USING NEUROCOMPUTING METHODS TO CLASSIFY BRITAINS RESIDENTIAL AREAS

Stan Openshaw, Marcus Blake and Colin Wymer

WORKING PAPER 94/17

SCHOOL OF GEOGRAPHY • UNIVERSITY OF LEEDS

1. Introduction

The explosion of spatial information occasioned by the GIS revolution and the ease by which individual databases can now be aggregated to small zones (*i.e.* census ED's, postcode sectors, wards) emphasises the importance of being able to simplify the resulting multivariate complexity. The Small Areas Statistics (SAS) can be regarded as providing a good example of a generic problem of world wide significance. An ability to apply a multivariate classification procedure to reduce census data, for the 145,716 1991 census EDs in the UK, to a relatively small number of major types of residential area is extremely useful. For instance, most commercial geodemographic classifications segment Britain's residential areas into about 50 types based on a mix of census and non-census data for large numbers of small areas as a means of adding value to data that otherwise could not be used in an area profiling and targetting context.. The objective here is somewhat simpler in that only census data are of interest, although the challenge is harder in that the aim is to provide for research and academic purposes the best possible classification of 1991 census EDs for Britain. Previously only the 1981 Super Profiles classification (obtainable from the Essex Data Archive) was freely available for academic purposes.

In general, the quality of any area based census classification reflects three major factors. First, the classification algorithm that is used, second, the manner and extent to which knowledge about socio-spatial structure is used and is represented in the classification, and third, the sensitivity of the technology to what can be condensed to be the geographical realities of the spatial census data classification problem. In much previous classification research the quality and utility of the end product has been regarded as being critically dependent mainly upon the performance of the classification algorithm that is employed. Under this assumption, it is quite reasonable to believe that better classifications can only be produced by the application of improved classifiers; see for instance, Openshaw and Wymer (1994) and Evans and Webber (1994). There is, however, one major flaw in this approach; namely, there are limits to the degrees of improvement likely to be possible solely by developing or using improved classifiers. Maybe the possibility exists to become more intelligent in the way these classifications are developed and thus seek a quantum leap in performance rather than just a marginal gain of dubious value; Openshaw (1994a). Indeed, this goal of seeking to inject more intelligence into the total classification process rather than to continue down a narrowly focused purely classifier algorithm route is very important. It may also help explain the anomaly whereby the differences in end-user performance between a hastily cobbled together classification and one rigorously produced after a massive expenditure of research effort are not particularly noticeable (see Charlton *et al*, 1985). In short, it has long been apparent, but perhaps not sufficiently clearly recognised, that the really critical limitation on the performance of small area census classification is not the classification algorithm per se but the other steps in the process. Developing Artificial Intelligence based classifiers is only part of a wider process of being more intelligent in how we go about building better spatial data classifications.

2. The Census Classification Process

The classification of census data usually involves the following steps:

Step 1. Decide upon the purpose of the classification

Step 2. Select variables and data to meet this purpose

Step 3. Apply a classifier to the data

Step 4. Evaluate the results and select the most appropriate number of clusters

Step 5. Label the clusters

Step 6. Embed the results in some easy to use end-user system that allows the classification to be linked to the postcode geography.

It is noted that the critical stages are highly subjective and operational decisions made there may substantially determine the utility of the results (Openshaw and Gillard, 1978). Clearly there is no easy way of turning the whole process into an optimisation problem since there is no single global function that can be optimised which would simultaneously meet all the goals. For example, there is no simple relationship between optimising a statistical measure of classification performance such as the within cluster sum of squares and the end-users perception of classification performance in a particular context. Indeed, the quality of the results is dependent in some little understood way on the performance of the classifier in step 3, on the usefulness of the results in step 6, on the extent to which the step 5 cluster labels "make sense" or correspond to what is known about socio-economic and demographic structure, and on the degree to which the variables and data selected in step 2 deliver results that are perceived to be useful at the end of the process. Note also that the perceivers of usefulness of the results may be a number of different users and not just one possessed of a homogeneous point of view.

One solution is to create 50 or 100 classifications based on different numbers of clusters (and variables), and then select whatever works best in a particular application. Various cross classification validation methods can be used to automate this decision, see Openshaw (1994b). However this 'intelligent geodemographic' approach is probably more relevant to applied commercial uses of census classifications than in a research context, although the principles are transferable. It is also clearly an extreme response and that before commending its universal adoption it is necessary to investigate whether or not the needs of many research users would be better satisfied by developing a single good classification. This again re-emphasises the design of the classification algorithm (Step 3).

The conventional best practice is to employ a K-means nearest neighbour classifier to spatial census data that has been orthonormalised. Table 1 outlines some of the problems associated with this approach. The next question now is whether or not it is possible to greatly improve this technology? There are three ways of becoming smarter in developing a census classification.

(1). Improve the classification procedure by switching to superior algorithms, for example, simulated annealing approaches and neuroclassification methods;

(2). develop a means of incorporating knowledge into the classification process; and

(3). discover how to make the cluster assignment or allocation stage more sensitive to the nature of the task.

Until fairly recently, insufficient computing power was available to allow much or any progress to be made on any of these themes; for instance, a key requirement in designing new classifiers is computational tractability when faced with 145,716 or so zones to process. There has been, therefore, a tendency to simply continue using old methods dating from the early 1960's on larger and larger data sets. As a result this has lead to a failure to evolve new approaches that are really needed to ensure that the census classification challenge is being properly addressed. This has also been a failure to properly appreciate the complexity of the problem. The relative ease by which virtually any multivariate classifier when fed with virtually any set of census variables produces plausible results has tended to disguise the inherent difficulty of the task.

