123
5	63	2	-0.033	0.080
6	33	2	-0.073	0.204
7	61	14	-0.433	0.065
7 cluster solution	408	26	-0.723	0.128
1	65	5	-0.066	0.119
2	39	4	-0.028	0.104
3	45	6	-0.238	0.172
4	37	1	-0.002	0.148
5	75	1	-0.004	0.152
6	53	8	-0.262	0.082
7	31	3	-0.037	0.197
8	63	0	0.000	0.098
8 cluster solution	408	28	-0.637	0.130

It could be argued that where cluster solutions have similar average silhouette values, as is the case here, the better solution would be the one with fewer values below 0. As Table 2.10 shows, whilst both cluster solutions have similar average silhouette values and similar counts of silhouette values below 0, the sum of the <0 silhouette values for 7 clusters is worse than it is for 8 clusters, indicating that where cases have weak associations with the clusters they have been assigned to, these weak associations are worse in a 7 cluster solution. Therefore taking all of this evidence into consideration – as well as the assertion by Milligan (1996 quoted in Shepherd, 2006) that where there is doubt, the higher figure should be taken – an 8 cluster solution will be chosen for the final classification.

As a postscript, it should be noted that this process was also carried out using Euclidean distance as the measure of distance between clusters. The main point of note, is that although the average silhouette widths were much higher when using Euclidean distance, the range of cluster sizes was also much higher – in some cases producing single case clusters with silhouette values of 1, or clusters with few cases, some heavily mis-specified (silhouette values very much in the negative). The work of Aggarwal *et al.* had already pointed to Manhattan distance providing better cluster solutions – the huge variation in cluster sizes and silhouette values using Euclidean distance confirms this.

2.4.2 The final cluster solution – an internal migration classification for Britain

A final *k*-means cluster run was carried out in MATLAB, this time using 1,000 replicates to ensure the final solution could be judged with certainty to be the best possible global cluster solution. As Figure 2.23 indicates, the final 1,000 replicates solution varies very little from the earlier 200 replicates solution in terms of the size and shape of the clusters. The clusters are in a different order to Figure 2.22, however the only small differences occur in the cases with values <0. This is not a surprise as these cases, by their very nature, could very well be assigned to other clusters.

Figure 2.23 Silhouette plot of final 8 cluster solution – *k*-means, 1,000 replicates, Manhattan distance.

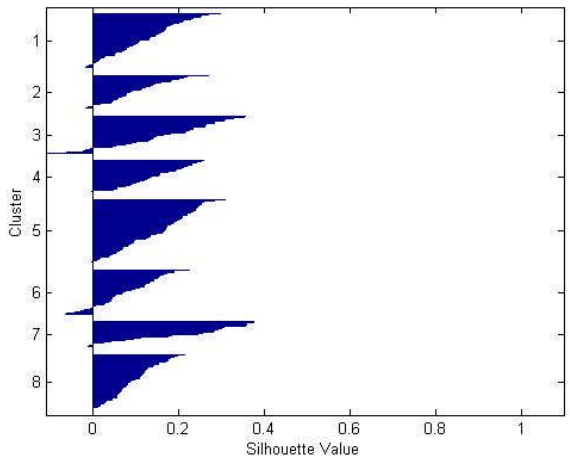


Figure 2.24 Internal Migration District Classification – 8 Cluster Solution

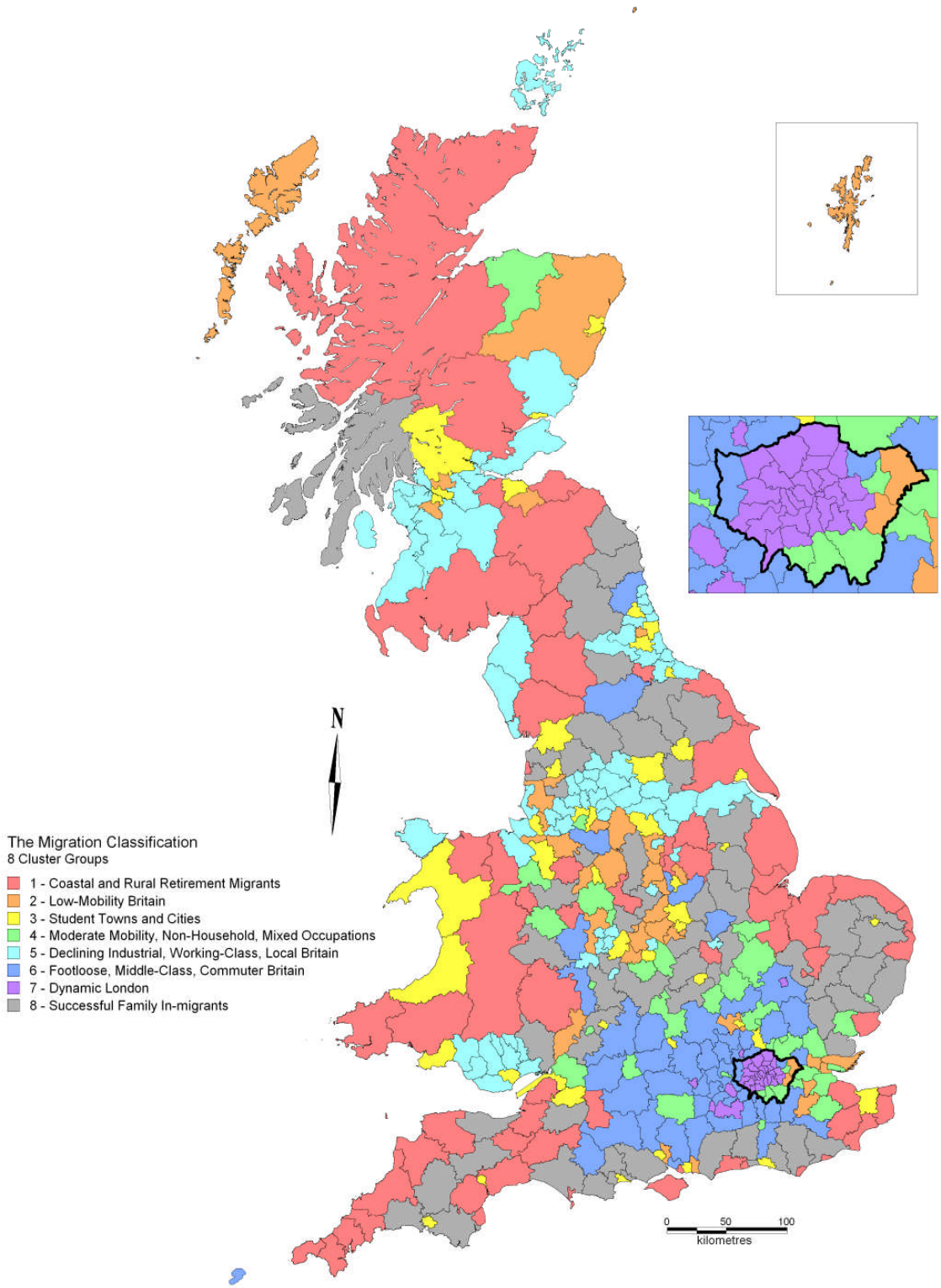


Figure 2.24 maps the final 8 clusters, revealing their spatial distribution across Britain. Although the map gives the impression that each area featured is a firm member of whichever cluster its shading corresponds to, this is a little misleading. Indeed, this is a problem that besets all classifications of this type (whether the end user is aware of the issue or not). The trouble is the degree of membership each district has with the cluster to which it belongs. The silhouette plot in Figure 2.23 shows clearly that each cluster features cases with a greater or lesser degree of membership. This means that where clusters have particular characteristics, the districts within will correspond with these characteristics to a greater or lesser degree depending on the silhouette value.

The potential limitations of the categorical nature of classifications are clear. If one is attempting to classify objects as either red or yellow, an orange object placed in either category will be incorrect to a greater or lesser extent. The solution in this situation would be that the orange object be given a degree of membership to either category – 60% red and 40% yellow for example. Applying this theory to cluster analysis are a family of techniques known as 'fuzzy cluster analysis'. Within this family various algorithms have been developed designed specifically to create 'fuzzy' partitions in data where 'hard' or 'crisp' partitions might unnecessarily constrain cases to particular clusters. Höppner *et al.* (1999) give a detailed overview of some of these including the 'fuzzy *c*-means' algorithm, which, given a number of clusters *c* to find in a dataset, will assign cases to clusters in a similar fashion to the *k*-means algorithm but with a membership grade or value determining the degree of membership to that cluster.

Whilst creating a set of fuzzy clusters is attractive as it avoids the rigid allocation that happens with hard clusters, one of the main aims of this exercise was to create a classification which could be used to analyse existing and future migration data. Analysis of flows between clearly defined areas is far more straightforward than the analysis of flows between areas with degrees of membership to a cluster (although the problem is not insurmountable). A parallel fuzzy classification will not be attempted here, although a degree of fuzziness can be added to the existing classification using the silhouette data already produced. The use of silhouettes in fuzzy clustering is advocated by Everitt *et al.* (2001) and can be easily applied to the rigid 8 cluster solution here so that cluster membership can be seen as stronger or weaker for each case in the cluster. As discussed earlier, the closer the silhouette value for a particular case is to 1, the more associated it is with that cluster; as values get close to 0 the

membership becomes more ambiguous; closer the -1 and the case could more easily be associated with another cluster (although which cluster is not apparent).

Figure 2.25 A ‘fuzzy’ representation of the internal migration district classification.

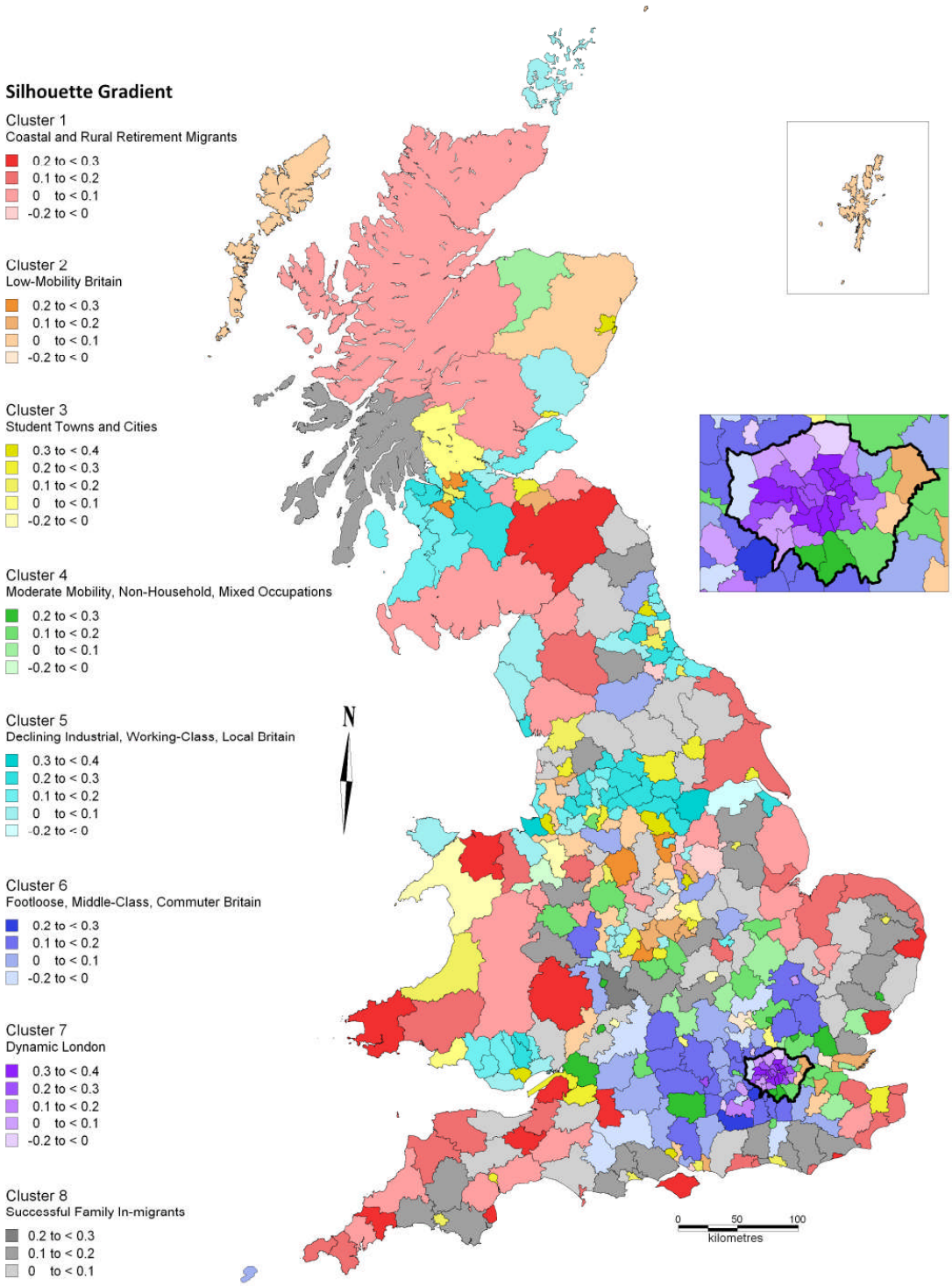


Figure 2.25 represents a fuzzy version of the more rigid classification in Figure 2.24. Here the strength of membership is represented by the heaviness of the shading, with more heavily shaded areas having a stronger association with that particular cluster. Details of the silhouette values associated with each case in each cluster, as well as the profiles of each of the clusters in the classification will be presented in the next section.

2.5 Cluster Profiles

The following section outlines the constituent districts of each of the 8 clusters. The key variables defining each cluster are identified by the bars in the associated charts representing the *z*-scores for each variable. By taking the average *z*-score value for each variable across the districts comprising each cluster it is possible to ascertain which variables are more and less important within the cluster. The first graph in each cluster portrait contains *z*-scores for in, out and within area migration rates and should be interpreted with scores >0 showing over-representation of a variable in this cluster, and scores <0, under representation. Larger bars equal greater under or over representation. The second graph in each portrait contains *z*-scores for migration efficiency rates for moving groups. This graph should be interpreted differently as efficiency rates are directional. So for this graph, a value of 0 means that in/out migration is in balance. A value of >0 represents in-migration for a given variable, and a value of <0 represents out-migration. Larger bars equate to a greater intensity of movement.

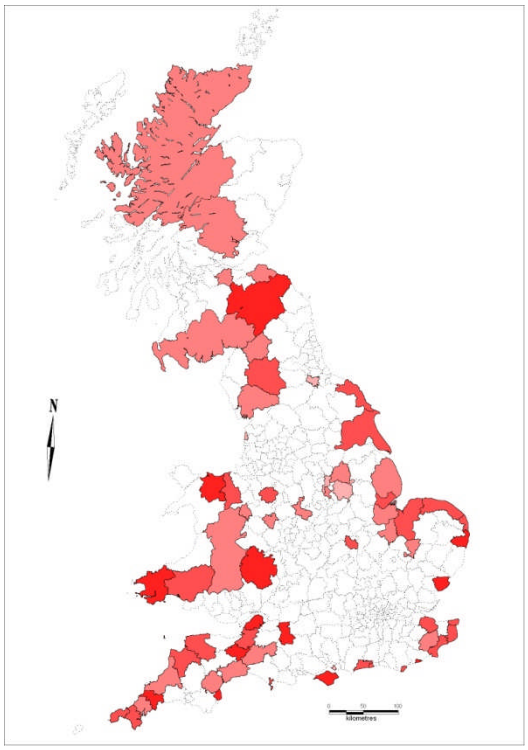
It was decided that given the distinct profile of each cluster, a representative name should be chosen. A name which summarises the key features of a cluster will aid in its identification and differentiation from others – something which is important if the typology is to be used in additional analysis. The obvious issue with giving a cluster, characterised by variation across a range of variables, a name, is that any short name is likely to be a subjective generalisation. As Vickers (2006) points out, naming of clusters in geodemographic classifications is often a very contentious issue and can be open to much criticism. In the development of the Output Area Classification for the Office of National Statistics – a classification which has the status of a 'National Statistic' (http://www.statisticsauthority.gov.uk/national-statistician/types-of-official-statistics) and thus is bound by an official code of practice, Vickers (2006) embarked upon an extensive quality assurance exercise through consultation with a range of stakeholders and experts.

