

How geodemographic classifications are built.

Richard Webber

Introduction.

In this chapter we consider the methods that are used to build geodemographic classifications. Though no two developers of classifications use exactly the same methods, their approaches are broadly similar. In this chapter we will therefore examine the methods used by Experian which has built more geodemographic classifications in more countries than any other commercial organisation. The methods used by Claritas, CACI and Euro-direct employ many of these techniques.

When scientists develop methods for classifying geological series, climatic zones or vegetation cover, they build classification systems that operate irrespective of political boundaries. By contrast whilst there is much in common between types of neighbourhood in different countries, there are many practical reasons why it better for geodemographic classifications to be optimised, initially at first, on a country by county basis. No two countries' data infrastructure is the same. The sources of data available for different countries differ; these data are available at different levels of geographic detail; regulations governing access are different; countries have different update frequencies; and whilst there is increasing harmonisation on the questions covered by national censuses, no two census agencies make available the same set of variables for use in building classifications.

It is important to understand that the categories of neighbourhood used in a geodemographic classification are not defined in advance. The builder does not start with the requirement to find a category called 'Rural Isolation' or 'Mortgaged Families'. Such categories may or may not emerge from the computer programmes designed to build the typology. However, from his or her prior experience in different countries, the builder of a geodemographic system will expect the computer programmes to identify a number of types of neighbourhood which will bear strong similarities to those identified in other markets. It is for this reason that Experian has been able, across the 18 different national markets where its systems operate, to identify a set of 13 global 'lifestyle groups' ranging from 'Agrarian Heartland' to 'Shack and Shanty'. Needless to say not every category is found in every market.

In a relatively small and homogeneous countries, such as Ireland, Hong Kong or Peru, a classification will be able to identify around 25 to 30 distinct types. In larger and more heterogeneous countries, such as the US and the UK, it is possible to recognise a larger number of types, as many as 60 to 65. However because it is difficult for users to remember each of these types individually, the more detailed neighbourhood 'types' (as they are conventionally called) are usually classed hierarchically into a smaller number of neighbourhood 'groups', typically ranging from 7 to 12.

Data sources

In many countries, such as China, Hong Kong and Peru, the census is the sole source of data used to build geodemographic classifications.

However in many other countries the census is supplemented with statistics from other sources. Examples of these other sources are electoral registers (UK, Australia, Spain), the files of mail order companies (Netherlands), car registration files (Italy), Property Registers (Germany, New Zealand, UK), registers of shareholders and of directors (UK), statistics on house prices and on council tax bands (UK) and registers of addresses (Australia, UK). In the Netherlands, where census statistics are not published at a small area level, market research respondent files are used, in the UK the results from lifestyle questionnaires.

These non census sources of information may be useful for a number of reasons. Questions in national censuses understandably tend to focus more strongly on measures of disadvantage than on measures of affluence, asking their populations about their literacy (Brazil, China), long term illness (UK) or unemployment (Australia). Information from non census sources, such as directors or shareholders registers, is often helpful in redressing this bias and in providing greater detail about the location of more privileged members of the community. A second advantage of using non census sources is that in many instances it is available at a finer level of geographic detail than that at which census statistics are published. A third advantage is that in many markets the use of non census sources makes it possible to update the classification codes given to existing areas as their population character changes over the ten year interval between censuses. Likewise by using non census data sources it is possible to assign classification codes to neighbourhoods built since the date of the previous census.

Where the census is the sole source of information used to classify neighbourhoods then the classification will be built at whatever level of geography the census authorities use for the publication of their small area statistics. However where information used originates from multiple sources, the unit of classification may be more detailed than that of the census. So in the UK, for example, where there are on average five postcodes for each census output area, different postcodes within a single census output area may be assigned to different types of neighbourhood.

In Australia and in the UK, Experian have pioneered the use of variables, derived from the census, which are calculated for the concentric circles around each of the smallest levels of geography used in the classification system. The purpose of this innovation was to differentiate suburban areas in very small service towns serving agricultural hinterlands from areas of similar demographics located in the inner areas of larger metropolitan suburbs. This innovation has proved to be very efficient in improving the ability of the classifications systems to predict variations in the level of factors such as risk of crime.

Where the information used in the classification system is created for more than one level of geography it is necessary to link together the data for these different levels into a

single 'rectangular' database, spreading data for larger geographical units across all the lower level geographic units that fall within them. Building the relationship between the larger and the smaller units can often take some time, especially where the postcode system used to report non census data sources does not mesh with the administrative geography in terms of which census statistics are usually reported. One key advance in the design of the 2001 UK census is that the areas for which census statistics are published now consist of combinations of whole postcodes.