3. A Neuroclassification Procedure

Some of these problems might be avoidable by using a more spatially sensitive census data classifier that can also provide a good representation of the natural levels of fuzzyness that seem to characterise spatial data in general and census data in particular. One of the key characteristics of census data is that EDs vary in size and thus the level of the precision and resolution of the data varies geographically. This aspect is often ignored. The problem is that this variation is not random but is spatially structured as it tends to reflect systematic urban rural differences and population density. This is partly due to a small number of problems which result in the most extreme results being found in the smallest areas which will tend to be more homogeneous and rural rather than urban. On the other hand, the largest EDs tend to be urban and often have very mixed characteristics, but their size produces data values which are much more accurate than is the case for small areas. The conventional classifier gives equal weight to each ED and thus will tend to focus on the more extreme results that will tend to represent best those areas for which the census data are least reliable. This is the opposite of what the geographer might wish to happen. As a result many of the larger urban EDs with mixed characteristics will tend to be "poorly" classified and there may well be a number of different possible allocations. Of course this may turn out to be a geographical fact of life with census data; however, it is worth investigating whether or not it might be reduced by using either much larger numbers of clusters in a conventional classifier or by switching to a more data sensitive classification process.

Another feature of UK census data concerns the mix of 100 and 10 percent data coding and the data blurring employed by the census agency to ensure confidentiality. These effects are partly handled by taking into account ED size variation but they also operate on an individual variable level. A small area may have highly accurate data for some variables and highly uncertain data for others. It would seem important that these data uncertainties are taken into account by the classifier rather than simply ignored.

It is with these factors in mind that Openshaw (1994) argues that the use of an unsupervised neural net based on Kohonen's self organising map (1984) provides the basis for a much more sophisticated approach to spatial classification that reduces the number of assumptions that have to be made and neatly incorporates many of the sources of data uncertainty; see Openshaw (1994a). A basic algorithm is as follows:

Step 1. Define the geometry of the self-organising map to be used and its dimensions. Here a grid with 8 rows and 8 columns is used.

Step 2. Initialise a vector of M weights (one for each variable) for each of the 8 by 8 neuronal processing units.

Step 3. Define the parameters that control the training process: block neighbourhood size, training rate, and number of training iterations.

Step 4. Select a census ED at random but with a probability proportional to its population size.

Step 5. Randomise the vector of M variable values to incorporate data uncertainty computed for each variable separately (optional).

Step 6. Identify the neuron which is "closest" to the input data.

Step 7. Update the winning neuron weights and those of all other neurons in its block neighbourhood or vicinity.

Step 8. Reduce slightly the training parameter and the block neighbourhood size.

Step 9. Repeat steps 4 to 8 a very large number of times.

If Step 4 is replaced by a sequential selection process and Step 5 is ignored then the algorithm is essentially the same as a K means classifier; with a few differences due to the neighbouring training which might well be regarded as a form of simulated annealing and it maywell provide better results and avoid some local optima. However, from a geographical perspective Step 4 is extremely important because it provides a means of explicitly incorporating spatial data uncertainty into the classification process. The method also provides a very natural means of handling cluster fuzzyness without having to impose an arbitrary metric; since the distance between the best and the next best neurons can be readily measured.

The simplicity of the self-organising map approach readily lends itself to ad hoc modification designed to improve the quality of the geographic representation offered by the classification. There are various ways of meeting this objective. the simplest is to select an ED, as in the standard algorithm described previously, but then to use a distance weighted average value for the *k* nearest ED's. This neighbourhood in geographic ED space is gradually reduced as the block neighbourhood in the self-organising map's topological space is also reduced, slowly over many millions of iterations. The logic is to incorporate some notion of local geographical neighbourhood structure into the classification. Here the geographic neighbourhood is limited to the 10th nearest neighbour of each ED.

Another way of attempting the same objective is to change the updating mechanism (see Steps 6 and 7 in the basic algorithm) to update neurons assigned to the *k* th nearest geographical neighbours of the ED being used for training at any particular instance, irrespective of whether these neurons are within the block neighbourhood of the winning neuron. Experimentation suggests that the 'OR' rule is slightly better than the 'AND' rule. Equally, restricting the neuron updating to only the geographical neighbour related neurons also yielded slightly poorer results. However, the resulting classifications seemed to offer levels of descriptive resolution equivalent to conventional cluster systems with many more cluster in them.

The principle disadvantage of neuroclassification concerns the computationally intensive nature of the method. If the technique is to properly handle and represent the 150,000 cases then large numbers of training iterations (Step 3) are required. In a census application an ability to represent the data is much more important than any generalisation to unseen data; since there is none. This requires many millions of training iterations; indeed runs of up to one billion iterations have been investigated. In practice this means that parallel implementations are required and a parallel supercomputing version is under development. However, it is worth noting that a conventional classification of 150,000 ED's may well require 200 passes through the data. This does not seem much but nevertheless it would be equivalent to 30 million training iterations and this conventional classifier is much harder to parallelise or vectorise in any worthwhile manner

Finally, Tables 2 and 3 briefly summarise some of the strengths and weaknesses of a spatial neuoclassification approach.

4. Some Empirical results

Following the census classification process as described in section 2, a set of 85 broadly representative 1991 census variables were derived; see Blake and Openshaw (1994) for a full description. These variables are listed in Appendix 1. A conventional iterative relocation procedure is then used to create flock of cluster systems with between 2 and 2,000 cluster in them. The CCP (Census Classification Program) software is described in Openshaw (1983) and is still available at MCC for research uses. Figure 1 shows a plot of the average percentage within cluster sum of squares for these classifications. The resulting is very smooth and would apparently confirm the general view that somewhere between 40 and 70 clusters is needed to provide a useful classification of Britain's residential areas; indeed most 1991 census geodemographic systems offer less than 60 cluster solutions. However this disguises the fact that some variables are much better represented than others; for example, variables such as older couples (35-54) without children and couples aged 55-74+ (denoted as A and B in Figure 1) are not well represented. It is this application or data specificity that Openshaw (1994a)'s

Intelligent Geodemographic Targetting System (IGT/1) attempts to exploit. In the present context it merely means that a general purpose census classification with a fixed number of clusters will not satisfy all purposes equally well, but maybe in a research context with the cluster codes being used as a simple index summarising multivariate complexity, it is of a little consequence.