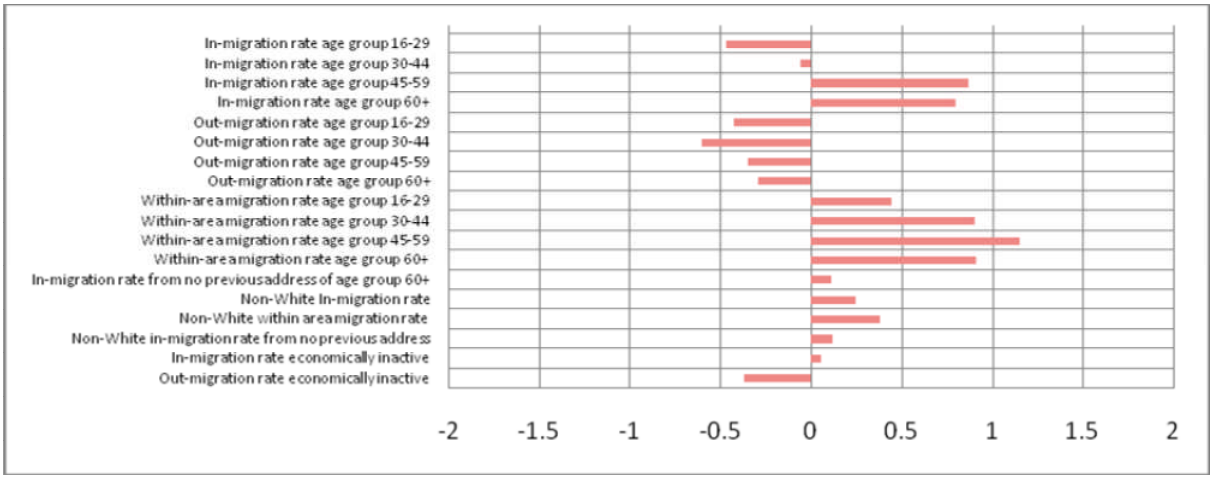
As the Migration Classification is not governed by such statutes a slightly exhaustive consultation exercise was undertaken, although that is not to downplay the importance of the exercise which was undertaken. An initial set of names for each cluster were decided upon using the information contained in these cluster profiles. This provisional set of cluster names was presented to a delegation of academics and other experts at an ESRC Census Programme workshop on Social and Spatial Classification (http://www.esds.ac.uk/news/eventdetail.asp?id=2455) where the cluster profiles below were presented along with their provisional names. Delegates were invited to critique existing names and offer alternative suggestions. At the end of the workshop, documents were collected and some minor alterations were made to the cluster names in the light

of the suggestions made. The names presented below are the combined efforts of judgements made originally by this author and suggestions made during the quality assurance exercise.

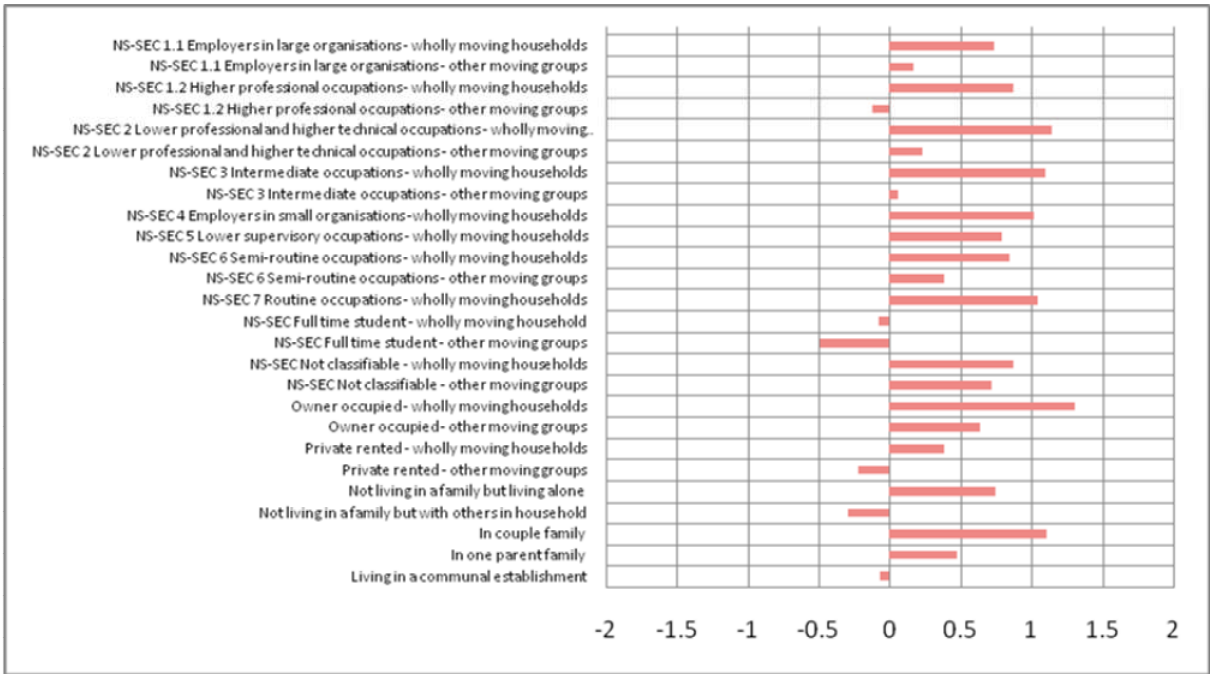
2.5.1 Cluster 1: Coastal and Rural Retirement Migrants

Cluster 1 is dominated by coastal and rural areas, particularly in the South-West, Kent, Norfolk, the South Coast, Wales and Scottish Borders and Highlands. The Isle of Wight is the district most representative of this cluster, with Blackpool the district most unrepresentative. The cluster is characterised by in-migrants and within-area migrants in the older age groups – 45 and above. Younger in-migrants are very much underrepresented. Migrants into these areas are from across the socio-economic spectrum, although the very high socio-economic groups are less common. Migrants preferentially move into owner occupied accommodation, and tend to be either or alone or in couples, far more than parent families.





2.5 Cluster Profiles



Cluster 1 contains 65 districts:

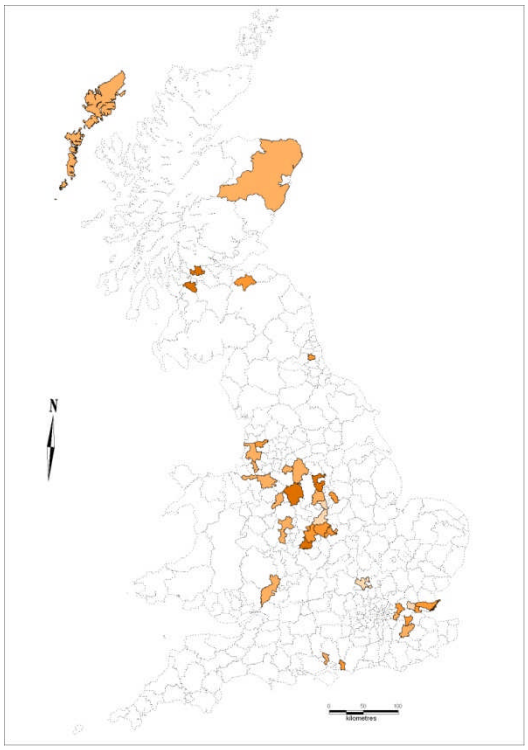
District	**Silhouette Value**
Isle of Wight	0.297
Conwy	0.276
Torbay	0.270
Herefordshire County	0.253
Waveney	0.252
Tendring	0.249
Scottish Borders	0.234
Taunton Deane	0.230
North Somerset	0.228
Hastings	0.212
Pembrokeshire	0.208
Restormel	0.202
West Wiltshire	0.201
Eastbourne	0.188
Penwith	0.182
Torridge	0.177
Thanet	0.173
Great Yarmouth	0.171
Kettering	0.171
Denbighshire	0.171
Weymouth and Portland	0.167
Shepway	0.164

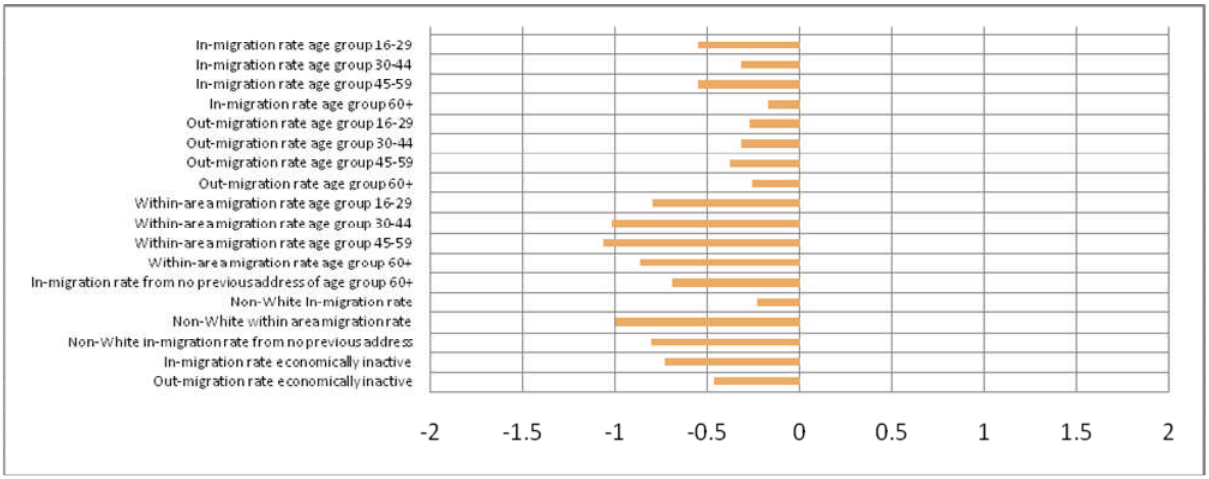
Worthing	0.162
Scarborough	0.160
Arun	0.153
Crewe and Nantwich	0.151
Kings Lynn and West Norfolk	0.148
North Devon	0.146
Carmarthenshire	0.135
Eden	0.132
Kerrier	0.131
Swale	0.131
Boston	0.114
Sedgemoor	0.109
Dover	0.106
East Riding of Yorkshire	0.103
North Norfolk	0.102
South Holland	0.097
North Cornwall	0.092
Oswestry	0.091
Powys	0.085
Bassetlaw	0.079
Carrick	0.078
Ashfield	0.076

Telford and Wrekin	0.073
Ashford	0.067
East Devon	0.053
East Staffordshire	0.052
East Lindsey	0.050
Fenland	0.050
Highland	0.045
West Lothian	0.028
Perth & Kinross	0.025
Forest Heath	0.025
Dumfries & Galloway	0.017
Teignbridge	0.013
South Lakeland	0.012
South Somerset	0.011
East Lothian	0.006
Carlisle	0.000
Darlington	-0.006
Bolsover	-0.007
Newark and Sherwood	-0.014
Gosport	-0.016
Blackpool	-0.023

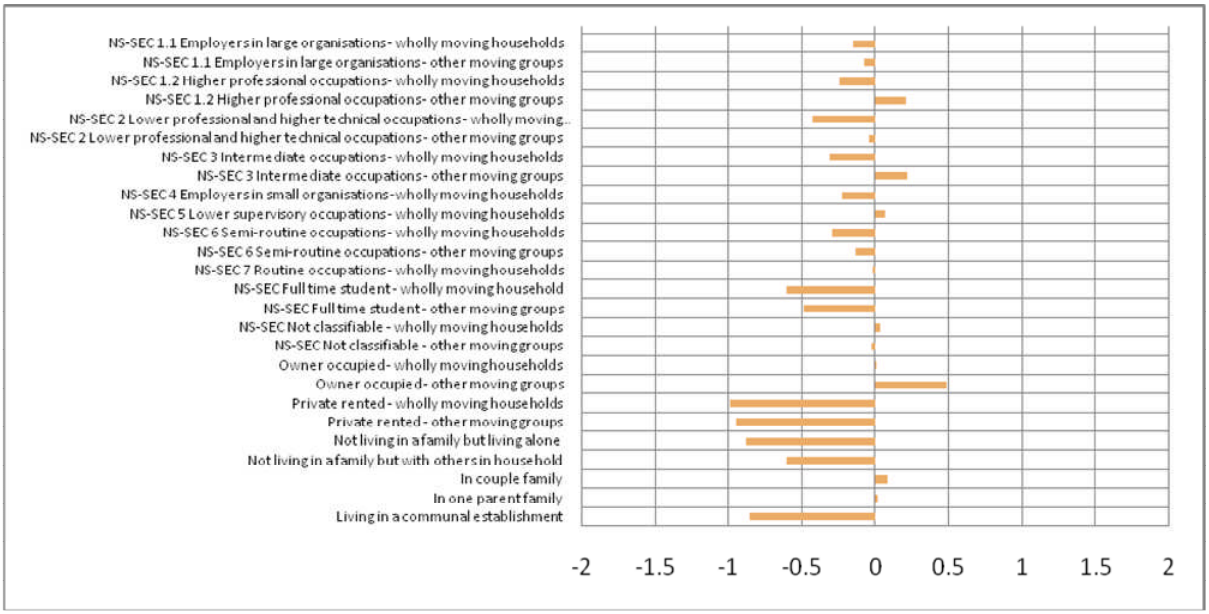
2.5.2 Cluster 2: Low-Mobility Britain

Cluster 2 is spread around Britain, although small concentrations exist in the Midlands moving into south Merseyside, and to the south and east of London. North East Derbyshire is the most representative district in this cluster, with Erewash and South Bedfordshire most likely to be misclassified. The cluster is characterised by very little internal migration activity, with in-migration and out-migration under-represented across all age groups. Within-area migration is particularly under-represented. Where in-migration does occur it tends to be into owner occupied housing and by migrants in slightly higher socio-economic groups.





2.5 Cluster Profiles



Cluster 2 contains 39 districts:

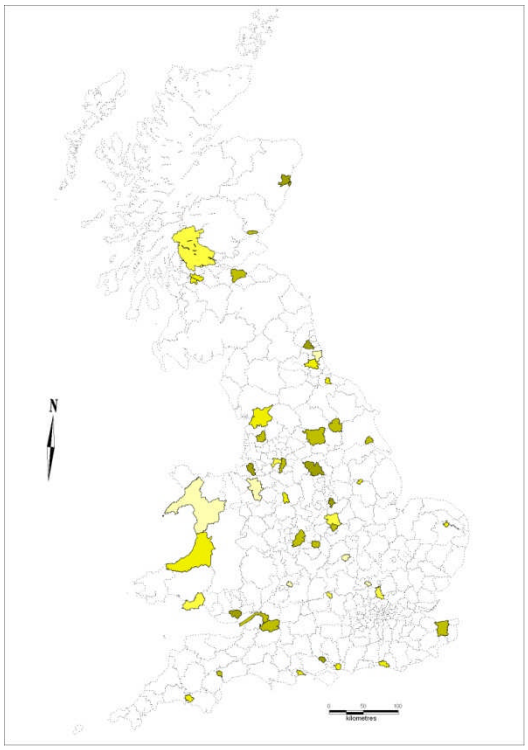
District	**Silhouette Value**
North East Derbyshire	0.270
East Dunbartonshire	0.225
Solihull	0.219
East Renfrewshire	0.212
Staffordshire Moorland	0.201
Hinckley and Bosworth	0.186
Havering	0.185
Castle Point	0.180
Gedling	0.170
Midlothian	0.158
South Ribble	0.144
Chester-le-Street	0.139

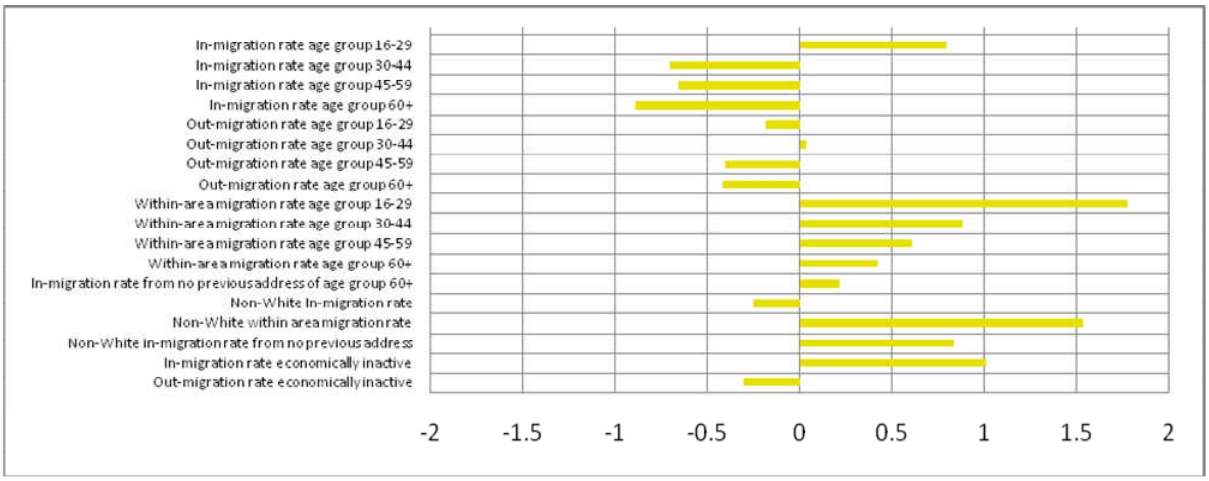
Gravesham	0.134
Tamworth	0.133
Rochford	0.122
Blaby	0.121
North Warwickshire	0.115
Havant	0.111
Newcastle-under-Lyme	0.092
Knowsley	0.086
South Staffordshire	0.086
Eilean Siar	0.079
Aberdeenshire	0.078
Oadby and Wigston	0.077
Amber Valley	0.076
Stockport	0.068

West Lancashire	0.067
Tonbridge and Malling	0.060
Ellesmere Port and Nes	0.057
Forest of Dean	0.054
Bexley	0.048
High Peak	0.046
Shetland Islands	0.041
Eastleigh	0.038
Vale Royal	0.016
Basildon	-0.001
North West Leicestershire	-0.002
South Bedfordshire	-0.011
Erewash	-0.015

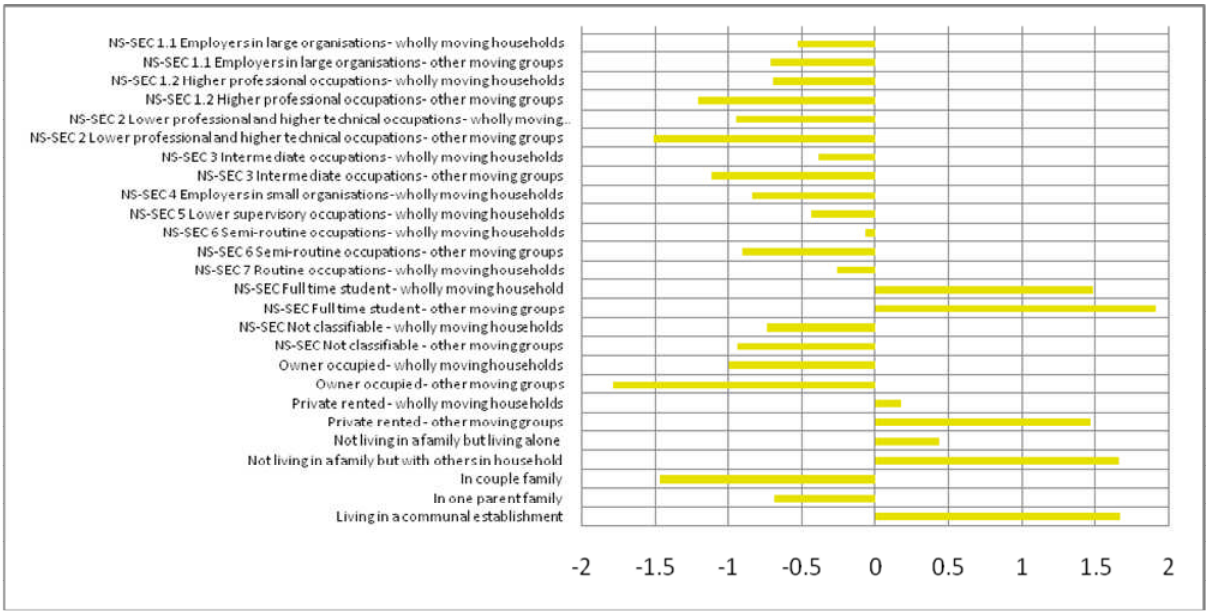
2.5.3 Cluster 3: Student Towns and Cities

Cluster 3 is comprised principally of larger towns and cities housing universities and higher education institutions. Newcastle Upon Tyne is the most representative district, with Luton being the least representative. Despite a strong average silhouette value, signifying a well defined cluster, 6 districts including Luton have very weak associations. The cluster is characterised by high levels of student in-migration, and young person within-area migration. Non-household moving groups into privately rented accommodation are common in this cluster, as are non-family households and individuals moving into communal establishments – all characteristics of a student population. In addition, non-white within-area migration is important, as is in-migration of economically inactive migrants.





