
Chapter Three - Making a Classification System: a Guide to Methods and Procedures

3.1 Introduction

The biggest question in many areas of investigation is how to organise observed data into meaningful structures? Clustering of data enables this to take place (Tyron 1939). Cluster analysis is not a typical statistical test, but a process in which a cluster algorithm is used to assign each object into a group of similar objects. Each object is represented by a point in multi-dimensional space; each dimension representing a different variable, the values of which fix the location of each object (Anderberg 1973). Unlike many statistical procedures cluster analysis does not require a prior hypothesis. It is very much a technique in the data exploration phase of research.

Although objects are clustered according to similarity, it must be noted that the variance between values within a cluster can be as large, or larger than the difference of values the between two classes. The view that the areas can be classified into mutually exclusive groups must therefore be challenged. Classifications must be viewed as object sets of fuzzy groupings where the outer points of each classification can overlap resembling clouds in a summer sky (Voas and Williamson 2001a). There are many different methods of classifying objects into groups of similarity. Geographic areas are the objects to be clustered in this instance.

Area Classifications are created by the clustering of geographical entities with the use of cluster analysis. The process of cluster analysis, although based on a fairly simple clustering algorithm, is much wider than the clustering of the objects themselves. To run a cluster analysis and therefore create an area classification requires a series of steps, with multiple decisions to be made at each stage (Milligan and Cooper 1987). Each decision has an incalculable, but real effect on the result of the analysis. This makes classification as much of an art form as a science. There is a great deal of skill and knowledge required to make these decisions with confidence. There are no right or wrong answers to any of the decisions that have to be made, they merely

produce different results. Consequently different decisions could be more or less suitable dependent on the purpose of the classification that is to be created (Lorr 1983).

The steps involved in cluster analysis are excellently summarised by Milligan (1996), who outlines the 'seven steps of cluster analysis'. Milligan's seven steps were further summarised by Everitt *et al.* (2001) who add their own comments and ideas to Milligan's framework. The steps are described as "fairly predictable" (Milligan 1996 p341). Each step represents a major or critical decision that has to be taken to successfully run a cluster analysis. While recognising that, dependent on application, some steps may be more or less important than others. Milligan suggests it is vital that the user recognises the critical decisions that need to be made, and the importance that they may have on the final results. A clear distinction needs to be made between cluster analysis and clustering method. The clustering method is simply the method by which the clusters are formed, while cluster analysis refers to the much wider sequence of steps that have to be followed to complete the whole analysis. Cluster analysis is much more than simply running a dataset through a clustering algorithm (Milligan 1996).

It is essential for users of cluster analysis, especially those hoping their classifications will be used by others, that they record and report decisions taken at each step of the cluster analysis and the reasoning behind each decision. This enables others to not only critically evaluate what the researcher has done, but also gives them the possibility of adding to, or extending the results of the analysis (Milligan 1996). There are many examples of authors who have failed to provide significant information about the decisions taken. Milligan sites Harrigan (1985) who failed to even name the clustering method that was used in the study. Although examples such as Harrigan (1985) can be found within academic literature, no one is as guilty of failing to provide information about the creation of classifications and the steps used in cluster analysis as the firms who create and license out commercial geodemographic classifications.

Harris *et al.* (2005) recognise that little is known about how geodemographic classifications are built or what information goes into them. While appreciating that the problem exists due to the constraints of commercial confidentiality, they fail to point out the implications for anyone who wishes to use these potentially rich data sources in an academic study. A link has been made between a commercial geodemographic firm and a geography department at a high ranking British university. Researchers at the institution have been given free access to a commercial geodemographic classification to aid their research, although there is no way of knowing exactly how much information the commercial firm has passed to the institution about the creation of the classification system. None of the information about the system is available outside the institution. The following question has to be asked. Can any of the research that they have

conducted be considered valid if no external party is privy to any of the information within the classification?

Milligan's seven steps are outlined below with comments comprising an amalgamation of the original description by Milligan (1996), additional comments by Everitt *et al.* (2001) and the addition of some further points that relate more directly to area classification.

Step 1. Clustering elements (Objects to cluster, also known as "operational taxonomic units")

- a. Should where possible be defined to give a 100% geographical coverage.
- b. Should be representative of the cluster structure believed to be present.
- c. Should be sampled properly if generalisation to a larger population is required.

Step 2. Clustering variables (Attributes of objects to be used)

- a. The variables represent the measurements taken on each entity/area that is to be clustered.
- b. Variables should only be included if there is a good reason for their presence such as adding definition to the clusters.
- c. Irrelevant or masking variables should not be included as they can hide more significant patterns within the clusters.

Step 3. Variable standardisation

- a. There is no requirement that standardisation must be performed on any set of data. It is up to the researcher to decide if standardisation is necessary and if so which method should be used.
- b. Standardisation over the range of each variable shows a good recovery of clusters (Milligan and Cooper 1988).

Step 4. Measure of association (Proximity measure)

- a. A measure of similarity or dissimilarity must be selected. This reflects the degree of closeness or separation between objects to be clustered. These can work in different ways. For example, Euclidean distance as a dissimilarity measure reports larger values as two entities become less similar, so that the distance between them in Euclidean space is greater. In contrast a similarity measure such as a Pearson correlation, assumes the opposite reporting larger values as two objects become more similar.
 - b. Either linear or non-linear measures can be used.
 - c. Few general guidelines. However, knowledge and context of the data may suggest an appropriate measure.
-

Step 5. Clustering method

- a. Methods used should be those designed to recover the type of clusters suspected to be present. This is important as different types of clustering method are better at finding different types of cluster structures.
- b. Robustness of method. Some are able to handle different amounts of data, and show different amounts of sensitivity to certain types of data.

Step 6. Number of clusters

- a. This is the most difficult decision to be made in cluster analysis. It is especially troublesome if there is no prior information as to the number of clusters expected to be in the dataset.
- b. There are several different rules that can be followed for the selection of the most suitable number of clusters. However, these can often be contradictory for the same application.
- c. If you can't choose between two solutions, then the larger number of clusters should be selected.
- d. You also need to consider if there are actually any clusters present with the data. When there is no obvious difference between the different solutions produced.
- e. There is no right answer to the selection of the number of clusters. The choice is not based on scientific theory and the solution selected should be judged on its usefulness rather than being a correct representation of the patterns within the dataset.

Step 7. Interpretation, testing and replication

- a. Interpretation of the results in the context of the applied problem and an assessment of whether the solution adequately meets the needs of the investigation should be undertaken. This requires knowledge and expertise in the discipline in which the investigation has been carried out.
- b. Re-run the analysis to make sure the same solution is found on all occasions.
- c. Test to determine whether there is a significant cluster structure within the data. Follow by cross-validation to instigate if the clusters are representative of data not originally included in the analysis.
- d. Perturbation: examine of the difference to the result by the removal of each of the variables included in the analysis.

Milligan's seven steps provide a good outline of what is involved in the creation of a classification. However, Milligan provides only general guidelines for all datasets; adaptations

will need to be made depending on the specificities of the dataset being clustered, in this case, spatial data and the creation of a general purpose area classification. The creation of an area classification can be seen as a combination of three general steps, inputs, processes and outputs (Harris 1999).