Creating area level variables

Assuming statistics have been sourced from available data bases and that they have been linked together into a single geography, the next stage in the process of building a classification is to create a set of variables for use in the clustering algorithm which is used to build the classification system.

As a general rule the more variables that are used in the clustering algorithm and the more different sources they come from the more meaningful the resulting clusters are likely to be. On the other hand it is important that variables should be included in this process only in they can be seen to be reliable indicators of what they purport to be. The evaluation process therefore is an important stage in the build process.

Very often the data items that have been sourced will arrive in the form of counts, such as total numbers of cars. Each of these counts needs to be related to a corresponding 'base' counts. For example total cars in a given zone (numerator) would be related to the base count 'total population', to the base count 'total adults' or to the base count 'total households' (denominators). Alternatively the count could be used against all three base counts to create three separate variables for use in the classification system.

The classification builder will also spend time at this stage deciding how to group some of the counts. For example he/she has available from the census the number of residents aged 0-4, 5-9, 10-14, 15-17, 18, 19-20 and so on. Should each of these separate counts be divided by total residents to create variables for each of these individual age groups? Or would the number of persons aged 18 be too small to create a variable in its own right? Would it be better to form fewer more robust indicators, for example 0-4, 5-9, 10-14, 15-20? Should even coarser bands be created? Alternatively should the clustering process use both finer five year bands as well as coarser 20 year bands? Likewise when we examine the statistics on employment by industry, should construction be included as a variable in its own right or should it be grouped with energy and transportation into a less specific but statistically more robust variable? Decisions of this sort need to be informed by the statistical reliability of the possible numerators and by knowledge of whether, by smoothing adjacent classes, the detail that is lost will be significant.

In general terms the classification builder, when in doubt, is likely to create a number of alternative measurements of any particular topic and to develop a set of strategies for

evaluating which of these alternatives are likely to be most appropriate for inclusion in the classification process.

Evaluation of input variables

In this stage of the build process the classification builder will apply various strategies for evaluating the appropriateness for including different variables in the clustering process. It may be that some variables, for whatever reason, are not deemed appropriate for using in the clustering process. But an equally important outcome of the evaluation process is a decision on the 'weight' that should be given to each variable in the forthcoming clustering process. Consider this by analogy with a recipe which sets out not only the ingredients to be included in a stew but also the relative quantities of these ingredients that should be used. Likewise with the cluster algorithm, the choice exists as to how much emphasis to be placed on each of the different input variables when calculating for each zone the cluster than it is statistically 'closest' to.

One important consideration is the extent which a variable is skewed. In an ideal world we would include as clustering variables only those which have a bell curved 'normal' distribution. In practice many important dimensions which need to be included in a classification are not 'normally' distributed. The residential location of people who work for the military tends to be very tightly concentrated in a small number of garrison locations. Newly arrived minority ethnic groups tend to cluster into ghettos. People who work in agriculture tend to be concentrated likewise in a limited number of census output areas. These groups of people are not 'normally' distributed.

The cluster analysis algorithm transforms all input variables, subtracting the national mean from each value and dividing by the national standard deviation. For this reason an extreme value in a zone on a single highly skewed variable can be sufficient to determine its cluster memberships on its own and without regard to any other data about the zone. This is clearly unsatisfactory and various strategies can be used to avoid this happening. For example the variable can be transformed using a log function. An upper limit can be applied to values, with all values above 25% being given a value of 25%. Both these methods are quite appropriate for variables, such as population density or distance from the coast, which are not demographic in nature. Experian classification builders normally refrain from using these methods with demographic variables, preferring instead to reduce the 'weight' given to very skewed variables to levels at which extreme scores do not override all other criteria when assigning zones to their best fit clusters.

Another important consideration is the extent to which variables have adequate samples. Supposing we had access from the census to the number of people by output area aged over 100. Although this might at first glance be a very interesting variable to include in the classification, we have to bear in mind that there is likely on average to be only one person aged over 100 in every twenty census output areas. When we consider that perhaps 95% of the people who were over 100 on census night, April 2001, are likely to have died by the time the census results are published, September 2003, it is evident that this variable, if used in a classification, is not going to be particularly useful for

describing a census output area in 2004 let alone in 2010. Small sample size also is a particular issue where variables are sourced from files which are not universal in their coverage, for example market research files in the Netherlands and lifestyle surveys in the UK. In these instances it is best to use variables which combine a large mean (such as people who smoke) and which have a large standard deviation at output area level (such as people who have a garden or live in a house built before 1945). One strategy for using these variables may be to use them only at a high level of geography. For example, though the proportions of smokers may be known at postcode level it may be safer to use the variable only at the much coarser output area level.