2.5 Cluster Profiles



Cluster 3 contains 45 districts:

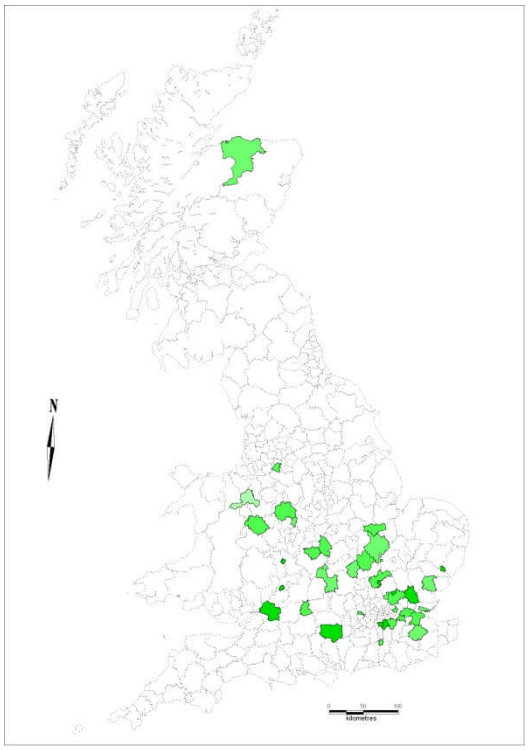
District	**Silhouette Value**
Newcastle upon Tyne	0.365
Sheffield	0.357
Cardiff	0.355
Nottingham	0.349
Southampton	0.329
Liverpool	0.310
Aberdeen City	0.306
Kingston upon Hull	0.294
Dundee City	0.285
Leicester	0.284
Leeds	0.271
York	0.264
Canterbury	0.261
Bristol	0.261

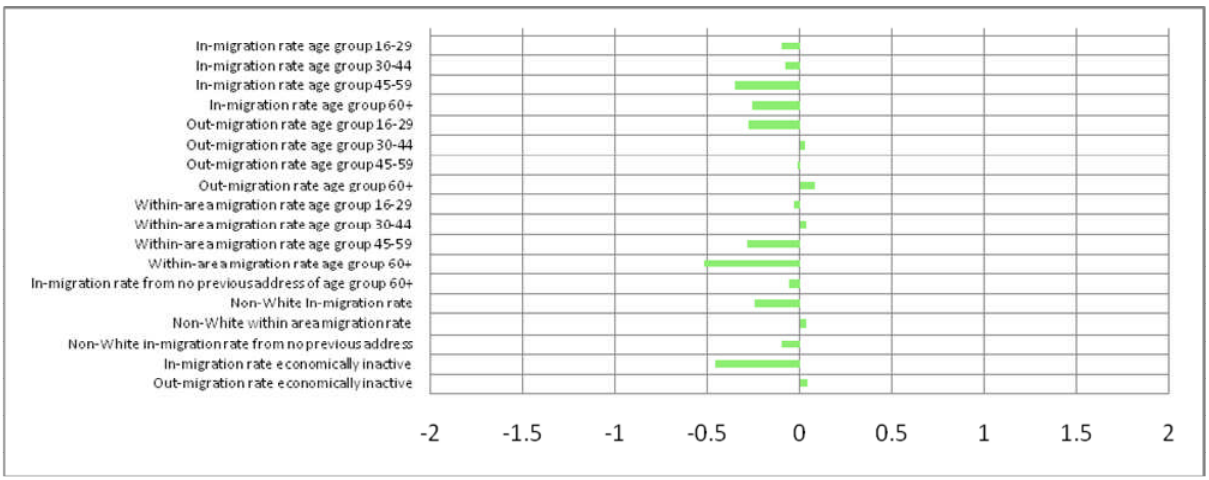
Preston	0.254
Manchester	0.235
Coventry	0.233
Exeter	0.226
Edinburgh	0.225
Birmingham	0.218
Bath and North East Somerset	0.214
Ceredigion	0.190
Glasgow City	0.180
Middlesbrough	0.165
Plymouth	0.151
Brighton and Hove	0.148
Lancaster	0.143
Durham	0.140
Portsmouth	0.137

Lincoln	0.135
Norwich	0.128
Bournemouth	0.104
Stoke-on-Trent	0.101
Swansea	0.090
Charnwood	0.077
Stirling	0.074
Oxford	0.052
Salford	0.046
Welwyn Hatfield	0.002
Gwynedd	-0.005
Sunderland	-0.016
Northampton	-0.022
Cheltenham	-0.026
Chester	-0.060
Luton	-0.109

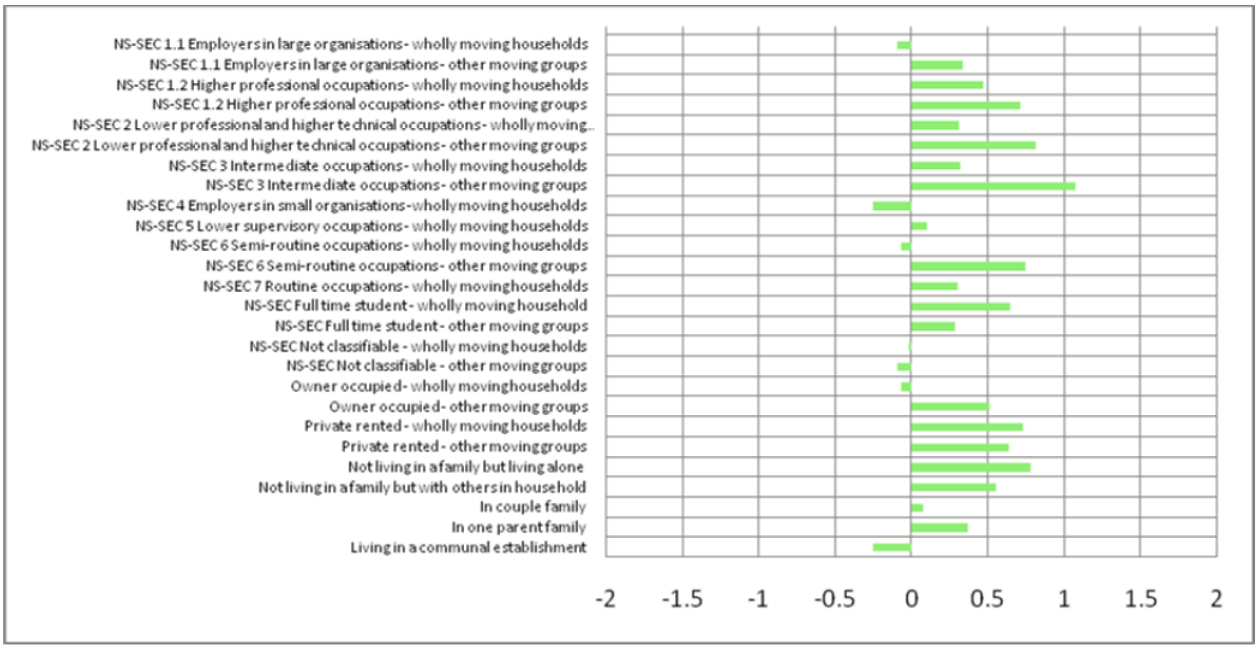
2.5.4 Cluster 4: Moderate Mobility, Non-Household, Mixed Occupations

Cluster 4 is the second smallest cluster, with districts tending to be found in the south and Midlands. Whilst small, it is reasonably well defined, with only Wrexham having a noticeably ambiguous membership. Croydon is the district with the characteristics most representing this cluster. The cluster features relatively low levels of migration in general, however migrants moving into these areas are more likely to be engaged in intermediate occupations. Migrants who move alone or in non-family households are also more common in areas in this cluster. Wholly moving households and owner occupiers are less common.





2.5 Cluster Profiles



Cluster 4 contains 37 districts:

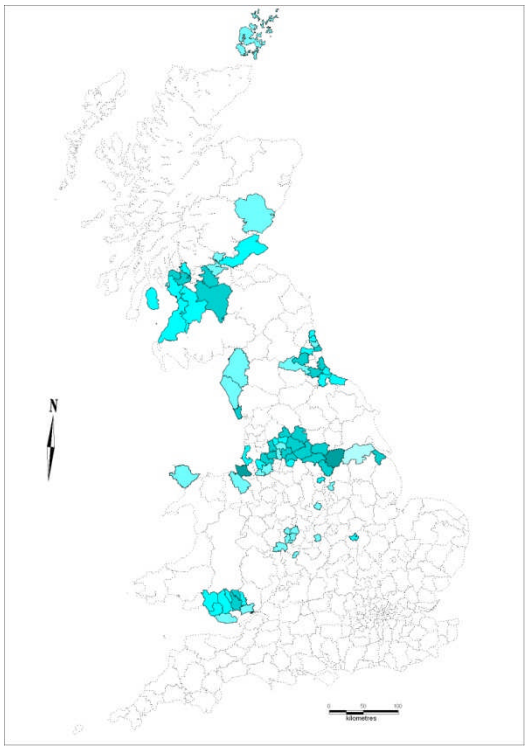
District	**Silhouette Value**
Croydon	0.260
Gloucester	0.253
Sutton	0.242
Stevenage	0.231
Harlow	0.231
Worcester	0.230
South Gloucestershire	0.227
Ipswich	0.221
Chelmsford	0.212
Basingstoke and Deane	0.205
Bromley	0.198

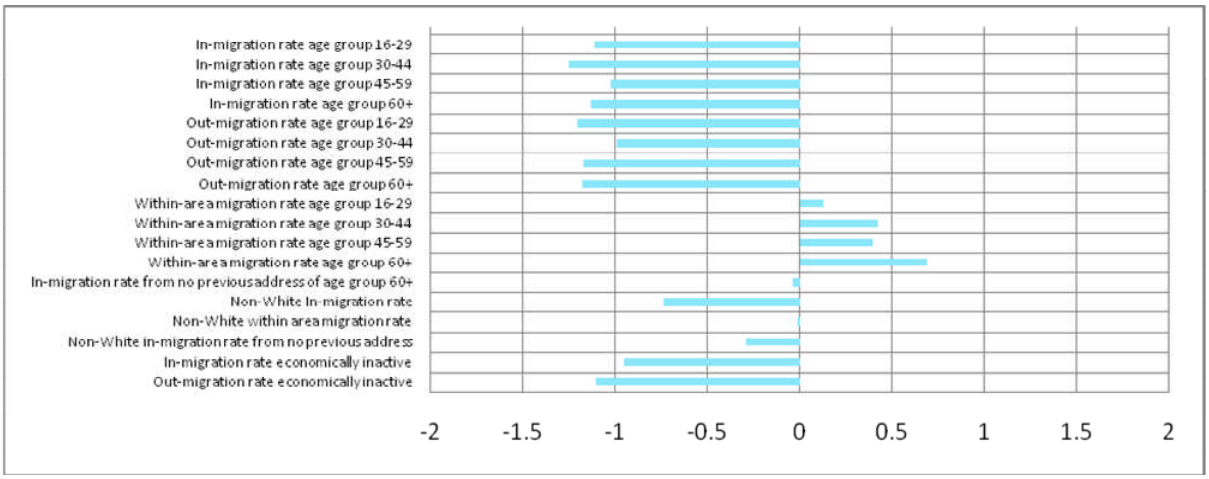
Thurrock	0.195
Peterborough	0.191
Bedford	0.187
Warwick	0.176
Medway Towns	0.157
Stafford	0.157
Barking and Dagenham	0.157
Shrewsbury and Atcham	0.152
Dartford	0.138
Swindon	0.136
Slough	0.130
Trafford	0.126
North Hertfordshire	0.120

Epping Forest	0.120
Crawley	0.109
Rugby	0.101
Southend-on-Sea	0.100
Maidstone	0.092
Broxbourne	0.087
Cherwell	0.076
Milton Keynes	0.063
Cannock Chase	0.057
Huntingdonshire	0.056
Colchester	0.053
Moray	0.048
Wrexham	-0.002

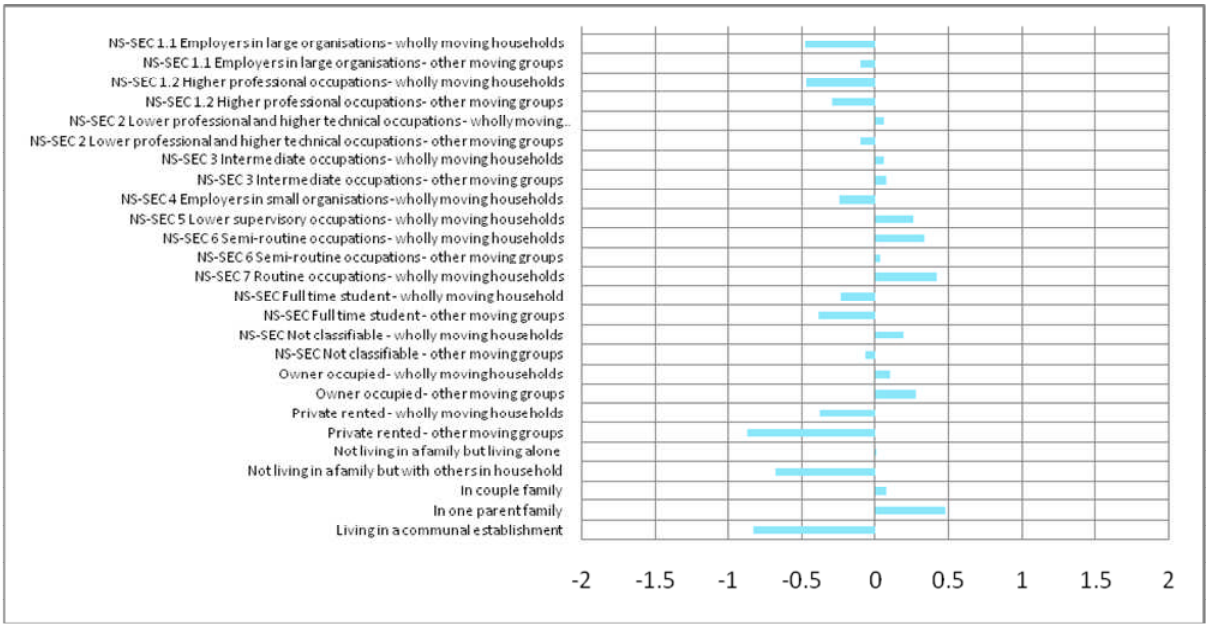
2.5.5 Cluster 5: Declining Industrial, Working-Class, Local Britain

Cluster 5 is the largest cluster and is concentrated mainly around the ex-industrial areas of South Wales, Yorkshire, Greater Manchester and Lancashire, the North-East and Scotland. The cluster is also well defined with only North Lincolnshire having a silhouette value lower than 0. Districts in this cluster have very much below average in-migration and out-migration for all age groups, signifying a degree of isolation from the rest of the clusters in Britain. Shorter distance, within-area migration is slightly above average. Moves into these areas come from individuals in the lower socio-economic groups, with moves of one-parent families being above average. Moves of economically inactive individuals, however, are very much below average.





