*UNIVERSITY OF SOUTHAMPTON*

*FACULTY OF SCIENCE*

*DEPARTMENT OF GEOGRAPHY*

# An Integration Of Aggregate And Disaggregate Census Data

**by Richard James Lamacraft Mitchell**

*Thesis Submitted for the degree of Doctor of Philosophy*

*August 1997*

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE

GEOGRAPHY

Doctor of Philosophy

# AN INTEGRATION OF AGGREGATE AND DISAGGREGATE CENSUS DATA

by Richard James Lamacraft Mitchell

Identifying the different social characteristics and location of residential populations is an important task (Shevky and Bell 1955), both for social theorists and for the machinery of state and administration. Currently, the data with which the relationship between social character and residential location is explored, are inadequate for the task. This inadequacy stems from the lack of detailed information about social structure at a fine spatial scale (Birkin and Clarke, 1995). The source most commonly applied to issues of socio-spatial structure is the Small Area Statistics (SAS), drawn from the 1991 UK Census Of Population. The SAS data are aggregate and do not provide detailed information on the characteristics of households or individuals for the small spatial areas which they describe. They are therefore often used to provide single classifications of areas, based upon their aggregate census characteristics.

In 1993 a new, disaggregate, data set was released from the 1991 Census of Population. The Samples of Anonymised Records (SAR) provide considerable detail about the characteristics of two samples of households and individuals. However, a restriction on the spatial resolution of these data is employed to assist in preserving the anonymity of sample members. Alone, the SAR is not suited to the detailed analysis of socio-spatial structure required.

Through the use of social theory (Harvey, 1989), an analogy may be drawn between the problems in using aggregate census data and those faced by Earth Observation (EO) researchers in the analysis of satellite imagery. EO researchers have developed solutions to these problems, through the fuzzy classification of small areas. Such an approach is adapted for use with census data, through deriving a classification of households from the SAR and inferring the proportional presence of these household classes in a small area, from the SAS.

The fuzzy classification represents new information at a level beneath which the aggregate data are released. This is shown to have advantages over traditional single area classification, both for academic research and more practical tasks. The principles of this technique may also be applied to other aggregate and disaggregate data sets.

Shevky, E., Bell, W. (1955) *Social Area Analysis*, Stanford University Press, Stanford, California

Birkin, M., Clarke, M. (1995) Using microsimulation methods to synthesise census data in Openshaw, S. (ed.) (1995) *Census Users' Handbook*, GeoInformation International, Cambridge

Harvey, D. (1989) *The Urban Experience*, Blackwell, Oxford

# *Contents*


Table of Tables ..... 1
Table of Tables (contd.) ..... 2
Table of Figures ..... 3
Table of Figures (contd) ..... 4
Acknowledgements ..... 5
Abbreviations Used ..... 5
Chapter 1 ..... 6
Chapter Overview ..... 6
1.0 - Research Themes and Aims ..... 6
1.1 - Thesis Structure ..... 7
1.1.1 - Chapter 2 ..... 7
1.1.2 - Chapter 3 ..... 7
1.1.3 - Chapter 4 ..... 8
1.1.4 - Chapter 5 ..... 8
1.1.5 - Chapter 6 ..... 8
1.1.6 - Chapter 7 ..... 9
1.1.7 - Chapter 8 ..... 9
1.1.8 - Chapter 9 ..... 9
1.2 - Chapter Structure ..... 9
1.3 - Text, Table and Figure Referencing ..... 9
1.4 - Summary ..... 10
Chapter 2 ..... 11
Chapter Overview ..... 11
2.0 - A Reason For Interest ..... 11
2.1 - Academic Interest ..... 12
2.2 - Government and Service Interest ..... 13
2.3 - Commerce ..... 14
2.4 - Sources of Data ..... 15
(i) Primary Data Generation ..... 15
(ii) Secondary Data ..... 16
2.5 - Aggregate Data ..... 17
2.6 - The Ecological Fallacy ..... 17
2.7 - Small Area Statistics ..... 18
2.8 - Geodemographics ..... 19
2.9 - Variation at a finer scale ..... 23
2.10 - The Samples of Anonymised Records ..... 25
2.11 - A List of Desires ..... 28
2.12 - Two Desires Satisfied ..... 29

2.13 - A New Analysis ..... 31
2.13.1 - The Result of An Integration of SAR and SAS Data ..... 31
2.14 - Conclusions and Summary ..... 31
Chapter 3 ..... 33
Chapter Overview ..... 33
3.0 - Introduction ..... 33
3.1 - Quantification and Science in Geography ..... 34
3.2 - From Regional Geography to Quantification ..... 35
3.3 - Quantification Crisis: The rise of humanism ..... 36
3.4 - Meanwhile ..... 36
3.5 - Geographic Information Systems ..... 37
3.6 - Criticisms of GIS ..... 38
3.6.1 - Remote ..... 39
3.6.2 - Reliance on Secondary Data ..... 40
3.6.3 - Data Treated as Unproblematic ..... 41
3.6.4 - Positivistic ..... 42
3.6.5 - Powerful ..... 42
3.6.6 - Description, not Explanation ..... 43
3.7 - Reflexive Action ..... 44
3.7.1 - Awareness of Power ..... 45
3.7.2 - Awareness of technological and personal influence ..... 45
3.7.3 - Beyond Awareness ..... 45
3.8 - Structural Class Schemes ..... 46
3.9 - A Concern with Cause ..... 47
3.10 - New Concepts of Class ..... 48
3.11 - Geodemography ..... 49
3.12 - Space, Socialisation and Process ..... 49
3.13 - Using This Understanding ..... 51
3.14 - Conclusions and Summary ..... 51
Chapter 4 ..... 53
Chapter Overview ..... 53
4.0 - Introduction ..... 53
4.1 - Foreign Microdata Sources ..... 53
4.2 - What Are Microdata Used For ? ..... 54
4.2.1 - How Attributes Relate to Each Other ..... 54
4.2.2 - Spatial Analysis ..... 55
4.3 - Integration ..... 57

4.3.1 - Commercial Attitudes to Census Microdata ..... 58
4.4 - Microsimulation Techniques ..... 59
4.5 - Genetic Algorithms ..... 63
4.6 - Data Fusion ..... 65
4.6.1 - How Does it Work? ..... 66
4.6.2 - What Can it be Used For and How Good Is It? ..... 68
4.6.3 - Why is Data Fusion Not Appropriate For This Research? ..... 69
4.7 - New Ways Forward ..... 70
4.8 - Conclusions and Discussion ..... 70
Chapter 5 ..... 73
Chapter Overview ..... 73
5.0 - Introduction ..... 73
5.1 - Theory and Empirical Information ..... 73
5.2 - Theory, Reality and a Way Forward ..... 75
5.3 - The Earth Observation Analogy ..... 77
5.4 - Earth Observation Solutions ..... 79
5.4.1 - Mixture models ..... 79
5.4.2 - Fuzzy classification ..... 80
5.4.3 - Neural network approaches ..... 81
5.5 - Artificial Neural Networks ..... 82
5.5.1 - Components and Mechanisms ..... 82
5.5.2 - Learning ..... 84
5.6 - Unmixing Pixels in Practice ..... 85
5.7 - Unmixing Enumeration Districts ..... 86
5.7.1 - Defining Social Groups ..... 86
5.7.2 - Generating Training Data ..... 88
5.8 - From Data to Model: A Process Structure ..... 90
5.9 - Summary and Conclusions ..... 90
Chapter 6 ..... 92
Chapter Overview ..... 92
6.0 - Introduction ..... 92
6.1 - Data Selection ..... 93
6.1.1 - Variable Choice In Practice ..... 94
6.2 - Data Extraction ..... 96
6.2.1 - Dealing With Error in the SAR ..... 96
6.3 - Exploring Univariate Schemes ..... 97
6.4 - Multivariate Classification ..... 99

6.4.1 - Conventional Approaches..... 99
6.4.2 - Neural Approaches ..... 100
6.4.3 - GroupMaker: A Custom Written Classifier..... 101
6.4.4 - Principles of Distance and Controlling Classification..... 103
6.5 - Analysing Group Schemes..... 104
6.5.1 - Visualisation ..... 105
6.5.2 - Numerical Analysis ..... 105
6.6 - Initial Exploration ..... 107
6.7 - Synthetic Small Area Statistics Generation ..... 108
6.7.1 - Recipes for Enumeration Districts..... 108
6.7.2 - Synthetic Enumeration District Generation ..... 108
6.7.3 - Small Area Statistics from Enumeration Districts ..... 109
6.8 - Neural Network Training ..... 110
6.8.1 - Network Architecture and Parameters.. 110
6.9 - Neural Network Testing ..... 112
6.10 - Using The Trained Network to Produce a Model..... 112
6.10.1 - Extraction of Real Small Area Statistics..... 112
6.10.2 - Model Creation ..... 113
6.11 - Summary and Conclusions..... 114
Chapter 7..... 115
Chapter Overview..... 115
7.0 - Introduction..... 115
7.1 - The Sequence of development..... 116
7.2 - Obtaining a Working Set of Variables ..... 117
7.2.1 - Variables with little specific influence in the classification..... 117
7.2.2 - Variables with Redundant Categories.... 117
7.3 - Production of a Six Class Test Classification ..... 119
7.4 - The First ANN Unmixing..... 120
7.5 - Results ..... 121
7.5.1 - Testing the six class model on EDs in Southampton ..... 122
7.6 - Production of a 12 Class Scheme ..... 123
7.6.1 - Production of a Basic Scheme..... 124

7.6.2 - Manipulating the Basic Scheme..... 125
7.6.3 - A Working Structure ..... 128
7.7 - ANN Development ..... 130
7.7.1 - Order of Experimentation..... 130
7.7.2 - Key Findings From Experiments with ANN Parameters ..... 131
7.8 - Testing on Southampton EDs..... 133
7.8.1 - Failure to Account for All Variation ..... 133
7.8.2 - Difficulty of Model Interpretation ..... 134
7.9 - Refinements..... 134
7.9.1 - Further Classification..... 135
7.9.2 - Experiments and Failure ..... 137
7.10 - K-Means Cluster Analysis..... 138
7.10.1- Cluster Results..... 138
7.11 - ANN Training with Fifteen Classes..... 139
7.12 - Forward to Testing ..... 141
7.13 - Conclusions and Summary ..... 142
Chapter 8..... 144
Chapter Overview..... 144
8.0 - Introduction..... 144
8.1 - How Accurate is the Unmixed Model?..... 144
8.1.1 - Comparison With A Single Area Classification..... 145
8.1.2 - Inspection on the ground..... 151
8.1.3 - Reconstruction of Small Area Statistics from Model Outputs..... 156
8.1.4 - Results..... 158
8.2 - New Information From This Research..... 163
8.2.1 - Analysis in Aggregate ..... 163
8.2.2 - Social Mixing and the Reliability of Single Area Classification..... 166
8.2.3 - Underlying Class Presence..... 168
8.2.4 - Information About Different Class Behaviours ..... 168
8.2.5 - Detailed Analyses of Small Areas..... 171
8.2.6 - How good is the model?..... 172
8.3 - Future Developments?..... 172
8.3.1 - Different SAR-based Classifications..... 173
8.3.2 - Variables of Interest Associated With SAR- based Classes..... 173

8.3.3 - Non Census Based Disaggregate Data.... 174
8.3.4 - Non Census Based Aggregate and Disaggregate Data ..... 174
8.4 - Conclusions ..... 175
Chapter 9 ..... 177
Chapter Overview ..... 177
9.0 - Introduction ..... 177
9.1 - Technical Achievements ..... 177
9.2 - Problems with Innovation ..... 179
9.2.1 - Limitations on Model Testing ..... 179
9.2.2 - Quantity and Nature of The New Information ..... 180
9.3 - Theory / Data / Method ..... 181
9.4 - Key Messages ..... 183
Appendix I: The Small Area Statistics ..... 184
Appendix II - The Samples of Anonymised Records .. 186
Appendix III: Backpropagation in SNNS ..... 190
Appendix IV - Experiments to Determine Optimum ANN Parameters ..... 191
Appendix V - GroupMaker Code ..... 196
Reference List ..... 211

## *Table of Tables*


| No. | Title | Page |
|---|---|---|
| 2.1 | A Description of The Geodemographic Classes In Figure 2.4 | 22 |
| 2.2 | Measures to Preserve Confidentiality in the Sample of Anonymised Records | 27 |
| 2.3 | Desirable Enhancements to Socio-Spatial Information | 28 |
| 3.1 | The Registrar General's Class Structure | 47 |
| 4.1 | Extract From A Crosstabulation Of The 14-Cluster Classification by Regions | 56 |
| 4.2 | Results From Bronner's (1983) Data Fusion Experiment | 69 |
| 4.3 | Distribution of Error in Bronner's (1983) Data Fusion Experiment | 69 |
| 6.1 | Initial Variable Set for Classification | 94 |
| 6.2 | Ensuring Variation in Training Data | 108 |
| 6.3 | Initial ANN Architecture and Formulae | 111 |
| 7.1 | Variable Adjustments Following Initial Experiments | 118 |
| 7.2 | Variables For Six Class Scheme | 119 |
| 7.3 | Six Class Scheme: Class Characteristics | 120 |
| 7.4 | ANN Configuration For Six Class Training | 120 |
| 7.5 | Results from ANN Training Using a Six Class Scheme | 121 |
| 7.6 | Proportional Presence of Each Class in EDs | 122 |
| 7.7 | Initial Variable Weights In The Classification Process | 125 |
| 7.8 | Variable Weights Used In An Interim Classification | 126 |
| 7.9 | An Interim Classification: Class Characteristics | 127 |
| 7.10 | Variable Weights Used To Produce 12 Class Scheme | 128 |
| 7.11 | Twelve Class Scheme: Class Characteristics | 129 |
| 7.12 | Optimal ANN Parameters For A Twelve Class Model | 132 |
| 7.13 | Typical Correlation Values Between ANN Output and Known Target Values, by Class | 133 |
| 7.14 | The Variable 'Household Space Type' | 136 |
| 7.15 | Variable Set Used for Further Classification Experiments | 137 |
| 7.16 | Descriptions of the Fifteen Classes Produced by Cluster Analysis | 140 |
| 7.17 | Numbers of Cases in Each of the Fifteen Groups | 139 |

## *Table of Tables (contd.)*


| No. | Title | Page |
|---|---|---|
| 7.18 | Optimum Parameters For Use With A Fifteen Class Model | 139 |
| 7.19 | Correlations Between Target Values and ANN Outputs | 141 |
| 8.1 | An Illustration Of Proportionally Dominant Class | 146 |
| 8.2 | Class descriptions for GBprofiler | 147 |
| 8.3 | Class Descriptions for the Fifteen Class Scheme | 149 |
| 8.4 | Information for ED Visits | 152 |
| 8.5 | SAS Re-building; A Summary of Results | 161 |
| 8.6 | Aggregate Analysis of Southampton from Unmixing | 164 |
| 8.7 | Aggregate Analysis of Southampton from GBprofiler | 164 |
| 8.8 | Comparison of the Aggregate Analyses of Southampton | 165 |
| 8.9 | The Distribution of Classes In Southampton | 170 |
| AII.1 | Variables Available in The SAR | 187 |
| AIV.1 | Correlations From Experiments With ANN Parameters | 194 |
| AIV.2 | Class By Class Correlations for Test 19; Test Data Set 1 | 194 |
| AIV.3 | Class By Class Correlations for Test 19; Test Data Set 2 | 195 |

## *Table of Figures*


| No. | Title | Page |
|---|---|---|
| 2.1 | Burgess's Urban Model | 12 |
| 2.2 | Origins Of Small Area Statistics | 18 |
| 2.3 | The Geodemographic Link Between Place and Identity | 20 |
| 2.4 | A Geodemographic Analysis of Southampton | 21 |
| 2.5 | (plate) Inside ED JPFM12 | 24 |
| 2.6 | Graph Representing a Geodemographic Classification | 24 |
| 2.8 | How SAR and SAS Data Are Related | 26 |
| 2.7 | A New Model: Social Mix in Southampton | 30 |
| 3.1 | Criticisms of GIS | 38 |
| 4.1 | Sequence of Attribute Simulation | 61 |
| 4.2 | Data Fusion in Principle | 66 |
| 5.1 | Social Groups in the City | 76 |
| 5.2 | A Mixed Pixel | 78 |
| 5.3 | A Mixed ED | 79 |
| 5.4 | Fuzzy Classification | 81 |
| 5.5 | ANN Architecture | 83 |
| 5.6 | Sub-Optimal and Optimal Learning | 84 |
| 5.7 | The Use of ANNs in Pixel Unmixing | 85 |
| 5.8 | Training Data Generation | 89 |
| 5.9 | Integration In Principle - An Outline | 90 |
| 6.1 | The Unmixing Process: Steps Explored in Chapter 6 | 93 |
| 6.2 | SAR Data Extraction Syntax | 96 |
| 6.3 | Univariate Classification | 98 |
| 6.4 | Flow Chart for the Custom Classifier 'GroupMaker' | 102 |
| 6.5 | An Example of PV-Wave Visualisation | 106 |
| 6.6 | Output from Numerical Analysis of Groups | 107 |
| 6.7 | Number of Groups Vs Distance Threshold | 107 |
| 6.8 | SPSS and Synthetic ED Generation | 109 |
| 6.9 | Typical ANN Training Curve | 111 |
| 6.10 | A Sample SASPAC91 Syntax File | 113 |
| 6.11 | Output Matrix from Unmixing | 113 |
| 7.1 | Integration in Principle - An Outline | 116 |
| 7.2 | Factors Influencing The Number of Classes to be Used | 123 |

## *Table of Figures (contd)*


| No. | Title | Page |
|---|---|---|
| 7.3 | Range of ANN Parameters With Influence Over Performance | 130 |
| 7.4 | The ANN Training Maze | 131 |
| 7.5 | Order of Adjustment In Search for Optimum ANN Parameters | 131 |
| 8.1 | Key to GBprofiler's Ten to Five Class Collapse | 146 |
| 8.2 | Key to The Unmixing Classification, Fifteen To Five Class Collapse | 148 |
| 8.3 | Five Class Images of the City | 150 |
| 8.4 | Location of EDs Visited in Model Testing | 153 |
| 8.5 | Rebuilding SAS Data From The Unmixed Model | 157 |
| 8.6 | A Potential Source of Error in the SAS Re-building Process | 158 |
| 8.7 | Mapped Error and Entropy | 160 |
| 8.8 | Entropy as a Measure of Social Mixing | 167 |
| 8.9 | Class 12 Proportional Presence vs. Dominant Presence | 169 |
| 9.1 | The Potential for Generating a Neighbourhood Surface | 178 |

## *Acknowledgements*

Firstly, I would like to thank Dr David Martin for the enormous amount of advice, support and supervision which he has provided during the last three (and a half) years. There would have been no thesis without him. Thanks too, to Prof. Giles Foody for introducing me to the mysteries of neural nets and for learning to speak 'census'.