Section 3.2 outlines the inputs into a classification system, including data issues and perspectives about selecting variables. Section 3.3 introduces the processes involved in cluster analysis, including: standardisation, weighting and clustering. Section 3.4 discusses issues relating to the output of area classification systems. Section 3.5 concludes with a summary of the chapter.

3.2 Inputs

The inputs to an area classification are spatial data, predominantly areal data at whatever scale of geography the classification is to represent, but any form of geo-referenced individual or point data can also be used. There are several view points on the data that should be included in an area classification. Areal data are generally more geographically comprehensive and show greater stability over time than point data, due to the aggregating effect of the areal units. Individual or point data can often provide additional information that is not available for areal units. However, individual data rarely have 100% geographic coverage and are much more susceptible to change over time.

3.2.1 Data, the More the Better?

Commercial companies advertise their geodemographic classifications as having hundreds of variables, suggesting that their classification is better than their competitors as it is built from more data. However, there is little evidence to suggest this is true. Milligan (1996) contends that the opposite to be true, suggesting that variables should only be included if there is a very good reason. The inclusion of less relevant variables can mask and reduce the effectiveness of more relevant variables within the clustering process. Multidimensional datasets are very difficult to understand and difficult to represent graphically. Adding further redundant information to the classification process serves no purpose other than to make the results of the analysis harder to interpret and unnecessarily complicated (Milligan 1996).

3.2.2 Census Data

The UK Census provides much of the data needed to create a geodemographic classification; on the 29th April 2001 the numbers and composition of the present day UK population were surveyed with the undertaking of the 20th Census of UK population. The Census of the UK has

taken place once every ten years since 1801 (with the exception of 1941 due to World War II). As well as the decennial census there was also a 'sample' census which took place once in 1966, but was not repeated. The territorial extent of the UK has not stayed constant during the history of the census: from 1801-1911 the United Kingdom represented Great Britain and Ireland, but following the Irish war of independence (1919-21) a north-south partition of Ireland was established and the South of Ireland gained independence from the UK. Therefore, from 1921 onwards the census of the UK represents Great Britain and Northern Ireland, but not the South of Ireland.

The Census is the most complete source of information on the number, characteristics and location of the UK population. The census collection and dissemination process requires three main components: firstly, the people who complete census returns; secondly, the census offices who are responsible for the collection, editing and production of data; and thirdly, the licensed census partners who disseminate census data and produce value added data products (Rees and Martin *et al.* 2002). The importance of the census should not be underestimated, as results from the census provide an input into a large number of the reports, findings and policies of both national and local government, research into decisions to open or close facilities such as schools, hospitals, and clinics use information revealed by the census (Boyle and Dorling 2004). Census information is also a valuable tool for marketing companies, business planners and academic researchers (Raper *et al.* 1992).

Superficially a census looks relatively simple as it is based on a form containing only 20-40 questions (Dale 1993). However, each of those questions has a number of categories, each of which can lead to an indicator (e.g. the percentage of the population aged under 5, the percentage of the population aged 5 – 14, the percentage of the population aged 15 -19 etc.). The categories of one question can be crosstabulated against several others. For example, the percentage of females in a particular age band who are married, and who work full time. This piece of information involves the answers to four separate questions. The number of possible crosstabulations and associated indicators is a number of gargantuan proportions; the number of sub-populations for which crosstabulations and indicator variables can be generated is also very large. There were 223,060 output areas generated in the 2001 Census of the United Kingdom. Even if the results are confined to a simple rectangular matrix of 223,060 rows by say 10^6 indicators, this will produce $2.23E11$ or 2.23×10^{11} cells of information.

Data available from the UK Census include such things as Population present, Population resident, Age, Living arrangements, Marital status, Country of Birth, Ethnic Group, Religion, Health and provision of unpaid care, Economic activity, Hours worked, Industry of employment, Occupation groups, Qualifications and students, National Statistics Socio-

economic Classification, Travel to work, Household spaces and accommodation type, Cars or vans, Tenure, Rooms and amenities, Household composition, Communal establishments, Migration and for Wales only, Knowledge of Welsh.

3.2.3 Issues of Census Data Quality

The census is a high quality and comprehensive dataset, but there are still several data quality issues, of which all users of census data need to be aware. It is all too easy to jump into using a dataset without examining issues of quality that would affect how the data should be best used. Knowledge of these issues can prevent even an experienced researcher from falling into some potentially serious traps.

The 2001 Census is the most comprehensive survey of the UK population ever undertaken and attempts to count everyone present in the UK on census day. However, this does not mean that everybody in the UK was counted in the enumeration process. A large number of people failed to fill in and return their census forms; even after follow up attempts to problem residences there were a large number of missing returns. Any people who failed to respond had to be imputed into the census using knowledge of who they expected to find at each non-responding residence (ONS 2003b). Not only was whole record imputation necessary, but answers to individual questions had to be imputed or changed where contradictory responses had been given. This would be enough of a problem, but the response rate differs greatly by geography and demography (ONS 2005a). People were much less likely to not respond to the census in the centre of large cities than in less urban areas. Irregular housing patterns in these areas did not help in this matter. Younger people, especially young males, were less likely to return their census form than other sections of society (Simpson 2002). Some people such as those who live in the UK illegally were unlikely to fill in their forms for fear the information given could be used against them (ONS 2005a).

Another issue relating to census data is its time reference or currency. The census is only carried out once every ten years and some of the data are not released until up to four years after the census enumeration. This means some 2001 Census data will still be in use in 2015, fourteen years after its capture. While population estimates are made for the intervening periods, this only helps with the actual number of people not their social make-up. Simply by examining the average migration rate of 12% a year, would mean that only 17% ($100 \times (1-0.12)^{14}$) of people will live in the same place in 2015 as they did in 2001. When births and deaths are also taken in to account, can we have any confidence in the long term value of census data? Well to a certain extent yes we can. When some body moves out of an area they are likely to be replaced by someone broadly similar in terms of socio-economic status. Residential social patterns change

only very slowly over time (Orford *et al.* 2002). Although census data will of course date over time, it should still be broadly representative of the population present at the time of the next census. However, with the passing of time users of census data must be aware that the data are constantly ageing. Precautions can be taken to limit the effect of this ageing. For example by using the largest feasible geographic scale of data.

The final major issue of census data quality is that of disclosure control. The census agencies have an obligation to anonymise the data that they produce to ensure that the characteristics of any one person are not disclosed. For aggregate statistics of large areas this is not a problem as the counts of people in each cell are likely to be numerous. However, for much smaller geographies such as Output Areas (OAs) the chances of finding a cell with a single count are much higher, therefore the data are altered to prevent the identification of people and the disclosure of information about them (Rees and Martin 2002). The most obvious evidence of disclosure control in the 2001 Census is the absence of any ones or twos in any of the census data. This presents itself as an abundance of zeroes and threes in the dataset, especially at the output area scale. Much comment has been made about the methods of disclosure control used in the 2001 Census as they have had a much greater effect on the data than methods used in previous censuses (Williamson 2005). The problem becomes especially apparent when multiple variables are used when small numbers acting together can produce values of over 100%. There is little that can be done to overcome these problems, although the problem is not a serious one for the creation of an area classification as it is the large numbers displaying distinctive patterns that are indicative, not the small numbers. The problem will come when a variable is considered for inclusion that only has a very small membership nationally. The difference between zero and three for these kinds of variables could have a significant impact on the classification especially at OA scale. Careful consideration will have to be given before including such variables in the classification.