2.5 Cluster Profiles

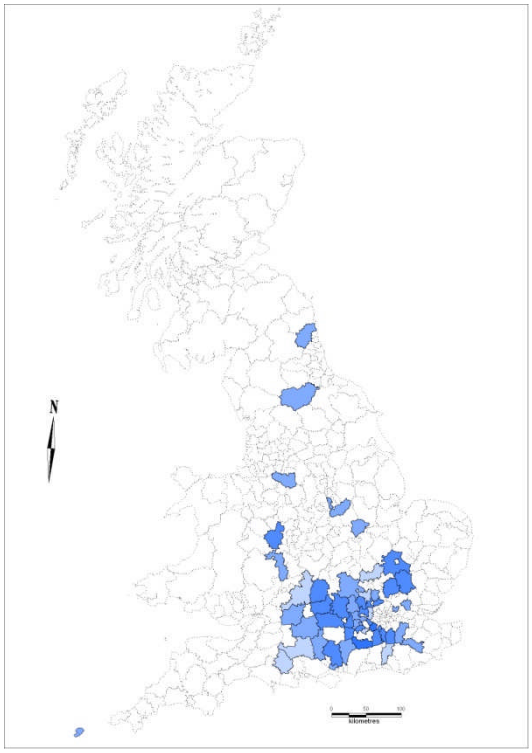


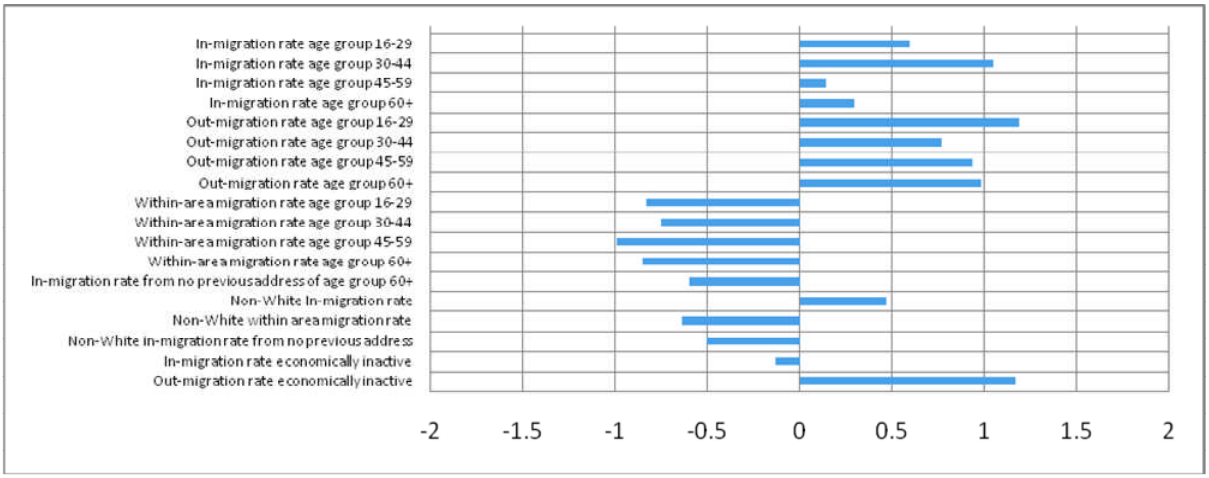
Cluster 5 contains 75 districts:

District	Silhouette Value
Doncaster	0.324
Wirral	0.310
Bolton	0.283
Rochdale	0.266
Bradford	0.260
South Tyneside	0.254
Barnsley	0.253
Blaenau Gwent	0.252
Rotherham	0.249
Inverclyde	0.246
Renfrewshire	0.245
Wigan	0.244
Barrow-in-Furness	0.240
Caerphilly	0.238
Oldham	0.238
Derwentside	0.234
South Lanarkshire	0.229
Blackburn with Darwen	0.222
Wakefield	0.220
Pendle	0.219
Easington	0.210
Sedgefield	0.210
Kirklees	0.209
North Lanarkshire	0.206
Calderdale	0.203
North East Lincolnshire	0.201
East Ayrshire	0.195
Burnley	0.190
West Dunbartonshire	0.189
Hyndburn	0.187
Bridgend	0.181
Rhondda Cynon Taff	0.178
Redcar and Cleveland	0.174
Neath Port Talbot	0.172
Wansbeck	0.172
Stockton-on-Tees	0.171
Torfaen	0.170
Tameside	0.167
St. Helens	0.167
North Ayrshire	0.165
South Ayrshire	0.161
Sefton	0.150
Blyth Valley	0.146
Hartlepool	0.134
Fife	0.133
Corby	0.129
Gateshead	0.129
Merthyr Tydfil	0.122
Chesterfield	0.108
Falkirk	0.098
Dudley	0.096
Angus	0.094
Allerdale	0.092
Mansfield	0.089
Wolverhampton	0.084
Copeland	0.083
Walsall	0.082
Flintshire	0.071
Bury	0.069
Wyre Forest	0.068
Nuneaton and Bedworth	0.066
Newport	0.061
Sandwell	0.060
Redditch	0.056
Isle of Anglesey	0.054
Orkney Islands	0.045
The Vale of Glamorgan	0.044
Halton	0.040
North Tyneside	0.036
Rossendale	0.026
Warrington	0.016
Clackmannanshire	0.013
Wear Valley	0.009
Derby	0.005
North Lincolnshire	-0.004

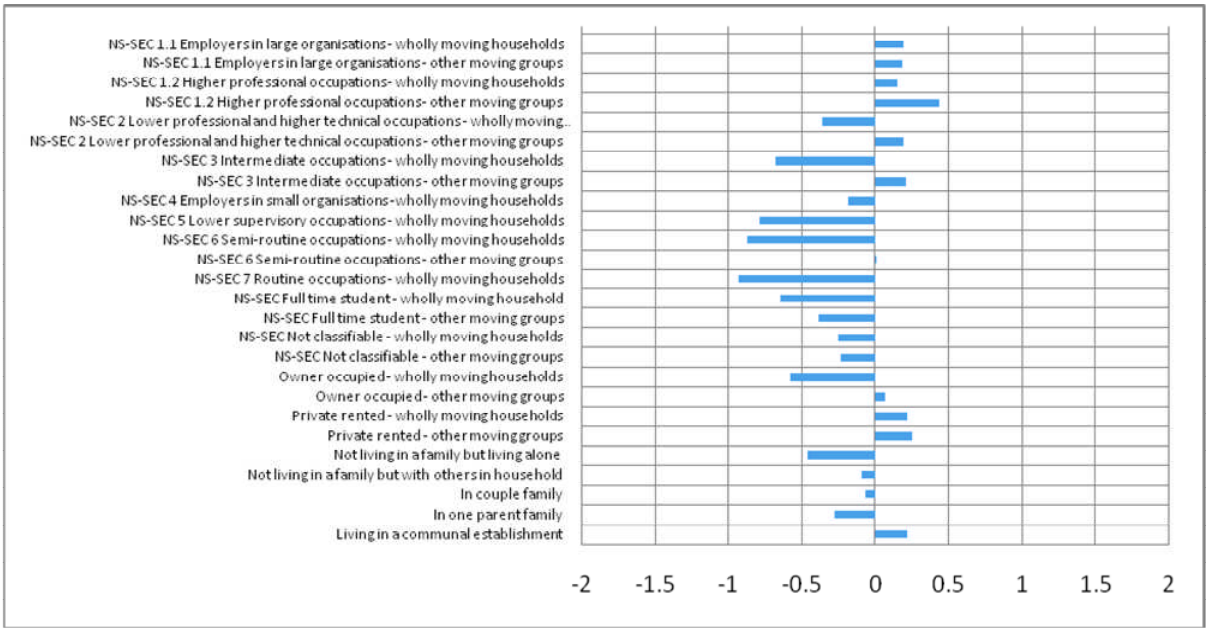
2.5.6 Cluster 6: Footloose, Middle-Class, Commuter Britain

Cluster 6 is the most poorly defined cluster, with the lowest average silhouette value. 8 districts out of the 58 have silhouette values lower than 0. The most representative district of the cluster is Waverley. In general, districts in this are concentrated just outside of London in the Home Counties and heading out west along the M3/M4 corridor. This cluster is characterised by higher rates of in and out-migration, particularly in the below 30 age groups. Within area migration is very much below average. Out-migration rates of economically inactive individuals are much higher than average. Migration efficiency rates are very negative for lower socio-economic groups, but positive for those in higher groups.





2.5 Cluster Profiles



Cluster 6 contains 53 districts:

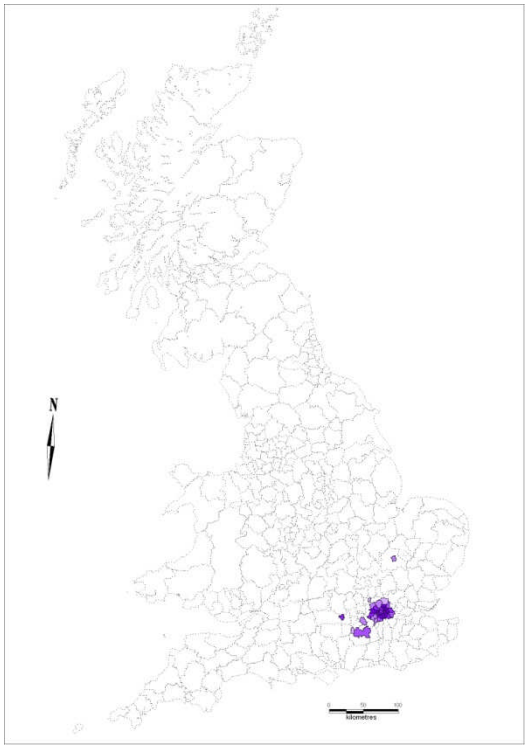
District	**Silhouette Value**
Waverley	0.249
Elmbridge	0.225
Hart	0.190
Wokingham	0.190
South Bucks	0.182
Epsom and Ewell	0.174
Winchester	0.164
Mole Valley	0.164
Uttlesford	0.155
Surrey Heath	0.153
Spelthorne	0.151
South Oxfordshire	0.148
Hertsmere	0.143
Tandridge	0.143
Vale of White Horse	0.142
Reigate and Banstead	0.133
Bridgnorth	0.131

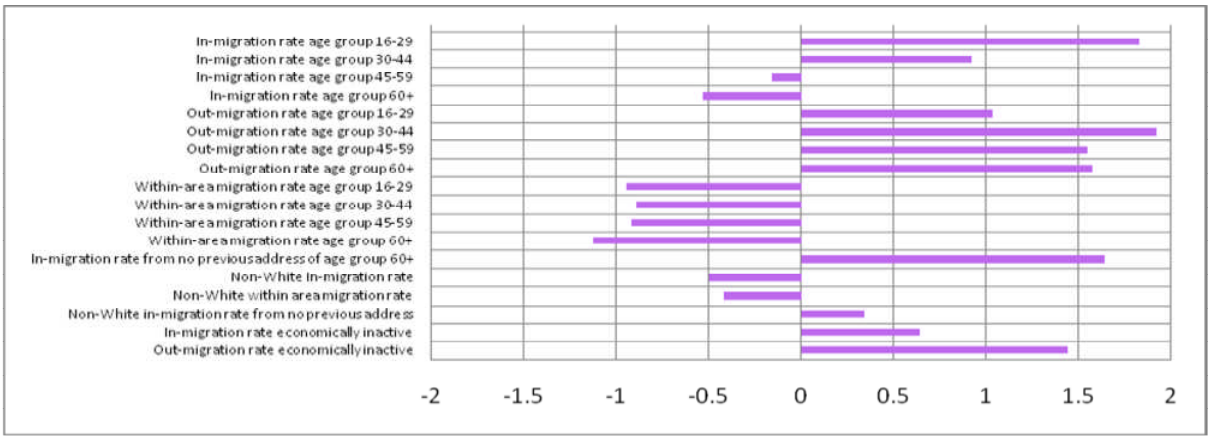
West Berkshire	0.123
South Cambridgeshire	0.121
West Oxfordshire	0.120
Chiltern	0.120
East Hertfordshire	0.112
Three Rivers	0.110
Brentwood	0.098
St. Albans	0.089
Windsor and Maidenhead	0.076
Test Valley	0.074
Malvern Hills	0.070
East Hampshire	0.069
Kennet	0.066
Rushcliffe	0.054
Castle Morpeth	0.051
Wycombe	0.048
Tunbridge Wells	0.044
Sevenoaks	0.044

Bracknell Forest	0.044
Broxtowe	0.043
Redbridge	0.042
Richmondshire	0.037
Aylesbury Vale	0.034
Isles of Scilly	0.027
North Wiltshire	0.025
Macclesfield	0.012
Rutland	0.006
Fareham	0.001
Woking	-0.005
Mid Sussex	-0.014
Cotswold	-0.018
Dacorum	-0.028
Mid Bedfordshire	-0.031
North Dorset	-0.042
Hillingdon	-0.061
Salisbury	-0.063

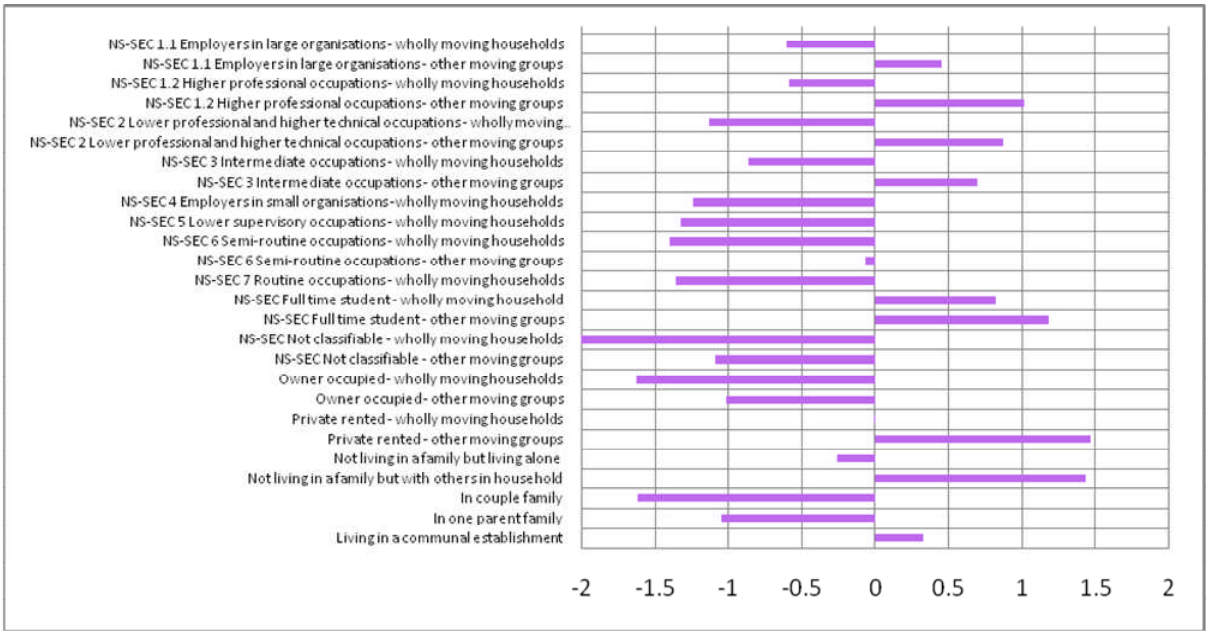
2.5.7 Cluster 7: Dynamic London

Cluster 7 is almost entirely concentrated within Greater London – only Cambridge, Reading, Guildford and Runnymede fall outside the M25. The cluster is defined by some of the highest and lowest *z*-score values, indicating it is the most dynamic cluster in the classification. It features very high rates of in-migration for the migrants under 30, but below average in-migration rates for those over 30. Out-migration rates are very high for all groups, but especially for those between 30 and 45. Within area migration rates are much below average, as are those of non-whites (except those with no-previous address). This cluster features the highest rates of movement of the economically inactive. Across the four highest socio-economic groups there are positive efficiency rates for other moving groups, but negative rates for wholly moving households, indicating if whole households move, they leave these areas, whereas non-households individuals tend to move in – especially into privately rented accommodation. Students are also an important group of in-migrants to this cluster. Families (both couples and single parents) are noticeably moving out of this cluster.