I would like also like to thank all the postgrads at Southampton for making the 'full-time' years so enjoyable and especially Daisy for keeping me vaguely sane during the 'part-time' era. Thanks to Wim, for being in the park when I needed a chat, to Carol for going on holiday so that I could actually finish this thing, to Laura for maths and of course, to M & D for everything.

Finally, I gratefully acknowledge the assistance provided by the ESRC. This research was funded under the Large And Complex Data Sets initiative, award number R00429334242.

---

## *Abbreviations Used*

ANN - Artificial neural network
ED - Enumeration district
EO - Earth Observation
SAR - Samples of Anonymised Records
SAS - Small Area Statistics
UM - Unmixed Model

# Chapter 1

## Chapter Overview

This chapter introduces the themes and structure of the thesis. The research questions are presented briefly, followed by a preview of the contents of each chapter. Practical matters, regarding the referencing of text, tables and figures are also dealt with.

## 1.0 - Research Themes and Aims

The main theme of this research was information about the differing characteristics of individuals and households, and their spatial location. The thesis describes the development of a methodology which enhances the information available for exploring the relationship between location and social character of residential populations. This relationship will be referred to as the *socio-spatial structure*.

Currently, publicly available data about socio-spatial structure are problematic. A user interested in the social characteristics of a population *and* the fine scale spatial patterning of those characteristics, faces a choice. They can use data which describe characteristics at the household or individual level in detail but have a poor spatial resolution, or they can use data which offer a poor description of household or individual character, at a reasonably fine spatial resolution. There are no publicly available data which provide detailed information about social characteristics at a fine spatial resolution.

Information about where different kinds of people live is extremely important for many aspects of our society (see chapter 2). The allocation of resources for example, is often tailored to areas in which a population has been shown to be in greater need. Determining 'need' and its spatial location requires information about the location of different kinds of people. The greater the accuracy with which need and spatial location can be matched, the greater the possibility for achieving justice in the distribution of resources. Much research into society and its processes also depends heavily on an understanding of the differences between people's characteristics and where or how they live together. The research presented in these chapters was about enhancing the information available with which to complete these studies and operations. Chapter 2 provides an extended discussion of the kinds of work

which use socio-spatial information and the reasons why this information needs improvement.

The research was not only about the production of a new kind of information. It was also very much concerned with an exploration of *how* research of this nature is carried out. The relationship between theories of socio-spatial structure and process and quantitative research techniques, was explored in some depth. The interface between theory and technical development was the focus of this exploration. Its aim was to gain a precise understanding of what each of these elements brought to the research process and how they influenced its final outcome.

These themes are explored in the following chapters. To orientate the reader, the next section provides a brief outline of the thesis structure.

## *1.1 - Thesis Structure*

The thesis contains nine chapters.

### 1.1.1 - Chapter 2

In chapter 2, the reader is introduced to the context for the research. The chapter explores the reasons for interest in socio-spatial structure and the variety of parties which have that interest. It highlights the limitations of the data which are suitable and publicly available and proposes a set of desirable enhancements. These data are introduced and described. They are the Samples of Anonymised records and the Small Area Statistics, both drawn from the 1991 UK Census of Population.

The research produced a new level of information about socio-spatial structure and an example of this information is presented for The City of Southampton in the UK. The juxtaposition of problematic models of socio-spatial structure and an enhanced version is intended to emphasise the two poles which the subsequent chapters will connect.

### 1.1.2 - Chapter 3

Chapter 3 situates the research within academic geography. Contemporary criticisms of quantitative geography are reviewed and addressed. In the light of these criticisms, a number of practical actions to enhance the quality of the research process are proposed and these are implemented in later chapters. This chapter also reviews theories of social and socio-spatial structure and the role of these theories in understanding criticisms of quantitative method is

made clear. Chapter 3 then acts as a starting point for the exploration of the relationship between theory and quantitative research, introduced in section 1.0.

### 1.1.3 - Chapter 4

Chapter 4 reviews relevant empirical work. There are two main themes to this review. Firstly, there is an exploration of work using *data sources* which might have facilitated the proposed enhancements. Some research of direct relevance is found, but none which had achieved the desired aims. Secondly, there is an account of existing *techniques* which might have been used to enhance the fine scale socio-spatial information. Again, no entirely suitable technique is found. The chapter concludes that a new technique will have to be developed, but uses the existing work to further refine a specification for that technique. The shortcomings of existing work become hurdles for this research to overcome and the output of one particular piece of research is proposed as a suitable outcome.

### 1.1.4 - Chapter 5

This chapter illustrates the use of the theoretical understanding gained in chapter 3, to reconceptualise the data sources described in chapter 2. A new approach to the research task is suggested by this reconceptualisation. The chapter describes how an analogy can be drawn between the difficulties faced by social scientists in exploring socio-spatial structures and those faced by earth observation scientists in satellite image processing. Earth observers have developed techniques to deal with their problems. By using a fuzzy classification model, these techniques enhance information about variation within aggregate data for small areas. The adaptation of one such technique is proposed for social geography. This adaptation is outlined in principle, together with an introduction to the neural technology which will be needed.

### 1.1.5 - Chapter 6

Each stage of the methodology proposed in chapter 5 is made operational in chapter 6. This represents an adaptation of the census data for use in the earth observation technique. At the close of chapter 6, the range of manipulations and software which will facilitate the modelling of fine scale socio-spatial structure has been presented. The work described in the chapter is akin to building a tool kit for the research task. The operation of the tools and their refinement is then presented in chapter 7.

### 1.1.6 - Chapter 7

This chapter presents an account of the development and refinement of the modelling technique. Presented in chronological order, the chapter illustrates a succession of improvements and adjustments to the technique and the data set it employed. Some testing of the model is described but these tests acted as a guide to development rather than a rigorous evaluation. During the chapter, the relationship between relevant social theory and the technical constraints on the research process is demonstrated. This relationship is then discussed in its concluding section.

### 1.1.7 - Chapter 8

Chapter 8 serves two purposes. It begins with a description of the testing which took place on the new model. Since the research had produced a new level of information, there were no comparable data with which to test the model. However, a variety of approaches to its evaluation were undertaken and are presented here. The chapter then presents an analysis of the City of Southampton to demonstrate the model's potential and to determine its shortcomings. This analysis is also used to provide suggestions for improving the model. The relationship between relevant socio-spatial theories and the new information provided by the model is discussed.

### 1.1.8 - Chapter 9

A summary and series of conclusions from the research are presented in chapter 9. The achievements are reviewed and the relative success of the focus on theory and practice is discussed. Potential developments for the model are suggested.

## *1.2 - Chapter Structure*

Each chapter starts with an overview which describe its contents and how these relate to other chapters. The text is then broken down into a number of sub-sections for ease of reference. Every chapter closes with a 'conclusions and summary' section to highlight the key points made in the chapter and to prime the reader for the next one.

## *1.3 - Text, Table and Figure Referencing*

Throughout the thesis, reference is made to text, tables and figures in different chapters. The first digit in every reference refers to the chapter in which the text, table or figure is found. Where the reference is to text, it will be simply the section number within which the reference is to be found. Tables and figures however, are numbered sequentially within chapters. For

example, “figure 2.3” refers to the third figure found in chapter 2. The tables at the front of the thesis provide the relevant page numbers.

## *1.4 - Summary*

This chapter has introduced the reader to the thesis structure and provided a very brief preview of each chapter. Chapter 2 now sets out the context for the research and its main questions:

# Chapter 2

## Chapter Overview

The principal aim of chapter 2 is to set the context of this research. It begins by considering the reasons for interest in socio-spatial structure and illustrating the variety of interested parties. The assumption of a link between residential location and social character is common to many of these parties and this assumption is briefly explored. Given this demonstration of interest in 'who lives where', the question of how users obtain and employ their information is raised. Aggregate data from the 1991 UK Census are explored and some difficulties with their use are highlighted. Limitations in the use of aggregate data are illustrated with a geodemographic analysis of Southampton in Hampshire. A second census data set (the Samples of Anonymised Records) is introduced and its own incompatibility with socio-spatial analysis is explained. A list of desirable improvements to the information available for socio-spatial research is subsequently proposed. An alternative socio-spatial model for Southampton, is then introduced and contrasted with the geodemographic analysis of the city. The two models illustrate the position between which the thesis will span. The new model achieves some of the desired enhancements to the information available for socio-spatial analysis. Finally, it is suggested that the production of this new model will have ramifications beyond geography. The chapter does re-visit some apparently obvious features of differing approaches to socio-spatial structure, however this is necessary in order that their assumptions and weaknesses can be clearly identified.

## 2.0 - A Reason For Interest

This research is primarily about trying to understand what kinds of people live in specific locations. It is important to be clear about why this topic is significant. Why is anybody interested in what kinds of people live where?

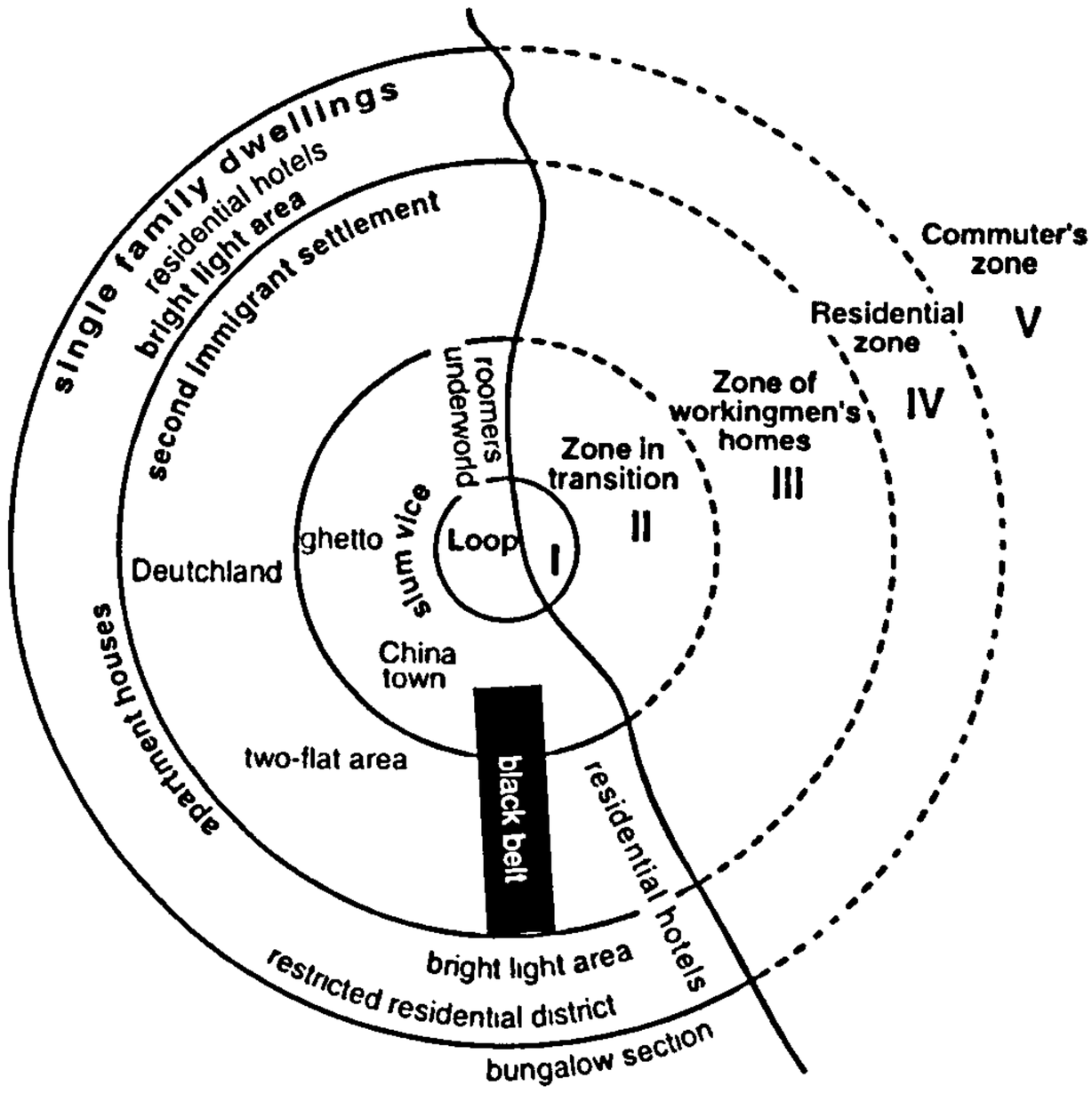
Interest in information about the nature and location of residential populations crosses the academic, commercial and public sector boundaries. The interest is sustained by belief in a robust connection between the location and the social character of residents. Interested parties note that people with similar aspirations, behaviours and achievements may often be found in close proximity to each other - forming a 'neighbourhood' (Robbins, quoted in Burnham 1983). The belief that if information is held on one or two of those

commonly associated elements, the others may be predicted or better understood (CDMS 1993) sustains the attention of interested parties. The fact that urban socio-spatial patterns appear to be relatively stable has led to considerable investment in understanding them (and exploiting this understanding). Who is interested in these patterns, and what is the nature of their interest?

## *2.1 - Academic Interest*

Geographers have long been interested in the residential patterns within urban areas. The location and differing character of city residents was first and perhaps most famously tied together by Burgess and the Chicago School in 1925 through the concentric rings model of Chicago (see figure 2.1). Their model located various social and ethnic groups within a broader ring structure. It also noted both a temporal component to the development of the city's socio-spatial structure, and a spatial differentiation of different social groups.

*Figure 2.1 - Burgess' Urban Model*




Whilst the model offered little in the way of explanation for the forms it identified, it served to focus attention on the fact that a pattern existed (Badcock 1984) and prompted a great deal of research into residential differentiation. Throughout subsequent years, the nature of research into residential pattern and its supporting theory has changed considerably. Some,

like Shevky and Bell, concentrated on developing the best techniques to illuminate patterns of residential differentiation. Although they contributed little to the theoretical understanding of how or why differentiation occurs, they clearly understood the implications of their findings; "We hypothesise that persons living in a particular type of social area would systematically differ with respect to characteristic attitudes and behaviours from persons living in another type of social area." (Shevky and Bell 1955, cited in Knox 1982). These ideas are at the heart of most academic interest in urban socio-spatial structure.