3.2.4 Other Data Sources

As well as using data derived from the 2001 Census, other data sources can be used to supplement the census information and hopefully add new dimensions to the classification. However, the principal role of adding such data to a census based classification should be to provide data that is not provided in the census. The most obvious topic not covered by the census is information on income and wealth. Commercial geodemographic firms add non-census data to their classifications to principally provide information on wealth and affluence that is the major weakness of the census, which adequately provides information on the very poorest in society but struggles to identify the best off quite so well. Commonly used non-census datasets that are used in commercial classifications include the electoral roll, county

court judgements, Land Use Surveys, Financial data such as share ownership, Monthly updated unemployment figures, Indices of Deprivation 2004, Land Registry data, DVLA - Car Registration, DfES - School results, Lifestyle data from consumer loyalty schemes, Credit referencing data and Companies House data (Sleight 2004). Additional data sources that can be used will depend heavily upon availability, confidentiality, and the spatial scale at which the data are produced. Some of the data are freely available or available on request and under licensed conditions. However, some data such as credit reference information and lifestyle data are only available for inclusion in commercial classifications as they are collected by the companies who make the classification or companies with whom they have data sharing agreements.

3.2.5 Data Quality Issues from Other Data Sources

There are obvious benefits to adding non-census data to a classification to enrich the pool of information from which the classification is built; additional data can provide information that is not available from the census. Another benefit of non-census data is that it is likely to be updated at much more regular intervals than the census data, often annually and in some cases monthly.

There are several dangers that should be taken into account when using different sources along with data from the census. Firstly, the accuracy of the data has to be assured even from reputable sources such as other government departments. It is important to know how and when all the data were constructed. Few datasets are as well documented as the census in terms of the enumeration and processing methods and it is unusual to find significant data support for any of these other data sources. The coverage of the data will not be the 100% that is available in the census. Most of these datasets will be based on a sample of the population unlikely to represent more than 10% of the country. Additionally these samples are unlikely to be representative either geographically or demographically; certain sections of society are likely to be over or under represented in the data. There will also be many other sources of uncertainty, it is impossible to know if the data contain undocumented and unidentifiable errors.

A classic example of errors that can be put into a classification system by adding data from other sources is the set of DVLA statistics on car registration, which are used by several commercial geodemographic systems. Car firms register their cars at the factory to facilitate a quick sale especially when a new model is released. This can be seen clearly in Experian postal sector data which shows the Swindon postal sector SN5 6 to have a population of 5,034 and 106,644 registered cars which works out at 21.2 cars per person including children and non-drivers. Postal sector SN5 6 is not home to several thousand car collectors; it is the location of a

Honda car factory (Vickers 2003). The dangers of using such data are obvious to see. Even with the removal of such anomalies from the dataset it would be very difficult to have any confidence in the rest of the data.

Data from other sources should only be used if they add further dimensions to the classification. By adding more variables to a classification that just reinforce trends already provided by the census data only hinders the formation of the classification through the added complexity that it brings to the variable list. It is important that any non-census data which is brought into the classification are at the same spatial scale and refers to the same geographical system. Although several methods of transferring data between systems of spatial registration have been formulated and indeed some are used widely, no system has yet been formulated that can transfer data between overlapping areal units to a satisfactory level of accuracy (Vickers 2003).

3.2.6 The Theory of Selecting Input Variables

The goal of the variable choice for the creation of an area classification is to select the minimum possible number of variables that satisfactorily represent the main dimensions of the 2001 Census and therefore, to get the most information possible in the fewest variables possible into the classification (Bailey *et al.* 1999a, 1999b and 2000). Although in the previous section the use of non-census data was discussed, for simplicity here only choices and comparisons between census variables will be discussed.

There are two main reasons why the minimum possible number of variables should be used to avoid co-linearity and to reduce computational demands. To prevent co-linearity, each variable that is included should add something that the other variables do not give to the classification. As the data are all from the same sources and about the same geographic areas, it is likely that any selection of variables will contain a host of interrelated variables. The problem that co-linearity gives is that it makes it difficult to assess the effect that a variable is having on the classification. It will not only have its own characteristics, but will be working with any correlated variables making it difficult to assess the strength of effect each is having on the classification. The more variables that are added to the classification the less likely they are to add any new information, and the more likely they are to be just repeating information which is covered by one or more variables already selected.

Voas and Williamson (2001a) go beyond suggesting that fewer variables should be used, calling for the increase in 'problem-specific' classifications. In response to comment by Harris (2001) on their paper the same authors suggest that "*By conflating a range of marginally correlated measures, such as income and newspaper readership, there is an inherent tendency to obscure*

the actual between-area differences on matters of specific interest” (Voas and Williamson 2001b p335). Voas and Williamson (2001b) make the important point that, by adding in every variable which the architect of a classification can get his/her hands on, patterns of interest can be clouded by other irrelevant variables. They see this as a reason to change tack and produce a series of ‘problem-specific’ classifications. However, one could take a different stance on this point to say that the process of variable selection is a very important one and great consideration should be given to the inclusion/exclusion of any variable. Variables should only be included if they inherently contain information that you wish to be displayed in the geographic patterns represented by the classification.

There is another, less intellectually based reason for selecting fewer variables, which is fundamental to the successful creation of any classification. The greater the number of variables used the more computer processing power is required to generate the clusters and therefore the more time it takes for the procedure to run. If the number of variables to be clustered goes beyond a certain level, it could go past the current capabilities of the computer.

Selecting the fewest possible variables is not an argument that is put across by everyone. Many people reason the more variables used the better. Harris *et al.* (2005) state that: “*As a general rule, but with limits, the more variables that are used in the clustering algorithm and the more different sources they come from the more meaningful (nuanced not idiosyncratic) the resulting set of clusters is likely to be*” (Harris *et al.* 2005 p151). This view is not supported by anyone except commercial geodemographics companies. Academic literature especially within mathematics suggests that the minimum possible number of variables should be used (Everitt *et al.* 2001).

It is not easy to gauge the opinion of the creators of commercial classification systems on this issue as they have traditionally been reluctant to disclose how their classifications are made. However, it is generally regarded that they have a “the more the better attitude” in terms of variables. The latest Mosaic UK brochure states that “*A total of 400 data variables have been used to build Mosaic*” (Experian 2005). A EuroDirect promotional leaflet for their Cameo classification claims it includes “*over 9,000 pieces of information for over 150,000 Census units*” (EuroDirect Unknown p5) while their current website states that the latest version contains “*over 2 billion items of data*” (EuroDirect 2005). So what evidence is there of the effect that too much data can have on a classification? Harris *et al.* (2005 p160) displays a table of the age data used in the creation of Mosaic UK, which contains 22 age groups as follows: Aged 0-4, Aged 5-9, Aged 10-14, Aged 15-19, Aged 18-24, Aged 20-24, Aged 25-29, Aged 25-44, Aged 30-34, Aged 35-39, Aged 40-44, aged 45-49, Aged 45-64, Aged 50-54, Aged 55-59,

Aged 60-64, Aged 65+, Aged 65-74, *Aged 75-84*, **Aged 85+**, **Aged 85-89** and Aged90+ (the highlighting in this list is explained below).