For current purposes the neuroclassifier is run with an 8 by 8 matrix of neurons for the 85 variables listed in Appendix 1. A total of 200 million training iterations were used. Step 5 was omitted to reduce the run-time on a workstation to 5 days. The labels that were derived for the resulting clusters are listed in Appendix 2. Comparisons with conventional classifications suggest that the differences appear to be slight in a qualitative sense. Quantitative comparisons are more difficult because it is not clear as to what the performance measure should be.

It seems then that any preference for a neuroclassifier requires both a significant amount of faith and a judgement about the relative merits along the lines of Table 1 to 3. This can be back-ed up by an assessment of whether the results are plausible. Figure 2 shows the distribution of the principle residential areas types in Sheffield. This stands up well to both local knowledge and previous research (Haining, Wise and Blake, 1992) on area types in Sheffield. For example, on a broad scale, the classic east-west division found in many industrial cities can be seen, with the affluent west and south-west of the city being dominated by the affluent and climbing categories while the city centre, and east of the city has more struggling and aspiring areas, see Figure 2.

5. Fuzzyness in the classification of small area census data

An illustration of the apparent complexity of the census data classification process is provided by allowing fuzziness to occur in the cluster assignment stage. Openshaw (1989a & b) suggests that there is a particularly easy way of incorporating spatial data uncertainty into the spatial classification process. This illustrates the two principal sources of uncertainty; fuzziness in the geography space and fuzziness in the classification space. Traditionally, neighbourhood effects in geography are regarded as a spatial phenomenon in that people who live "near" to each other tend to share some behavioural characteristics despite other differences. In geodemographic classifications these effects have been implicitly exploited at the enumeration district scale; hence why these classifications are sometimes referred to being neighbourhood classifications. However this is an extremely crude representation of a highly complex and high variable spatial phenomenon. Geographers in the GIS era should really be able to do better than this and regard spatial neighbourhood effects in an elastic fashion rather than at a discrete ED geography space. Similarly fuzziness in the classification space should be exploited rather than ignored. Areas may differ by only very small amounts in the classification but be assigned to very different clusters. This is particularly important with census data because of lack of social homogeneity of the census ED and the tendency of the classification process to focus on highly distinctive minority characteristics of areas due to small number effects. As a result, it is likely that in many classifications the distinguishing cluster descriptions are minority features that are either created by aggregation effects at the ED scale or represent a profile based on the mixture of different individual household types. It is with great regret that in the UK there is currently no data available which can be used to measure these effects. The ecological fallacy problem needs to be handled rather than ignored in census classifications. Openshaw (1994a) provides a specification of a fuzzy geodemographics system to try and handle these problems. This can be demonstrated by using the results of the neuroclassification procedure.

The first aspect to consider is the structure of the Kth nearest neighbour distances in the classification. Figure 3 shows the histogram of the number of different clusters "near" to each ED. It suggests that perhaps a surprising number of EDs are "near" to more than one cluster and could in fact be assigned with only a relatively small degree of error to a different cluster all together. Figure 4 shows a map that identifies the location of these "uncertain" EDs in Sheffield. Relatively few areas seem to be without some classification uncertainty. This measure of fuzziness is however only partial in that geographic neighbourhood or distance effects are excluded. It perhaps matters less if an ED can belong to two or more different clusters if these clusters are located nearby than if they are a long way off. The converse may also be important; that is neighbouring EDs should perhaps tend to belong to the same or similar cluster types.

To illustrate the further effects of fuzziness, Table 4 provides a cross tabulation of the census EDs in Britain by different levels of uncertainty in both the geography space and the classification space for a few illustrative cluster types. It is immediately apparent that a small amount of fuzziness soon introduces a number of other EDs that could be considered as belonging to each of the clusters. In fact it seems that the all or nothing nature of the conventional census classification is hiding considerable degrees of uncertainty. A surprisingly large numbers of EDs can in fact be assigned to different clusters. This may well reflect the heterogeneity of the census ED as a geographical entity, Openshaw (1984). However, not all this fuzziness is harmful to the classification as it can be used to improve the local fit of a classification by using geographic neighbourliness as a kind of smoothing operator. In fact, the first column in Table 4 shows the distribution of nearest neighbour geographic distances for ED's in the selected cluster. The distribution varies according to the nature of the cluster. Some are very closely related; for example council multi-storey housing and others much less so; for example poor semi-detached.

6. Disseminating the Results: GB Profiles '91

Finally, one of the objectives of the present research is to constructed a geodemographic profiling system that researchers can easily use. Using Microsoft Visual Basic an easy-to-used windows based system called GB Profiles '91 has been developed. This allows the classification of the underlying ED of every unit postcode in Great Britain to be accessed. Its primary use is to allow the academic community easy access to the results of the neuroclassification research.

It has two modes of use, an interactive single postcode search(or Single Search Mode - SSM) which instantly provides the cluster information on the screen and a multiple postcode search (Multiple Search Mode - MSM) which allows the user to batch process postcodes stored in a file. Some of the Windows associated with the MS mode are shown in Figure 5. The Search Setup Window allows the user to select a particular classification and determine which mode of operation to use, SS or MS mode. If MS mode is selected and a file loaded then this is stored in the list box of the Search Window where postcodes can either be added or removed. When these are processed a record is kept of postcode which have failed to be found and those which are duplicates. These statistics are provided in the Search Statistics Window. The results of the search are stored in a set of arrays which can be viewed on screen or saved to a file. Further information on the frequency distribution of the clusters found and a more detailed description of the clusters is also provided.

The underlying data structure is modular and this will allow different classifications to be loaded and then selected from the interface. Modules that provide photgraphic images and summary statistics are also developed.

7. Conclusions

The paper has argued that the use of a neuroclassifier provides a much more flexible and potentially superior means of generating census classifications. However, the substantially improved results are unlikely until it is possible to improve all aspects of the classification process so that the classification better represents both the complex nature of spatial data and incorporates meta knowledge that exists about the nature of residential areas in Britain. A start has been made but the really definitive results have yet to be produced.