Gradually, geographers have turned away from concern with identifying pattern to concern with its origin and have become concerned with the role of space as a constituent part of those social processes (Willis 1977, Harvey 1989). Contemporary class theory postulates that neighbourhood character plays a considerable role in the socialisation of its residents. Residents are thought to adopt the aspirations and attitudes of those around them and these characteristics may be reflected in the physical appearance of their immediate neighbourhood (Harvey 1989). As people gravitate socially towards those similar to themselves, so they gravitate in space and it is thought that this process helps to perpetuate both social and spatial identities through time; "you are where you live". If this is the case then clearly "where you live" and "who you are" are key concerns for those interested in a vast range of topics from urban form to class relations. Such widespread interest demands comprehensive information. Although this research has not focused on a specific urban research question, it is important to recognise that the work does rest within an urban geography context. Situating the work within geography in this way has allowed the flow of conceptual and empirical information both into the research and out from it.

Although academics may note with interest the fact that residential segregation and spatial pattern exists in urban areas, other bodies must deal with it and its implications on a daily basis.

### *2.2 - Government and Service Interest*

Governments have been interested in 'who lives where' since Old Testament times (see for example, Exodus 30: 11-16). The processes of collecting revenue and allocating resources are inherently population oriented and if, according to urban theory, physically distinct areas contain socially distinct groups, the needs and resources of these groups can be expected to differ too. A government's demand for accurate information about its residential

population is therefore correspondingly large. Countless systems and indices have been utilised which seek to solve the equation between resource and need at a variety of levels. In more recent times in the UK the de-centralised, autonomous service providers have become responsible for proving their case to receive their sustaining resources. This is especially the case in health care provision; a general practice in Waterlooville, Hampshire for example is currently investigating the use of social indicators on its catchment area to assess the possibility of claiming increased resources to match the demands of its deprived patients (Saville, 1996). Whilst some standard indices try to equate population character with resource need or workload required (e.g. The Jarman Underprivileged Area Index (Jarman 1983)) there is considerable interest in what form the best index should take (see Payne et al. 1993, Ward et al. 1994). As the divide between all public and private provisions of service becomes blurred, the impetus for efficient provision of service increases. Where market forces are unleashed, efficient provision can be viewed either as a question of allocating plentiful resources where need is greatest or perhaps more realistically as the spreading of an inadequate resource in the least damaging way. Whatever the concept behind the process is, it requires information about the location and character of the client population.

### *2.3 - Commerce*

The commercial world has been dealing with the same set of priorities as the new service providers for much longer and it has learnt to be more aggressive and ruthless in its dealings with supporting information. Understanding the location and character of markets is an absolutely essential foundation for competitive performance (Wrigley 1988, CDMS 1993). In fact, like academic interest in 'who lives where', there is a considerable history behind contemporary uses of locational information. Today's information industry, serving those who need more detail on their client base (present and future), has its origins in the nineteen fifties with a W. Applebaum (Davies 1984). Applebaum pioneered his own techniques after growing frustrated with the central place theories of his time and their lack of utility for the retailer trying to locate new stores. His techniques included profiling a successful store catchment in terms of its social groups and then searching for similar populations not yet served by a store (Davies 1984). As a cartographer, Applebaum naturally chose the map as his medium of analysis and communication. Today's market information industry has maintained this belief in clarity of presentation together with a disregard for the politics behind exploitation of the purchasing population. Contemporary marketing services sell an understanding of 'who lives where' and 'how much they have to spend'

- together with 'what they like to spend it on'. Their role is often to suggest a more precise marketing strategy, both in terms of location and in terms of customer type. Hence the heightened value of the kinds of information focused upon here (Beaumont 1991). Although academics often criticise commercial practitioners for their poor understanding of society (Goss 1995, Hacker 1984), closer inspection often finds those selling understanding of the social and spatial make up of a market have a very good grasp of the geography of their task. Here, geography includes an understanding of the culture and behaviour of various strata of society, the constraints under which they live and the aspirations they have. The key difference between the academic gaze and the commercial or service provider gaze is that ultimately the academic is concerned with the origins of what they see. The practitioner applies their knowledge for financial gain.

Whether it is ultimately a financially driven desire or an academically driven desire, the thirst for information about 'who lives where' is clearly considerable. Most importantly the variety of activities which are fuelled by information about where people live and who they are *affects everyone*. It is therefore not unreasonable to ask questions about the nature and the quality of the information that is being used. Contemporary geography is just beginning to grapple with some of these questions (see Pickles 1995) although the origins of reviews found in literature like "Ground Truth" (i.e. stemming from those who are non-quantitative geographers) mean that the questions being asked are often not of a technical nature. The issues of primary concern in this research are the mechanisms by which people are represented and their accuracy.

## *2.4 - Sources of Data*

In the light of these concerns the first question to consider is this; if there is such widespread use of socio-spatial data, from where are these users acquiring their data?

### (i) Primary Data Generation

Users with specific needs from a data set most often generate their own. The data are correspondingly task specific and might, for example, be generated through repeated contact with a client base. The advent of store loyalty cards is an excellent example of a user generating their own data on a large scale. When a customer uses their Tesco ClubCard at the end of their shopping, Tesco acquires a record of what they have bought, which store the goods were purchased at, how much was spent and importantly where the goods are

going, (i.e. the home address of the customer). That data can then be incorporated into models of store catchment areas or site searching systems as well as other operational systems. There is also an inter-business market for restricted versions of these kinds of data. Companies can purchase lists of names and addresses, not necessarily with data attached to them, but often with a common context. For example, "all these households have purchased a car from us within the last year," might be all the information an insurance company needs from a retailer to help target its marketing campaign. So whilst the actual numbers of variables within the data are very small, placed in a context which ties location and some kind of behaviour, the data assume a considerable value (CDMS, 1993). This kind of data set is by nature disaggregate, or individual based. Its spatial resolution will be at least at unit postcode level and more likely the data will offer the user a specific postal address for each person of interest. For the academic user, fieldwork is the obvious source of primary data.

**(ii) Secondary Data**

Whilst task specific data can be highly valuable, the user often requires a more general understanding of the wider population. This data might help to construct a background population against which to compare their client base, or perhaps be used as an aid to expansion into unknown markets (both geographical and social). Since the cost of collecting this kind of data is prohibitively high, there is a general reliance on state generated data sets. For the academic user, generating large field samples is enormously expensive in terms of time and resources. Often they will resort to utilising state generated data, perhaps in combination with field gathered data to enhance the information that they have to work with. For both groups of users, the UK decennial census is the most common source of state generated data. It is unmatchable in terms of spatial coverage, population coverage and accuracy. The Census Of Great Britain and Northern Ireland is a survey of the entire population, held every ten years. The last census was held on April 21st 1991 and despite controversy over who was and who was not enumerated, the census data sets are regarded as being as near to a complete record of the population at that time as is humanly possible[1]. The census questionnaire is completed on a household basis and it includes questions about the number of residents, family structure, employment of residents, age and sex of

---

[1]The number of residents 'missed' by the 1991 Census is estimated to have been 1.2 million in Great Britain (CMU 1993). For further details on nature of the missing population see CMU 1993 and OPCS 1991.

residents, recent migration, education level, house size and type, ethnic identity, health status and travel practices. The census is organised within a nested hierarchy of spatial units covering the entire land area. Generally, the data are made available to the public in such a way that as their spatial origins are made more precise, the level of detail they convey is reduced. The majority of the data sets publicly available from the census are *aggregate*.

## *2.5 - Aggregate Data*

In many societies, information about individuals or their families is rightly regarded as being sensitive (Hoffman 1980, Savage & Edwards 1984). In order to protect the individuals whose details are captured by the census, many of the data sets released from it are aggregate in nature. In simple terms, aggregation is akin to summarising data for a group of subjects, most often through presenting a description of the group in terms of its collective features. Consider for example a class of children; an aggregate data set about them might describe the total number of boys and girls, the number of A, B and C grade students respectively, the number who take the bus to school and the number who travel by car. When interest is focused on 'who lives where' as it is in the census, the unit of information is not a group, but a spatial zone. The residential population of a designated *area* are the subjects of the aggregate data set.

It takes little insight into this kind of data to comprehend the fundamental flaw with working in groups or spatial units. The process of aggregation masks subject to subject variation. From the school class data set there can be no understanding of whether the children who travel to school by bus are all A grade students or whether more boys get Cs than girls. There can be no understanding of how individuals have contributed to the collective appearance of a group provided by an aggregate data set, nor can there be any understanding of which categories an individual actually falls into.

## *2.6 - The Ecological Fallacy*

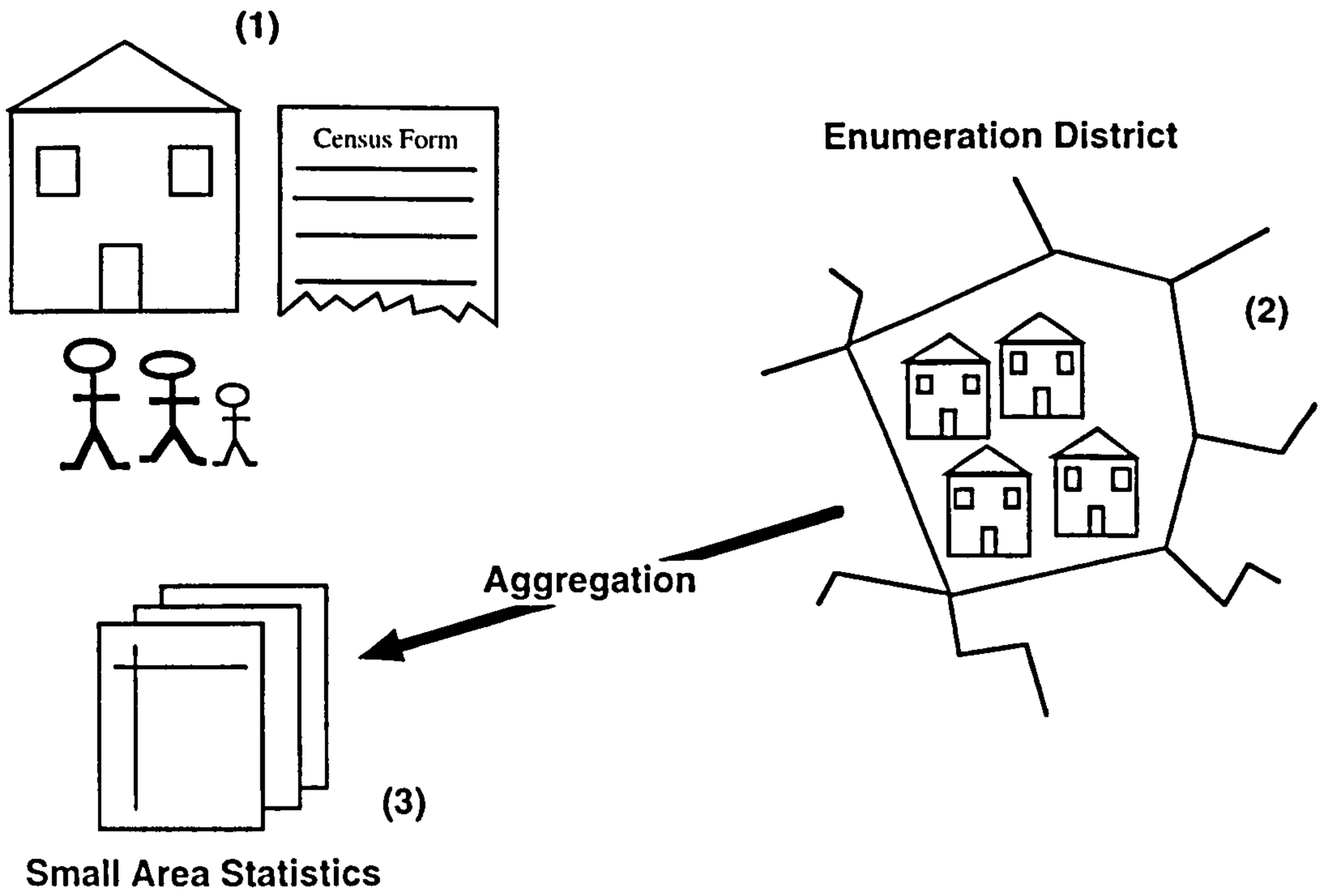
In using aggregate data a researcher is forced into either ignoring the fact that variation in a population of interest may be occurring beneath the level at which the data are held, or (despite recognising the problem) they must make the assumption that the underlying population is of uniform character. Such a difficulty is part of a wider set of problems stemming from aggregation known as the ecological fallacy (Blalock 1964). The problems arise from the fact that there are numerous possible ways in which data about a set of individuals can be grouped together and that the way in which grouping takes

place will affect how observed relationships between variables behave (Martin, 1991). This means that in addition to the fact that relationships between variables may not appear to be the same at the aggregate level as at the individual level, variation is also possible at *different levels of aggregation* (e.g. for spatial units of differing size or shape). Therefore, when inferences about individuals are drawn from aggregate data, their validity must be in doubt. All of the data users discussed above are confronted by this difficulty and its implications. Some will be more aware of it than others. However, the main implication of the ecological fallacy is that users may be placing faith in relationships observed in a data set which are to some extent a feature of its construction.

## *2.7 - Small Area Statistics*

With the ecological fallacy in mind, it is important to consider an aggregate data set in which the majority of people appear and which is utilised by a very wide spectrum of those interested in knowing 'who lives where'. This will draw together some of the concepts presented thus far and introduce a data set which is at the heart of this research. Figure 2.2 shows the origins of the Small Area Statistics (SAS). The SAS data set is the output from the UK decennial census with the finest spatial scale.

*Figure 2.2 - Origins of Small Area Statistics*




On census night, every household in the country (in theory) fills in a census form, giving details about their house and those who live in it (1, figure 2.2).

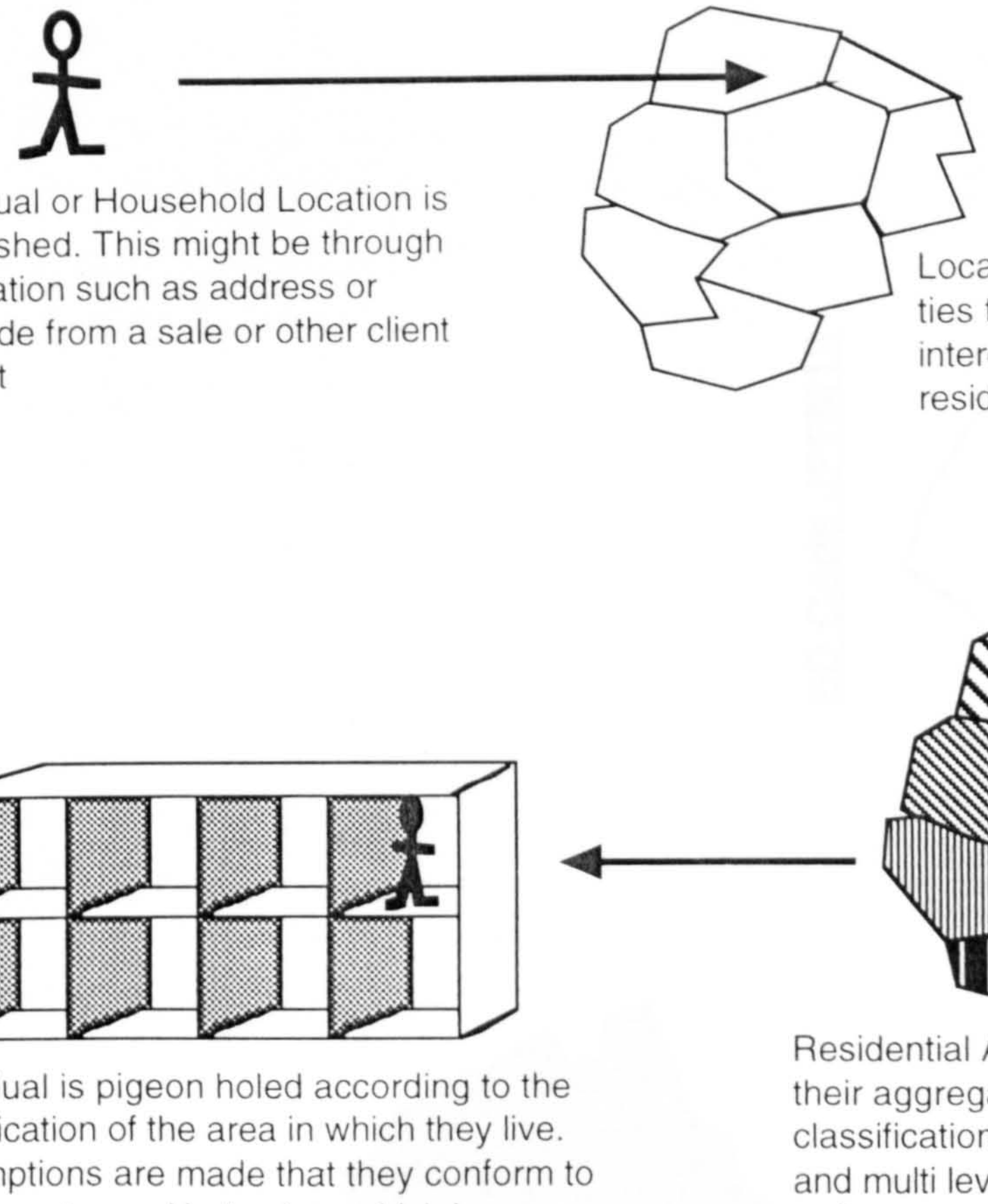
Each household is situated inside an Enumeration District (ED) which is the smallest bounded area within a nest of census areas covering the UK (2, figure 2.2). A typical ED contains about 100-150 households (Denham 1993). All the households in the ED contribute to its resultant census data set. The data are available as a series of tabulations and cross tabulations after the census returns have been aggregated (3, figure 2.2). This set of tabulations is called the Small Area Statistics (SAS) and represents the largest and most wide ranging socio-demographic data set that is publicly available. It also offers a complete spatial coverage which makes it the first choice of data for those interested in 'who lives where'. Up to four items may be cross tabulated within the SAS (Birkin & Clarke 1995). This means that the maximum amount of information available for individual households from the SAS is four items or less because there is no way of knowing how a specific household fits into each tabulation. In this way and through further data modifications, household identity is protected. (For more detail on the SAS see appendix I). SAS form the base for a very wide range of 'who lives where' applications, but also exhibit all the features of aggregate data and thus their problems. Since the SAS tabulations are so widely used, any successful attempt to raise the level of detail that can be obtained from them could prove to be highly significant.