2.5 Cluster Profiles



Cluster 7 contains 31 districts:

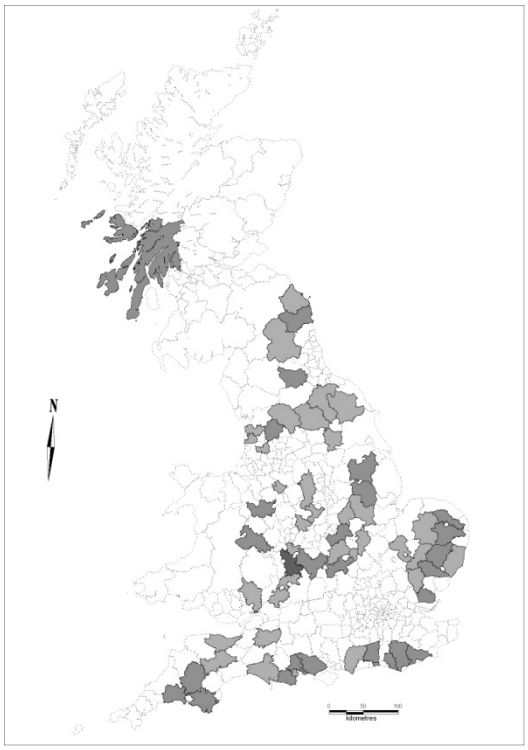
District	**Silhouette Value**
Islington	0.376
Haringey	0.375
Hammersmith and Fulham	0.363
Southwark	0.358
Wandsworth	0.358
Lambeth	0.357
Hackney	0.329
Tower Hamlets	0.307
Ealing	0.301

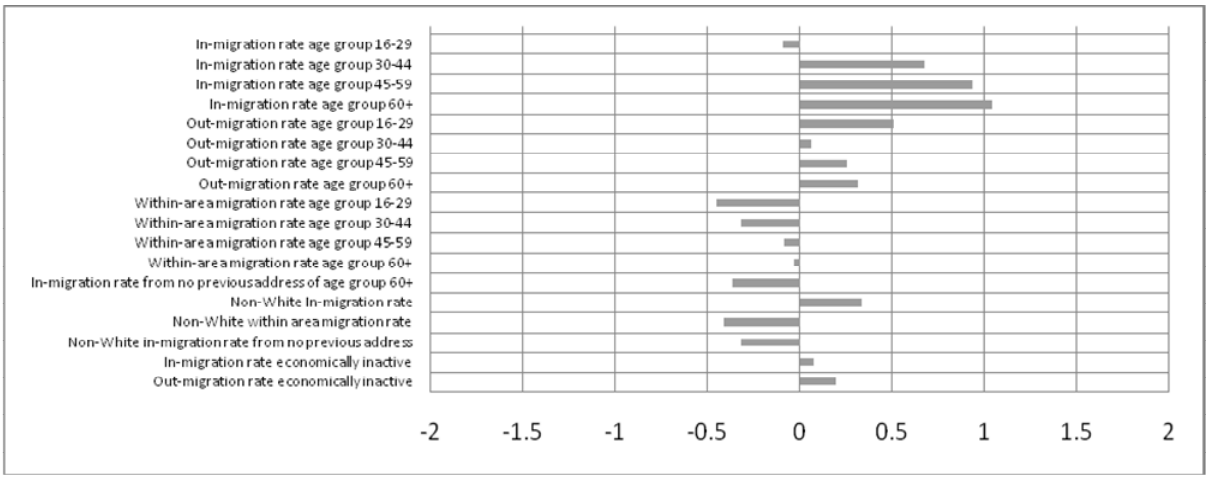
Camden	0.294
Newham	0.292
Lewisham	0.285
Brent	0.281
Hounslow	0.263
Kensington and Chelsea	0.256
Reading	0.245
Westminster	0.226
Merton	0.183
Greenwich	0.177
Waltham Forest	0.111
Guildford	0.107

Kingston upon Thames	0.094
Richmond upon Thames	0.075
Barnet	0.050
Cambridge	0.041
City of London	0.025
Harrow	0.000
Runnymede	0.000
Rushmoor	-0.008
Watford	-0.011
Enfield	-0.018

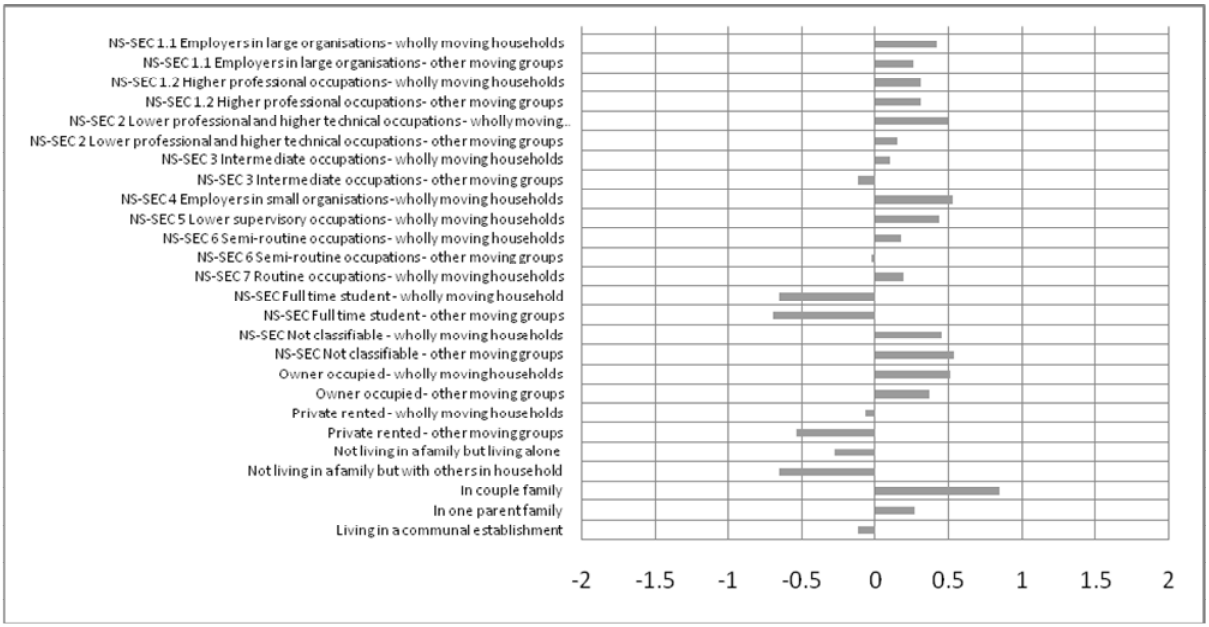
2.5.8 Cluster 8: Successful Family In-migrants

Cluster 8 is located mainly in rural areas of England, and is distributed quite evenly across the country. The cluster is relatively well defined as no districts have negative silhouette values. Wychavon is the most representative district in this cluster. In-migration of all age groups above 30 is above average, with importance increasing as age increases. Within-area migration is less significant. There are positive, in-migration balances across all socio-economic groups, although there are noticeable out-migrations of students from this cluster. When migrants move into this cluster, the preference is to move into owner occupied accommodation, with couple families being more important than any other group.





2.5 Cluster Profiles



Cluster 8 contains 63 districts:

District	**Silhouette Value**
Wychavon	0.215
Alnwick	0.198
West Lindsey	0.197
South Norfolk	0.189
Harborough	0.181
Christchurch	0.178
South Hams	0.176
South Shropshire	0.174
East Dorset	0.171
West Devon	0.159
Mid Suffolk	0.154
Babergh	0.153
Argyll & Bute	0.145
Wealden	0.142
Rother	0.139
Lewes	0.130
Caradon	0.128
Broadland	0.127
New Forest	0.124
Purbeck	0.121

North Kesteven	0.120
Ribble Valley	0.120
Teesdale	0.118
South Northamptonshire	0.118
Stratford-on-Avon	0.114
Adur	0.112
North Shropshire	0.110
Horsham	0.110
Maldon	0.108
Bromsgrove	0.099
Craven	0.096
Monmouthshire	0.090
West Somerset	0.086
East Cambridgeshire	0.082
Chichester	0.082
Hambleton	0.081
West Dorset	0.078
Daventry	0.078
St. Edmundsbury	0.075
Breckland	0.070
Ryedale	0.069

Fylde	0.068
Tynedale	0.066
Congleton	0.064
Braintree	0.062
Poole	0.062
South Kesteven	0.060
Tewkesbury	0.059
Berwick-upon-Tweed	0.057
Derbyshire Dales	0.056
Wellingborough	0.056
Chorley	0.050
Mendip	0.045
Suffolk Coastal	0.044
Lichfield	0.033
Wyre	0.031
Selby	0.028
Mid Devon	0.025
East Northamptonshire	0.020
Stroud	0.016
South Derbyshire	0.016
Melton	0.015
Harrogate	0.011

2.6 Classification Evaluation and Comparison

Now that a final classification of districts has been achieved, the last stage in the classification building process is to evaluate the final solution. There are a number of ways in which the results of a classification can be examined and tested, all dependent on exactly what the classification is being tested for. For example, it may be desirable to test that the partitions created are robust and represent actual clusters rather than being artefacts of particular algorithms – as in the example earlier where the *k*-means algorithm as implemented in SPSS could produce completely different classifications from the same data, merely through re-ordering the objects being clustered. Alternatively the test may be to see whether a classification is comparable or different to an existing classification – for example, that a Migration Classification offers something different to a general purpose classification. Or it may be that one wants to test a classification to see whether it is more successful at predicting behaviours than chance – particularly useful in marketing contexts where classifications are used for customer targeting. It could be that the classification needs to be tested to confirm that the variables selected were indeed most appropriate and important for the final solution. Or perhaps it may be that one wishes to assess how well the classification represents the real-world.

All of these reasons for testing classifications are valid in particular contexts, however, it may not be necessary or appropriate to test for all (or indeed any) of them all of the time. So the question that arises is what is the most suitable way of assessing the Migration Classification? The process of variable and algorithm selection was very thorough in this classification, as already described. The *k*-means algorithm as it was implemented in Matlab means that there can be confidence that the partitions created are robust, given the variables selected for inclusion. Vickers (2006), in creating the OA Classification from the 2001 Census, employed several different techniques to assess and 'quality assure' his classification. One of these methods to which much time was devoted was sensitivity analysis. In sensitivity analysis, variables are selectively removed from the classification and the algorithm run again. Through examining the change in the average distance to the cluster centre for objects in each cluster, an appreciation of the impact each variable has on the classification can be ascertained, therefore pointing to whether it was wise to include that variable in the first place (in theory). Interestingly, after extensive sensitivity analysis, Vickers (2006, p173) concludes that *"as long as the reasoning for the original variable selection was sound, removing a*

variable from the analysis cannot really be justified". Vickers argues that even where variables have an apparent small effect on the clusters in a classification, removing them may not be the most sensible solution as whilst they may have a small effect overall, they may be *"vital to the formation of an individual cluster"* (Vickers, 2006 p173). Thanks to the work of Vickers, it is therefore possible to conclude that sensitivity analysis is not something which will be necessary for the Migration Classification – the original variables were chosen carefully and because they were deemed to add value to the classification. Removing variables will reduce the amount of information present in the classification – something which is important where one benefit of the Migration Classification could be to add value to the attribute poor data available between censuses.

Vickers (2006) also devotes a significant amount of time to assessing the reduction in variability within individual variables afforded by the clusters within the OA classification; the idea being that the better the classification, the greater the reduction in variability there will be within each variable. The rationale for this type of validation for the OA classification was that work by Voas and Williamson (2001) indicated that with very few variables, an ad hoc general-purpose classification system can offer much of the discriminatory power that a more carefully constructed classification can – i.e. a similar level of reduction in the variability of variables can be achieved. Vickers demonstrated that this was not necessarily the case with the OA classification, and where this type of validation was needed to challenge the assertions of Voas and Williamson when constructing a general-purpose classification, a similar exercise is not necessary here. The variables selected for the Migration Classification situate it very much as a 'purpose-built', bespoke framework for analysing migration data, so there is not the need to justify its existence as other, similar classifications do not exist.

The Leeds Classification for Community Safety (LCCS) (Shepherd, 2006), a bespoke classification more akin to the Migration Classification than the OA classification in that its intended use was very specific, employed a very different validation technique to those used by Vickers (2006). Shepherd constructed a series of 'Gains Charts' to assess relative advantages of the LCCS over general purpose classifications for predicting particular types of crime - a comparison technique also used by See and Openshaw (2001). This type of validation was appropriate, given that the principal purpose of the classification was to improve community safety through predicting patterns of crime. A similar type of analysis is not necessary or appropriate for the Migration Classification. Firstly, it is not the intention that the classification be used for predicting behaviours in the same way that a small area

geodemographic classification might – rather the Migration Classification is intended for use as a monitoring and complexity reduction tool. Secondly, where the Migration Classification is based entirely on migration variables, and other general purpose district classifications exclude migration variables, it would be illogical to test to see whether the Migration Classification is better at predicting migration behaviours than classifications which largely exclude migration variables.

Another common method used in the evaluation of small area geodemographic classifications is 'ground-truthing' (Vickers and Rees, 2009). This involves assessing the final solution by examining some of the small areas present in the different cluster groups to see if the real world situation bears any resemblance to that described by the classification group to which the area is assigned. Whilst this approach has some obvious flaws (the extent to which an entire cluster covering many areas can be validated through examining the physical environment of a small number of places is debateable), it is an approach that can only be adopted successfully where the areas being classified are relatively small and self contained (output areas or unit postcodes for example). Ground truthing areas where the smallest spatial unit is a local authority district is not feasible.

So as none of the above approaches to validation appear to be appropriate, an alternative method of validation is necessary. Referring back to the original rationale for developing the Migration Classification, one of the reasons for creating it was that in the context of monitoring migration between censuses, it was argued that migrants do not necessarily exhibit the characteristics of the underlying population and therefore a general purpose classification does not make sense if studying migration flows. It follows, therefore, that an appropriate method of validation would examine this assertion – are we are getting something new from this classification? Is it really an entirely new way of classifying districts, or merely a surrogate for one of the other available district classifications? And whilst sensitivity analysis, whereby individual variables are removed from the classification to assess their individual impact on the final cluster solutions, has been dismissed for reasons already explained, one interesting alternative would be to assess the impact of removing whole groups of variables from the classification and comparing the results with the original to see how different the cluster solutions are (i.e. how different are the clusters created if all socio-economic, or housing or age variables are dropped?). Could fewer variables have produced a similar solution? For example could a classification based entirely on age variables (where age will also be a feature of any migrant exhibiting another feature such as ethnicity) produce

a result very similar to the final classification arrived at here? If it does, then this will tell us as much about the relationship between migration variables as it does about the validity of the classification.

2.6.1 Mathematical methods for comparing clusters

A number of techniques have been developed to compare the results of different classification solutions where the same objects have been clustered differently. However, broadly speaking, they all operate similarly in that they assess the extent to which two different classification solutions agree and provide a statistic which quantifies the strength of this agreement. As noted by Everitt *et al.* (2001), one straight-forward way of measuring the association between two solutions with equal number of clusters is through calculating either a simple percentage agreement or Cohen's Kappa coefficient.

All cluster comparison solutions work on a contingency table of cluster agreement – a cluster 1 x cluster 2 matrix, termed $N = n_{ij}$. Consider Tables 2.11a and 2.11b below which represent a hypothetical dataset and related contingency table (notation adapted from Hubert and Arabie, 1985; example adapted from Yeung and Ruzzo, 2001):

Table 2.11a Dataset containing 10 objects to cluster and two different classification solutions – class u and cluster v

object	A	B	C	D	E	F	G	H	I	J
class (u)	1	1	2	2	2	2	3	3	3	3
cluster (v)	1	2	1	2	2	3	3	3	3	3

Table 2.11b Contingency table n_{ij} representing the agreement between the two classifications u and v

class/cluster	v1	v2	v3	sum	max	%
u1	1	1	0	2	1	0.2
u2	1	2	1	4	2	0.4
u3	0	0	4	4	4	0.4
sum	2	3	5	10	7	1
max	1	2	4	7		
%	0.2	0.3	0.5	1		

The similarity between the two classifications u and v can be calculated as the average similarity of u to v and v to u:

Where:

$$similarity\ u\ to\ v = \sum_{Ni+} max\left[n_{ij}\right] \times \frac{100}{N} \quad (5)$$

$$similarity\ v\ to\ u = \sum_{N+j} max\left[n_{ij}\right] \times \frac{100}{N} \quad (6)$$

Which, in the case of contingency Table 6.11b would be:

$$similarity\ u\ to\ v\ and\ v\ to\ u = 7\ \times \frac{100}{10} = 70\% \quad (7)$$

Cohen's kappa coefficient *k* (Cohen, 1960) uses the contingency table in a slightly different way, this time considering the probability of random agreement between the two classifications. In doing this it can be seen as a more robust measure than the percentage agreement between the two solutions.

Cohen's kappa can be calculated thus:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (8)$$

Where Pr(a) is the observed percentage agreement between class u and cluster v (or v and u), and Pr(e) is the probability of random agreement. In the contingency table example:

$$\Pr(a) = \frac{1+2+4}{10} = 0.7 \quad (9)$$

Pr(e) = the probability of being u1*v1 + u2*v2 + u3*v3 – or

$$\Pr(e) = (0.2 \times 0.2) + (0.4 \times 0.3) + (0.4 \times 0.5) = 0.36 \quad (10)$$

Therefore *k* = 0.53.

The range of *k* is between 0 and 1; 1 representing complete agreement between the classifications, 0 representing no agreement. In the case of this example, we might interpret that there is moderate agreement between the classifications.

Using metrics such a percentage agreement and Cohen's Kappa are fine when the two solutions being compared have the same number of clusters, these techniques fall down

where the numbers of cluster differ. This is important in the context of evaluating the Migration Classification, as other district level general purpose classifications such as the ONS classification of local authorities and the Vickers *et al.* classification have different numbers of clusters. Where the number of clusters differ, the Rand index (Rand, 1971), can be used as *'it is based on the agreement or otherwise of every pair of n objects'* (Everitt *et al.*, 2001 p182) rather than a cross-tabulation of the frequencies of agreement.