References

Blake, M. & Openshaw, S., 1994, 'Selecting census variables for use in classification research', *Working Paper, School of Geography, Leeds University.*

Charlton, M., Openshaw, S., & Wymer, C., 1985, 'Some new classifications of census enumeration districts in Britain: a poor man's ACORN', *Journal of Economic and Social Measuremnt*, 13, 69-98.

Evans , N., and Webber, R., 1994, 'Advances in geodemographic classification techniques for target marketing', *Journal of Targeting, Measurement and Analysis for Marketing*, 2, 313-321.

Kohenon, T., 1984, *Self-organization and associative memory*, Springeer-Verlag, Berlin.

Openshaw, S., 1983, 'Multivariate analysis of census data: the classification of areas', in D. Rhind (ED) *An Census User's Handbook*, Methuen, London, 243-264.

Openshaw, S., 1984, 'Ecological Fallacies and the analysis of areal census data', *Environment and Planning A*, 16, 17 - 31.

Openshaw, S., 1989a, 'Learning to live with spatial databases', in M. Goodchild & S. Gopal (ED's) *The Accuracy of Spatial Databases*, Talyor & Francis, London, 264 - 276.

Openshaw, S., 1989b, 'Making geodemographics more sophisticated', *Journal of the Market Research Society*, 31, 111 - 131.

Openshaw, S., 1994a, 'Developing smart and intelligent target marketing systems: part I', *Journal of Targeting, Measurement and Analysis for Marketing*, 2, 289-301.

Openshaw, S., 1994b, 'Developing smart and intelligent target marketing system', *Working Paper 94/3*. School of Geography, Leeds University.

Openshaw, S., 1994c, 'Neuroclassification of spatial data', in D.C. Hewitson and R.G. Craneleds, *Neural Nets: Applications in Geography*, Kluwer, Boston, 53-70.

Openshaw, S. and Gillard, A. A., 1978, 'On the stability of a spatial classification of census enumeration district data', in P.W.S. Batey (ED) *Theory and Methods in Urban and Regional Analysis*, Pion, London, 101-119.

Openshaw, S., & Wymer, C., 1994, 'Classification and regionalisation', in S. Openshaw (ed), *Census User's Handbook*, Longmans, London.

Table 1: Problems with a conventional classification procedure

1.	Use of a correlation matrix which acts as a linear filter.
2.	Use of principal component scores which use Z score transformation of the data, emphasizing non-normal distributions and affected by spatial dependency.
3.	All or nothing nature of the classification assignment.
4.	Single move heuristic which might become stuck in sub-optimum locations.
5.	Global function that is being optimised but with no basis for knowing whether the results are better than random.
6.	Imposes arbitrary structure (viz. minimum variance) on the data.
7.	No way of handling data outliers and variations in data precision.
8.	No means of including prior knowledge into the classification process

Table 2: Some of the benefits of a neurocomputing spatial classifier

1.	Use of raw data removes the need for an orthonormalising linear filter.
2.	The self-organising nature of a Kohonen map allows structure to emerge rather than be imposed from the top.
3.	Incorporation of data uncertainty into the classification.
4.	Simplicity and greatly reduced number of source code lines.
5.	Possible to incorporate prior knowledge into the classification process making it more intelligent.
6.	Fuzziness of the results are preserved in a particularly easy to use form.
7.	Reduction in importance of knowing precisely how many clusters are needed.
8.	Cluster interpretation is easier because the classification takes place in the data space rather than in some transform space.
9.	Non-linear technology.
10.	Less likely to be trapped in a local sub-optimum.

Table 3: Some of the problems of a neurocomputing spatial classifier

1.	You need to prove that the potentially superior technology yields improved results by comparison with conventional benchmarks.
2.	Extensive computer run times are needed requiring the use of parallel supercomputing to adequately train on large data sets.
3.	A number of design aspects are entirely subjective, in particular; the number of training iterations, the architecture of the net, the updating process, and the choice of metric for the classification.
4.	The current absence of an intelligent framework for using the results.
5.	Lack of experience with the technology.

Table 4: Neural net classification analysis of fuzziness

Cluster No 1: Multi-ethnic council tenants

Members = 2869

	Cluster Similarity Distances							
Geog. Dist.	0.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
0.000	1	0	0	0	0	0	0	0
100.000	84	106	63	33	18	24	7	1
200.000	393	448	320	221	129	138	48	9
300.000	526	669	617	458	330	438	143	42
400.000	422	577	758	578	417	587	274	91
500.000	232	484	698	607	542	817	377	144
750.000	366	775	1461	1533	1405	2344	1211	541
1000.000	141	604	1095	1330	1403	2476	1525	633
2000.000	309	1313	2971	3908	4352	8432	5555	2601
3000.000	127	684	1841	2499	2568	4895	3486	2039
.gt.3km	269	3396	12412	12605	9490	12150	6648	4081

Cluster No 5: Poor semi detached housing

Members = 2920

	Cluster Similarity Distances							
Geog. Dist.	0.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
0.000	0	1	0	2	0	0	0	0
100.000	27	45	42	29	13	15	4	1
200.000	104	187	181	127	70	78	18	3
300.000	192	228	293	267	172	181	67	9
400.000	195	285	399	339	295	361	115	18
500.000	155	228	364	374	362	463	153	52
750.000	339	516	881	1011	1003	1433	629	217
1000.000	155	365	656	936	1015	1731	927	357
2000.000	505	1107	2003	2738	3387	7123	4640	2337
3000.000	351	1011	1680	2116	2586	5793	4313	2650
.gt.3km	897	3808	7585	8437	8883	20244	16240	10546

Cluster No 6: Council multi-storey housing

Members = 3958

Geog. Dist.	Cluster Similarity Distances 0.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
0.000	7	3	2	0	0	0	0	0
100.000	1374	1269	621	253	150	82	27	16
200.000	994	1692	1486	765	483	449	161	106
300.000	445	1172	1390	916	635	746	343	255
400.000	248	684	1054	858	621	794	467	334
500.000	151	511	821	685	597	794	484	433
750.000	209	786	1418	1364	1260	1825	1336	1381
1000.000	121	375	751	896	877	1516	1226	1516
2000.000	190	556	1087	1546	1717	3700	3395	5074
3000.000	83	196	467	700	845	1708	1755	3115
.gt.3km	136	525	1340	2265	2997	8032	9813	20093