The following section ties together all of the issues discussed so far. It will also provide a useful contrast for an illustration of the techniques developed in the research.

## *2.8 - Geodemographics*

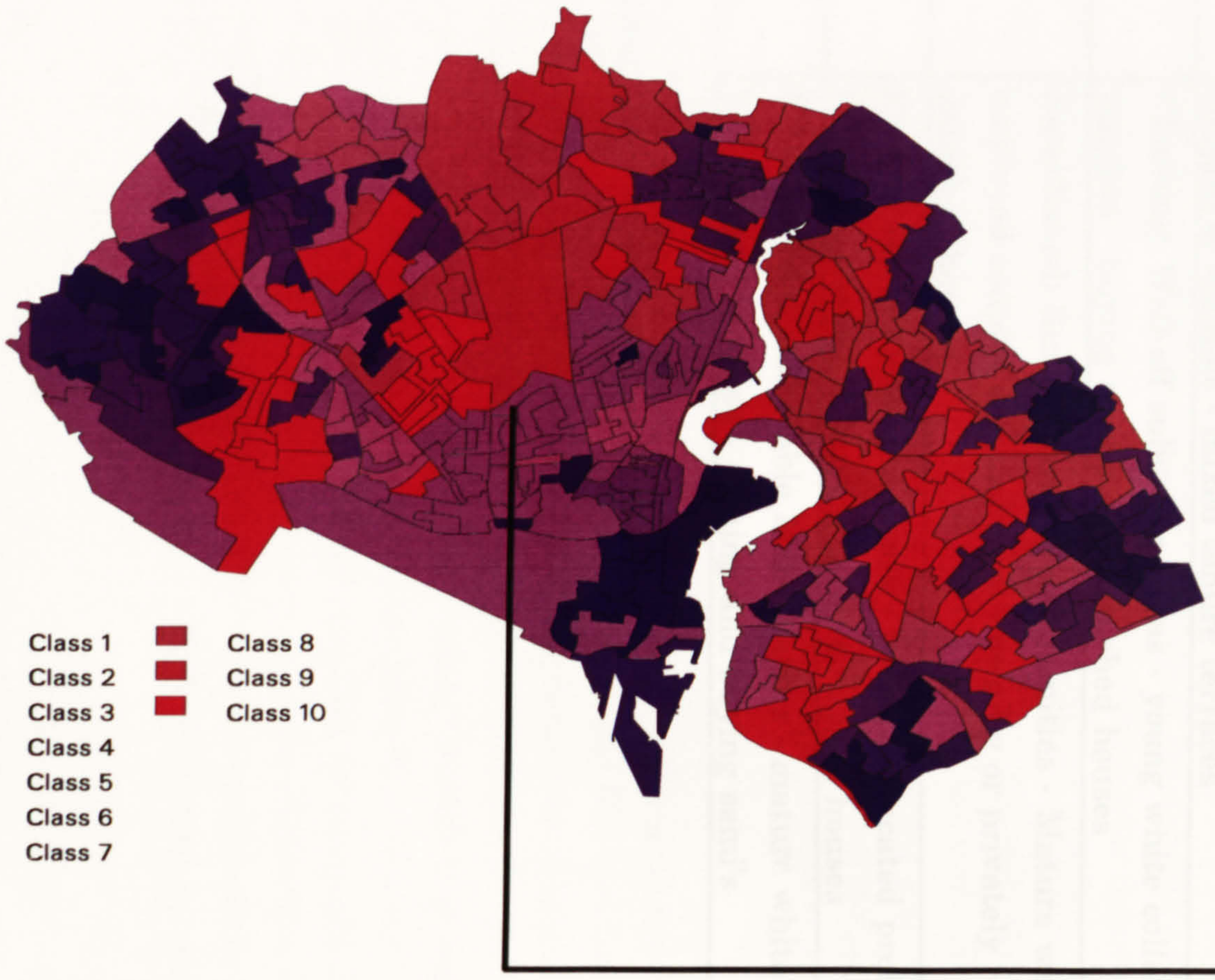
Geodemography is a process used in many socio-spatial analyses. Small residential areas are assigned a 'character' as a result of the analysis of appropriate descriptive data sets. SAS are usually at the heart of the process, though they may be supplemented by other sources such as consumer survey or credit rating data. The aim of the analysis is to place the area into one of a number of classes based on its socio-economic character (Beaumont 1991), and to complete a chain linking an individual's residential location to a personal profile dependent on the type of area in which they reside. The profile aims to describe their wealth, aspirations and often their likely receptiveness to various marketing campaigns or need for services (Goss 1995). Figure 2.3 illustrates the elements which contribute to this linkage between residential location and identity.

*Figure 2.3 - The Geodemographic Link Between Place and Identity*

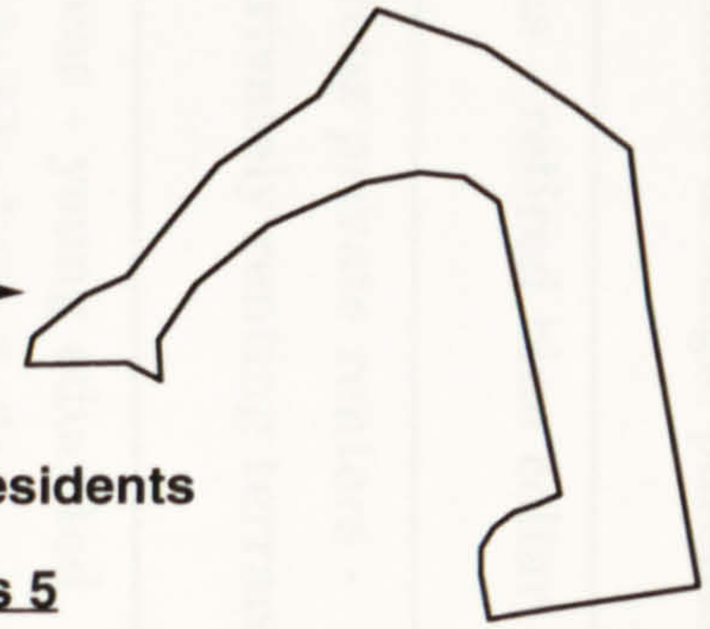



As noted at the outset of this chapter, there is a belief that individuals live in areas which contain populations of broadly uniform aspirations, achievements and affluence. Geodemographic analysts often promote this concept to their clients through the adage that "birds of a feather flock together" (CDMS 1993). Whilst this is clearly a saying of some validity, it does not offer a robust guarantee that residents who live close together always possess a similar set of characteristics.

Although geodemography was initially developed from academic work, its employment in academic research has been reasonably limited. The literature on geodemographic developments has tended to focus on reviewing the actions of the geodemography industry, rather than exploring its possible uses. These reviews tend to be rather polarised; either positive (see for example Flowerdew & Goldstein 1989, Beaumont & Inglis 1989) or more recently, negative (see Goss 1995). As noted earlier, considerable investigation into the possible benefits of geodemographics has been carried out in the field of health care and this is increasingly joined by research into




**ED Code JPFM12**



**ED Contains:**
**187 Households, 412 Residents**

**Area Classified as Class 5**

***"Aspiring: Academic Centre and Student Area***
***Young educated white collar singles & couples***
***Privately rented bedsits and Flats"***

*Table 2.1 - A Description of The Geodemographic Classes In Figure 2.4*

| Group | Description |
|---|---|
| 1 | Struggling; multi-ethnic areas - pensioners & single parents - high unemployment - LA rented flats |
| 2 | Struggling; Council tenants - blue collar families & single parents - LA rented terraces |
| 3 | Struggling; Less Prosperous pensioner areas - retired blue collar residents - LA rented semi's |
| 4 | Struggling; Multi ethnic areas: less prosperous private renters - young blue collar families with children - privately renting terraces and bedsits |
| 5 | Aspiring: Academic centres and student areas - young educated white collar singles & couples - privately rented bedsits & flats |
| 6 | Aspiring; Young married suburbia - young well-off blue collar couples & families - mixed tenure terraces |
| 7 | Climbing: Well-off suburban areas - young white collar couples and families - buying semi's and detached houses |
| 8 | Established; Rural Farming communities - Mature well-off self employed couples & pensioners - owning or privately renting large detached houses |
| 9 | Prospering; Affluent achievers - mature educated professional families - owning & buying large detached houses |
| 10 | Established; Comfortable Middle agers - mature white collar couples and families - owning and buying semi's |

Source: Blake and Openshaw 1994

urban crime (P. Brown, 1993) as the most significant of the academic uses. Some of the reasons behind this apparent lack of academic interest in geodemography will be explored in Chapter 3.

Figure 2.4 is an example of a geodemographic analysis of Southampton. It was constructed using a ten class system (rather fewer than a commercial consultancy might employ), and the classes were defined entirely from SAS using a neural classification technique[2] (see section 8.1.1). The Southampton area seen in the image will form the study area for illustrations throughout the research presented here. Southampton is city on the southern coast of England, with a population of approximately 200,000. It is a light industrial city, with one of England's busiest industrial ports. There is also a very large student population, concentrated at the city's physical centre.

The EDs in the image have been shaded according to the geodemographic class to which they have been assigned , based upon their SAS characteristics (see table 2.1 for a description). One ED has been expanded at the side of the map. It has a classification and the number of households and residents inside it has been established. If there was a particular interest in finding other residents with the same profile (perhaps, for example, research had shown a high propensity to purchase particular products amongst this group), it would be possible to determine which of the other EDs were assigned the same class and hence determine a potential market size. However, the . actual information about the residential population provided by the classification, beyond conjecture about their behaviour, is of limited depth.

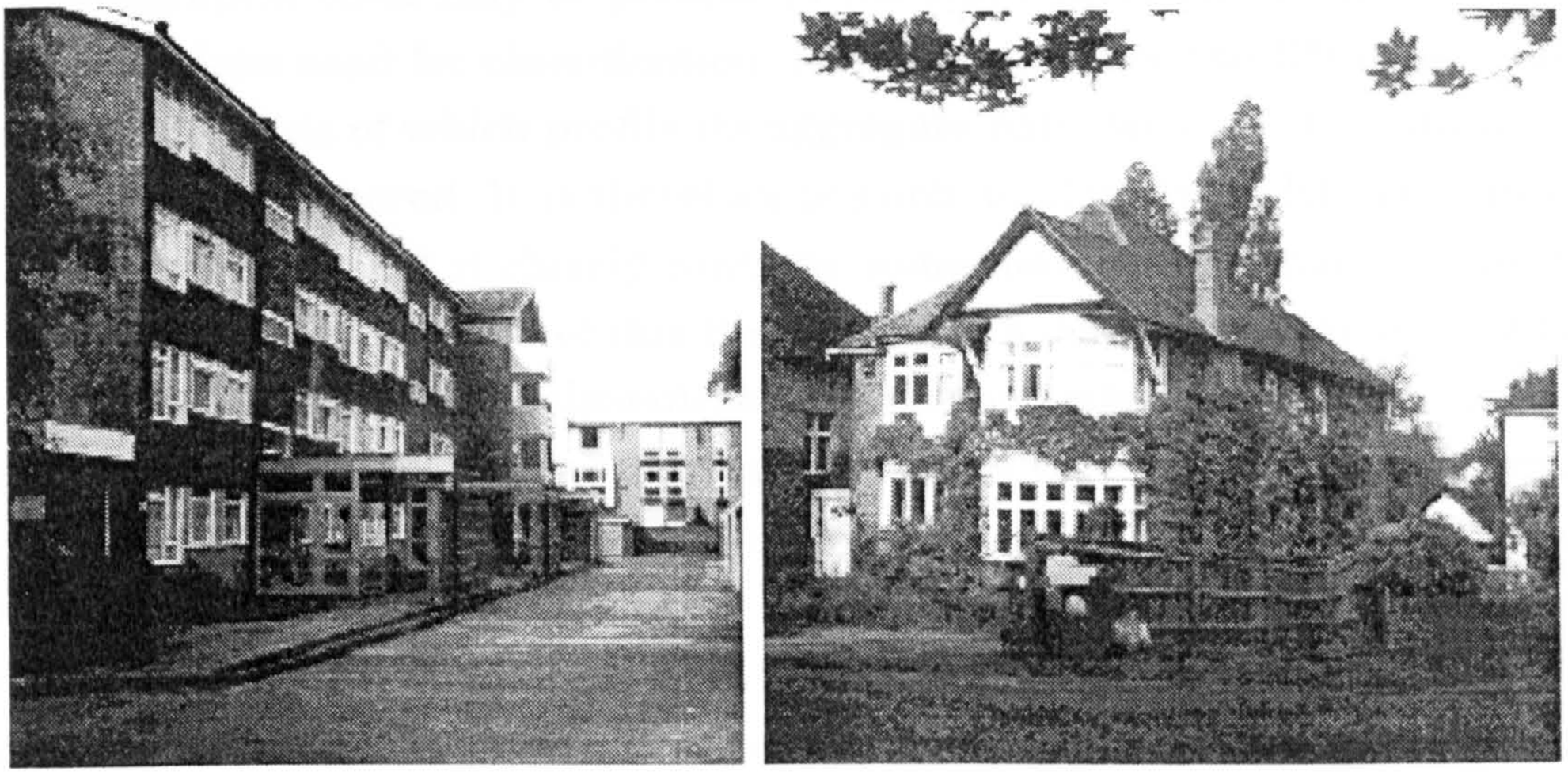
## *2.9 - Variation at a finer scale*

Figure 2.5 is a photograph taken from inside the ED in question. It illustrates clearly visible differences in house type within the ED, and from this a difference in some aspects of household character can reasonably be presumed. This image is typical of the kind of evidence for the fact that variation is occurring on the ground at a scale beneath that measured by the SAS, or at least beneath that to which a blanket classification has been assigned. This then, is evidence for the fact that an assumption of ED homogeneity is incorrect and an illustration of the difficulties of dealing with aggregate data. Some of the households in the image would be incorrectly classified and any subsequent analysis would be limited as result.