Given two different partitions of the same objects (class U and cluster V), the Rand index (R) can be described as:

$$R = \frac{(a+b)}{(a+b+c+d)} \tag{11}$$

Where:

a = number of pairs of objects in the same class U and the same cluster V.

b = number of pairs of objects not in the same class U or the same cluster V.

c = number of pairs of objects that are in the same class U and a different cluster V

d = number of pairs of objects that are in a different class U and the same cluster V

Some have criticised the Rand index as it varies, increasing as the number of clusters being compared increases. It is also does not take into consideration the chance agreement between the cluster allocation. Consequently, Hubert and Arabie (1985) proposed an adjusted Rand index which deals with these problems, with the form as follows:

$$AR = \frac{2(ab-cd)}{((a+d)(b+d)+(a+c)(b+c))} \tag{12}$$

Where the Rand index alone may cause problems in comparing classifications with differing numbers of clusters, here both the Rand index and the adjusted Rand index (Hubert and Arabie, 1985) will be computed to compare the solutions. To exemplify the calculation of the Rand and adjusted Rand indices for the example data in Table 2.11a, consider Table 2.12 Below:

Table 2.12 – Pairs of objects and associated cluster linkages for calculating Rand and adjusted Rand indices.

		a	b	c	d
Object pair		same u same v	not u not v	same u not v	same v not u
A	B	1	1	0	0
A	C	1	0	0	1
A	D	1	0	0	0
A	E	1	0	0	0
A	F	1	0	0	0
A	G	1	0	0	0
A	H	1	0	0	0
A	I	1	0	0	0
A	J	1	0	0	0
B	C	1	0	0	0
B	D	0	0	0	1
B	E	0	0	0	1
B	F	0	1	0	0
B	G	0	1	0	0
B	H	0	1	0	0
B	I	0	1	0	0
B	J	0	1	0	0
C	D	0	0	1	0
C	E	0	0	1	0
C	F	0	0	1	0
C	G	0	1	0	0
C	H	0	1	0	0
C	I	0	1	0	0
C	J	0	1	0	0
D	E	1	0	0	0
D	F	0	0	1	0
D	G	0	1	0	0
D	H	0	1	0	0
D	I	0	1	0	0
D	J	0	1	0	0
E	F	0	0	1	0
E	G	0	1	0	0
E	H	0	1	0	0
E	I	0	1	0	0
E	J	0	1	0	0
F	G	0	0	0	1
F	H	0	0	0	1
F	I	0	0	0	1
F	J	0	0	0	1
G	H	1	0	0	0
G	I	1	0	0	0
G	J	1	0	0	0
H	I	1	0	0	0
H	J	1	0	0	0
I	J	1	0	0	0
45		7	25	6	7

So for Table 2.12:

$$R = \frac{(7+25)}{(7+25+6+7)} = 0.71 \tag{13}$$

$$AR = \frac{2\times((7\times25)-(6\times7))}{((7+7)\times(25+7)+(7+6)\times(25+6))} = 0.31 \tag{14}$$

Similarly to Cohen's Kappa coefficient, both Rand indices range between 0 and 1, 0 indicating no agreement between the two matrices and 1 indicating total agreement. However the adjusted Rand index also gives an indication of how likely the agreement is by chance. For the example used here, the Rand index indicates relatively high agreement between the two indices, with the adjusted Rand index indicating a similarly high probability that the agreement between the two matrices is by chance (0 indicating greater likelihood of chance). Given the ability of Rand and adjusted Rand indices to compare classifications with different numbers of clusters, it is these measures that will be adopted to compare the Migration Classification with other district level classifications.

2.6.1 Comparison with other district level classifications

There are two principal district level classifications available for the UK currently: the ONS classification of local authorities (ONS, 2004), and the Vickers *et al.* (2006) local authority district classification. Other district level classifications do exist, such as the DEFRA rural/urban classification (DEFRA, 2009). But this particular classification only covers districts in England, so will not be used in the comparison with the Migration Classification. Both of these general purpose classifications are hierarchical, with three tiers of clusters, so the Migration Classification will be compared with each tier.

Table 2.13 Comparison of district level classifications

Classification pair					
Classification	Number of clusters	Classification	Number of clusters	Rand Index	Adjusted Rand Index
Migration classification	8	Vickers Family	4	0.64	0.06
Migration classification	8	Vickers Group	12	0.78	0.05
Migration classification	8	Vickers Class	24	0.82	0.03
Migration classification	8	ONS Super-group	7	0.68	0.05
Migration classification	8	ONS Group	12	0.77	0.06
Migration classification	8	ONS Sub-group	23	0.82	0.05
Vickers Family	4	ONS Super-group	7	0.69	0.28
Vickers Group	12	ONS Group	12	0.85	0.36
Vickers Class	24	ONS Sub-group	23	0.92	0.40
Migration classification	8	Age only classification	8	0.86	0.39
Migration classification	8	NS-SEC only classification	8	0.84	0.30
Migration classification	8	Family status only classification	8	0.81	0.20
Migration classification	8	Housing tenure only classification	8	0.83	0.25

Table 2.13 reveals the results of the comparisons between the classifications using both the Rand and adjusted Rand indices. The first three examples compare the Migration Classification with the three tiers of the Vickers et al. classification. As might be expected with the Rand index, as the number of clusters increases through the Vickers et al. hierarchy, so too does the index. At the Family level there is only moderate agreement between the two classifications, whereas at the class level agreement is relatively high. However, the usefulness of also including the adjusted Rand index is apparent as for all tiers the figure is very low, and reduces as the number of clusters increases. This suggests that the majority of agreement between the clusters shown by Rand index can be explained by chance. It is a very similar story when comparing the Migration Classification with the ONS classification; a moderate increasing to relatively high association between the two classifications according to the Rand index, mitigated by a very high likelihood of chance agreement demonstrated by the adjusted index. In comparison, when the Vickers et al. classification is compared with the ONS classification, as also shown in the table, the Rand index indicates a moderate to very high agreement between the two classifications which is maintained far more convincingly when the adjusted Rand index is also taken into consideration.

The results of the comparison between the Migration Classification and the two general purpose classifications are encouraging. They show that there is significant difference in the way that the alternative schemas group the districts in Britain. This helps to confirm the original hypothesis that a migration-specific classification will offer something different to a general purpose classification typology. And in answer to the question posed earlier, we can conclude that the Migration Classification *is* a new way of classifying districts, and not merely a surrogate for one of the other general purpose classifications already available. Certainly it offers a more different way of classifying districts than the two general purpose classifications do when compared with each other.

It remains then that the last task in the validation of the classification is to ensure that the final selection of variables indeed produced a solution distinct from any sub-sets of the same variables. Four alternative classifications were produced using the exact same methodology used in the original Migration Classification, however, this time sub-sets of the original suite of variables were chosen to be clustered. Classifications using only the variables associated with age, socio-economic status, housing tenure and family status were chosen. These classifications were then compared with the results of the original using the Rand and adjusted Rand indices, with the results also displayed in Table 2.13. As would be expected

using sub-sets of the original variables, the agreement between the Migration Classification and the sub-classifications is relatively high. The age-only classification achieves the highest Rand index and adjusted figures, followed by socio-economic status, housing tenure and then family status. Whilst in all cases agreement is quite high, it is not so high that one may be attempted to adopt one of the more parsimonious classifications in place of the original. For example, if the age only classification when compared with the Migration Classification scored a Rand index higher than 0.9 and an adjusted Rand index well over 0.5, then the value of including other variables in the classification might be brought into question; it could be argued that age variables explain enough of the final classification for other variables to be dropped. As it is, however, the adjusted Rand index is still relatively low for all of the alternative classifications and with Rand index agreement not above 0.87 for any alternative, then it can be concluded that including all variables and keeping the original classification is offers the best solution.

Concluding remarks and future research

At the beginning of this paper, a broad research hypothesis was set out which suggested that the complex processes of internal migration within Britain could be simplified and made easier to comprehend through the development of a migration based-classification. The hypothesis was developed on the premise that certain places would exhibit distinctive profiles formed by the migrants moving in, out and within, and that these profiles may be similar for some groups of places. It is clear from the research detailed in this paper, that this is indeed the case. After a long development process, the final classification outlined in section 2.5 reveals that eight distinct clusters of areas can be discerned in Britain, each cluster exhibiting particular attributes, defined by the migrants associated with them.

- Cluster 1 – Coastal and Rural Retirement Migrants – featuring districts around the periphery of Britain which attract older, often retirement age, migrants seeking the physical and social characteristics associated with these coastal and rural areas.
- Cluster 2 – Low-Mobility Britain – characterised by lower levels of migration activity across the board.
- Cluster 3 – Student Towns and Cities– with very high levels of young in-migrants and non-household migrants moving into privately rented accommodation.
- Cluster 4 – Moderate Mobility, Non-Household, Mixed Occupations – featuring low levels of migration, but where migration is occurring, it tending to involve single migrants and those in more intermediate occupations.
- Cluster 5 – Declining Industrial, Working-Class, Local Britain – a very distinctive cluster located in ex-industrial areas, where in and out-migration is uncommon, but local, short distance moves are not.
- Cluster 6 – Footloose, Middle-Class, Commuter Britain – almost the antithesis of the previous cluster where in and out-migration is very common and the migrants tend to be in the higher socio-economic groups.
- Cluster 7 – Dynamic London – located almost entirely within the M25, where levels of in and out-migration are very high across the board. And finally

- Cluster 8 – Successful Family In-migrants – a clear destination for family migrants and frequent origin for student migrants.

Whilst the final results produced are both useful, and justify the creation of a migration classification, one of the key findings of this piece of research is that there is not a short-cut to the development of a reliable, robust classification. The creation of a plausible initial classification proved that with relatively little effort, a result can be produced. This classification could easily have been adopted as a final classification, with the cluster profiles offering believable portraits of areas. What was also demonstrated, however, was that with so many elements in the classification building process, the final result can only be truly robust when each individual decision or choice in the process has been fully explained and justified both in the context of accepted theory and, perhaps more importantly, the context of the data being used. An exhaustive process of testing variables and methodology with this classification has meant that the final product is as reliable as it can be, insofar as any such classification will always have its limitations.

Within the sphere of methodology, the issue of using the appropriate piece of software for the task became important in this research. SPSS – a very widely available and commonly used piece of software – was shown to be inadequate for the construction of this classification. The limitations in relation to the selection of distance measures and the inability of the software to find a global minimum meant that it was not suitable for this task. The search for software which did not suffer from these limitations lead to the use of MATLAB for the creation of the final classification. The flexibility of MATLAB, which allowed different distance measures to be used, and a ‘replicates’ parameter which offered the possibility of a global minimum solution to be found, meant that it was far more suited to the classification building process.

Another finding coming out of the classification building process is that there are some areas which are much easier to classify than others. Some areas which appeared in clusters in the initial classification remained in very similar clusters in the final classification even after some variables had been dropped or a different clustering methodology used. For example, the ex-industrial areas defined in cluster 5 were also present in their own cluster in the initial classification; as were many of the coastal and rural areas, areas in London and commuter areas around the perimeter of London. There were also those areas which consistently proved more difficult to classify. These areas, whilst always allocated to one cluster or another, were also the areas which frequently changed cluster at different steps of the process. These areas

are defined by low silhouette values, and perhaps can be seen as being distinct, precisely because the migrants moving in, out and within these areas are indistinct – these are, if you like, migration 'grey areas' and may well benefit from increased attention in the future, perhaps at a sub-district level with a ward classification, to see if there are clearer patterns visible at a higher resolution. It may well be that these areas suffer from the smoothing and cancelling effects of coarser geographies.

This research has now constructed both a framework for new research and has paved the way for continuing developments in the classification of areas by migration. Going forward, there is now the opportunity to use a very similar methodology to construct a similar classification for wards in Britain (or the whole UK). Whilst offering less potential for additional analysis – ward level migration data generally not being available (perhaps with the exception of student migration data from the Higher Education Statistics Agency) – a ward level classification would offer the opportunity to examine movements at a much finer spatial scale and may well improve the definition of some of the harder to define districts in this classification. It also offers the chance to incorporate distance information in the guise of data derived from functional regionalisation experiments – something which would add an additional dimension to the work, and increase our understanding still further. One of the original perceived drawbacks of a ward level classification was that the level of variable detail offered was inferior to that at the district level. This issue is no longer founded as experiments at the district scale resulted in more generalised variables being used anyway in order that the increased geographical scale did not cancel out variation in more nuanced variables.

This district level classification itself though, can now be used as a framework for analysing internal migration in Britain since 2001. Patient register data, available annually from 2001, can now be examined in the context of this classification. Patient register data suffer from being much less detailed than the census – migration by age is the limit of the detail that is offered. The value of this classification is that where areas are classified by migrants of a particular ethnic origin, socio-economic status, tenure or family status, we can infer these details onto the less detailed data we have – building a much richer picture than would have been possible before.

In addition, it should also be possible to project future migration patterns for these classified areas using spatial interaction models. Where flows between origins and destinations are

known, and where some measure of distance between these origins and destinations can be derived, and where the attractiveness of origins and destinations can be inferred, it is possible to represent the spatial interactions – in this case migration flows – between origins and destinations in the form of a simplified model. This model can then be used to predict the flows between origins and destinations where the attractiveness of origins and/or destinations is estimated in the future. The exact form of any spatial interaction model is not known at this stage, for example it could be for a matrix of 8 cluster origins by 8 cluster destinations, or, for a series of matrices for each cluster, with district origins and destinations in each one. The potential for such research, however, is there.

It is clear then that this classification has been useful for developing an understanding of the complex internal migration patterns occurring in Britain. It is also clear that this new classification of districts provides a solid framework for a wealth of additional research into internal migration from 2001 onwards.

References

Aggarwal, C.C., Hinnenburg, A. and Keim, D.A. (2001). "On the surprising behaviour of distance metrics in high dimensional space." *Lecture Notes in Computer Science* **1973**: 420-434.

Aldenderfer, M.S. and Blashfield, R.K. (1984). *Cluster analysis*. Sage Publications, Inc., London.

Bailey, S., Charlton, J., Dollamore, G. and Fitzpartick, J. (2000). "Families, groups and clusters of local and health authorities: revised for authorities in 1999." *Population Trends* **99**: 37-52.

Bales, K. (1999). "Popular reactions to sociological research: the case of Charles Booth." *Sociology* **33**(1): 153-168.

Bates, J. and Bracken, I. (1982). "Estimation of migration profiles in England and Wales." *Environment and Planning A* **14**(7): 889-900.

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 International, Glasgow.

Boden, P. and Rees, P. (2009). "International migration: the estimation of immigration to local areas in England using administrative data sources." *Paper submitted for consideration by the Journal of the Royal Statistical Society*.

Boden, P. and Rees, P. (Forthcoming). *Improving the reliability of estimates of migrant worker numbers and their relative risk of workplace injury and illness*. Health and Safety Executive.

Boden, P. and Stillwell, J. (2006). "New migrant labour in Yorkshire and the Humber." *Yorkshire and Humber Regional Review* **16**(3): 18-20.

Bohara, A.K. and Krieg, R.G. (1998). "A simultaneous multinomial logit model of indirect internal migration and earnings." *Journal of Regional Analysis and Policy* **28**(1): 60-72.

Böheim, R. and Taylor, M.P. (2002). "Tied Down Or Room To Move? Investigating the Relationships between Housing Tenure, Employment Status and Residential Mobility in Britain." *Scottish Journal of Political Economy* **49**(4): 369-392.

Boyle, P. (1998). "Migration and housing tenure in South East England." *Environment and Planning A* **30**: 855-866.

Boyle, P., Cooke, T., Halfacree, K. and Smith, D. (1999). "Intregrating GB and US census microdata for studying the impact of family migration on partnered women's labour market status." *International Journal of Population Geography* **5**: 157-178.

Boyle, P.J., Flowerdew, R. and Shen, J. (1998). "Modelling inter-ward migration in Hereford and Worcester: The importance of housing growth and tenure." *Regional Studies* **32**(2): 113 - 132.

Brown, L.A. and Holmes, J.H. (1971). "The delimitation of functional regions, nodal regions and hierarchies by functional distance approaches." *Journal of Regional Science* **11**: 57-72.

Burgess, E.W. (1925). "The growth of the city: an introduction to a research project". 156-163. In LeGates, R.T. and Stout, F., (eds.) *The city reader*. Routledge, London.

Cameron, G., Muellbauer, J. and Murphy, A. (2005). "Migration within England and Wales and the housing market". *Royal Economic Society Annual Conference*. http://www.nuffield.ox.ac.uk/users/Muellbauer/MIGRATION-APRIL06.pdf

Carstensen, B. and Keiding, N. (2005). *Age-period-cohort models: statistical inference in the Lexis diagram*. Institute of Public Health, Copenhagen. http://staff.pubhealth.ku.dk/~bxc/APC/notes.pdf.