This is a more than comprehensive list of age groups. Some of them even overlap (e.g. Aged 85+, Aged 85-89 highlighted in bold), therefore covering the same information twice. How strongly are these two variables related? How much new information does including them both give us? By running a simple Pearson's Correlation on the two variables the level of redundancy can be established. The analysis produced a statistically significant correlation of 0.941, a very strong relationship. By multiplying the result by one hundred then squaring that the percentage of one variable which is associated with the other can be calculated, that's $(0.941 \times 100)^2 = 88.5\%$. This shows that by having just one of the variables in the selection we also get 88.5% of the other or another way of looking at it by adding the second variable the amount of new information that is gained is only 11.5% of the information that is given by the first variable. This is not a surprising result as the two variables overlapped so shared information, but they were both included in the Mosaic system.

How much of a relationship do two contiguous variables show? To test this Aged 85-89 will be correlated with Aged 75-84 (*Italics*). The analysis produced a statistically significant correlation of 0.682, not as strong a relationship as last time which is not unexpected as the variables are no longer overlapping, but the relationship is still statistically significant. How much new information does the variable Aged 75-84 give? $(0.682 \times 100)^2 = 46.5\%$, this shows that only just over 50% of the variable is new information. The rest is associated with the other variable. It may seem intuitive that these variables are as highly correlated as they are similar age groups will have similar residential needs/preferences. An important point to note is that in datasets with a very large number of observations such as this, even very low correlations will be significant simply because of the large number of data points involved.

Is there then any relationship with a variable that is not as similar? Aged 0-4 (underlined) is at the other end of the age scale and therefore the reasons of high correlation because of similarity should not be present. By correlating Aged 85-89 and Aged 0-4 the result gives a statistically significant correlation of -0.305, with a variance of $(-0.305 \times 100)^2 = 9.3\%$. The correlation is much less than those previously seen but it is still statistically significant. We have almost 10% redundancy, why is this? The reason for the 10% data redundancy between these two variables is for a different reason from the previous variables. The variables do not overlap and they are not contiguous, so why does one explain 10% of the other? The clue is the negative nature of their relationship. The previous relationships were positive; this is because the two variables were inherently very similar. These two variables are, however, not very similar and are at different ends of the life course. However, the reason for the redundancy is a fairly simple one,

in that they are both age variables. The redundancy is caused by each person only being able to be in one age category all the age variables will show a certain level of inter-redundancy because of this.

Negative redundancy between categories within the same variable will always be experienced and is something that cannot be avoided. However, there are some things that can be done to reduce its effect. If age variables are going to be used in the classification and we have n different age groups to get all the information about age into the classification we only need use $n-1$ groups. Why is this? For the same reason that 10% of the prevalence of 0-4 year olds is explained by 85-89 year olds' negative inter-correlations. By using $n-1$ all the information is still being used even though all the groups are not present, because of inter-dependency. Despite dropping any one of the groups within a variable the data are still there. A simple example of this is households with access to a car and households who do not have access to a car. The inter-dependency here should be clear to see: by having a car, you cannot also not have a car and can therefore only be in one group. Adding both variables into the classification does not give any more information than just using one. In fact, a variable has been inadvertently double weighted as the same data would have been used twice. If you look back at the list of age variables used in the Mosaic classification you can see that there is no gap in the age categories each one has been used, they have used n not $n-1$. This is unnecessary as it does not yield any more information.

The examples given show the type of considerations that need to be addressed when selecting variables for the classification. Inter-correlation and inter-dependency between variables are to be avoided. It is harder to interpret and understand the patterns that are produced and the reasons behind them.

3.2.7 Correlation Analysis

Correlation techniques have some idiosyncrasies that are important to keep in mind before using such techniques. Variables which share the same denominator (i.e. calculated as a percentage) have a natural tendency to produce a negative correlation (Miles and Shevlin 2001). There is for example a strong inverse relationship between the married and single population. In some cases this effect can be difficult to untangle from a genuine negative correlation. The relationship between the people who have no car and those who have numerous cars has not only a technical negative correlation, but one that also shows social importance (Voas and Williamson 2001a). Variables that don't share the same denominator can also show a close relationship. The number of married men shows significant relationship to the number of women who are married. This

can be traced back to the fact that they are derived from the 'marital status' question on the original census form, which has then been sub divided by gender (Voas and Williamson 2001a).

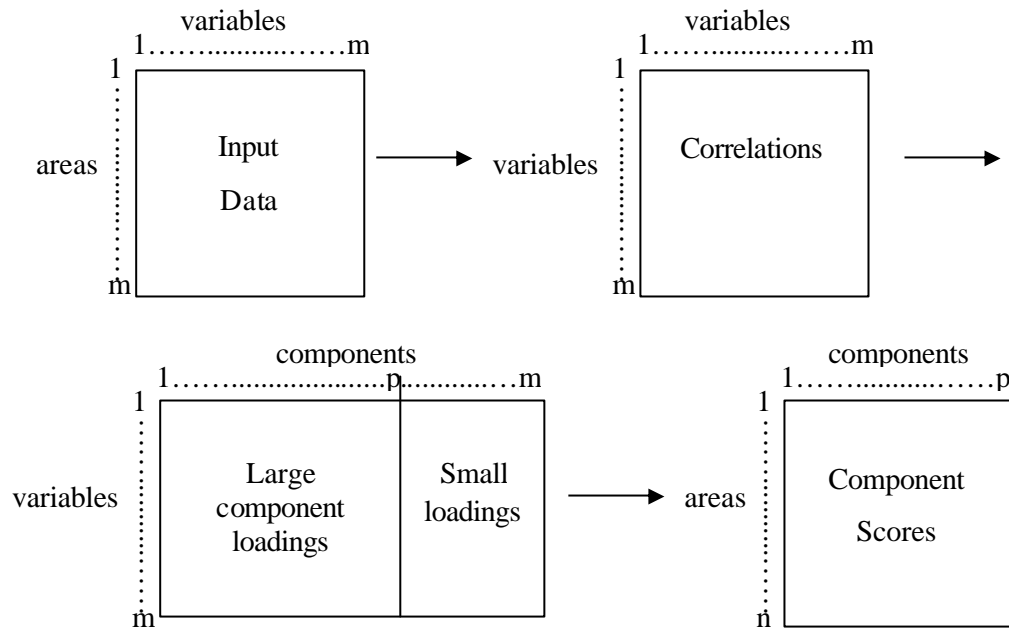
3.2.8 Data Reduction

In the past, mainly due to computational limitations, data reduction techniques have been employed on the variable list prior to clustering. Principal Components Analysis (PCA) is not a method of classification, but a preparatory technique used to remove redundancy from the variable list. By calculating the correlation of each variable with all others, redundancy can be reduced by removing one of a pair of variables that are highly correlated (Voas and Williamson 2001a). By removing redundancy from the dataset this not only makes classification techniques quicker and easier to run, but also enhances the effect of the less correlated variables on the classification.

The basic assumption made by PCA is that a few underlying components or factors within the data can be used to explain the complex relationships within the whole dataset (Norusis 1985). Correlations between the data show that the correlated variables share a dimension of commonality. The prime aim of PCA is to identify these non-directly observable factors based on a set of chosen observed variables. PCA has been widely used by geographers since the 1960's although its complexity is rarely appreciated by researchers who make use of it via off the shelf computing statistics packages (Robinson 1998). It is based on the application of Pearson's product-moment correlation to a standard geographical matrix of places versus a series of observed statistics about those places (Rummel 1970).