The variable 'aged over 100' is a good example not just of a variable which has too small an average value for the unit of geography for which it is calculated by also of a variable whose value is unstable over time. Given the longevity of a classification, it makes sense other things being equal to select variables whose values at a small area level change only slowly over time (such as the percentage of households living in apartments, the percentage of buildings constructed before 1945). It makes less sense to use variables whose rank order of zones is likely to be volatile from year to year. Examples of the latter are variables relating to change itself, such as the number of houses sold in the previous year, the proportion of households who have moved, changes in the level of house prices, unemployment etc. High values on any one of these variables in any one year are seldom reliable indications of high values at any subsequent point in time.

Another consideration is whether or not the variable is one which can be reliably updated over time. The value of census variables to the classification process is that they tend to have a high level of reliability and universal coverage. The disadvantage of census variables is that they are updated only as often as the census is which, in most countries, is only once in ten years. Whilst non census data sources are seldom as robust and reliable as the census, updatability may justify giving them a greater weight in the classification process than otherwise would be the case.

Once the various measures have been evaluated in this way the candidate variables are then correlated with each other and the results expressed in the form of a minimum spanning tree. The value of the minimum spanning tree, or single linkage analysis as it is often called, is that it identifies sets of variables which have particularly high correlations with each other, whether positive or negative. The tree will therefore highlight 'duplicate' or 'near duplicate' variables. In these circumstances the classification builder may wish to choose from among duplicate variables the one(s) which have links with the larger number of other variables, signifying a higher level of correlation – this being an indication of the extent to which the variable is likely to be reliable and robust. Alternatively more than one near duplicate variables may be included in the classification process but with each one being given a lower 'weight' than would otherwise have been the case.

The ability of the minimum spanning tree to group variables into 'islands' can also be helpful for the classification builder since it will alert him/her to 'domains' which are under or over represented in the data set. Classification builders may want to ensure that

rural areas are clearly identified by the classification. Taking 'rurality' as a domain we would see from a Chinese tree that it is covered by variables from a number of different topics covered by the census, for instance by occupation (farmer), by industry (farming) by civil status (hukou) and by mode of heating (wood). Knowing the extent to which the domain is covered by the variable set is important when decisions are taken about the weight given to individual variables.

One approach which is not used in the Experian methodology is principal components analysis. From experience it is found that the use of this technique tends to blur rather than clarify fine distinctions between the cluster types when used in this context. The use of principal components is likely to be more appropriate in contexts where the variables to be clustered relate to unit records rather than to statistical aggregates, where the level of accuracy of the measurement process is more uneven and where there is a greater degree of randomness due to infrequent occurrences of some of the attributes used in the classification process.

Selecting weights

In the absence of the use of principal components analysis, considerable reliance is placed on the appropriateness of the weights given to different variables in the clustering algorithm.

As we have seen there are many considerations which need to be taken into account when setting these weights. However a useful check on their appropriateness is to calculate the share of the total weight that is being accorded to variables belonging to different 'domains'. In the Australian classification fixed proportions of the total weight were assigned in advance to variables pertaining to housing, to variables pertaining to measures of social and economic status, to variables pertaining to age, household composition and cultural identity and, finally, to variables relating to accessibility. Likewise decisions were made as to the relative weight to be accorded to the three levels of geography used in the classification, street segments, census collection districts and concentric circles. The weights initially given to even variable after the evaluation process were then adjusted to ensure appropriate overall weights for these domains and geographic levels.

Clustering

Once the variables have been defined, robust ones selected and given appropriate weights, the clustering process begins.

At this stage all that the classification builder has to do is to specify the number of clusters that he/she thinks would be appropriate for the market being clustered bearing in mind the range of data available, the level of geography used and the complexity of the market.

The first stage in the clustering process involves the calculation of the means and standard deviations of the input variables and the normalisation of the data. That is to say that the values for each zone on each variable are expressed in terms of standard deviations from the average for all zones. An important feature of this process is that these, and all subsequent, computations are 'population weighted'. That is to say that when calculating the means and standard deviations the algorithm gives correspondingly more attention to the values of zones with high populations than to those with low populations. This 'population weighting' is particularly important in so far as the populations of rural zones and of inner city zones will in many markets tend to be lower than the populations of zones containing new housing estates on the edge of urban centres. Without using population weighting the clustering process would be biased to identifying more clusters in these types of areas than is warranted by the situation on the ground. In practice it could be equally appropriate to select households or adults rather than population for weighting.