---

[2] Neural Classification is discussed at length in Chapter 4

*Figure 2.5 - Inside ED JPFM12*



Part of the reason why a geodemographic analysis is particularly vulnerable to this kind of error is illuminated by the graph in figure 2.6;

*Figure 2.6 - Graph Representing a Geodemographic Classification*

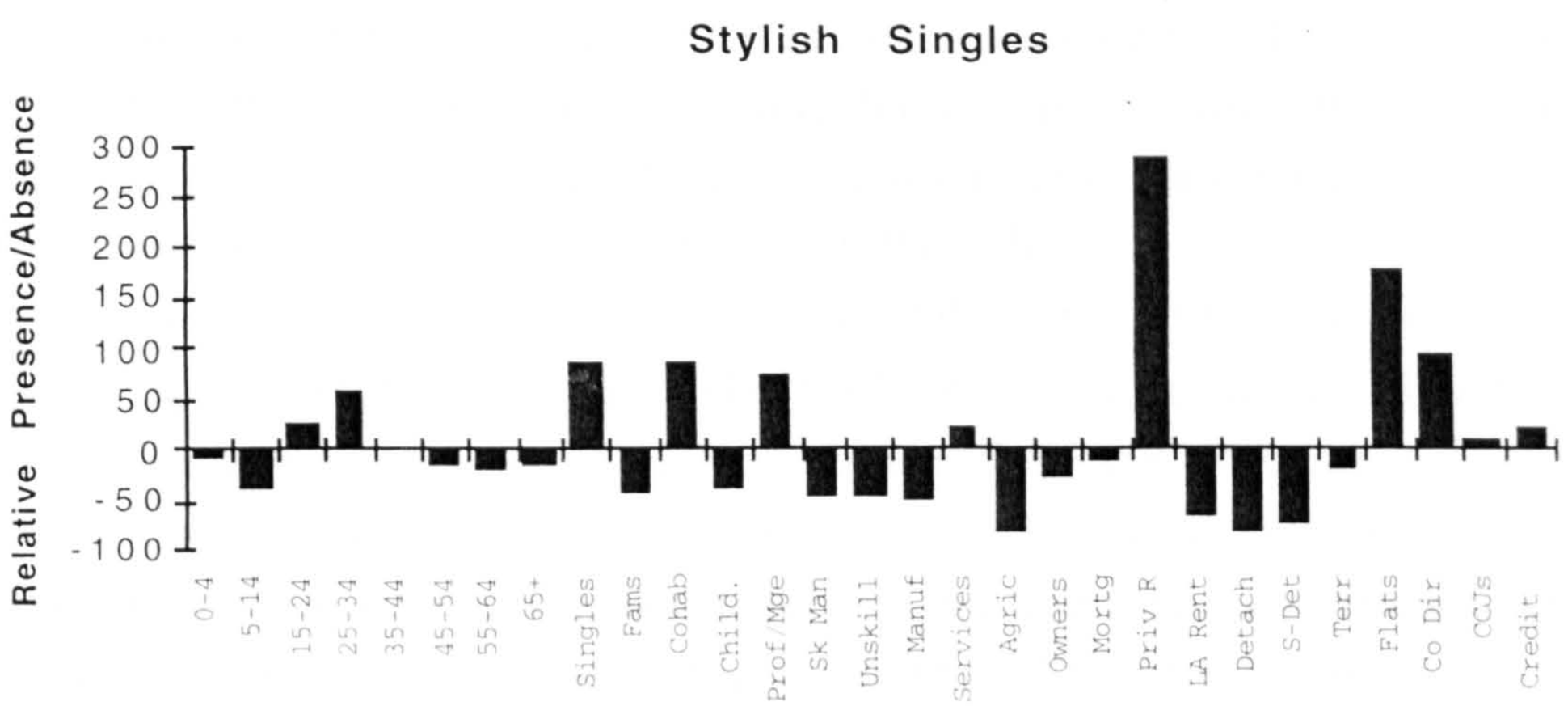



Figure 2.6 shows a graphical representation the geodemographic class described as 'stylish singles' (CCN 1993). The information is expressed in a form which shows whether a certain feature is present in the class at a level above or below the national average (shown as 0 on the Y axis). The graph

shows that whilst the class is clearly dominated by singles in rented flats other household types might still be represented, albeit at a low level. The implication of this is that residents who do not match the profile of an area's geodemographic class may be present in ED and clearly visible in the aggregate data used for classification. However, because the ED is assigned its class on the basis of which profile its aggregate data are *closest* to, these households are ignored. It is therefore possible to classify an ED as containing 'singles' where in fact it clearly contains some married or cohabiting couples. Whilst it is not the purpose of this thesis to offer a detailed explanation of the reasons for maintaining a classification technique which allows this to happen, it is more likely a result of attempting to keep the ED classifications straightforward enough for boardroom minds to comprehend than a lack of technical expertise on the part of the geodemographics industry (CDMS 1996). Nevertheless, the problem of hidden variation amongst residential populations when employing classification based on aggregate data is clear.

Thus far, the focus has been on aggregate data - specifically the SAS drawn from the 1991 census. Given the problems associated with using these data, is there any alternative source which might be suitable for socio-spatial analysis?

There is now a new data set in the public domain which is also drawn from the 1991 census, but is not aggregate:

## *2.10 - The Samples of Anonymised Records*

The samples of anonymised records (SAR) were released by OPCS in October 1993 (CMU 1993). This date marked the culmination of a spirited campaign by a cross disciplinary group of academics to gain access to a disaggregate data source from the 1991 census (Marsh et al 1991). They are a microdata source (i.e. based at the individual and household level) and offer the user an almost complete copy of responses made to the 1991 census questionnaire. The SAR are available as two, non-overlapping, samples of the UK population. One sample details households in their entirety and the other details individuals, together with some contextual household information (Dale and Marsh 1993). For a comprehensive technical description of the SAR data sets see appendix II.

The household level data are a 1% sample of census returns and details 215,789 households together with the 541,894 persons resident in them. The individual level data are a 2% sample containing 1,116,181 people (CMU 1993).

Whilst the concept behind the SAR is to offer the user details of how individual characteristics covary, ONS can clearly not make such sensitive information available to the public without safeguarding the anonymity of those who appear in the sample. The body charged with releasing and maintaining the data set has insured against disclosure of an individual's identity through a number of protective measures summarised in table 2.2.

The information which the SAR provide is unrivalled for answering questions on the characteristics of households and individuals. However, the spatial resolution restrictions to which it is subject render the source virtually useless for investigating the 'who lives where' question (Marsh 1991). Nonetheless, it is certainly worth exploring how this new data set relates to the SAS data, if only to understand the differences between the two more clearly. Since the two data sets stem from the same census questionnaire, the interface between them is potentially very large.

*Figure 2.7 - How SAR and SAS Data Are Related*

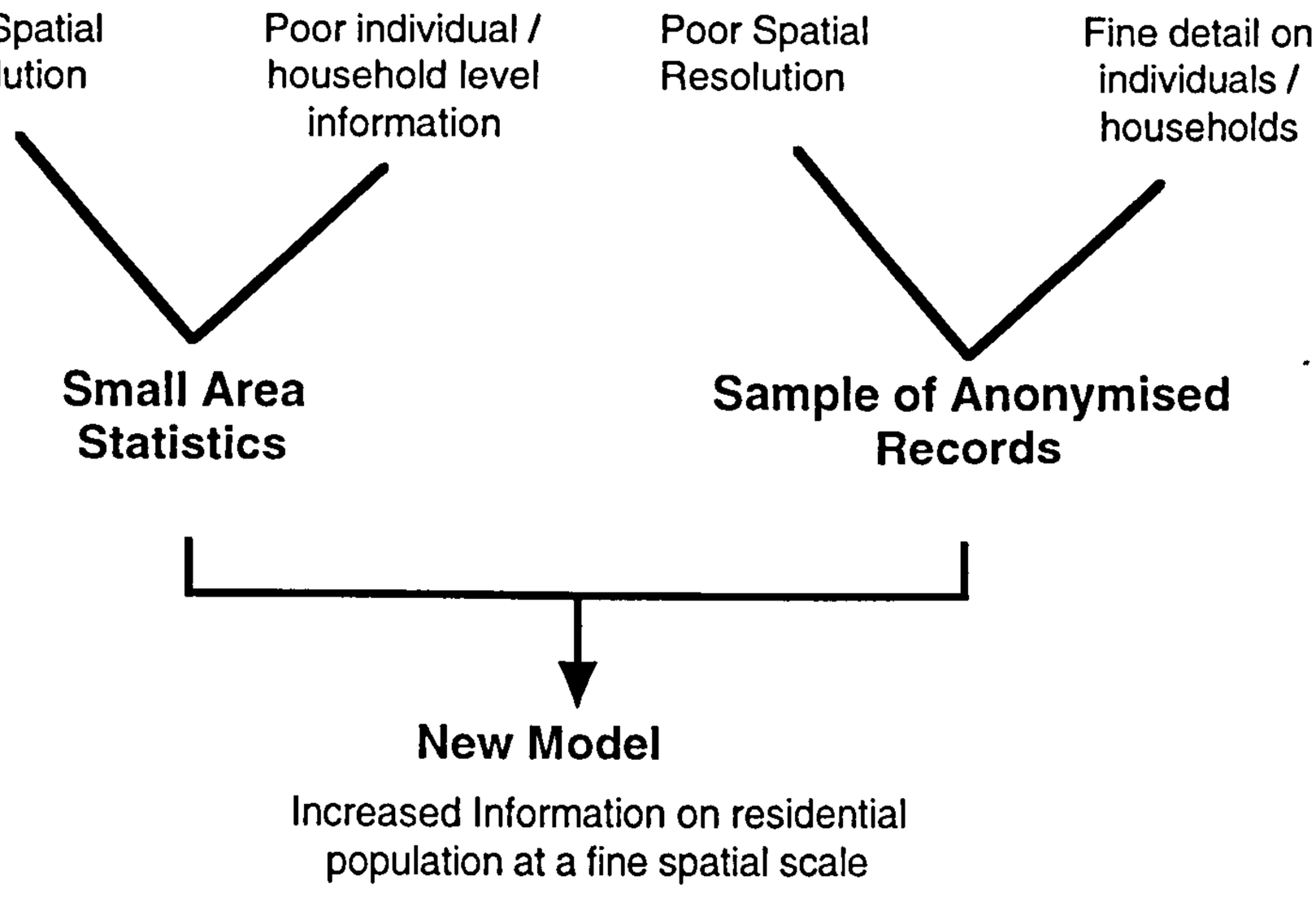


*Table 2.2 - Measures to Preserve Confidentiality in the Samples of Anonymised Records*

| Measure | Effect |
|---|---|
| Sampling | Low sample fraction reduces the chance of an individual of interest being included in the SAR data set |
| Suppression of Data | The data are altered to reduce the number of rare or unique cases. Addresses and dates of birth are suppressed |
| Grouping of Categories | Categories with fewer than 25,000 expected cases in the population are collapsed together or into other, larger categories. Some variables with few numbers at extreme ends of a variable scale have been top-coded e.g. age has a blanket 95+ top code |
| Re-ordering of Records | The record order in the SAR has been scrambled within the region of origin to remove the possibility of tracing an individual by district or area |
| Restriction of Geographical Detail | The geographical origin of 2% individual SAR records can be determined to within areas of approx. 120,000 people. 1% SAR records can be located within one of the Registrar General's Standard Regions. The South East is further split into inner and outer London and the rest of the South East. |
| Legal Controls over Use | License to access the SAR makes it clearly illegal to attempt disclosure of a sample member's identity |

Source: CMU 1993

When viewed in this manner it becomes clear how these two data sets complement each other, the SAS providing considerable detail about the location of households, the SAR providing detail on household characteristics. Agonisingly, there is no direct relationship between the two sets. The SAR are a small sample of census respondents and cannot be directly reconciled with the SAS, since the anonymity of the sample members must be maintained. Thus, for the user interested in 'who lives' where, *neither* data set is particularly suited to their analysis. With the inadequacies of both the SAS and SAR data sets for socio-spatial analysis now clear, the task for this research can be more precisely described:

## *2.11 - A List of Desires*

Since the available data are not well suited to socio-spatial analysis, a list of desirable enhancements is proposed. These enhancements would seek to overcome some of the difficulties outlined above.

*Table 2.3 - Desirable Enhancements to Socio-Spatial Information*

| Number (Increasing Detail) | Enhancement |
|---|---|
| 1 | A measure of social mix within an area for which aggregate data are being employed |
| 2 | Enhancement (1) together with details on the nature of the social mix |
| 3 | Detailed information about the individuals and households within the area of interest |

At the top of the list would be new information about how uniform the population in an area of interest actually is. If it was possible to derive a measure of social mix (defined as the relative homogeneity of individual and household types), it would be possible to assess the validity of assumptions about underlying homogeneity which single area classification makes. Such an indicator might allow a measure of confidence to be attached to a geodemographic analysis and illuminate any geography of social mix with implications for the research task. The next item (2), would be an extension of the first to include an understanding of the nature of the variation. If an area is flagged as being socially mixed, of what does that mixture consist? Are the households and individuals broadly similar or radically different in their character? Is an ED, for example, predominantly one type of household with a

small pocket of others? At this level precise knowledge of *who* lives in the ED would not be required. Rather the question could be conceptualised in terms of what *kinds* of people live in the ED.

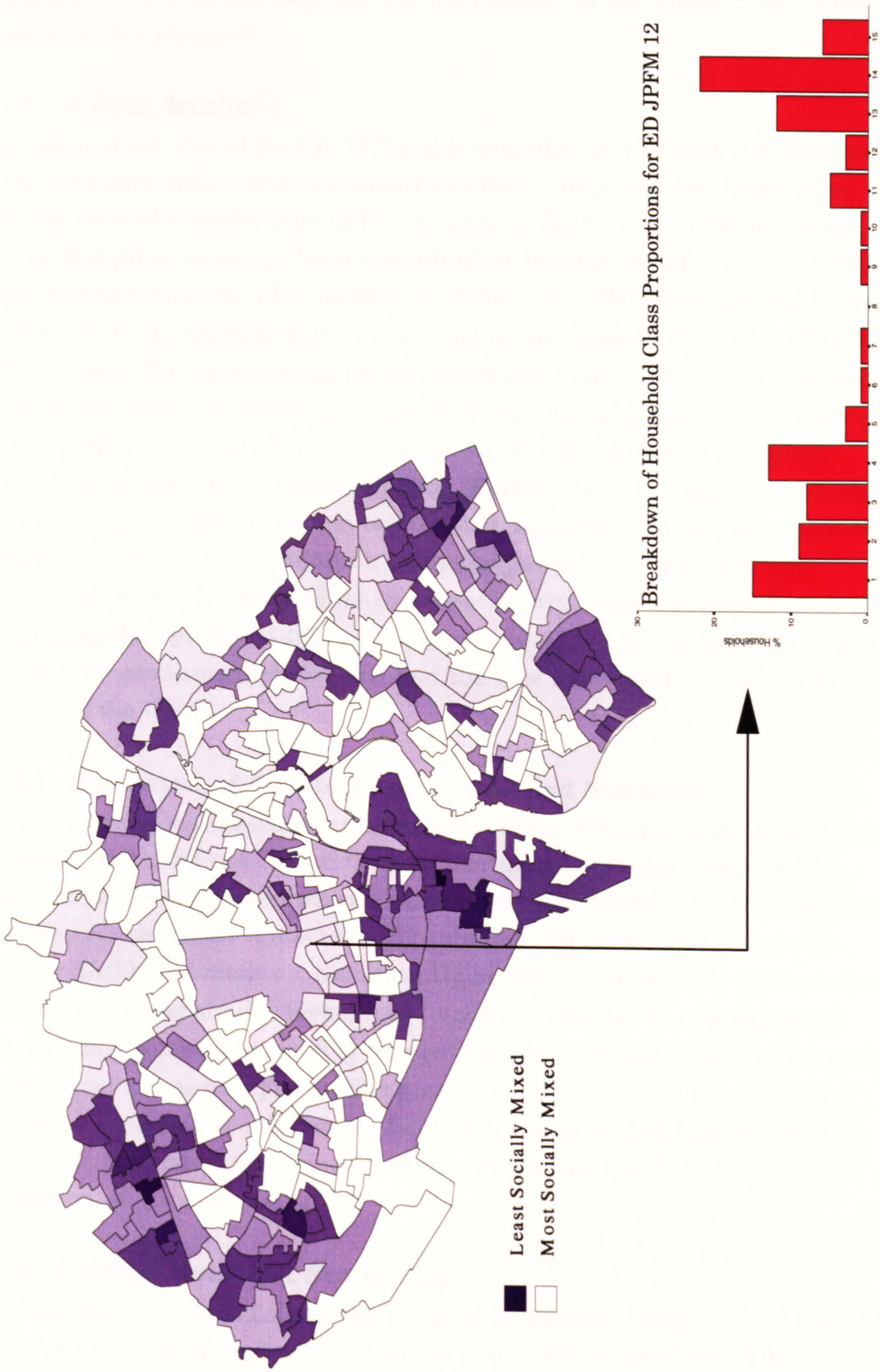
One step beyond this level of information brings a detailed knowledge of *precisely* who lives in the area of interest (3). This information might include names and addresses, but would consist primarily of socio-demographic and economic details at the household and individual level. As has been discussed, elements of this kind of information already exist in a fragmented form, both as primary data and secondary data. The electoral register for example provides the user with personal information on name and address and other sources such as common housing registers, provide information on property and tenancy. As has also been discussed however, detailed personal information is regarded as sensitive. The most detailed socio-economic data sets which couple locational precision with detailed data are not publicly available. Attempts to synthesise that kind of data from readily available sources are either illegal (Barr 1993, Savage & Edwards 1984) or as with the SAR, are rendered impossible through the employment of anonymity protection measures (Dale and Marsh 1993). Although the aim of this research is to generate more information about a residential population than can at present be obtained from publicly available data sets it must remain within the bounds of the law. The law specifically prohibits attempts to synthesise data which could lead to the identification of an enumerated household or individual.