Starting with a matrix of n areas by m variables, the aim is to reduce this to a matrix of n areas by p important components, where the p components are combinations of the original m variables. The number of p is less than m because the components that are associated with only a small amount of the variance of the input dataset are ignored. The matrix scores of the p components are then inputted into the cluster analysis as a substitute for the original variable set. This sequence is outlined in Figure 3.1.

Figure 3.1: The sequence of Principal Component Analysis



The advantages of using PCA are that it removes variable redundancy from the dataset, focusing on the main patterns. The disadvantage is that not using the original variable set makes the resulting cluster profiles difficult to interpret as the component scores are composites. This results in the additional labelling problem of naming each of the component. It is sensitive to the magnitude of correlations between the variables. It is sensitive to outliers, missing data, and poor correlations between variables. Outliers must be screened out due to their influence upon the calculation coefficients, which in turn has a strong influence on the calculation of factors and components. In effect, it is a variable reduction technique, equivalent to reducing the number of variables based on correlations between them. However, examining and assessing the correlations between variables is a much more transparent way of reducing redundancy within the dataset.

With the continued increase in computational power, it is no longer essential to run data reduction techniques such as PCA on the variable list prior to clustering. So does PCA have any intrinsic value on top of its function data reduction or does it no longer need to be used? PCA still has value as a useful tool for assessing the predictive power of variables prior to clustering. The values of each component can be used to assess the likely discriminatory power of each variable prior to clustering. The variables which have high values for the early principal components represent those that are likely to have the most discrimination within the clustering process. PCA can therefore be used to make an assessment of the predictive power of variables prior to clustering. However, clustering on principal components rather than the variables themselves is an outdated and unnecessary course of action.

3.3 Processes

The processes involved in the creation of an area classification are more than just the procedure of clustering itself. The data must be prepared for clustering. Clustering algorithms are sensitive to difference in scale (e.g. tens and thousands) and types (e.g. ordinal, ratio or interval). If these issues are not attended to before clustering begins then it is likely that any clusters produced will be a feature of the format of the data rather than the actual data values.

3.3.1 Methods of Standardisation

Before any clustering can be done the variables need to be standardised. This ensures that each variable has the same weighting in the classification. This is especially important when there are different types of data e.g. *population density* will give number of people per unit area, whereas *detached housing* is a percentage of all households. The range of the *population density* is only limited by the number of people who can fit into a specified area. In the UK at OA scale population density ranges from just above 0 to 12,715 people per hectare for OAs whereas housing type can only range between 0 and 100%. These variables are not on the same scale. If left un-standardised the population density would completely control the classification because of the larger range over which the data are stretched. This would also create a large number of outliers based solely on the population density variable. Therefore if these variables were clustered without being standardised it would add bias to the clusters.

All clustering techniques are based on the similarity or dissimilarity of the cases to be clustered. This is measured by constructing a distance matrix reflecting all the variables in the data set for each case. It is clear that problems will occur if there are differing scales or magnitudes among the variables. In general, variables with larger values and greater variation will have more impact on the final similarity measure. It is necessary to therefore make each variable equally represented in the distance measure by standardising the data. The process involved in calculating each type of standardisation is outlined in the following sub-sections.

3.3.2 Z-Score Standardisation

This is the most common form of standardisation. To create z-scores or 'standard normal variate' the standard deviation is calculated. The z-score is then calculated by taking the mean value of the variable away from the value for that variable for each area, squaring the difference, adding over all areas, square root the result and then dividing them by the standard deviation of the variable across all areas. This should be repeated for all variables to standardise them over the same range. Let x_i be the value of a variable for area i and x_{mean} the average value of the variable across all n areas.

The standard deviation is defined as:

$$S_x = \frac{\sqrt{\sum_i (x_i - x_{mean})^2}}{n} \quad (3.1)$$

The standard normal variate or z-score is defined as:

$$Z_i = \frac{x_i - x_{mean}}{S_x} \quad (3.2)$$

3.3.3 Range Standardisation

This method was implemented in the ONS 1991 classification of Local Authorities; see Wallace and Denham (1996). The data were standardised by the method of range standardisation between 0 and 1 for each variable. The range standardisation method is defined as:

$$R_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3.3)$$

where x_{max} is the maximum value of x and x_{min} the minimum value of x and R_i is the range standard variate. After the data have been standardised as above each variable has a range of 1 with the maximum value being 1 and minimum value being 0.

3.3.4 Inter-decile Range Standardisation

This method is a slight variation of the range standardisation method, standardising the data over a smaller range.

This method is defined as:

$$D_i = \frac{x_i - x_{med}}{x_{90^{th}} - x_{10^{th}}} \quad (3.4)$$

The inter-decile range standardised variate D_i compares each value of a variable, x_i to the median, x_{med} which is then divided by the distance between the 90th percentile, $x_{90^{th}}$ and the 10th percentile $x_{10^{th}}$.

3.3.5 Weighting of Variables

Unintentional weighting of variables was touched upon in § 3.2.6. The weights being referred to here are intentionally given to certain variables. In the Mosaic system all variables are given weights defined by the creators of the classification. What is meant by weighting variables? If you have two variables a and b to put into a simple cluster analysis, but you think that, although both variables should be used, b is twice as important as a . If this is so the variables can be weighted by multiplying the value of b by 2 (after standardisation), it will have twice the effect on the clustering procedure than variable a .

The variable choice in itself is the start of the weighting process as in effect all the rejected variables are simply being weighted zero in the clustering process, and those that are chosen are given a weighting of one (Everitt *et al.* 2001). However, weighting goes much further and is much more complex than this. There are many different opinions on how variables should be weighted for cluster analysis. There are those who would simply use their knowledge and experience to weight variables. This maybe a satisfactory way of getting a good result, but is something that is hard to explain and therefore difficult to pass on to others. Others have experimented with weighting algorithms that are designed to reduce the influence of variables which are irrelevant to the clusters present within the data (Milligan 1989). As well as reducing the effect of variables that have little effect on the cluster structure it is also considered advantageous to weight the contributory variables to enhance the cluster structure that is present in the dataset (DeSarbo *et al.* 1984). Investigations into the success of weighting schemes have shown that, weighting schemes based upon carefully chosen estimates of within-cluster and between cluster variability are generally more effective than weightings based on standard deviation or range (Gnanadesikan *et al.* 1995).

Nearly all of the research into the success of different forms of weighting has been carried out on task specific classifications, where it is simple to assess how well the weighting has improved the classification by the partitioning of the clusters for that purpose. However, for a general purpose classification such as an area classification the task is all the more difficult, as the success of the partition cannot be compared against another specific application.

The weight given to a variable reflects the investigator's view of the importance of that variable to the task of the classification (Everitt *et al.* 2001). Therefore a question that arises from this is: is it sensible to weight variables in a general purpose classification? Applications of the classification are not known at the time the classification is created. Weighting variables and testing how well they perform against another dataset can not give a good indication of how the weighting has affected the performance of the classification, as an improvement of discrimination against one dataset could reduce its discriminatory powers against another.

The difficulty, impracticality and doubt over the benefit of weighting were considerations that were not lost on Romesberg (2004) who suggests that the weighting of variables for cluster analysis only makes sense when the research goal is clearly defined. For a general purpose classification the research goal is fairly loose as it will be used for multiple applications.