The next stage of the clustering process involves the selection of a set of 'seed' zones. If the classification builder has required the cluster algorithm to create a 45 cluster 'solution' then 45 zones in the data set will at this stage be selected to form the nucleus of each cluster. The zones will be selected on a random fixed interval basis with a probability proportionate to their population. The first zone to be selected would be once half of the sampling interval is reached.

The algorithm will now examine every zone in the database and, taking into account the normalised data for each variable and the weight assigned to each variable, calculate a measure of similarity between that zone and each of the 45 seeds. This is known as a k-mean squared distance. It is calculated by taking the square of differences in the standardised scores of the zone and the seed zone, summed across all variables and weighted by the weight accorded to each variable. The zone is then 'assigned' to the seed for which this measure of distance is lowest, in other words the zone to which it is most similar.

Once this process has been repeated across all zones, the situation will arise when each zone is assigned to its nearest or 'best fit' cluster. The programme will then take all the zones which were found to be closer to seed 1 than to any other seed and calculate the average value of this set of zones on each of the input variables used in the clustering process. This set of calculation is repeated for all the other 44 seeds.

At this point the algorithm will commence a second loop (or 'iteration') during which it reviews for each of the zones in the database which of the 45 seeds it is now closest to in terms of similarity. For many of the zones the result of this computation will be the same as it was on the previous iteration. However, since the averages of the seed zones have now been replaced by the averages of the zones assigned to them, a number of zones will now find themselves shifting from the cluster to which they were first assigned to a different cluster.

At the end of this iteration the averages are recalculated for each of the 45 seed clusters and a further iteration is done. This process continues often for more than 20 iterations. However each subsequent iteration will cause a smaller number of zones to move from one cluster to another and progressively smaller changes in the values of the variable averages for each cluster. Eventually, when an iteration generates no further changes, the process comes to an end. The algorithm has reached a local optimum beyond which it can not improve.

Optimisation process and manual intervention

Once the cluster algorithm has optimised the classification, it finds an efficient way of ordering the resulting clusters such that clusters that are broadly similar are given consecutive numbering. It then reports to the cluster builder a number of diagnostics from which he/she can assess the effectiveness of the solution. After all the algorithm has no way of knowing whether what it has created is a 'local' or a 'global' optimum. It is not unlikely that by skilful modification the classification builder can improve on the result.

At this stage the classification builder will consider and probably implement a number of manual adjustments.

One particularly important diagnostic that the classification builder will consider is whether the two most 'similar' clusters are so similar that the user is unlikely to be able to tell them apart, in which case the system should have fewer clusters, or whether there are clusters which are so heterogeneous that a better solution could be achieved by being set to run with more clusters. Particular diagnostics support these decisions but, as a general rule of thumb, a good solution is likely to be one where the two most similar clusters could be merged for a loss of variance of 0.22% of the total variance in the dataset. This consistently appears to be a limit below which further divisions of clusters are indistinguishable.

The classification builder will now use a number of functions to try to improve the classification. These include ones which cause two or more specified clusters to be fused together into a single cluster, which allow individual clusters to be deleted and which allow individual clusters to be split. Sometimes it appears sensible to use all three of these functions. When these modifications are made, the cluster algorithm, having remembered the previous solution, makes the specified modifications and then undertakes as many further iterations as are necessary to come up with a revised solution. Often this optimal solution will account for more variance in the original data set than the previous one, even for no change in the total number of clusters. In this case it is likely to be a 'better' 'local' solution though not necessarily to be the 'optimal' solution.

At the same time as he or she merges, splits or deletes clusters from an old solution the classification builder may want to review the weights that were assigned to the different variables. If, for example, the clusters appear to be over influenced by population density to the exclusion of other important differences, the classification builder may want to

reduce the weight given to that variable and re-run from the old solution but with new weights.

Typically it may take five to a dozen different runs of the cluster algorithm before an ideal solution is finally agreed and much time can be spent evaluating which of the alternative solutions is the best one. One method of evaluating the solutions is to see which one 'explains' the most variance in the input variables. Other things being equal this is a sensible approach. However it is not an entirely fair adjudication procedure if the number of clusters has been changed or if there have been alterations in the variable weights between the two solutions. A fairer evaluation is to compare the extent to which alternative classifications are effective in predicting variations in behaviour on data sets which were not used to build the classification. Tests against 100 external files enabled Experian to improve the efficiency of their 2001 census based UK Mosaic by 3% compared with what it would otherwise have been. This process of evaluation is particularly effective in determining the most appropriate set of weights to give to variables in different domains and, perhaps even more important, the set of weights to give to variables at different levels of geographic aggregation.