Cluster No 19: Well off metro singles

Members = 3849

Geog. Dist.	Cluster Similarity Distances 0.00	0.25	0.50	0.75	1.00	1.50	2.00	3.00
0.000	4	1	1	0	1	0	0	0
100.000	1641	883	477	177	146	85	16	3
200.000	1185	1156	1068	637	429	328	129	51
300.000	309	701	971	800	565	579	270	163
400.000	158	505	689	781	544	646	331	245
500.000	91	313	577	683	546	675	343	307
750.000	136	449	1128	1316	1151	1510	960	977
1000.000	58	295	689	884	906	1341	937	1055
2000.000	116	395	1113	1738	1918	3252	2921	3711
3000.000	47	161	393	629	812	1769	1690	2823
.gt.3km	104	416	1415	2590	3532	9139	11076	23331

Figure 1: A Plot of the average percentage within cluster sum of squares

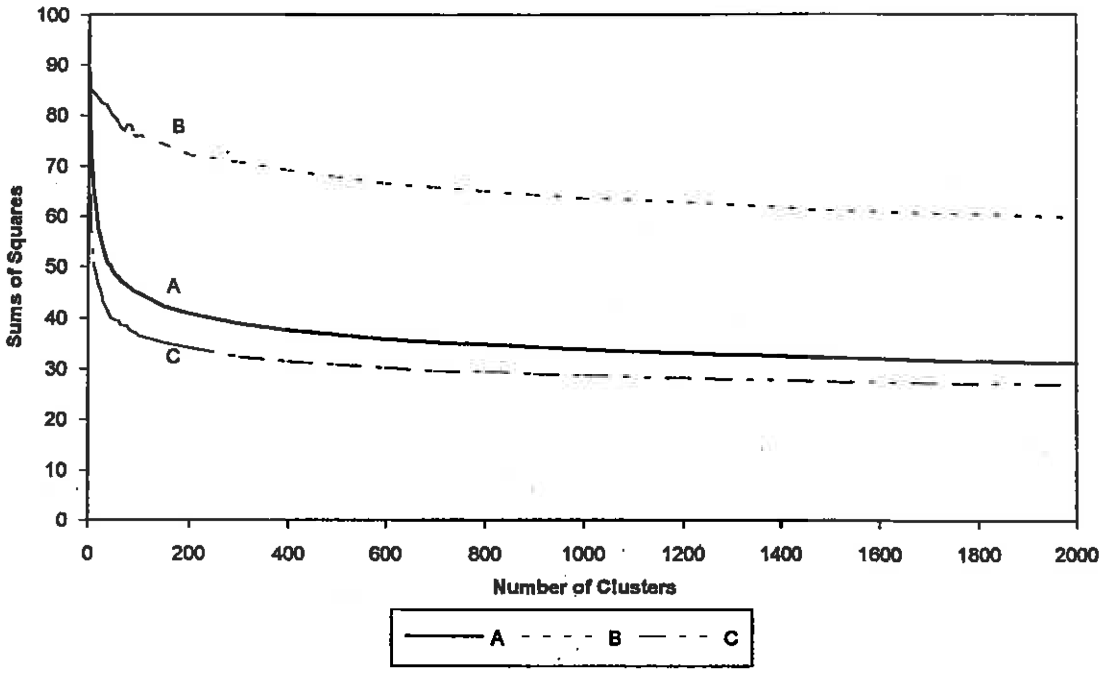


Figure 3: The Distribution of the Number of Different Clusters "Near" to each ED

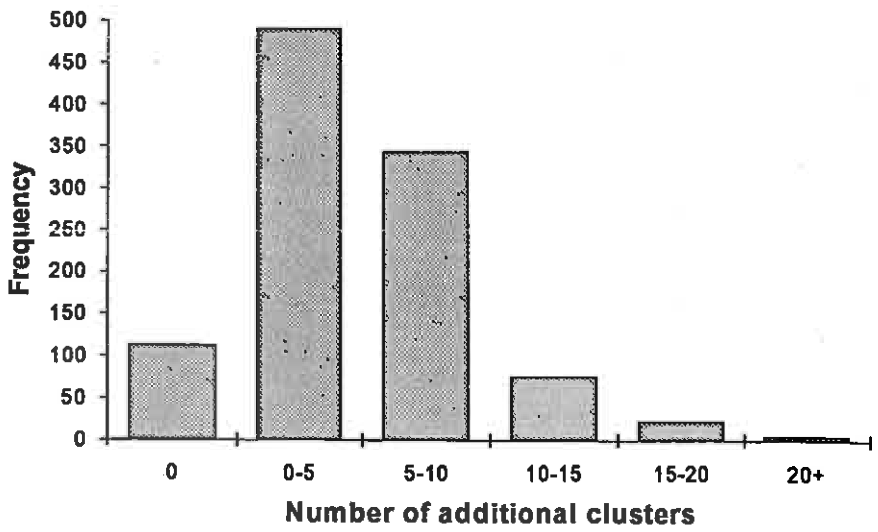


Figure 2: The Distribution of Major Classes identified using the Neuroclassification procedure