Castro, L.J. and Rogers, A. (1981). *Status-specific age patterns of migration: family status*. IIASA Working Paper WP-81-060 International Institute for Applied Systems Analysis, Laxenburg.
Champion, A.G. (2005). "Population movement within the UK". 92-114. In Chappell, R., (ed.) *Focus on people and migration*. Palgrave Macmillan, Basingstoke.
Champion, A.G., Coombes, M., Raybould, S. and Wymer, C. (2007). *Migration and socio-economic change: a 2001 census analysis of Britain's larger cities.* Joseph Rowntree Foundation, Bristol.
Champion, A.G., Fotheringham, A.S., Rees, P., Boyle, P. and Stillwell, J. (1998). "The determinants of migration flows in England: a review of existing data and evidence.". http://www.geog.leeds.ac.uk/publications/DeterminantsOfMigration/report.pdf
Clark, W. and Huang, Y. (2004). "Linking Migration and Mobility: Individual and Contextual Effects in Housing Markets in the UK." *Regional Studies* **38**(6): 617 - 628.
Cohen, J. (1960). "A coefficient of agreement for nominal scales." *Educational and psychological measurement* **20**(1): 37-46.
Cooke, T.J. (2008). "Migration in a family way." *Population Space and Place* **14**(4): 255-265.
Cooke, T.J. and Bailey, A.J. (1999). "The effect of family migration, migration history, and self-selection on married women's labour market achievement". 102-113. In Boyle, P. and Halfacree, K., (eds.) *Migration and Gender in the Developed World*. Routledge, London.
Coombes, M. (2000). "Defining locality boundaries with synthetic data." *Environment and Planning A* **32**: 1499-1518.
Coombes, M. (2002). *Travel to work areas and the 2001 census*. Centre for Urban and Regional Development Studies, University of Newcastle, Newcastle.
Coombes, M., Green, A. and Openshaw, S. (1986). "An efficient algorithm to generate official statistical reporting areas: the case of the 1984 travel-to-work areas revision in Britain." *Journal of the Operational Research Society* **37**(10): 943-953.
Coombes, M., Raybould, S. and Wymer, C. (2004). *Analysis of census migration data to assist in defining housing market areas for Tyne and Wear*. Centre for Urban and Regional Development Studies, University of Newcastle Upon Tyne, Newcastle Upon Tyne.
Crowson, R.A. (2006). *Classification and biology*. Aldine Transaction, New Brunswick.
Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
Debenham, J., Clarke, G. and Stillwell, J. (2003). "Extending geodemographic classification: a new regional prototype." *Environment and Planning A* **35**(6): 1025-1050.
DEFRA (2009). *Defra Classification of Local Authority Districts and Unitary Authorities in England - An Introductory Guide*. Department for Environment Food and Rural Affairs.
Dennett, A., Duke-Williams, O. and Stillwell, J. (2007). *Interaction data sets in the UK: an audit*. Working Paper 07/05. University of Leeds, Leeds.
Dennett, A. and Stillwell, J. (2008a). *Internal migration in Great Britain - a district level analysis using 2001 Census data*. University of Leeds - Working Paper 01/08.
Dennett, A. and Stillwell, J. (2008b). "Population turnover and churn - enhancing understanding of internal migration in Britain through measures of stability " *Population Trends* **134**: 24-41.
Dennett, A. and Stillwell, J. (2009). "Internal migration in Britain, 2000-01, examined through an area classification framework." *Population Space and Place* (DOI: 10.1002/psp.554).

Dennett, A. and Stillwell, J. (2010). "Internal Migration Patterns by Age and Sex at the Start of the 21st Century ". In Stillwell, J., Duke-Williams, O. and Dennett, A., (eds.) *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*. IGI Global.

Dennett, A. and Stillwell, J. (Forthcoming). "Internal Migration Patterns by Age and Sex at the Start of the 21st Century ". In Stillwell, J., Duke-Williams, O. and Dennett, A., (eds.) *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*. IGI Global.

Dixon, S. (2003). "Migration within Britain for job reasons." *Labour Market Trends* **April**: 191-201.

Duke-Williams, O. and Blake, M. (2003). *Database fusion for the comparative study of migration data*. gisca, Adelaide. http://www.geocomputation.org/1999/068/gc_068.htm. 7/408

Eliasson, K., Lindgren, U. and Westerlund, O. (2003). "Geographical labour mobility: migration or commuting?" *Regional Studies* **37**(8): 827-837.

Everitt, B.S. and Dunn, G. (2001). *Applied multivariate data analysis - second edition*. Hodder Arnold, London.

Everitt, B.S., Landau, S. and Leese, M. (2001). *Cluster Analysis*. Arnold, London.

Faggian, A., McCann, P. and Sheppard, S. (2006). "An analysis of ethnic differences in UK graduate migration behaviour." *Annals of Regional Science* **40**: 461-471.

Falkenauer, E. and Marchand, A. (2001). "Using K-means? Consider ArrayMiner." *Proceedings of the 2001 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*.

Farr, M. and Webber, R. (2001). "MOSAIC: From and area classification system to individual classification." *Journal of targeting, measurement and analysis for marketing* **10**(1): 55-65.

Feldman, O., Simmonds, D., Troll, N. and Tsang, F. (2006). *Creation of a system of functional areas for England and Wales and for Scotland*. MVA Consultancy, London. http://www.mvaconsultancy.com/papers/Creation%20of%20a%20system%20of%20functional%20areas%20for%20England%20and%20W%85.pdf.

Field, A. (2005). *Discovering Statistics Using SPSS*. Sage, London.

Findlay, A., Mason, C., Houston, D., McCollum, D. and Harrison, R. (2009). "Escalators, Elevators and Travelators: The Occupational Mobility of Migrants to South-East England." *Journal of Ethnic and Migration Studies* **35**(6): 861-879.

Finney, N. and Simpson, L. (2008). "Internal migration and ethnic groups: evidence for Britain from the 2001 census." *Population Space and Place* **14**: 63-83.

Finney, N. and Simpson, L. (2009). "Population Dynamics: The Roles of Natural Change and Migration in Producing the Ethnic Mosaic." *Journal of Ethnic and Migration Studies* **35**(9): 1479-1496.

Geist, C. and McManus, P.A. (2008). "Geographical mobility over the life course: motivations and implications." *Population, Space and Place* **14**(4): 283-303.

Gordon, A.D. (1999). *Classification - Second Edition*. Chapman and Hall, London.

Grayson, D. (2004). "Some myths and legends in quantitative psychology." *Understanding Statistics* **3**(1): 101-134.

Harris, C.D. and Ullman, E.L. (1945). "The nature of cities." *Annals of the American Academy of Political and Social Science* **242**: 7-17.

Harris, R. (2005). "Considering (mis-)representation in geodemographics and lifestyles". http://www.geocomputation.org/1998/82/gc_82.htm

Harris, R., Sleight, P. and Webber, R., (eds.) (2005). *Geodemographics, GIS and neighbourhood targeting*. John Wiley & Sons Ltd, Chichister.
Heasman, D. (2008). "A local area model for emigration." Paper presented at GSS Conference, 23 June 2008. http://www.ons.gov.uk/about/newsroom/events/thirteenth-gss-methodology-conference--23-june-2008/programme/a-local-area-model-for-emigration-paper-by-dick-heasman.doc
Höppner, F., Klawonn, F., Kruse, R. and Runkler, T. (1999). *Fuzzy Cluster Analysis*. Wiley, Chippenham.
Hoyt, H. (1939). *The structure and growth of residential neighbourhoods in American cities*. Federal Housing Administration, Washington.
Hubert, L.J. and Arabie, P. (1985). "Comparing partitions." *Journal of Classification* **2**(1): 193-218.
Hussain, S. and Stillwell, J. (2008). *Internal migration of ethnic groups in England and Wales by age and district type*. Working paper 08/3, School of Geography. University of Leeds.
Inter-departmental Migration Task Force (2006). *Report of the Inter-departmental Migration Task Force on Migration Statistics, National Statistics, London*. http://www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/TaskForceReport151206.pdf.
Kaufman, L. and Rousseeuw, P.J. (2005). *Finding groups in data - an introduction to cluster analysis*. John Wiley & Sons, New Jersey.
King, G., Tanner, M.A. and Rosen, O., (eds.) (2004). *Ecological inference: new methodological strategies*. Cambridge University Press, Cambridge.
Kline, P. (1994). *An easy guide to factor analysis*. Routledge, London.
Longley, P. (2005). "Geographical information systems: a renaissance of geodemographics for public service delivery." *Progress in Human Geography* **29**: 57-63.
Martin, D. (1998). "Optimising census geography: the separation of collection and output geographies." *International Journal of Geographical Information Science* **12**(7): 673-685.
Martin, D. (2000). "Towards the geographies of the 2001 UK Census of Population." *Transactions of the Institute of British Geographers* **25**: 321-332.
Martin, D. (2002). "Output areas for 2001". In Rees, P., Martin, D. and Williamson, P., (eds.) *The census data system*. John Wiley and Sons, Chichester.
MathWorks (2009). *MATLAB Statistics Toolbox 7 - User's Guide*. The MathWorks Inc., Natick.
Milligan, G.W. (1996). "Clustering validation: results and implications for applied analyses". 341-375. In Arabie, P., Hubert, L.J. and De Soete, G., (eds.) *Clustering and Classification*. World Scientific, Singapore.
Milligan, G.W. and Cooper, M.C. (1985). "An examination of procedures for determining the number of clusters in a dataset." *Psychometrika* **50**: 159-179.
Milligan, G.W. and Cooper, M.C. (1987). "Methodology Review: Clustering Methods." *Applied Psychological Measurement* **11**: 329-354.
Milligan, G.W. and Cooper, M.C. (1988). "A study of standardization of variables in cluster analysis." *Journal of Classification* **5**(181-204).
Murphy, A., Muellbauer, J. and Cameron, G. (2006). *Housing market dynamics and regional migration in Britain*. Discussion Paper Series. Department of Economics, University of Oxford, Oxford.
Norman, P., Boyle, P. and Rees, P. (2005). "Selective migration, health and deprivation: a longitudinal analysis." *Social Science & Medicine* **60**(12): 2755-2771.

O'Sullivan, D. and Unwin, D.J. (2002). *Geographical information analysis*. John Wiley and Sons, London.
ODPM (2003). *English indices of deprivation 2004 (revised)*. ODPM Publications, London.
ONS (2004). "National Statistics 2001 area classification for local authorities". http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/la/
ONS (2008). *International Migration*. Series MN No 33, 2006 Data. Office for National Statistics. http://www.statistics.gov.uk/downloads/theme_population/MN33.pdf.
Openshaw, S. (1989). "Making geodemographics more sophisticated." *Journal of the market research society* **31**: 111-131.
Openshaw, S. and Blake, M. (1996). "GB Profiler 91". School of Geography, University of Leeds. Leeds.
Openshaw, S. and Taylor, P.J. (1979). "A million or so correlation coefficients: three experiments on the modifiable areal unit problem". 127-144. In Wrigley, N. and Bennet, R.J., (eds.) *Statistical applications in spatial sciences*. Pion, London.
Openshaw, S. and Wymer, C. (1995). "Classifying and regionalizing census data". In Openshaw, S., (ed.) *Census Users' Handbook*. Geoinformation International, Cambridge.
Orford, S., Dorling, D., Mitchell, R., Shaw, M. and Smith, G.D. (2002). "Life and death of the people of London: a historical GIS of Charles Booth's inquiry." *Health & Place* **8**(1): 25-35.
Owen, D. (1997). "Migration by minority ethnic groups within Great Britain in the early 1990s". *28th annual conference of the British and Irish section of the Regional Science Association International*. Falmouth College of Arts.
Rand, W.M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* **66**(336): 846-850.
Ravenstein, E.G. (1885). "The laws of migration." *Journal of the Statistical Society of London* **48**(2): 167-235.
Ravenstein, E.G. (1889). "The laws of migration." *Journal of the Royal Statistical Society* **52**(2): 241-305.
Raymer, J., Abel, G. and Smith, P.W.F. (2007). "Combining census and registration data to estimate detailed elderly migration flows in England and Wales." *Journal of the Royal Statistical Society Series a-Statistics in Society* **170**(4): 891-908.
Raymer, J., Bonaguidi, A. and Valentini, A. (2006). "Describing and projecting the age and spatial structures of interregional migration in Italy." *Population Space and Place* **12**(5): 371-388.
Raymer, J. and Giulietti, C. (2009). "Ethnic migration between area groups in England and Wales." *Area* **41**(4): 435-451.
Raymer, J., Smith, P.W.F. and Giulietti, C. (2008). *Combining census and registration data to analyse ethnic migration patterns in England from 1991 to 2007*. Working Paper M08/09. University of Southampton, Southampton Statistical Sciences Research Institute, Southampton. http://eprints.soton.ac.uk/63739/.
Rees, P. (1977). "The measurement of migration from census and other sources." *Environment and Planning A* **9**: 257-280.
Rees, P., Denham, C., Charlton, J., Openshaw, S., Blake, M. and See, L. (2002a). "ONS classifications and GB profiles: Census typologies for researchers". In Rees, P., Martin, D. and Williamson, P., (eds.) *The Census Data System*. John Wiley and Sons Ltd., Padstow.
Rees, P., Martin, D. and Williamson, P., (eds.) (2002b). *The census data system*. John Wiley & Sons Ltd. , Padstow.

Rees, P., Stillwell, J., Boden, P. and Dennett, A. (2009). *A review of migration statistics literature*. UK Statistics Authority, London.
Rogers, A. and Castro, L.J. (1981). *Model migration schedules*. Research Report-81-30,. International Institute for Applied Systems Analysis, Laxenburg, Austria.
Rogers, A., Raymer, J. and Willekens, F. (2002). "Capturing the age and spatial structures of migration." *Environment and Planning A* **34**(2): 341-359.
Rousseeuw, P.J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics* **20**: 53-65.
Schaffer, C.M. and Green, P.E. (1996). "An empirical comparison of variable standardisation methods in cluster analysis." *Multivariate Behavioural Research* **31**(2): 149-167.
See, L. and Openshaw, S. (2001). "Fuzzy geodemographic targeting". 269-281. In Clarke, G. and Madden, M., (eds.) *Regional science in business*. Springer, Berlin.
Shepherd, P.J. (2006). "Neighbourhood profiling and classification for community safety". *School of Geography*. University of Leeds. Leeds.
Simpson, L. and Finney, N. (2009). "Spatial patterns of internal migration: evidence for ethnic groups in Britain." *Population, Space and Place* **15**: 37-56.
Sjaastad, L.A. (1962). "The costs and returns of human migration." *The journal of Political Economy* **70**(5): 80-93.
SPSS (2005). *SPSS Base 14.0 User's Guide*. SPSS Inc., Chicago.
SPSS (2006). *SPSS 14.0 for Windows*. SPSS.
Stillwell, J. and Duke-Williams, O. (2005). "Ethnic population distributions, immigration and internal migration in Britain: What evidence for linkage at district scale". *British Society for Population Studies Annual Conference*. University of Kent at Canterbury.
Stillwell, J. and Duke-Williams, O. (2007). "Understanding the 2001 UK census migration and commuting data: the effect of small cell adjustment and problems of comparison with 1991." *Journal of the Royal Statistical Society Series A (Statistics in Society)* **170**(2): 425-445.
Stillwell, J. and Hussain, S. (2008). *Ethnic group migration within Britain during 2000-01: a district level analysis*. Working Paper 08/2. University of Leeds, Leeds.
Stillwell, J., Hussain, S. and Norman, P. (2008). "The internal migration propensities and net migration patterns of ethnic groups in Britain." *Migration Letters* **5**(2): 135-150.
Stillwell, J., Rees, P. and Boden, P. (1992). *Migration patterns and processes volume 2; population redistribution in the United Kingdom*. Belhaven Press, London.
Uren, Z. and Goldring, S. (2008). "Migration trends at older ages in England and Wales." *Population Trends* **130**: 31-40.
Vandeschrick, C. (2001). "The Lexis diagram, a misnomer." *Demographic Research* **4**: 97-124.
Vansershrick (1992). "Le diagramme de Lexis revisite." *Population* **92**(5): 1241-1262.
Vickers, D. (2006). "Multi-level integrated classifications based on the 2001 Census". *School of Geography*. University of Leeds. Leeds.
Vickers, D. and Rees, P. (2009). "Ground-truthing Geodemographics." *Applied Spatial Analysis and Policy*.
Vickers, D., Rees, P. and Birkin, M. (2003). "A new classification of UK local authorities using 2001 Census key statistics". *University of Leeds - Working Paper 03/03*. University of Leeds - Working Paper 03/03. Leeds. http://www.geog.leeds.ac.uk/wpapers/03-3.pdf
Vickers, D., Rees, P. and Birkin, M. (2005). *Creating the national classification of census output areas: data, methods and results*. Working Paper 05/2. University of Leeds, Leeds. http://www.geog.leeds.ac.uk/wpapers/05-2.pdf.

Voas, D. and Williamson, P. (2001). "The diversity of diversity: a critique of geodemographic classification." *Area* **33**(1): 63-76.
Ward, J. (1963). "Hierarchical grouping to optimize an objective function." *Journal of the American Statistical Association* **58**: 236-244.
Yano, K. (2001). "GIS and quantitative geography." *GeoJournal* **52**: 173-180.
Yeung, K.Y. and Ruzzo, W.L. (2001). *Details of the Adjusted Rand index and Clustering algorithms Supplement to the paper "An empirical study on Principal Component Analysis for clustering gene expression data"*. University of Washington. http://faculty.washington.edu/kayee/pca/supp.pdf.
Založnik, M. (2006). "Geographical variation of geodemographic classifiability". University of Liverpool.