There have only been limited suggestions as to how to weight variables for a general purpose classification. Harris *et al.* (2005) suggests that variables should be weighted so each domain receives the same total weight when the weights of variables within each domain are summed together, but do all domains deserve or require equal weighting? Some domains may have more relevance than others. The domains with the least variables would receive the highest weightings, but exactly which of the variables should be weighted the highest? It is difficult to give an answer to any of these questions. Variables with strange or unreliable distributions could be weighted lowly (Everitt *et al.* 2001). However, it could be argued that they should not be included at all unless they are absolutely vital variables. It has been suggested that those variables which show the most discrimination should be weighted highest (Everitt *et al.* 2001). Is this really necessary though as these variables are already the most discriminatory? This would be fine in a specific purpose classification, but in a general purpose classification as much of a case could be made for weighting the least discriminatory variables higher to try and get more discrimination out of them.

Makarenkov and Legendre (2001) suggest that weighting procedures should be used to eliminate noisy variables that do not contribute relevant information to the classification structure. However, can any variable be seen as noisy or masking in a general purpose classification when the application is not known and therefore the value of each variable cannot be assessed for all possible uses? The noise that seems to be produced by a certain variable may actually improve discrimination for certain applications of the classification.

It would seem that for a general purpose classification the weighting of variables is as likely to confuse as it is to improve the classification. The main problem being as the classification is not task specific there is no way of knowing if the weightings chosen are any better than an alternative selection as all possible uses that the classification will be put to cannot be known. It is probably more sensible to spend extra time and effort in selecting the list of variables to go into the classification, which is in itself a form of weighting.

3.3.6 Methods of Clustering

The process of classifying information is one that many people have made attempts at redesigning and reinventing. There are positives and negatives to most of the procedures, from the more traditional clustering algorithms to more sophisticated techniques such as neural networks. This section will briefly review alternative clustering methodologies and then a more detailed description of the clustering methods that were used in the project is given in the subsequent sections.

There are many different clustering algorithms. However, many algorithms are either very similar to each other or unusual or designed for a specific purpose, and therefore rarely used. There are a few commonly used types that are favoured mainly due to their reliability and transparency. The most commonly used can be grouped into two broad types: hierarchical agglomerative and iterative relocation (Harris *et al.* 2005). Everitt *et al.* (2001) and Gordon (1999) give excellent descriptions of many different clustering methods.

Hierarchical agglomerative or stepwise clustering methods are top-down approaches to clustering. This is one of the conceptually simpler approaches, where each object starts separately and is joined together one at a time creating a cluster hierarchy, of every cluster number from n to 2 (Harris *et al.* 2005). The main advantage of this method being that it produces multiple cluster solutions with just a single running of the algorithm. The hierarchical nature of the system enables more than one solution to be selected and used simultaneously without contradiction (Everitt *et al.* 2001). The major disadvantage of this is that the top down approach takes a long time to compute and is consequently difficult to implement on datasets containing more than about one thousand objects (Harris *et al.* 2005). Although multiple solutions are created, because of the hierarchical nature of the classification one optimal solution is unlikely to be produced. Each new cluster level is created by the merging of two clusters from the previous level. The hierarchy does not allow objects to move between clusters with the increase in the number of clusters (Romesburg 2004). There are multiple forms of agglomerative clustering; available alternatives in the SPSS system are between-groups linkage, within-groups linkage, nearest neighbour, furthest neighbour, centroid clustering, median clustering, and Ward's method (SPSS Inc. 2001). Ward's method is the most commonly used of these methods and appears to work well, although can sometimes impose a general spherical cluster where one does not necessarily exist (Everitt *et al.* 2001).

Divisive or 'de-agglomerative' methods work in the opposite way to agglomerative clustering methods, with all objects starting in one large cluster and successively splitting into more and more clusters (Everitt *et al.* 2001). Often used with binary data for which the method is efficient on a simple presence/absence basis. Less commonly used than agglomerative methods,

their main advantage being that the main structure of the dataset is revealed from the start of the clustering rather than towards the end of the clustering as in agglomerative methods (Kaufman and Rousseeuw 2005). The method is computationally demanding if all possible sub-divisions are considered at each stage of division (Everitt *et al.* 2001).

The k-means iterative relocation algorithm is the most commonly used method of classification. The primary benefit of this method is that the clusters it produces retain a high proportion of the variance of the input variables. K-means also produces clusters that are relatively even in terms of membership, especially with a large number of objects and a small number of clusters (Harris *et al.* 2005). The main drawback of the method is that the number of clusters has to be specified before the process is run. Although this does provide a saving in terms of computational processing, it means that if the ideal number of clusters is not known before clustering. The process has to be run many times and a choice has to be made between solutions (Gordon 1999).

There has been a great deal of attention paid to artificial neural networks as a method of clustering in recent years (Everitt *et al.* 2001). Artificial neural networks are computing algorithms that attempt to emulate the capabilities of large networks of simple elements, originally introduced as models of neural activity in the brain (Openshaw and Wymer 1995). A neural network contains three main features: the neurons or basic computing elements, the design of connections between computing units and the training algorithm used to establish the parameters for performing the set task (Everitt *et al.* 2001). Openshaw (1994) describes how an artificial intelligence technique, known as a Self Organising Map developed by Kohonen (1984) was used to create the GB profiles geodemographic system, which clustered the EDs from the 1991 Census. An excellent overview of 'neural networks for clustering' is provided by Murtagh (1996). Artificial neural networks have been shown to successfully classify data especially unsupervised versions such as the Self Organising Map that do not require the number of output clusters to be pre-specified (Kohonen 1998). However, neural network methods have a major drawback. The black box nature of their hidden layer(s) makes the operations that take place within the system difficult to understand, repeat and describe. It is essential to understand exactly what is happening during the clustering process and using neural network methods makes this almost impossible.

The vast majority of clustering methods are based around mean values, although this is not essential. Kaufman and Rousseeuw (2005) describe a method called 'Partitioning Around Medoids' (PAM), which clusters data based on median values. The method is generally robust, but has a number of drawbacks, not least that a two equally valid medoids can be calculated around which a partition can be made and a tendency for atypical objects to produce singleton

clusters even when a relatively small number of clusters are specified (Kaufman and Rousseeuw 2005).

A further clustering procedure is included in the SPSS statistical package. The TwoStep Cluster Analysis procedure is an exploratory tool. The algorithm employed by this procedure the ability to handle both categorical and continuous variables (SPSS Inc. 2001). The ability to incorporate categorical data into the clustering process is very useful in certain instances, but is not necessary for this study.

The partitioning of objects into classes is considered by many as an oversimplification of the structure of many complex datasets (Gordon 1999). This is especially relevant to objects that appear towards the edge of clusters, or objects that display attributes of more than one cluster. Fuzzy clustering is a method that is put forward to provide additional details about the properties of each object; they give proportional membership of a number of clusters rather than total membership of one (Everitt *et al.* 2001). Fuzzy versions of most clustering methods have now been developed, and have many advocates such as Feng and Flowerdew (1998). However, there is one major drawback to using a fuzzy classifier. The whole point of clustering data is to simplify a complex system to aid understanding and a fuzzy classifier adds more complexity to what was once simple. Despite undoubtedly more accurately representing reality, this should not be the main objective of a classification. The simplicity of a classification system has been its greatest asset and therefore fuzzy classifications will not be considered for use as the main output from this project.