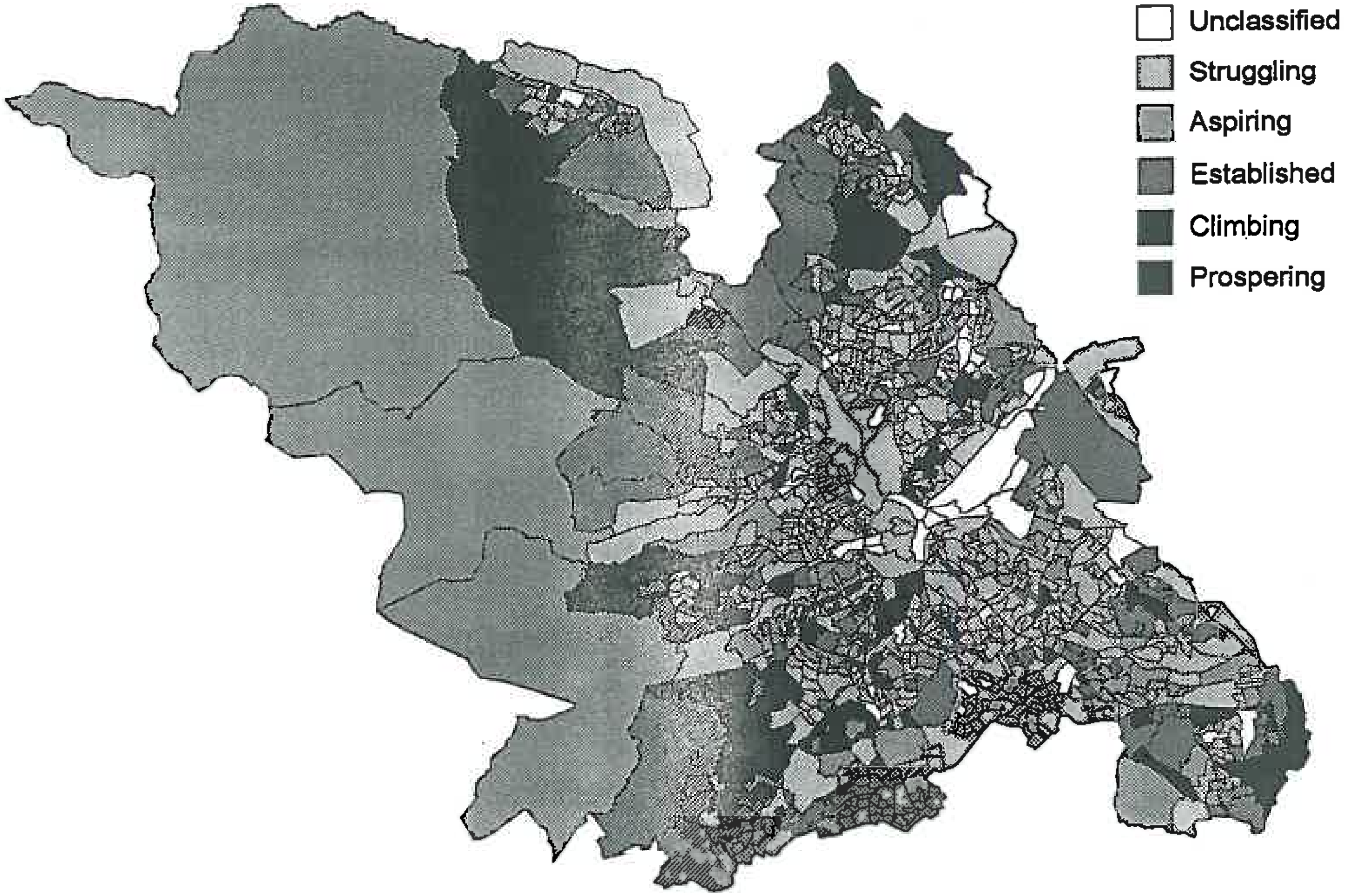


Figure 4: Distribution of "Uncertain" ED's within Sheffield

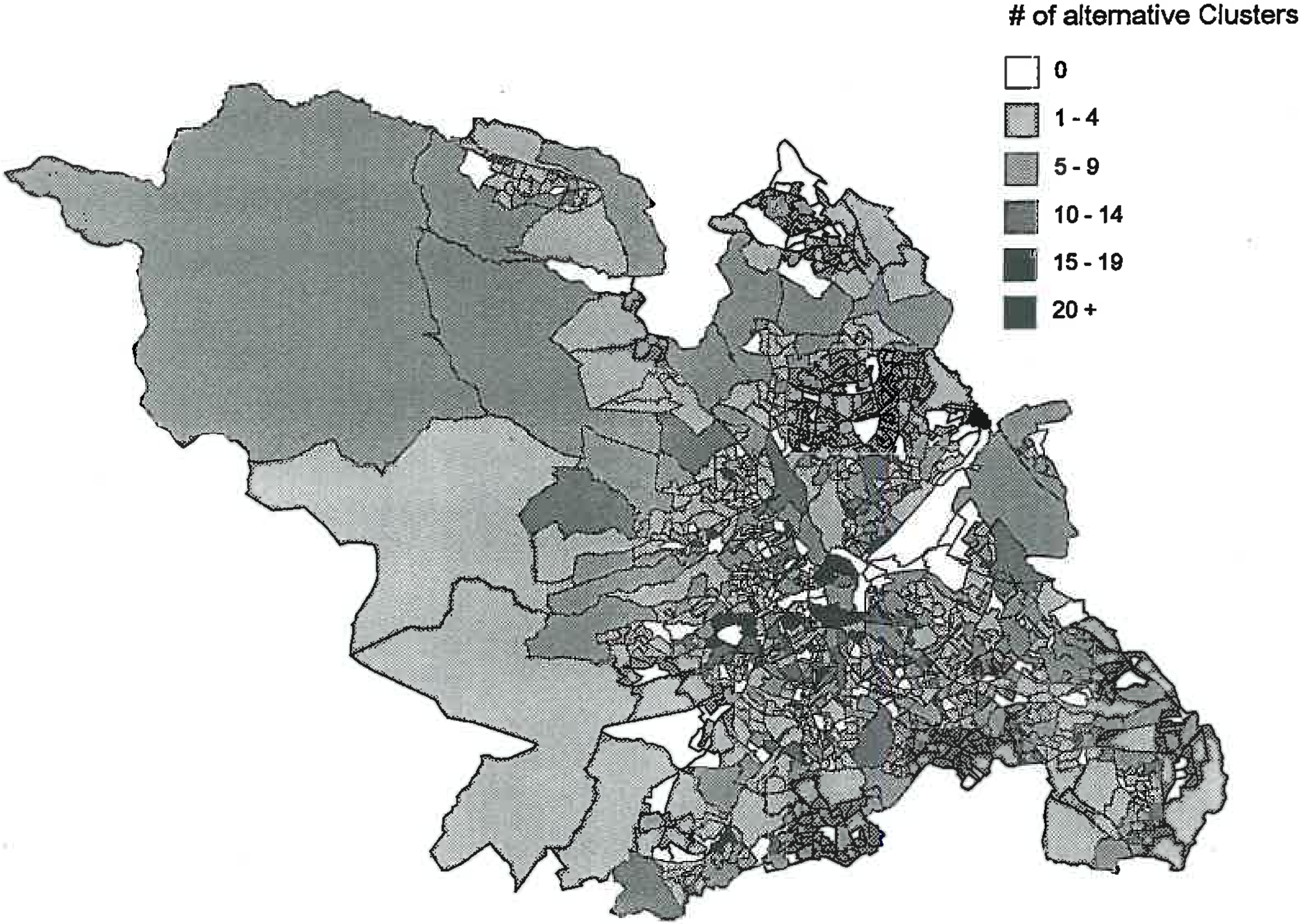
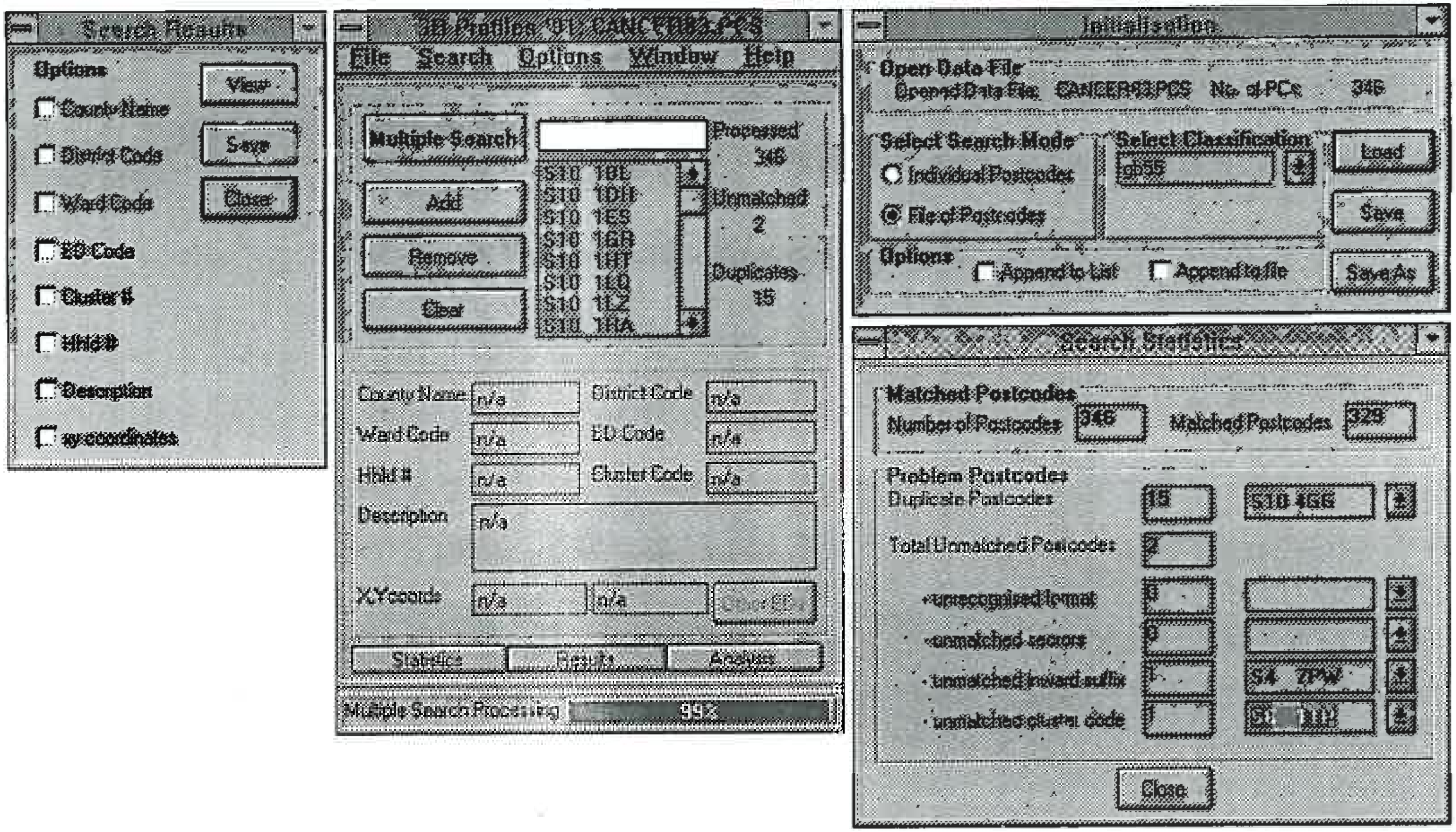


Figure 5: Layout of GB Profiles '91 in Multiple Search Mode



Appendix 1: The Variables used in both the Classification Procedures

Demographic Variables

Ref#	Description	10%
1	resident persons in the 0-4 age grp.	
2	resident persons in the 5-14 age grp.	
3	resident persons in the 15-24 age grp.	
4	resident persons in the 25-44 age grp.	
5	resident persons in the 45-64 age grp.	
6	resident persons in the 65-74 age grp.	
7	resident persons in the 75-84 age grp.	
8	resident persons in the 85+ age grp.	
9	resident persons who are single	
10	hhlds(with residents) with children, that have two or more adults	
11	female residents who are between 16 & 45	
12	resident persons that are married	
13	residents who are single parents	
14	resident persons who are of pensionable age	
15	persons aged 16+ who are students	

Ethnic Variables

16	residents who are white	
17	residents who are black	
18	residents who are Indian	
19	residents who are Pakistani	
20	residents who are Bangladeshi	
21	residents who are Chinese & others	