Thus far, the shortcomings of aggregate data have been explored, the unsuitable nature of the SAR for fine scale socio-spatial analysis has been made clear and a number of desirable developments have been outlined. Given that the third item on the list is intensely problematic, the task of this research was to satisfy the first two items. Figure 2.8 illustrates the solution to these difficulties and desires which the research described in this thesis has produced.

## *2.12 - Two Desires Satisfied*

Figure 2.8 illustrates an analysis of Southampton which satisfies the first two desires listed in table 2.3. The map illustrates a measure of social mix for every ED in the study area. This index has been derived through a process of modelling the proportions in which fifteen pre-defined household classes are present within the ED. Where the ED is relatively mixed (i.e. a number of different classes are present in reasonable numbers) the shade is white and

*Figure 2.8: A New Model Social Mix In Southampton*



where the ED is relatively pure (i.e. one or two classes dominate it) the shade is blue. The graph to the right hand side of the map illustrates the proportional presence of each of the fifteen classes for ED JPFM12. The model which produced both the map and the information in the graph is the main product of this research.

## *2.13 - A New Analysis*

The information offered for ED JPFM12 is available in a similar way for every ED in the study area. Figure 2.8 illustrates that a step has been taken away from the kind of classification of EDs as seen in figure 2.4. Geodemography can be thought of as using 'hard classification' because it offers no flexibility in its dealings with the ED's residential population. The model behind figure 2.8 enables 'fuzzy classification', in contrast to that seen in figure 2.4 (Klir & Folger, 1988). The idea of fuzzy classification is to remove the sharp boundary between the classes to which an object or group might be assigned. It allows an ED to simultaneously belong to a number of different social groupings rather than simply being assigned to one. Furthermore, the extent to which the ED belongs in each class can be specified, reflecting the mix of household types within it. Fuzzy classification offers the possibility of a host of new analyses of the study area and its socio-spatial structure. The aim of introducing the 'results' at this early stage is to aid the illustration of the path between the geodemographic model and the new fuzzy model which will form the bulk of the thesis.

### 2.13.1 - The Result of An Integration of SAR and SAS Data

The fuzzy model was based on an integration of the SAS and SAR data sets, in a manner which did not breach the confidentiality constraints imposed by ONS. The model can be reproduced for any area covered by the UK census that is desired which makes it a potentially valuable tool for all those currently utilising SAS based analysis. Significantly however, it also sheds light on the possibility of using similar techniques to add information to *any aggregate data set*, thus potentially tackling elements of the ecological fallacy at a variety of scales and across disciplines. The process which led to the fuzzy model involved a synthesis of information held at differing scales. Understanding that this is the underlying principle of the integration will help in the exploration of its complexities.

## *2.14 - Conclusions and Summary*

The following chapters describe the concepts behind the integration of the SAS and SAR and explain, in technical terms, how it was carried out. They

present an analysis of the quality of the integration and seek to set it in its wider context. Considerably more insight into the significance of the research will be gained by staying conscious of its context, conceptualised here as being composed of a number of elements.

- The integration is a response to dissatisfaction with current techniques and data utilised in research into 'who lives where'.

- The integration has used SAS and SAR data from the 1991 census. Used independently, neither set provides a satisfactory medium through which to explore socio-spatial structures.

- Whilst this research has focused on census data sets, the principles that are explored and possibilities that are put forward need not necessarily be tied to that data source. Where appropriate, consideration of how the concepts considered and techniques utilised might work in other circumstances will be offered.

- The integration technique provides a model of the kinds of household resident within an ED and their relative proportional presence. It will not identify specific households, as to do so would be illegal, unethical and probably impossible.

- The real product of the research, towards which the following chapters will guide the reader, is a methodology and suite of techniques for undertaking the integration. It is replicable for any area of the UK.

- This research is *geography*. It therefore needs to be situated within that discipline and note taken of the contemporary debates which impact on this kind of work. The next chapter seeks to carry out that task.

# Chapter 3

## Chapter Overview

This chapter situates the research within geography. The discussion is divided into two themes; the first an exploration of a debate over the place of scientific method and quantitative analysis within the discipline, the second concerned with the nature of social group or class. This research utilises numerical representations, computer manipulations and complex modelling of social phenomena, all of which have been contested in the last thirty years. Through a review of the origins and nature of these discussions, the position of this research within the debate is established. The chapter also reviews theories of social class and socio-spatial structure. Much of the class debate is an extension of issues raised in the quantification debate. For this reason, the two themes are pursued in turn before the broader findings and practical implications of both debates are drawn together. Specific illustrations of how the knowledge gained from these debates has been utilised in this research will follow in later chapters. The primary aim of the chapter is to illustrate the origins of relevant work and to explore the possibilities for enhancing quantitative research practice through considerations of theoretical debates.

## 3.0 - Introduction

Two key themes are explored in this chapter; the place of quantitative method in geography and debates over the nature of social class or group. The place of quantitative research will be explored first, since it raises a number of issues of direct relevance to the social class debate. An exploration of the quantification debate is presented for two reasons. Firstly, at the time of writing, many of its themes are being revisited by geographers at conferences and in publications. The ontology and methodology adopted in this research have been (indirectly) challenged by these events. As this is research undertaken by a geographer, it is essential that the various arguments are addressed and a position within the debate is adopted. Secondly, and of greater significance, it is hoped that a review of the issues within the quantification debate will enhance the quality of the research. By attempting to address the criticisms raised against work of this kind, a robust, coherent and contemporary methodology will result.

## *3.1 - Quantification and Science in Geography*

It is thirty years since geography rose to the challenge of adopting formal scientific method and underwent its 'quantitative revolution' (Johnston 1986). In recent times there has been a resurgence of interest in such a 'meta-debate'. This reflects the fact that since grasping the scientific nettle, geography has never been able to put to rest the issues that it raises. That the nature of geography should be disputed is of little surprise to anyone with a grasp of its breadth. Consider these two abstracts as an illustration of this breadth. Both have been taken from the same issue of a journal whose title proclaims it as 'geography':

*"This essay compares the imaginative geographies of Egypt produced by Florence Nightingale and Gustave Flaubert as they travelled up the Nile Valley and back to Cairo in 1849-50. Their experiences are used to emphasise the physicality (rather than merely the textuality) of travel writing. The differences between their imaginative geographies and in particular between their representations of landscape, space and people, illuminate the complex and fractured formation of Orientalism as a constellation of power, knowledge and spatiality, and its entanglements with patriarchy, sexuality and various colonialisms." Gregory (1995)*

*"Dijkstra's shortest path algorithm is widely used in spread simulation. While Dijkstra's algorithm is robust in solving classical network problems, its implementation in raster-based geographical information systems (GIS) has not been satisfying. A major reason for the inadequate simulation is the restricted number of raster cells (generally the 4 to 8 adjacent cells) included as links in the algorithm. We propose several alternative methods that incorporate non-adjacent cells as additional links, thus overcoming the limitation of a raster data structure. The simulation accuracy is increased from less than 60 per cent to above 95 per cent for spatially homogenous elliptical spread and appears to be increased significantly for spatially heterogeneous conditions." Xu & Lathrop (1995)*

Geographers are people of a variety of ages, nurtured and tutored by those from differing eras. In this way, the variety of schools of geographical thought is maintained through time and tends to be revisited as new researchers delve into their subject's past. This constant attention to history keeps many of the issues discussed by geographers in the past, fresh and relevant today. The growth of computational geography (Openshaw, 1995) is the contemporary development which has helped to rekindle widespread interest in this particular debate. As will be shown, previous encounters with quantitative

techniques and technology seem to have left many geographers with a prejudice against all things numerical. Together with more valuable criticisms, these prejudices are now directed towards contemporary quantitative geography (a category into which this research falls). In order to effectively address these criticisms their origins must be understood. The Johnston (1986) 'excursion through time' will serve as a means to gain this understanding.

## *3.2 - From Regional Geography to Quantification*

Consider academic geography in the sixties. The paradigm known as 'regional geography' held the mainstream position within the discipline. It conceptualised the world as a mosaic of unique areas in which physical and political landscapes were closely linked (Johnston 1984). This required a case study based approach to the organisation and comprehension of the empirical world. Whilst many pursued this descriptive, textual, travel oriented mode of study, others felt that geography was being left behind in the race to establish its identity as a serious and useful academic discipline. Some geographers began to adopt positivistic practices, inferential statistics and a geometrical approach to spatial issues, possibly influenced by the status awarded to other disciplines through their adoption of rigorous, scientific methodology. The advocates of scientific method suggested that they might illuminate human behaviour and spatial phenomena in terms of laws and physical principles through robust experimental and numerical approaches. The development of numerical models, adapted from other numerate disciplines, pushed geography towards a new identity as a 'spatial science'. This development built steadily into a quantitative era lasting twenty years and substantially changing the theory, practice and image of geography. Quantitative techniques were eventually adopted by the mainstream of the discipline.

By the late seventies however, quantitative geography had reached a crisis point and found itself under attack from a number of directions. Primarily, there was increasing attention on the failure of spatial science to adequately predict or account for human behaviour. In many cases the inaccuracy of models was due to poor quality data or computing power (Openshaw 1995), however it was also due to the lack of understanding that geographers of the seventies had about their world. As Openshaw comments "it was...clear that much more attention was needed to developing a basic descriptive understanding of how many people-related systems actually worked, why people behave the way they do, and how they interact with state and various other organisations that impinge on their social, political and economic spaces and environments" (1995, p. 160). This lack of understanding was at

the root of the dissatisfaction felt by many and perhaps epitomised by Sayer (1978). Not only did quantitative geography not have this social understanding, its approaches offered little hope of obtaining it. Geography had become technique rich, but theory poor. It had a powerful means of illuminating and describing phenomena, but no means of explaining them.

Coinciding with these difficulties was a shift towards a more politically motivated platform for geography. There was a new focus on the individual and their own world. The secondary data led analyses of the quantitative era and the search for social laws and generalisations were unsuited to this new focus as this acidic comment from Sayer highlights: "Using methods perfected in physics for the study of people may be as absurd as trying to interview an atom." (1989, p. 122)

### *3.3 - Quantification Crisis: The rise of humanism*

The leading edge of social geography moved away from quantitative methods and a positivistic philosophy into a series of '-isms' broadly bound together by a humanist theme. The importance of context in knowledge and particularly in empirical information became highlighted. Context refers here to a consideration of the origins of information, in particular the 'gaze' through which information is being constructed. The advocates of these new approaches no longer saw the world as detectable, readable and fixed. It was thought to be dynamic and individual. This conceptualisation of society and its processes explains the dissatisfaction felt with state created secondary data. Many geographers were no longer satisfied with a reliance on data which presented a pre-determined, *pre-conceptualised* version of the world.

The complex and rapid turnover of approaches to spatial phenomena which characterise this period need not be related here (for details see Johnston 1979, chapter 5). Whilst some historical accounts of this development give the impression that quantitative geography was stagnant at this time (Zelinsky 1970, Harvey 1973), this was certainly not the case:

### *3.4 - Meanwhile...*

As computational power increased and better quality data were constructed, many quantitative geographers pursued the development of their techniques. In the applied atmosphere of the eighties, some found an increasingly receptive commercial market for their expertise whilst others continued to utilise quantitative methods in more conventional research. It was during this period that the advent of *geographic information system(s)* (GIS) took place.

Establishing a clear definition of GIS provides the potential for an entire thesis. The term is confusingly applied to the physical systems, consisting of hardware, software and data, to the research process which employs these systems as a tool and to an (imagined) paradigm (Pickles, 1995). Goodchild (1995) states "(t)hough many definitions of GIS exist, most identify a database in which every object has a precise geographical location, together with software to perform functions of input, management, analysis and output." (p.35). For a comprehensive discussion of the difficulties of definition the reader is referred to Pickles (1995), Martin (1996) and Maguire, Goodchild and Rhind (1991).

The first collection of components widely accepted as a GIS was The Canada Geographic Information System (see Tomlinson 1988). From these beginnings a curious academic/commercial partnership grew up. The automated analyses of spatial phenomena have a vast market outside academic geography, as well as extensive use within it. The cross-sector nature of GIS development might well account for certain aspects of its character, since the priorities and concerns of the commercial world may be far removed from those of academia. This point will be explored more fully in section 3.11.

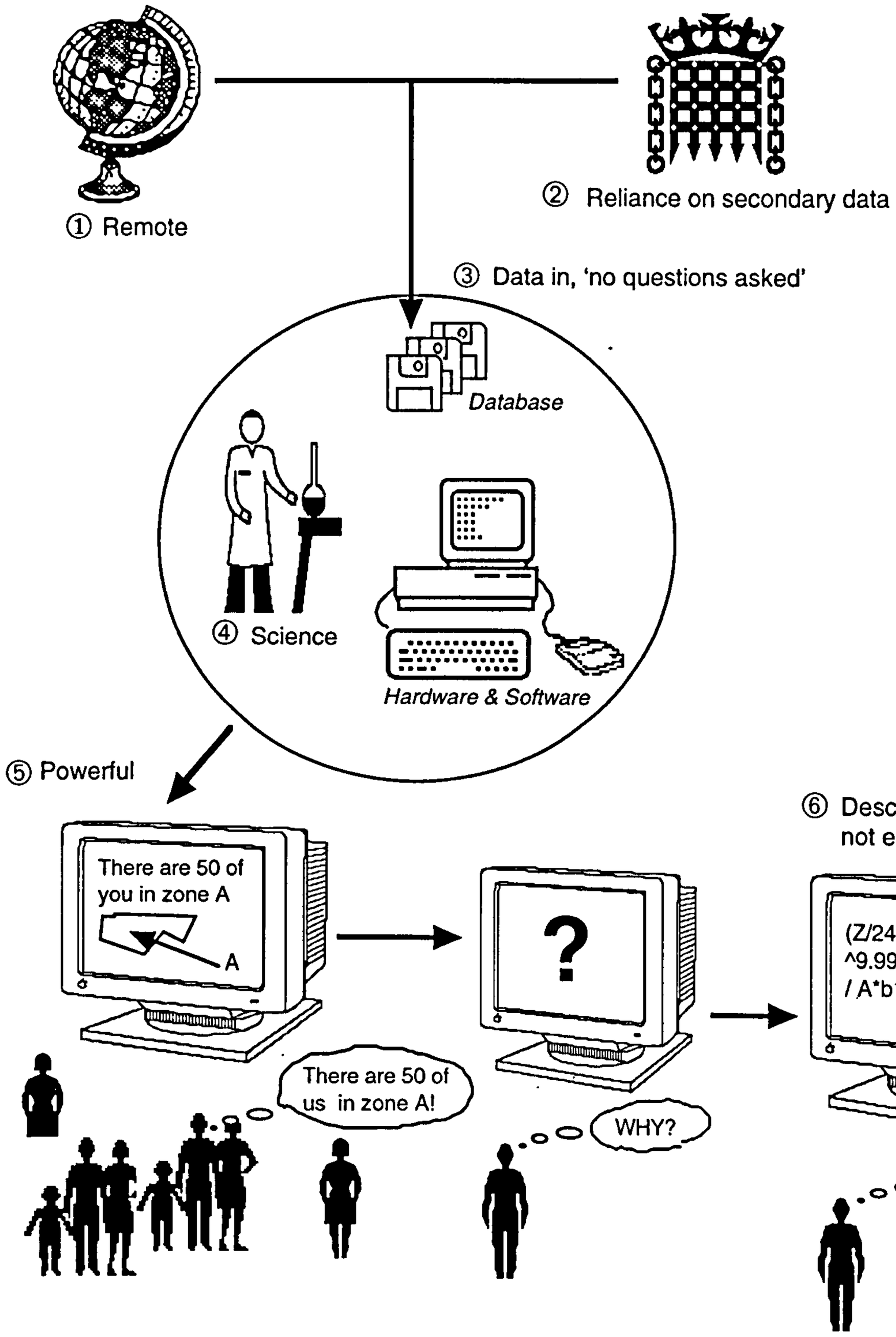
## *3.5 - Geographic Information Systems*

Perhaps more than any other field within geography, that employing GIS has acquired its identity from those who do not practice it. This identity has much to do with impressions of the quantitative revolution, (now consigned to history by some as a "20 year diversion" (Johnston 1985)). Many, perhaps especially those who did not live through that period of change, apply the 'quantifier' label and its associated derision to all those who use or develop GIS. The field is seen as the resurgence of a quantification movement and thus as something to be resisted at all costs for the reasons previously highlighted. It is important to understanding the shock and bewilderment that geographers experienced at the close of the quantitative era. The literature of the time is still a powerful influence on the critics of today's quantitative movement. It is as though the discipline has been traumatised by its previous experience of quantitative geography and has not yet recovered enough to face it again. The publication of 'Ground Truth' (Pickles 1995) marked the first contemporary collation of these criticisms. Whilst some of Ground Truth's later chapters (notably McHaffie's) add little to the debate through their pursuit of abstract concepts, the volume is valuable because it raises a number of key issues which GIS researchers have failed to address. It also provides an excellent example of the equation of GIS with quantitative geography and that field's past mistakes.