Ward's and k-means algorithms were chosen for the methodology for this project (as described in Chapters 4 and 5). More details about how Ward's and k-means work are outlined in § 3.3.7 and § 3.3.8. Although Ward's and k-means have been chosen for use all the methods reviewed above are valid for this form of analysis. Some methods have been used in the creation of previous classifications; for others there is no recorded evidence of their use for area classification. The choice of clustering method can be a point of great debate as each has its strengths and weaknesses. The most important factor is that the researcher is both comfortable with and confident in the algorithm they are using and that they have a good understanding of how they work.

3.3.7 Ward's Hierarchical Clustering Algorithm

Developed by and named after Joe H. Ward of the Aerospace Medical Division, Lockland Air Force Base, Ward's hierarchical clustering algorithm was first published in the Journal of the American Statistical Association in 1963. It was developed as a method "to cluster large numbers of objects, symbols or persons into smaller numbers of mutually exclusive groups, each having members that are as much alike as possible" (Ward 1963 p236). The aim was to join objects together into ever increasing sizes of cluster using a measure of similarity of distance. At the start of the process each object is in a class by itself. Then in small steps the criterion by which the objects are clustered is relaxed to produce fewer but larger clusters at the next step up the hierarchy. This process continues until all the objects being clustered fall within a single cluster. The process of linking more and more objects together means that they are amalgamated into larger and larger clusters of increasing dissimilarity (Ward 1963). The number of clusters does not have to be pre-specified. The technique produces n clusters to 1 cluster inclusive, giving the user the ability to choose the most suitable number of clusters after the clustering process.

The process of hierarchical clustering is an agglomerative or stepwise approach beginning with n groups each containing 1 object then after merging them together ending with 1 group containing n objects. The process of getting from n to 1 groups can be summarised as below (following Ward 1963):

1. Place each object into its own cluster C , creating the cluster file f :

$$f = C_1, C_2, C_3, \dots, C_{n-2}, C_{n-1}, C_n \quad (3.6)$$

2. Compute a measure of similarity between every pair of clusters in the cluster file f to find the closest cluster to each cluster $\{C_i, C_j\}$
3. Remove C_i and C_j from f
4. Merge C_i and C_j to create a new cluster C_{ij} which will be the parent of C_i and C_j in the hierarchical cluster tree.
5. Return to step 2 until there is only one cluster left.

Methods of hierarchical clustering have been incorporated into the statistical packages for the social sciences and are frequently used to cluster census type information. There are several different distance formulae that can be used as the criterion in a hierarchical grouping procedure. The most common are Euclidean or Squared Euclidean measures, although others are used (discussed in § 3.3.9).

3.3.8 K-means Classification

The k-means algorithm is a simple non-parametric clustering method, where k stands for the number of clusters created. The objective of the k-means algorithm is to minimize the within cluster variability. If the number of clusters within the dataset has already been pre-specified, a k-means classifier can be used, for example, to form five clusters that are as distinct from each other as possible. The k-means clustering function in a statistical package such as SPSS will move objects between clusters with two specific purposes, firstly to minimise variation within clusters, and secondly to maximise variation between clusters. K-means is one of the most commonly used methods in the geodemographics industry (Harris *et al.* 2005). It is an iterative relocation algorithm based on an error sum of squares measure. The basic premise of the algorithm is to move a case from one cluster to another to see if the move would improve the sum of squared deviations within each cluster (Aldenderfer and Blashfield 1984). The case will then be assigned/re-allocated to the cluster to which it brings the greatest improvement. The next iteration occurs when all the cases have been processed. A stable classification is therefore reached when no moves occur during a complete iteration of the data. After clustering is complete, it is then possible to examine the means of each cluster for each dimension (variable) in order to assess how distinctiveness of the clusters (Everitt *et al.* 2001). The k-means clustering algorithm is comparatively simple and works as follows in its SPSS implementation (Everitt *et al.* 2001, pp. 99-100 and SPSS Inc.1999):

1. Choose an initial grouping of objects into the desired k clusters; compute the means for the groups over all variables and the sums of squared deviations of objects from group means.
2. Move each object from its own group to each other group and re-compute the sums of squared deviations (the clustering criterion).
3. Choose the change which leads to the greatest improvement in the clustering criterion.
4. Repeat steps 2 and 3 for all objects until no transfer of an object to a new group results in improvement in the clustering criterion.

The clustering criterion is to minimize the Euclidean sums of squared deviations of objects from the cluster mean, E_c which is defined as:

$$E_c = \sum_{i=1}^{n_c} \sum_{j=1}^m (Z_{ij} - Z_{cj})^2 \quad (3.6)$$

where Z_{cj} is the mean value for cluster c of variable j and Z_{ij} is the value for object i of variable j .

3.3.9 Distance Measures

The way in which clusters are identified is by a measurement of how close objects are in multidimensional space. This can be calculated by either a similarity or dissimilarity measure. A similarity measure (proximity) will report the largest value for the two objects that are closest together and the smallest value for the two objects that are furthest apart. Conversely a dissimilarity measure (distance) will report the smallest value for the two objects that are closest together and the largest value for the two objects that are furthest apart (Everitt *et al.* 2001). There are many different measures of both similarity and dissimilarity that can be used within cluster analysis. The suitability for use of a measure of similarity or dissimilarity depends on the specificities of the individual dataset. For example, different measures are more suited for different types of data. It would not be sensible to use the same measure for continuous, discrete and categorical datasets as these different types of data need to be treated in different ways. Things can get more complicated than this as there is no reason why continuous, discrete and categorical data cannot be used together in the same cluster analysis.

The remainder of this section will give a brief description of the distance measures considered for use in the project. However, there are many others available. Everitt *et al.* (2001) or Gordon (1999), give excellent overviews and descriptions of many different distance measures. Like clustering algorithms there are numerous distance measures, but only a few are commonly used and most are similar. Others are particular to specific applications. Probably the most commonly used distance measure is the Euclidean distance measure (Aldenderfer and Blasfield 1984). The Euclidean distance function measures the ‘as-the-crow-flies’ distance between a point $x(x_1x_2...x_n)$ and a point $y(y_1y_2...y_n)$. Calculating the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values. This is simply an extension of the Pythagoras theorem which give the distance between two points in n -dimensional space (Gordon 1999):

$$\text{distance}(x, y) = \left\{ \sum_i (x_i - y_i)^2 \right\}^{1/2} \quad (3.8)$$

Squared Euclidean distance uses the same equation as the Euclidean distance metric, but does not take the square root. As a result, clustering with the Euclidean Squared distance metric is faster than clustering with the regular Euclidean distance. The distance scores are squared which enables the use of Increase in Sum of Squares, which minimizes the Euclidean Sum of Squares. The squared Euclidean distance measure is therefore better at handling larger and increasing numbers of objects (Everitt *et al.* 2001). Squared Euclidean distance helps convergence in large datasets. Euclidean distance does not always converge despite reaching the maximum number of iterations allowed within a clustering package or algorithm.

$$\text{distance}(x, y) = \sum_i (x_i - y_i)^2 \quad (3.9)$$

K-Means clustering is not affected if Euclidean distance is replaced with Euclidean distance squared. However, the output of hierarchical clustering is likely to change (Everitt *et al.* 2001). It is therefore beneficial to use a Squared Euclidean distance measure when using kmeans clustering especially when clustering large datasets as this will increase the speed of the analysis and increase the chances of the analysis reaching convergence. However, a Euclidean distance measure is preferable when using a hierarchical method of clustering. The classifications created in this project require only straight line distances to be measured between points; therefore a choice will be made between Euclidean distance and Squared Euclidean distance, dependent upon method of clustering.