Other important evaluation methods used at this stage of the build process involve the mapping of the clusters in towns familiar to the classification builder. Likewise, when building UK Mosaic, Experian undertook a photographic tour of the UK which generated over a thousand pictures of different postcodes. Alternative solutions were then partly evaluated with reference to whether the new cluster allocation seemed to provide a better representation of the photographed postcodes which had been assigned to different clusters in the different solutions than did the previous cluster allocation.

Forming groups from the types.

Once the solution is finally signed off at the cluster level, attention shifts to the process of arranging the clusters into 'groups'.

This process is undertaken initially using a 'stepwise fusion' algorithm which is integrated within the main cluster programme. This process starts by considering which pair of clusters could be merged together whilst contributing least loss of variability in the original data set. The pair of clusters which could best be merged in this way is likely to be a pair which on the one hand is very similar in terms of their average scores on the input variables. But they are also likely to be ones with relatively small populations. Once the first pair is fused the algorithm treats the merged pair as a single cluster and searches for the next pair of clusters to be merged. This process is repeated one time fewer than the total number of clusters until all the clusters have been joined together, at which point all of the original variance has been lost.

Whilst one of the purposes of this process is to renumber the clusters so that early joining clusters have consecutive numbers, the process also proposes an ideal set of clusters for any number of groups. In other words it tells us that say ten would be a good number of groups into which to organise our 45 clusters and how the 10 groups would be made up.

Whilst this process may maximise the efficiency of the solution, there can be other important considerations that it may not adequately address. For example one might want to ensure that the percentage of the population in each of the groups exceeded a threshold of 4% whilst not exceeding 20% and one might reasonably wish to ensure that all groups contained more than one cluster but not more than seven.

This process of manual intervention is facilitated by the drawing of a 'family tree'. This diagram, which is not unlike the minimum spanning tree of variables in form, links each cluster to the cluster within the set which it is most similar to. However, unlike the situation with the step wise fusion process, linked clusters are not combined and within each linked pair of clusters the tree will identify the one (of the two) which has the highest degree of similarity with a cluster belonging to another grouping.

This device has consistently proved helpful in identifying the major dimensions of differentiation within the system and these are often set out around the outside of the tree to orient the user. The final decision on how the clusters should be best organised into groups is seldom undertaken without some reference to such a diagram.

When the clusters have been arranged into groups it is necessary to undertake a final re-ordering in which it is customary to assign the category 'one' to the cluster with the highest status. Next begins the complex process of assigning labels to the categories. The labelling process is often contentious if only because the expectations and requirements of marketers, public sector users and academics can differ. Marketers typically want labels (such as 'Suburban Mock Tudor') which are both recognisable and memorable. The better the label provides insight into the mind set of residents of the cluster the better. Public sector users are more mindful of the political correctness of the labels, bearing in mind that they may be used in reports to elected representatives, and would ideally like labels which focussed on demographics. Academics on the other hand are most likely to want labels whose descriptive attributes can be verified by research evidence. When confronted with a label such as 'Low Horizons' they would ask on the basis of what evidence could such an attribute be selected. Label selection is a very important part of the build process. Good labels are ones which can only be true of the type they are given to, which can be accurately applied to virtually all postcodes within that type, which are memorable and not politically offensive. 'Fledgling Nurseries' is a good example of a label which meets all these criteria. An important rule applied by Experian is that type labels should be no more than 20 characters in length. This precludes the tendency for labels created by committee to become long, bland and lacking in insight.

The more people in the organisation involved in the labelling of the categories the more intelligible they are likely to be. However writing up the descriptions of the categories is a more solitary task. Typically the clusters and groups will each be subject to a textual portrait. This portrait will look at the clusters from different dimensions : their physical appearance; their historical origins; the types of people who live in them; their values; their consumption patterns; the patterns of movement into and out of the types; ways in which it is likely that they will change over time.

The evidence on which these portraits are created clearly includes a wide range of census and non census indicators, often a wider range than the restricted set that were used in the build process. Tables showing the regions of the country, the local authorities and the constituencies in which they occur are also helpful in being able to uncover subtle distinctions between otherwise similar types. Photographs are invaluable as are maps of places one knows well. The process of producing the portraits involves a fusion of art and science, of the qualitative and the quantitative and is best undertaken without interruption and with copious supplies of caffeine.