It is thus vital to explore and understand the criticisms of GIS, as a form of quantitative geography and thus in the light of what has gone before. To do this, the most important criticisms have been distilled into the six themes shown in figure 3.1.

## *3.6 - Criticisms of GIS*

*Figure 3.1 - Criticisms of GIS*



The following sections explain the figure:

### 3.6.1 - Remote

The purpose of the globe at section ① is to illustrate the criticism that GIS operate under conditions of remoteness from the individual's world. This criticism is of no surprise given the humanistic elements which are embedded in much contemporary geographical thought. It is argued (Curry 1995) that in being remote from its subject, GIS miss the individuality of processes formed by and impacting on people's lives. The computational model of the social world is a static one, or at best formed from fragments of time. It has been recognised since early Greek times (by Heraclitus in fact (Gaarder 1991)) that life *flows* and is not comprised of discrete parcels of time or events. It is therefore argued that extinguishing the dynamic from a model of society renders it inaccurate. For others, it is the reduction of aspects of life to a numerical value which is unacceptable. As Hegel comments, number is an "entirely in-active, inert and indifferent characteristic in which every movement and relational process is extinguished" (quoted in Georgescu-Roegen 1971, p. 72). In some of today's geography, especially that from which many criticisms of GIS arise, qualitative data gathering techniques take the researcher into direct contact with their research subjects. Remoteness from the subject for these geographers is unthinkable. Given that this research is heavily based on secondary data and has no contact with 'subjects', this is an important criticism to deal with.

For a researcher who believes that empirical information can only be gathered or made useful in one way, these are valid criticisms. Making GIS compatible with social science thus requires the adoption of two standpoints.

Firstly, there must be a belief that snap shots of society held in numerical representation *can* be valuable in illuminating its ongoing processes. Heraclitus suggested that it is not possible to stand in the same river twice, since it is continually flowing and changing. The extension of this argument is that secondary data sets, such as the census, are invalid attempts to capture the nature of society. They fail because society is in continual motion. However, the flow of a river can be understood in relation to its confining channel, its depth speed and load. The flow and its physical characteristics are inseparable. So it is with the snap shot image of society. Data such as the census reveal a great deal about the physical characteristics of a population. Guided by theory and understanding, much can be inferred about a society from those characteristics. The key element in the analysis is to understand its limitations. As one could not infer the *reason* for the flow from physical

measurements, so the flows and reasoning in society cannot be directly inferred from secondary data.

Secondly there must be a recognition that GIS *are* remote from the real empirical and experienced world and that this will be problematic. However *no* form of data construction is unproblematic, and effective use of a GIS thus requires a suitable research practice to deal with these problems. It is again a matter of knowing the limitations which a methodology brings to research. As a focus group based researcher should not attempt to generalise their findings to a wider population, so a secondary data set user should not attempt to infer the reasoning of individuals.

### 3.6.2 - Reliance on Secondary Data

Point ② is marked with the symbol of government. Contemporary critics often comment that (i) GIS is heavily reliant on state constructed, secondary data to build its models and that (ii) such a reliance should make GIS the objects of suspicion, or even that they are the "handmaiden[s] of the state" (Taylor and Johnston 1995).

To suggest that employing the data in GIS makes them some kind of political servant is problematic. GIS are a technology and like all technologies they can be employed for both positive and negative purposes. There are many examples of research which uses GIS and state constructed data to criticise the government (see for example Beatty and Fothergill 1996).

As a powerful data analysis tool, GIS have been utilised in the allocation processes and war machines of the establishment. As attention is often drawn to the losers in these processes, or the unpleasant roles of technology, GIS have become tainted by their involvement (see for example Smith 1992, Goodchild 1995). There is the possibility that a model built by this research could be used to alter allocation processes leading to the loss of resources for some residents. Others may gain however, and ideally the model should help to direct resources towards those in greatest need. In developing a tool which may assist in designing a more socially just society, GIS research has the capacity to *benefit* residents. An awareness of the possibilities for misuse of GIS and taking responsibility for work that is personally undertaken, is as far as one may go in reacting to this criticism.

The issue of what preconceptions lie behind the state's construction of the data is of more importance. Data drawn from the UK Census were utilised to

construct social groups or classes for this work (see chapter 5). Households and individuals have already been assigned into categories in the publicly accessible census data, using their responses to pre-determined questions. The process through which this occurs from questionnaire conception, implementation and answering, to data processing and release, thus requires careful consideration. In fact, as will be seen in chapter 6, the nature of the data used did have a significant impact on the nature of the groups derived from them. This does not however make this research the handmaiden of the state. It demands a more careful understanding from the researcher of their data and how they are used.

### 3.6.3 - Data Treated as Unproblematic

At stage ③ the data enter the GIS itself. The GIS is depicted as being composed of three important elements; database, hardware / software and user. Whereas criticisms highlighted at stages ① and ② hinge on beliefs about the best way to carry out research and are essentially rhetorical, stage ③ marks a specific shortcoming in research practice.

For quantitative researchers, issues of data quality have most often been associated with the introduction of error to a model, and its subsequent propagation through manipulations of the data (see Goodchild and Gopal, 1989, especially chapter 4). An understanding of the sources and possible ramifications of error is essential for GIS which may have an input to policy, or in which considerable manipulation of data might take place. Understanding the potential for error can lead to the calculation of confidence limits around key numerical values, but it should also focus the attention of the user on the potential fragility of their data and any model built from them. Remembering that the validity of models is dependent on data quality can be difficult when faced with apparently authoritative visualisations of it. Visualisation is considered a key strength of GIS, and they usually employ maps which have authoritative connotations (Harley 1990). Addressing the task of cautioning those who use the outputs from GIS is a significant mountain, as yet unclimbed by the GIS-using community. Active consideration of potential error in the data and models will be presented for this research.

Of greater importance is understanding the model of society which lay behind the construction of the data in question. Addressing this issue has profound implications for the model being produced in GIS and for the research as a whole. The identification of a 'void' in contemporary GIS research where

consideration of these issues ought be, is closely linked to another common criticism. This is that GIS as a research field is 'theory-less', both in understanding the world it is trying to model and in a conception of the modelling process itself. Whilst formal identification of stages in the transformation from empirical reality to computer model may have taken place (Martin 1991) little consideration of wider issues, such as the ontology underpinning these representations can be found. For GIS focused on socio-economic issues theoretical considerations of the world they are modelling has yet to move to the fore despite a number of calls for it to do so. It is unlikely that all GIS research papers will ultimately include a substantial section outlining the particular theoretical beliefs of the authors, however the critics suspect that no consideration of the issues is presently in existence. The acceptance of data into GIS 'no questions (of any sort) asked' is apparently current practice for many GIS users. Such a practice prevents the influx of many ideas and issues which might otherwise enhance the work that is done.

This research will attempt to address these issues by exploring the constructs within the census data it uses, and the social theories relevant to the work. It will attempt to illustrate how social theory and technology have influence the methodology adopted and results obtained. Discussion of these issues will be found in all the following chapters.

### 3.6.4 - Positivistic

Stage ④ of this model is represented by a stereotypical scientist and computer. For many critics, GIS are a thin new disguise for positivistic philosophies. The direct equation of quantification with positivism (by Sayer 1984 for example) has been clearly shown to be incorrect (see for example Pratt 1989, Johnston 1986) and this research refutes any such link. A positivistic methodology is designed to gain understanding of empirical appearances by conceptualising them as examples of general laws (Giddens 1974). The fact that this research requires a re-conceptualisation of what it is that census data reflect and that the model results are shown to be dependent on how the data are manipulated, illustrates that this use of GIS is certainly not about passive detection of phenomena. It is about actively constructing a model of the empirical world as conveyed by a specific data set. No hypothesis is set or tested, no general laws are sought.

### 3.6.5 - Powerful

Stages ⑤ and ⑥ indicate those criticisms to which this research and that using GIS as a whole need to pay most attention. Stage ⑤ attempts to illustrate

the potency of communicating something which may be perceived as scientific. Despite all its criticisms, quantification and its associated scientific-gloss undoubtedly carries considerable weight in the communication process. Griffin and Phoenix effectively illustrate this point in their account of using both quantitative and qualitative techniques:

*"the nature of the knowledge provided by the qualitative study was discounted by even more 'sympathetic' councillors and council officers because they needed to be able to quote 'statistics' or quantitative evidence, in council meetings, in order to bring about political change such as improving the council provision for young black people." Griffin & Phoenix (1994)p. 292*

It is unfortunate that much of the decision making and policy formulating architecture of our society relies on 'facts and figures' which are easily manipulated. However it is undeniable that we are conditioned into a reliance and to trust this kind of information. Controversy over the lack of scientific 'fact' in the 'Mad Cow' disease scare (1995-6) provides potent evidence for society's dependency on such information. There was genuine surprise that scientists were unable to provide precise or reliable indicators of the extent of infection and likely developments. The dependency of rigorous scientific method on a reasonable quantity and quality of information is not discussed in the media.

GIS couple the potency of number with that of the map or image. Whilst cartography has explored the potency of its images and the constructions behind them through the work of Brian Harley (1989;1990), the public lack such a healthy cynicism. Maps carry an air of authority. They legitimate their creator's view of the world, prioritising the spatial arrangement of features and the attributes with which those features are associated. The map hides the processes and relations which join elements, presenting an unproblematic association in the viewers mind (Harley 1990).

Visual image is an effective mass communicator. Hence, at stage ⑤ the audience receives the information, which becomes 'truth'. Only one audience member asks the question 'why?' and introduces the final criticism of GIS presented here (stage ⑥).

### 3.6.6 - Description, not Explanation

A GIS analysis provides description rather than explanation. Without a theory of what is represented and how the representation has been brought

about, GIS users will never be able to advance beyond description. This is the same criticism of quantitative methods that Sayer and others advanced in the late seventies. It is most certainly as true now as it was then.

Many GIS users are well aware that description does not equate with explanation and would never seek to make such a link. GIS research, like that presented in this thesis, is often concerned with developing new visualisations of phenomena. The development of alternative modes of empirical information through the use of quantification and mapping, can be beneficial to theoretical endeavour. New views of spatial phenomena, alternative analyses and new relationships between variables can prompt the development of theory. The type of model and image introduced in figure 2.8 may well be able to contribute to debates over residential differentiation. The author believes that theoretical progress can only be made in conjunction with empirical understanding agreeing with the faction within geography which believes that the deeply theoretical 'explanations' offered by certain kinds of geographers are counter productive. As Ley (1989) comments, "when explanation is referred to deep levels beyond the reach of empirical verification, it is possible for self-referential assertions to be presented as self-evident truths... The result is an ever increasing abstraction and formalism of discourse until the divorce from the empirical world is complete." p.238. The path from data to theory is problematic but worth pursuing.

The following section sets out a range of *practical* responses to the criticisms outlined above. It is the first of many attempts, throughout the thesis, to address the criticisms and subsequently enhance the research quality.

## *3.7 - Reflexive Action*

Throughout the discussions in section 3.6, relevance can be found for the most tangible development geography has made in response to its humanist critics; the introduction of reflexivity. It is paramount to be clear that whether scientific or not, quantitative or not, all research is affected by its practitioners and their choice of methodology. Although geographers have been exploring this issue since the eighties it is in fact a much older concept. In the nineteen fifties, Huxley (1959) was attempting to discuss this very issue in a series of lectures and essays to American students. He invoked Shakespeare to make his point: "But thought's the slave of life and life time's fool" (King Henry IV, Part I, V). This research is 'the slave of life' in a multitude of ways; in its philosophy, methodology, techniques, data and even in its reason for being. In recognition of this a number of practical actions have been taken which

attempted to introduce reflexivity to the work and to counter some of the criticisms explored in 3.6. There were three main principles to act on.

### 3.7.1 - Awareness of Power

The aim of the research was to produce a model and accompanying images of the socio-spatial structure of urban areas. An awareness of the power of image was therefore required, both to focus attention on the quality of the model and as a reminder that an image or an image-producing technique can be utilised for a variety of different purposes. There had to be an awareness of the implications of that image for the households and people it depicts. This implied a commitment to determining the accuracy of the representation in both numerical and conceptual terms. Although an evaluation of how individuals and households have been represented can only be subjective, in presenting it the work is open to scrutiny.

### 3.7.2 - Awareness of technological and personal influence

The available technology and data have had a significant impact on what research was undertaken and how it was approached. The research was prompted by the release of the Sample of Anonymised Records and by the availability of funds. The approach adopted to achieving an integration of small area data and disaggregate household information was influenced by the environment in which the work took place, notably the advice and expertise that was available. Interestingly, it is possible to trace the direction in which the work was pursued in conjunction with the attitudes of the researcher and the concepts to which he was introduced during the work. More specifically, the adoption of techniques outlined in chapter 6 (which were borrowed from a different field of geography), was suggested by the proximity of researchers utilising them. Indeed, the discussion of these issues is a direct result of exposure to certain ideas and individuals in the research training process. All of these factors have ultimately conditioned the nature of the research and its outputs. The extent to which this kind of information should be included in the presentation of research is however difficult to judge. Perhaps the declaration that these factors did have an influence will suffice.

### 3.7.3 - Beyond Awareness

Beyond an awareness of these issues lay some day to day practical actions. Many of the criticisms of GIS research have centred on poor research practice. Specific action was proposed to enhance this research, having identified these areas of weakness.

Building an understanding of the data sets used (both in technical and conceptual terms) was of primary importance. Understanding the data in conceptual terms required a grasp of the nature of the enumerated population and an exploration of theories of class and socio-spatial structure was needed. The acquisition of theoretical understanding alone however would not have been sufficient. It needed to be carefully exploited during the technical development process. Later chapters will illustrate how this has been carried out.

The following sections explore these theories and what can be gained from understanding them. If the debate previously outlined can be described as being between quantitative based, descriptive geography and theoretically based, explanatory geography, the development of approaches to socio-spatial structure will fit comfortably into the same model. Certain aspects of this development will however, add complications. There are two elements to any consideration of socio-spatial structures; that society is structured and that this structure is etched in space. The whole of social science has been concerned with issues of social structure and the development of ideas has not been a consistent, co-ordinated one. The multi-disciplinary environment in which interest in social 'class' and spatial structure exists has not led to co-ordination in the progress of theory and understanding. Ideas have bloomed and died in some fields before they have even penetrated others. Thus the temporal axis which underpinned our excursion through the geographical debate will be absent here–replaced with an axis describing the shift in class concept from structure, to process and from a-spatial to spatial.

## *3.8 - Structural Class Schemes*

'Class' is a term confusingly applied both to structural models of social hierarchy and to theoretical accounts of the processes by which social inequalities are produced and maintained. Although, as will be shown, the two have ultimately been dialectically linked, this was by no means always the case. The focus of interest in notions of class is the recognition that society is unequal and that this inequality has become a hierarchy. A vast range of indicators such as life expectancy, morbidity (Townsend and Davidson 1988), income (Coates 1989) and educational achievement (Halsey 1988) illustrate that some people are 'better off' than others, and that it is systematically the same groups of people who are best off, whichever indicator is being considered. Schemes which describe and arrange the "broad contours" (Crompton 1993, p.49) of this inequality are perhaps the most widely utilised models of class. In the UK, as in many societies, such a structure is often built

using groups of occupations to outline its various levels. The most common of these uni-variate schemes is that of the Registrar General (Crompton 1993):

*Table 3.1 - The Registrar General's Class Structure*

| | |
|---|---|
| I | Professional Occupations |
| II | Intermediate Occupations (Managerial / Technical) |
| III | (N) Skilled Non-Manual Occupations |
| III | (M) Skilled Manual Occupations |
| IV | Partly Skilled Occupations |
| V | Unskilled Occupations |

This classification scheme, which has consistently appeared in the UK Census reports, is often employed by policy makers, journalists and researchers. Such widespread use has lead it to become a common language through which groups may be identified and the population discussed (Crompton 1993). Attempts to modify it, notably the 1991 Standard Occupational Classification (SOC) still emphasise occupation. Whilst "the main aim of the new SOC was to classify jobs... as opposed to classifying persons" (OPCS 1991) it has the same implicit assumptions as the Registrar General's scheme; social status is equated with occupation. The basis of these ideas is a positive relationship between the status of an occupation and its material rewards. Whilst perhaps these two variables used to be quite closely aligned, as society has developed the alignment has become distorted. University lecturers for example, are regarded as professional in these schemes and are therefore at the top rank. Today's lecturers command neither the material rewards nor the widespread respect which might be suggested by their position in the class structure. Despite anomalies such as these, Crompton notes "the wide measure of recorded agreement concerning the relative prestige of different occupations corresponded to the distribution of material rewards and power attached to the occupations in question." (1993 p. 55).