3.4 Outputs

The outputs of an area classification are not just which cluster each area belongs to, but also a large amount of descriptive and explanative information, that is also required to produce a useful classification.

3.4.1 Selecting the Cluster Numbers and Classification Structure

One of the most difficult tasks in creating a classification is deciding what number of clusters will be the most suitable for use. This is especially difficult if there is no specific target number of clusters to be created, or if little or no information about the number of clusters expected to be present in the dataset. However, before the task of discerning how many clusters are present within the data it is important to consider the possibility that there are no naturally occurring clusters within the dataset (Milligan 1996).

There are several different rules of thumb that have been formulated to select the most suitable number of clusters. However, these can contradict each other within the same cluster analysis. Examples include:

- If you can't choose between two solutions then the larger number of clusters should be selected.
 - Select the cluster which shows the greatest reduction in the average distance from the solution with one fewer clusters, in a non-hierarchical system.
 - Select the solution that shows the greatest increase in the average distance between the most dissimilar objects within merged clusters, in a hierarchical system.
-

-
- Select the solution which has the most suitable number of clusters for purpose.
 - Select the solution which is most homogeneous in terms of the number of objects within each cluster, for example the solution which has the smallest difference between the number of objects in the smallest and largest clusters.

There is no right answer to the selection of the number of clusters; the selected solution will just be one of a number of possible representations. Therefore the solution selected should be judged as much on its usefulness in terms of cluster numbers, as being a correct representation of the patterns within the dataset. Hierarchical systems will require a structure to be given to the classification where the number of groups and the places where clusters are split to create another level of the hierarchy needs to be decided upon, as well as the initial number of clusters.

3.4.2 Naming the Clusters

The next step in the clustering process is to profile and name the clusters. The naming of the clusters is a near impossible task and one that always provokes much debate. However, it is a very important job, as if it is done wrongly it can give a false impression of the areas within a cluster.

Names and descriptions are a very contentious issue in geodemographic classifications. They can become an increasingly sensitive subject as the scale gets smaller and the classifications appear to be more person than area based. The names could and maybe should be seen as very much a side issue to the whole classification process as no matter what each cluster is called it does not alter the variable values of the cluster. However, many users of classifications use only the name to get an idea of what the clusters are like ignoring any additional information that is provided. The cluster names have also been easily picked up on by the media as they provide striking, but not always accurate headlines. Much of the criticism of geodemographics has been focused on the names of the groups. Make the name too specific and they only represent those areas very close to the centre of the same cluster. One could think of this as a form of the ecological fallacy. Users would think of the classification as being wrong as they find the very specific descriptions unrepresentative of the areas they are studying. Alternatively make the names too broad in an attempt to represent all of the areas that fall within a cluster and they become too vague and start to sound alike; a healthy balance needs to be found.

The commercial classifications available in the UK were slower than their American counterparts in giving their clusters catchy names. However, some systems have now embraced the use of snazzy eye catching names while others still have a very British way of naming their clusters. This can be seen clearly in the difference between the names in the Mosaic and Cameo systems.

Mosaic's names include such things as: *Global Connections*, *Fledgling Nurseries*, *Coronation Street*, *University Challenge* and *Pastoral Symphony* (Experian 2005). The Cameo names include the following: *Affluent Singles in Quality Rented Flats*, *Well off School Age Families in Semi-detached Properties*, *Younger Couples in Smaller Terraced Housing* and *Young Student Areas* (EuroDirect 2005). The distinction between the two in terms of their approach to naming clusters is clear. The Mosaic profiles (Experian) are designed to be creative, provocative and are perhaps a little inaccurate. The Cameo (EuroDirect) are more factual and duller. The names suggest little about the quality of the product, but they are indicative of the market each company is targeting. While the Mosaic names will be loved by a more style than substance advertising executive, the Cameo names would appeal to the more analytical minded spatial analyst. Whether this is a deliberate tactic from the two companies to target opposite ends of the market is unclear. What is clear is that the names matter and the two different approaches taken by Experian and EuroDirect in naming their clusters reflects not only on individual products, but on their businesses as a whole.

3.4.3 Pen Portraits, Maps, Photos and Visual Descriptions

The idea behind Pen Portraits is to create a short description, using text and variable information, which significantly expand on the names but can be understood simply in only a few minutes. Pen Portraits are intended to significantly expand the users understanding of the group without them having to trawl through the large amounts of variable information for each cluster. The profiles often include graphs, photos of typical homes or neighbourhoods and some statistical information along with an extended description of the clusters. Some of the recent releases of commercial systems have interactive portraits with sound as well as visuals and the ability for the user to find out an almost limitless number of statistics about each cluster. The most recent release of Experian's Mosaic system is particularly impressive in this regard see Experian (2004).

An area classification is a representation of areas and places and therefore the mapping of area classification should be seen as an essential output of the classification process. A vital part of understanding an area classification is mapping it, to see how the classification looks in reality. Mapping the classification brings it to life and the geography of the area classification can be truly fascinating. Unfortunately the geography of geodemographic classifications is often overlooked when put to such uses as profiling consumer records.

3.5 Conclusions

This chapter has provided a sequential overview of inputs, processes and outputs that form the backbone of cluster analysis, taking the researcher from census data in its rawest form to an area classification. Any classification, areal or otherwise, is never the correct answer nor is it the incorrect answer. There is no correct answer in creating a classification just a near infinite number of possible outcomes, based on decisions made during its creation.

Making a classification is a thankless task; criticisms are offered from many directions, by people who see the classification as a rival to their own product or somebody who has little understanding of the processes involved, or a misunderstanding of what the classification represents. However, to create a classification is of far more value than criticising an existing one. Milligan (1996) formulated a seven step approach to cluster analysis, which provides an excellent overview of the main decision points that have to be made in cluster analysis. Milligan's seven steps can be further summarised into a three stage sequence of inputs, processes and outputs as used by Harris (1999).

The discussion of inputs into a classification system covered many issues, discussing the availability and quality issues of both census data and other sources of demographic data. The contrasting theories of how much data to use in a classification system were discussed. The view portrayed by commercial geodemographics firms that the greater number of variables used in a classification system the better, this contrasts with the view portrayed in much of the cluster analysis literature that the fewest possible number of variables should be used, as adding more variables can cloud important patterns within the dataset. The different methods for comparing and reducing the number of variables in the dataset were described and assessed.

The processes involved in creating a classification were reviewed in detail. The descriptions began with the different forms of variable standardisation and the issue of variable weighting. The methods of classification used in the project were described in full, along with a review of other clustering methodologies not utilised. The distance measures that were used in the classification are reviewed and explained.

The chapter closed with a discussion of the outputs from an area classification. The importance of selecting a representative and practical number of clusters was stressed. The significance of the naming of the cluster solutions and the effect that they can have on the image of the classification were identified as important issues. The value of additional outputs such as pen portraits, maps and photographs to the user's understanding of what each of the clusters represent were drawn out. After understanding all the issues described in this chapter, a researcher is ready to start classifying!