Migration Variables

22	residents that moved last year	
23	residents that are pensioner migrants	

Housing Variables

24	all permanent hhlds that are owned outright	
25	all permanent hhlds that are mortgaged	
26	all permanent hhlds that are HA rented	
27	all permanent hhlds that are LA rented	
28	all permanent hhlds that are unfurnished rented	
29	all permanent hhlds that are furnished rented	
30	all hhld spaces that are detached	
31	all hhld spaces that are semi-detached	
32	all hhld spaces that are terraced	
33	all hhld spaces that are purpose built flats	
34	all hhld spaces that are converted flats	
35	all hhld spaces that are bedsits	
36	all permanent hhlds with no central heating	
37	all permanent hhlds with no/shared bath/shower/WC	
38	hhlds with residents which are overcrowded	

39	hhlds with residents which are very overcrowded	
40	hhlds with residents which have more than 6 rooms	
41	Number of rooms per hhld	
42	Rooms per person	
43	Average hhld size (rooms per hhld)	
44	hhlds with residents with 2 or more cars	
45	Average number of cars per hhld	

Household Composition Variables

46	hhlds with residents with 2 or more e.a. persons and no children	
47	hhlds with residents with a single e.a. person and no children	
48	hhlds with residents with a married couple	
49	hhlds with residents with children	
50	hhlds with residents with children and no car	
51	hhlds with residents with a single pensioner	
52	hhlds with residents with a single non-pensioner	
53	hhlds with residents with more than three adults	
54	residents aged 16+ in hhlds who are aged 16-24 and are without children	
55	residents aged 16+ in hhlds who are aged 16-24 and have children	
56	residents aged 16+ in hhlds who are aged 25-34 and are without children	
57	residents aged 16+ in hhlds who are aged 25-34 and have children	
58	residents aged 16+ in hhlds who are aged 35-54 and are without children	
59	residents aged 16+ in hhlds who are aged 35-54 and have children	
60	residents aged 16+ in hhlds who are aged 55-74 or more	

Socio-economic Variables

61	residents aged 16+ and over (employed & self-employed) that are in SEG 1,2,3 & 4	yes
62	residents aged 16+ and over (employed & self-employed) that are in SEG 5 & 6	yes
63	residents aged 16+ and over (employed & self-employed) that are in SEG 8, 9 & 12	yes
64	residents aged 16+ and over (employed & self-employed) that are in SEG 7 & 8	yes
65	residents aged 16+ and over (employed & self-employed) that are in SEG 11	yes
66	residents aged 16+ and over (employed & self-employed) that are in SEG 16 & 17	yes
67	residents aged 16+ and over (employed & self-employed) that are in manufacturing & mining	yes
68	residents aged 16+ and over (employed & self-employed) that are in agriculture	yes
69	residents aged 16 and over who are self- employed	
70	residents aged 16 and over who are unemployed	
71	residents aged 16 and over who are permanently sick	
72	residents aged 16 and over who are working (employers or employees) women	
73	residents aged 16+ and over (employed & self-employed) that are women working in manufacturing (metal etc. not other manuf.)	
74	residents aged 16+ and over (employed & self-employed) that are women working more than 41 hours per week	
75	residents aged 16 and over who work part-time	
76	male workers	
77	residents aged 16+ in hhlds who are female, married and working	
78	Proportion or residents aged 18 and over with a (higher) degree	yes
79	hhlds with residents with 2 or more adults in employment	

Health Variables

80	residents (S02) with LLI	

81	residents (S02) economically inactive with LLI	

Travel-to-work Variables

82	residents aged 16+ and over who work at home	yes
83	residents aged 16+ and over who go to work by car	yes
84	residents aged 16+ and over who go to work by train/bus	yes
85	residents aged 16+ and over who walk to work	yes

Appendix 2: Labeling System for the 64 clusters identified using the Neuroclassification Procedure

Group	Sub-group	Name	Cluster #
Struggling	Council Tenants with multiple social problems	Multi-ethnic council tenants	1
		LA rented Semis	24
		Overcrowded Council Housing	33
		Council tenants in Tower Blocks	6 & 7
		Single Parents Council tenants	29 & 34
		Single Parents in Tower Blocks	28 & 30
		Unskilled Council tenants	45
	Multi-ethnic, low income areas	Bangladeshi Areas	4
		Indian Areas	38
		Multi-ethnic Bedsit Areas	8 & 27 & 32
		Poor multi-ethnic singles	62
	Less Well-off Terraces	Terraces	2 & 36
		LA rented terraces	10 & 35
	Fading Industrial Areas	Industrial terraces	43 & 61
		Industrial Council tenants	51
	Less Well-off Pensioners	Pensioners Council tenants	17& 25 & 31
		Pensioners in converted flats	18 & 26
		Pensioners in HA rented terraces	57
Aspiring	Young Singles in Flats	Poor young singles & Students	3, 55 & 60
		Singles in PBFs	53
		Better-off singles	14 & 54
	Better-off Council Tenants	Council Semis	13
	Rural Communities	Rural areas	44 & 52
	Armed Services	Young Armed Services Families	12
Established	Semi-detached Suburbia	Semis	56
		Mortgaged Semis	63
		Owner occupied Semis	5
	Better-off Pensioners	Pensioner Migrants	15, 16, 23 & 59
	Comfortable Middle Agers	Middle Class Suburbia	37
		Wholly owned Semis	21
		The average	20
Climbing	Metro Singles	Well-off singles in Bedsits	14
		Well-off singles in PBFs	19
		Well-off singles in converted flats	50
	Academic centers	Students in Bedsits	41
Prospering	Wealthy Achievers	Middle aged Managers	46 & 58
		Well-off Middle Aged Managers	9 & 47
		Self-employed Managers	48
		Educated Professionals	22
	Wealthy Rural Communities	Rich Agriculturalists	11, 49, 39 & 64
Unclassified			40
			42

Produced By
School of Geography
University of Leeds
Leeds LS2 9JT
From Whom Copies May Be Ordered

August 1994