Appendix 1

Cluster lookup table

label	code	cluster	silhouette value
City of London	00AA	7	0.025393298
Barking and Dagenham	00AB	4	0.15670542
Barnet	00AC	7	0.050161683
Bexley	00AD	2	0.048213533
Brent	00AE	7	0.280990783
Bromley	00AF	4	0.19809761
Camden	00AG	7	0.294256001
Croydon	00AH	4	0.259965483
Ealing	00AJ	7	0.300621733
Enfield	00AK	7	-0.017966802
Greenwich	00AL	7	0.176857227
Hackney	00AM	7	0.32943061
Hammersmith and Fulham	00AN	7	0.363418695
Haringey	00AP	7	0.3751089
Harrow	00AQ	7	0.000115779
Havering	00AR	2	0.185186132
Hillingdon	00AS	6	-0.060816441
Hounslow	00AT	7	0.263225815
Islington	00AU	7	0.375722473
Kensington and Chelsea	00AW	7	0.256208419
Kingston upon Thames	00AX	7	0.094114769
Lambeth	00AY	7	0.356882836
Lewisham	00AZ	7	0.285107587
Merton	00BA	7	0.183006522
Newham	00BB	7	0.29235288
Redbridge	00BC	6	0.041641363
Richmond upon Thames	00BD	7	0.075031126
Southwark	00BE	7	0.35807335
Sutton	00BF	4	0.242440039
Tower Hamlets	00BG	7	0.307044071
Waltham Forest	00BH	7	0.110944037
Wandsworth	00BJ	7	0.357674215
Westminster	00BK	7	0.225586773
Bolton	00BL	5	0.283075255
Bury	00BM	5	0.068826463
Manchester	00BN	3	0.235157503
Oldham	00BP	5	0.237594266
Rochdale	00BQ	5	0.265970166
Salford	00BR	3	0.045709379
Stockport	00BS	2	0.068073327
Tameside	00BT	5	0.166765709
Trafford	00BU	4	0.125918563
Wigan	00BW	5	0.243586395
Knowsley	00BX	2	0.086468887
Liverpool	00BY	3	0.309700528
St. Helens	00BZ	5	0.166692277
Sefton	00CA	5	0.149908987
Wirral	00CB	5	0.309781571
Barnsley	00CC	5	0.252965797
Doncaster	00CE	5	0.323877162
Rotherham	00CF	5	0.248741869
Sheffield	00CG	3	0.356618595
Gateshead	00CH	5	0.129188721

Newcastle upon Tyne	00CJ	3	0.36477393
North Tyneside	00CK	5	0.035594504
South Tyneside	00CL	5	0.253717051
Sunderland	00CM	3	-0.015760535
Birmingham	00CN	3	0.217560163
Coventry	00CQ	3	0.233297895
Dudley	00CR	5	0.095690653
Sandwell	00CS	5	0.059891756
Solihull	00CT	2	0.219087363
Walsall	00CU	5	0.081666391
Wolverhampton	00CW	5	0.084278956
Bradford	00CX	5	0.259988345
Calderdale	00CY	5	0.202666212
Kirklees	00CZ	5	0.208718567
Leeds	00DA	3	0.270613895
Wakefield	00DB	5	0.21993725
Hartlepool	00EB	5	0.134374443
Middlesbrough	00EC	3	0.164999165
Redcar and Cleveland	00EE	5	0.173533141
Stockton-on-Tees	00EF	5	0.17100771
Darlington	00EH	1	-0.005630314
Halton	00ET	5	0.040231359
Warrington	00EU	5	0.016155509
Blackburn with Darwen	00EX	5	0.222280658
Blackpool	00EY	1	-0.022688256
Kingston upon Hull city of	00FA	3	0.293816614
East Riding of Yorkshire	00FB	1	0.103225066
North East Lincolnshire	00FC	5	0.200926627
North Lincolnshire	00FD	5	-0.003576125
York	00FF	3	0.264116336
Derby	00FK	5	0.005486873
Leicester	00FN	3	0.283671596
Rutland	00FP	6	0.00585765
Nottingham	00FY	3	0.349451115
Herefordshire Coun	00GA	1	0.25345775
Telford and Wrekin	00GF	1	0.072634521
Stoke-on-Trent	00GL	3	0.10111149
Bath and North East Somerset	00HA	3	0.214101185
Bristol City of	00HB	3	0.260840941
North Somerset	00HC	1	0.227925537
South Gloucestershire	00HD	4	0.227234024
Plymouth	00HG	3	0.150598526
Torbay	00HH	1	0.270069736
Bournemouth	00HN	3	0.104263538
Poole	00HP	8	0.062071267
Swindon	00HX	4	0.136499548
Peterborough	00JA	4	0.191388461
Luton	00KA	3	-0.10946281
Southend-on-Sea	00KF	4	0.099596572
Thurrock	00KG	4	0.195066005
Medway Towns	00LC	4	0.157376417
Bracknell Forest	00MA	6	0.043569476
West Berkshire	00MB	6	0.123454666
Reading	00MC	7	0.244830492
Slough	00MD	4	0.129905339
Windsor and Maidenhead	00ME	6	0.076059434
Wokingham	00MF	6	0.18972462
Milton Keynes	00MG	4	0.063454568

Brighton and Hove	00ML	3	0.148311006
Portsmouth	00MR	3	0.136642938
Southampton	00MS	3	0.328820462
Isle of Wight	00MW	1	0.297308669
Isle of Anglesey	00NA	5	0.054178364
Gwynedd	00NC	3	-0.00468716
Conwy	00NE	1	0.276002068
Denbighshire	00NG	1	0.170824428
Flintshire	00NJ	5	0.071117894
Wrexham	00NL	4	-0.002137701
Powys	00NN	1	0.085138871
Ceredigion	00NQ	3	0.189666939
Pembrokeshire	00NS	1	0.207843252
Carmarthenshire	00NU	1	0.134655796
Swansea	00NX	3	0.09007968
Neath Port Talbot	00NZ	5	0.171903446
Bridgend	00PB	5	0.18077939
The Vale of Glamorgan	00PD	5	0.044484642
Rhondda Cynon Taf	00PF	5	0.178424699
Merthyr Tydfil	00PH	5	0.122311543
Caerphilly	00PK	5	0.238146479
Blaenau Gwent	00PL	5	0.251975903
Torfaen	00PM	5	0.169925306
Monmouthshire	00PP	8	0.090249048
Newport	00PR	5	0.061095755
Cardiff	00PT	3	0.354744037
Mid Bedfordshire	09UC	6	-0.031084846
Bedford	09UD	4	0.186796375
South Bedfordshire	09UE	2	-0.010554306
Aylesbury Vale	11UB	6	0.033749514
Chiltern	11UC	6	0.119902406
South Bucks	11UE	6	0.181543723
Wycombe	11UF	6	0.047667754
Cambridge	12UB	7	0.041072818
East Cambridgeshire	12UC	8	0.082330637
Fenland	12UD	1	0.049868875
Huntingdonshire	12UE	4	0.055561491
South Cambridgeshire	12UG	6	0.120655035
Chester	13UB	3	-0.05950111
Congleton	13UC	8	0.063742601
Crewe and Nantwich	13UD	1	0.15118949
Ellesmere Port and Nes	13UE	2	0.056824621
Macclesfield	13UG	6	0.011995797
Vale Royal	13UH	2	0.01602211
Caradon	15UB	8	0.128145941
Carrick	15UC	1	0.078224338
Kerrier	15UD	1	0.131227103
North Cornwall	15UE	1	0.092039714
Penwith	15UF	1	0.182013981
Restormel	15UG	1	0.201649513
Isles of Scilly	15UH	6	0.02745009
Allerdale	16UB	5	0.092470332
Barrow-in-Furness	16UC	5	0.2396111
Carlisle	16UD	1	0.000441642
Copeland	16UE	5	0.082935216
Eden	16UF	1	0.131575619
South Lakeland	16UG	1	0.011850816
Amber Valley	17UB	2	0.076190472

Bolsover	17UC	1	-0.007250253
Chesterfield	17UD	5	0.108011091
Derbyshire Dales	17UF	8	0.056457405
Erewash	17UG	2	-0.015285131
High Peak	17UH	2	0.045502748
North East Derbyshire	17UJ	2	0.26988616
South Derbyshire	17UK	8	0.016065164
East Devon	18UB	1	0.052869959
Exeter	18UC	3	0.225595802
Mid Devon	18UD	8	0.024862659
North Devon	18UE	1	0.146169411
South Hams	18UG	8	0.176049307
Teignbridge	18UH	1	0.013045245
Torridge	18UK	1	0.177411182
West Devon	18UL	8	0.159445895
Christchurch	19UC	8	0.177584315
East Dorset	19UD	8	0.170674627
North Dorset	19UE	6	-0.042225557
Purbeck	19UG	8	0.120923145
West Dorset	19UH	8	0.078007536
Weymouth and Portland	19UJ	1	0.166577697
Chester-le-Street	20UB	2	0.138980972
Derwentside	20UD	5	0.234241593
Durham	20UE	3	0.140375628
Easington	20UF	5	0.209788312
Sedgefield	20UG	5	0.20969636
Teesdale	20UH	8	0.11836435
Wear Valley	20UJ	5	0.009255023
Eastbourne	21UC	1	0.187852797
Hastings	21UD	1	0.212336311
Lewes	21UF	8	0.129542037
Rother	21UG	8	0.139370836
Wealden	21UH	8	0.141845724
Basildon	22UB	2	-0.000662916
Braintree	22UC	8	0.062212179
Brentwood	22UD	6	0.098424477
Castle Point	22UE	2	0.179750891
Chelmsford	22UF	4	0.211990314
Colchester	22UG	4	0.052641964
Epping Forest	22UH	4	0.119914061
Harlow	22UJ	4	0.230625869
Maldon	22UK	8	0.107647687
Rochford	22UL	2	0.122264869
Tendring	22UN	1	0.249470907
Uttlesford	22UQ	6	0.154808535
Cheltenham	23UB	3	-0.026252673
Cotswold	23UC	6	-0.018196008
Forest of Dean	23UD	2	0.05389088
Gloucester	23UE	4	0.252575986
Stroud	23UF	8	0.016456292
Tewkesbury	23UG	8	0.058821601
Basingstoke and Deane	24UB	4	0.205142023
East Hampshire	24UC	6	0.069465014
Eastleigh	24UD	2	0.038498808
Fareham	24UE	6	0.000516419
Gosport	24UF	1	-0.016131758
Hart	24UG	6	0.190066148
Havant	24UH	2	0.110939094

New Forest	24UJ	8	0.123783017
Rushmoor	24UL	7	-0.007674778
Test Valley	24UN	6	0.074337677
Winchester	24UP	6	0.164138463
Broxbourne	26UB	4	0.087189633
Dacorum	26UC	6	-0.027828527
East Hertfordshire	26UD	6	0.111854159
Hertsmere	26UE	6	0.143418025
North Hertfordshire	26UF	4	0.120168384
St. Albans	26UG	6	0.089270354
Stevenage	26UH	4	0.230695228
Three Rivers	26UJ	6	0.109632779
Watford	26UK	7	-0.010992775
Welwyn Hatfield	26UL	3	0.002031356
Ashford	29UB	1	0.067441526
Canterbury	29UC	3	0.261176618
Dartford	29UD	4	0.137763142
Dover	29UE	1	0.105560267
Gravesham	29UG	2	0.133534168
Maidstone	29UH	4	0.092329034
Sevenoaks	29UK	6	0.044021223
Shepway	29UL	1	0.164426502
Swale	29UM	1	0.13100283
Thanet	29UN	1	0.173381644
Tonbridge and Malling	29UP	2	0.059804804
Tunbridge Wells	29UQ	6	0.044159657
Burnley	30UD	5	0.189613704
Chorley	30UE	8	0.050131095
Fylde	30UF	8	0.068392233
Hyndburn	30UG	5	0.186543708
Lancaster	30UH	3	0.142928898
Pendle	30UJ	5	0.218720445
Preston	30UK	3	0.254195219
Ribble Valley	30UL	8	0.120035197
Rossendale	30UM	5	0.026335622
South Ribble	30UN	2	0.143577041
West Lancashire	30UP	2	0.066924572
Wyre	30UQ	8	0.03127123
Blaby	31UB	2	0.120914187
Charnwood	31UC	3	0.076674001
Harborough	31UD	8	0.180690054
Hinckley and Bosworth	31UE	2	0.186027742
Melton	31UG	8	0.015416378
North West Leicestersh	31UH	2	-0.001524679
Oadby and Wigston	31UJ	2	0.077390965
Boston	32UB	1	0.114300261
East Lindsey	32UC	1	0.049967172
Lincoln	32UD	3	0.134752798
North Kesteven	32UE	8	0.120364293
South Holland	32UF	1	0.096776098
South Kesteven	32UG	8	0.059947973
West Lindsey	32UH	8	0.196725723
Breckland	33UB	8	0.070453843
Broadland	33UC	8	0.126781676
Great Yarmouth	33UD	1	0.17130184
Kings Lynn and West Norfolk	33UE	1	0.147634276
North Norfolk	33UF	1	0.10168363
Norwich	33UG	3	0.128449812

South Norfolk	33UH	8	0.189187022
Corby	34UB	5	0.129222216
Daventry	34UC	8	0.077768451
East Northamptonshire	34UD	8	0.02003937
Kettering	34UE	1	0.171223258
Northampton	34UF	3	-0.022385812
South Northamptonshire	34UG	8	0.118072125
Wellingborough	34UH	8	0.055615821
Alnwick	35UB	8	0.198098923
Berwick-upon-Tweed	35UC	8	0.056574889
Blyth Valley	35UD	5	0.146452503
Castle Morpeth	35UE	6	0.051278443
Tynedale	35UF	8	0.066068791
Wansbeck	35UG	5	0.171715487
Craven	36UB	8	0.096192259
Hambleton	36UC	8	0.080553833
Harrogate	36UD	8	0.010815925
Richmondshire	36UE	6	0.036797363
Ryedale	36UF	8	0.068566594
Scarborough	36UG	1	0.159745925
Selby	36UH	8	0.027924849
Ashfield	37UB	1	0.075913564
Bassetlaw	37UC	1	0.078642602
Broxtowe	37UD	6	0.042885267
Gedling	37UE	2	0.169740669
Mansfield	37UF	5	0.089157199
Newark and Sherwood	37UG	1	-0.014001597
Rushcliffe	37UJ	6	0.053695228
Cherwell	38UB	4	0.075981605
Oxford	38UC	3	0.052277212
South Oxfordshire	38UD	6	0.14775152
Vale of White Horse	38UE	6	0.141882343
West Oxfordshire	38UF	6	0.120257717
Bridgnorth	39UB	6	0.130610439
North Shropshire	39UC	8	0.110053135
Oswestry	39UD	1	0.090601774
Shrewsbury and Atcham	39UE	4	0.152304685
South Shropshire	39UF	8	0.174451873
Mendip	40UB	8	0.044798134
Sedgemoor	40UC	1	0.108649252
South Somerset	40UD	1	0.011137046
Taunton Deane	40UE	1	0.229695742
West Somerset	40UF	8	0.086029437
Cannock Chase	41UB	4	0.057278699
East Staffordshire	41UC	1	0.052471441
Lichfield	41UD	8	0.032631606
Newcastle-under-Lyme	41UE	2	0.091968207
South Staffordshire	41UF	2	0.085706562
Stafford	41UG	4	0.156863105
Staffordshire Moorland	41UH	2	0.200724761
Tamworth	41UK	2	0.13290622
Babergh	42UB	8	0.152761877
Forest Heath	42UC	1	0.025110002
Ipswich	42UD	4	0.221304029
Mid Suffolk	42UE	8	0.153850776
St. Edmundsbury	42UF	8	0.07502529
Suffolk Coastal	42UG	8	0.043970787
Waveney	42UH	1	0.252168068

Elmbridge	43UB	6	0.225493024
Epsom and Ewell	43UC	6	0.17392088
Guildford	43UD	7	0.106629546
Mole Valley	43UE	6	0.163801809
Reigate and Banstead	43UF	6	0.132999491
Runnymede	43UG	7	6.27949E-05
Spelthorne	43UH	6	0.150753032
Surrey Heath	43UJ	6	0.153275703
Tandridge	43UK	6	0.142928097
Waverley	43UL	6	0.248980314
Woking	43UM	6	-0.005119353
North Warwickshire	44UB	2	0.115220672
Nuneaton and Bedworth	44UC	5	0.0664007
Rugby	44UD	4	0.100899664
Stratford-on-Avon	44UE	8	0.113599495
Warwick	44UF	4	0.175781053
Adur	45UB	8	0.11177529
Arun	45UC	1	0.153038987
Chichester	45UD	8	0.081627159
Crawley	45UE	4	0.109305666
Horsham	45UF	8	0.110028743
Mid Sussex	45UG	6	-0.013932692
Worthing	45UH	1	0.162170689
Kennet	46UB	6	0.066312659
North Wiltshire	46UC	6	0.024589573
Salisbury	46UD	6	-0.063255491
West Wiltshire	46UF	1	0.20076191
Bromsgrove	47UB	8	0.098665152
Malvern Hills	47UC	6	0.07045531
Redditch	47UD	5	0.056413285
Worcester	47UE	4	0.230120793
Wychavon	47UF	8	0.214879529
Wyre Forest	47UG	5	0.068376854
Aberdeen City	60QA	3	0.305642485
Aberdeenshire	60QB	2	0.078034384
Angus	60QC	5	0.093767281
Argyll & Bute	60QD	8	0.145183283
Scottish Borders	60QE	1	0.233578716
Clackmannanshire	60QF	5	0.013312449
West Dunbartonshire	60QG	5	0.188509445
Dumfries & Galloway	60QH	1	0.016538813
Dundee City	60QJ	3	0.285121219
East Ayrshire	60QK	5	0.194916782
East Dunbartonshire	60QL	2	0.225341919
East Lothian	60QM	1	0.006123397
East Renfrewshire	60QN	2	0.212338418
Edinburgh City of	60QP	3	0.225291169
Falkirk	60QQ	5	0.097501027
Fife	60QR	5	0.133306724
Glasgow City	60QS	3	0.18014424
Highland	60QT	1	0.045093192
Inverclyde	60QU	5	0.246154943
Midlothian	60QW	2	0.157533082
Moray	60QX	4	0.047631082
North Ayrshire	60QY	5	0.164624582
North Lanarkshire	60QZ	5	0.205735013
Orkney Islands	60RA	5	0.045122478
Perth & Kinross	60RB	1	0.025483738

Renfrewshire	60RC	5	0.245167757
Shetland Islands	60RD	2	0.041386715
South Ayrshire	60RE	5	0.160636151
South Lanarkshire	60RF	5	0.229154007
Stirling	60RG	3	0.074403602
West Lothian	60RH	1	0.028223804
Eilean Siar	60RJ	2	0.079058923