## *3.9 - A Concern with Cause*

Utilising univariate class schemes in GIS would be easy. The data could be accepted 'no questions asked' (see 3.6). The purpose of undertaking this exploration of theory is however to *enhance* research practice and there is a need to know the context and nature of such data (see 3.7.3). If such a structure does indeed reflect some difference in society, what brings it about and why should it be sustained or tolerated? The politicised academy of the

seventies, tired of static, descriptive, quantitative accounts of social structure, proposed some explanation. Capitalism was held up as the driving force behind social inequality (for a mechanistic explanation see Johnston 1986 pp. 25-36). As Marxist theory became popular, social inequality became re-conceptualised in terms of relations to capital and suddenly the explanatory weakness of the census was highlighted. "In the social classes of the census, the owners of the capital are lost to sight", explained Nichols (1979). As Marxian and soon, Weberian (Weber, 1970, Hamilton and Hirszowicz 1993) conceptions of society were explored, attempts were made to produce 'theoretical' class schemes (see for example Wright 1979 or Goldthorpe 1974). Theoretical class schemes tried to offer an empirical account of social structure which could be wholly tied to a theoretical account of its existence. Classes were conceptualised in terms of their relation to a variety of factors including power, capital and property. Unfortunately, two problems persisted. Firstly, as Drudy (1991) highlighted, whichever theoretical or political standpoint was adopted the schemes produced essentially the same structure as before. Secondly, despite the theoretical underpinnings of these schemes there was still no mechanism for their continued existence. These shortcomings were enhanced by two related problems with the hierarchical accounts:

## *3.10 - New Concepts of Class*

The first was that the data used in these schemes were starting to become inappropriate for the task. During the eighties shifts in society had begun to challenge the primacy of economics in determining social class. As social differences narrowed in economic terms, ideas were put forward for the basis of *new* classes: "The emphasis has shifted from groups to individuals, from classes to categories and from producers to consumers" (Pahl 1989 p. 714). Pahl suggested that as society became richer, the categorisation of individuals or households by their socio-economic attributes needed to be replaced by one based on attitudes and values. Offe (1985) suggested a classification based upon consumption behaviour would be appropriate. Such a shift would have presented no difficulty for analysts if a direct link between affluence and attitudes could have been established (Pahl 1993). Unfortunately no such match could be made.

Secondly, whilst Marxist theories were well developed, theories of attitudinal or consumptive class were not. Having been striving to find an effective connection between economics and social characteristics with appropriate data, class theorists suddenly found themselves trying to deal with new and

less tangible concepts, without appropriate data. State generated data continued to be constructed as before.

## *3.11 - Geodemography*

In the meantime a different group of researchers had been carefully exploiting their own understanding of social grouping and its implications. The geodemographic analysts (see 2.8) had established strong links between the collective characteristics of groups of residents and their consumer behaviour. Although geodemography avoids explicit ordering or hierarchy in its social group schemes, that there is a hierarchy at work is undeniable. Since geodemography is essentially a tool for stereotyping and then targeting the consumer with money to spend, it is difficult to avoid the lauding of the wealthy and a disregard for the 'poor' (as they are described in Euro-Direct 1992). MOSAIC (a geodemographic scheme) discusses *clever* capitalists, *rising* materialists and *problem* families or *graffitied ghettos*. Since geodemographic analysis holds an important role in directing the location of investment and in the construction of consumer identities (Goss 1995), such conceptions of society hold considerable power (see stage ⑤ fig. 3.1). Geodemographics are also distinct from other static class structures in that they have a strong spatial component. The profiling of small areas indicates a link between location and social character. Geodemography successfully isolated and located homogenous groups, often with common backgrounds, status, aspirations, worries and opinions. Given their success, why are geodemographic schemes so unpopular in academic circles?

Firstly, the areal rather than individual basis of geodemographics considerably reduces their value to a researcher interested in class process. During the era in which geodemography was developed, geographers and other social scientists increasingly focused on the individual and their experiences. Secondly, although geodemography is spatial in that it offers a categorisation of areas, space is simply a means of identifying the locations of target populations. Academia had already noted the spatial differentiation of social groups and wished to progress beyond its description, to its explanation.

## *3.12 - Space, Socialisation and Process*

The link between residential location and social character has already been touched on in chapter 2. There is an extensive history of research into urban residential differentiation, with explorations of the role of the built environment, socio-economic environment, housing systems, managerialism and urban perceptions. These elements all contribute to a somewhat

fragmented understanding of urban residential patterns. Ley (1987) provides a thorough exploration of the history of research into this phenomenon.

The joining of residential differentiation theory with theories of class structure and process has been the catalyst for advances in class theory as a whole. Stark (1980) comments that in class analysis "the object of study is not the elements themselves, but the relations between them" (p. 97) and the key element in understanding how those relations occur is their *location*. In his seminal book 'Learning to Labour', Willis (1977) recounts an ethnographic study of 'working class' children in the final stages of their schooling and their transition into employment. He reports that the children's social group plays a key role in the construction of their class identities. Residential location is also a key factor in the generating opinions about others and about each other. As Larry (one of the boys in the study) illustrates, he is regarded as a snob "not only 'cos we go to a grammar school, but just because we're from here in the first place, which is regarded as a snob area." (Willis 1977, p. 58). Willis explores the contextual nature of class processes as they unfold in front of him and notes the manner in which meanings, styles, languages and behaviours are being constantly taken up, reproduced and reinforced by 'up and coming' working class children. "The home, family, neighbourhood, media and non-productive working class experience in general are equally vital for its continuous reproduction." p.106.

The connection between residential differentiation, arguably a spatial manifestation of class structures (Castells 1977), and social differentiation is clearly a valuable one. Whilst urban theorists had pursued research into the mechanisms behind residential differentiation, "analysts began to realise that the urban residential mosaic was the result of not one but several separate, though not usually independent, processes which allocated different groups to different areas" (Herbert and Johnston 1978, p. 19). They had also not explained the origin of inequalities in the ability to compete for urban locations despite these being the driving force behind their mechanisms. A combination of ideas of social structure and spatial structure was needed.

The socio-spatial dialectic, described by Soja (1980) but made relevant to these debates by Harvey (1989) knits the two elements together. "Models of how to behave are provided by the home, but also in school and the neighbourhood itself. Home is under direct control of the parents, but the other two reflect the nature of the area in which the home is situated" (Herbert and Johnston, 1978, p 25). Thus a household and its residents are shaped in thought, word and

deed by their residential area and in turn contribute to the shaping of that area. Individuals originating from a specific class of household and neighbourhood carry that socialisation with them into the processes and systems governing their occupation and place of residence. The neighbourhoods in which they settle are in turn affected by their presence and so the cycle continues.

## *3.13 - Using This Understanding*

With this conception of residential area and class processes in place, the actual nature of a data source like the census may become more clear. The data are a snapshot of a very complex process and cannot possibly incorporate all of the elements which control and are controlled by social class and residential location. This snapshot is taken with a large number of distorting and filtering 'lenses' in place which can affect the appearance of the subjects. The origin of these distortions lies both in the choice of variables to include on the census questionnaire and in the preconceptions embedded in the structure of the questions. The data thus require some decoding to counter the effects of these lenses, or at least to understand their impact. A grasp of how urban class processes may work will help this decoding process, and has been used in this research.

At this stage it is not practical to explore in detail how the discussions above have been used in the research process. Where possible this will be woven into the text describing specific technical development. In some cases this information has been used very directly, (for example, in the selection of census variables). Elsewhere it is has also had a less tangible, but significant influence on the research process. Through a consideration of how residents acquired their social character, and their residential location the author feels that the remoteness in section ① of figure 3.1 has been reduced. Data has not been used "with no questions asked" (stage ③, figure 3.1). By focusing on the limitations of census data and the potential power of a visual image, a sense of responsibility is heightened.

## *3.14 - Conclusions and Summary*

This chapter has reviewed the origins and implications of the relationship between geography and quantification. It has specified a number of criticisms of quantitative geography which, it is hoped, may be addressed through practical actions throughout the research process. One of those criticisms is that too much quantitative geography research takes place without a strong theoretical input. Given that this research is focused on socio-spatial

structures, relevant social theories have also been reviewed. The result has been a better understanding of the data involved in the research, and of how these data relate to social processes.

It is hoped that the inclusion of a theoretical chapter will achieve a number of things.

- To illustrate that through a consideration of the contemporary debate about the place of GIS in geography, *practical* action can be attempted which may enhance the quality of the research. This enhancement will come by answering reasonable criticisms made against GIS research in general. There is also value in stating, and drawing some reassurance from, the fact that certain criticisms are either invalid or stem from alternative schools of ontology and epistemology. Refuting a criticism, having spent time considering it, is a better means of strengthening the position of a piece of research that ignoring it altogether.

- To illustrate that an in-depth analysis of relevant social theory has *practical* benefits and will enhance the quality of the research too. Technical decisions become a negotiated choice based on an understanding of data strengths and weaknesses rather than just a matter of doing 'what works'. Choices made in the light of what is known about socio-spatial processes may be more robust and more easily justified.

Finally, the chapter stands as a mark of the author's belief that GIS research must become more theoretically aware and informed than it is at present. Such a development can only enhance its value and further safeguard the interests of those who are represented in its models. The chapter has illustrated that a quantitative researcher has engaged in the debates about GIS and class and usefully drawn from it whilst maintaining a quantitative approach. Valid methodologies can lie alongside one another in co-existence. If the reader finds this a new position and one that they can adopt, the chapter has served the best possible purpose. The following chapter re-focuses the reader on the specific research task by reviewing relevant empirical work and exploring techniques which might assist in enhancing fine scale socio-spatial information.

# Chapter 4

## Chapter Overview

In this chapter relevant empirical work is reviewed, complementing the theoretical explorations of Chapter 3. Work on foreign microdata resources is explored and integration of aggregate and disaggregate data is highlighted. No work of direct relevance for the research tasks (see 2.11) is found. Empirical work from the UK is also reviewed, covering microsimulation approaches to providing fine scale socio-spatial information and work which actually does attempt to integrate Samples of Anonymised Records (SAR) and Small Area Statistics (SAS) data. This work is found to be problematic and does not offer a technical solution to the research questions. Finally, data fusion is explored as a possible means to combining SAR and SAS data. The technical reasons why data fusion cannot be employed to do this are highlighted and it is concluded that an innovative approach, developing new techniques will be required.

## 4.0 - Introduction

Chapter 3 discussed the importance of acquiring and using theoretical views of the world which this research is trying to model. Theoretical literature is not however, the only source of reference. In this chapter a review will be offered of relevant empirical work. Although chapter 2 made it clear that there is no widely available technique which comprehensively accomplishes the stated aims, useful information may be gained from an awareness of what has been tried (and may have failed) to obtain the information of interest.

## 4.1 - Foreign Microdata Sources

The research presented in this thesis was prompted by the release of a census microdata source for the UK (the SAR) because the relationship between the SAR and SAS data hinted at the possibility of an integration (see 2.14). Other countries have had publicly available census microdata for considerably longer than the UK (Marsh et al. 1991). It is therefore appropriate to determine the kind of research which has been undertaken with those data and in particular whether any researchers in foreign countries have had similar ideas about integration. For this purpose census microdata will be defined as "a sample of anonymised individual and/or household responses to a national census, publicly available in a disaggregate format". Although the

precise format of the data varies from country to country, most conform to this core definition.

The American and Canadian microdata sources dominate the geographical literature surveyed, although other countries including Australia, Sweden, Italy and several South American nations also have census microdata sources. The microdata generally share a number of features. Firstly, they tend to consist of very small samples of the national population. In Canada for example the Public Use Sample Tape (PUST) is a 1% sample, whilst in the USA the Public Use Microdata Sample (PUMS) is released at a variety of levels from 5% to 0.1% depending on the detail required (Johnson and Roseman 1990). Secondly, the data are deliberately released at a poor spatial resolution in order to protect the identities of the individual sample members. The smallest area of origin given for a microdata sample is one whose population is a least 100,000 (Wilson and Marta 1989) in the USA. This is comparable with a value of 120,000 for the smallest area in the UK (Marsh 1991).

## *4.2 - What Are Microdata Used For ?*

Microdata sources have been available in some countries for over twenty years. In the USA and Canada, the majority of published research papers mentioning the PUMS or PUST are concerned either with the study of specific ethnic groups or with migrants. Rather than dissect this work in detail, which would provide little of direct relevance for this research, the *nature* of the work that has been carried out will be considered. The work may be separated into a number of categories:

### 4.2.1 - How Attributes Relate to Each Other

The question to which microdata are most commonly applied is 'do individuals in possession of a particular attribute also tend to possess other specific attributes?'. Determining the co-variances of attributes often leads to the question; 'which attribute may account most for a specific behaviour?'. For example, Breton et al. (1990) explored racial discrimination in Canadian urban job markets. Having identified evidence for discrimination in aggregate data, they attempted to ascertain whether ethnic status was the determining factor in employment differences, or whether the differences were based on a factor associated with it. Income was utilised as an indicator of job status. Having determined the statistical effect of seven factors on income, multiple classification analysis was used to examine the effect of ethnic status with each of the seven factors held constant. Breton et al. discovered that the inequalities were nearly all accounted for by differences in training and

educational attainment, which appeared to cluster strongly according to ethnic status. Only one ethnic group, (West Indian) appeared to be suffering discrimination explicable solely in terms of racial identity.

Another research methodology which employs microdata is the identification of a target group through their possession of a specific characteristic. Once a list of target individuals is assembled, other features of interest can then be explored. Grigsby and McGowan (1986) for example, used the PUMS to identify male adults still living at home and subsequently explored their economic and demographic characteristics. Work of this kind has also started to appear in the UK, following the release of SAR (see for example, Crompton, 1995). Whilst this kind of research is of no direct relevance to the work presented in this thesis, there are two important themes to note: (i) the use of census microdata to investigate how individuals and households actually form features detected in aggregate census data and (ii) the notion of examining how individual attributes group together according to some degree of person type. These themes are important because they indicate the use of microdata sources to investigate the relationship between a population and the data describing it. An understanding of that relationship was thought likely to be significant in the search for a means to integrate aggregate and disaggregate data.

This research is focused on the spatial differences in phenomena described in census data. The following section places the existing research in that context:

### 4.2.2 - Spatial Analysis

Although microdata sources tend to have poor spatial resolutions, they are often adequate for exploring migration at a large spatial scale, or a city's socio-spatial structure at a district scale. Johnson and Roseman (1990) for example, used PUMS data to explore a number of factors in out migration of Black groups from Los Angeles. They looked particularly at household composition and its change through time, together with the relationship between migration and household structure. Others have followed a very similar pattern of analysis, classifying migrants by a range of attributes including occupation (Ellis et al. 1993), country of birth (Belanger and Rogers 1992) and race (Gregory 1995). It should however, be noted that not all the work reviewed concerned the topics of ethnicity and migration. Lichter et al. (1989) for example, used the PUMS to identify geographical variation in female marriage rates in the USA. Preston et al. (1993) examined differences in

commuting behaviour according to location and other socio-demographic factors. It is likely that the dominance of research into ethnicity and migration is a result of a self-sustaining awareness of the microdata sources within those particular fields, and the suitability of the topics covered in the data.

The work of most direct relevance for this research, was also concerned with using microdata for spatial analysis. Openshaw et al. (1985) used Italian microdata to determine the social composition of regions of Italy. They were granted access to a 100% microdata set from the Italian census, (i.e., anonymised records for *every household in the country*) which was not publicly available. Although their main task was to classify the households into a manageable scheme, an analysis carried out with the new household classification is of interest:

A clustering process produced two classifications from the 15.6 million records. One scheme used 14 household groups, the other 30 household groups. Since each household's spatial origin was identifiable at the equivalent to enumeration district (ED) level, it was possible to explore social composition of any area at ED level or above in terms of the percentages of households belonging to each cluster that were present.

*Table 4.1 - Extract From A Crosstabulation of the 14-Cluster Classification by Regions (row percentages)*

| Group: Area | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Piemonte | 5.1 | 6.5 | 1.9 | 9.2 | 2.7 | 20.3 | 8.7 | 0.5 | 9.7 | 7.5 | 3.7 | 15.1 | 0.7 | 8.4 |
| Val d'Aosta | 6.0 | 7.7 | 2.2 | 10.3 | 2.4 | 17.8 | 6.9 | 0.4 | 8.6 | 4.7 | 4.4 | 17.9 | 1.0 | 9.7 |
| Lombardia | 2.2 | 5.8 | 1.9 | 10.3 | 4.5 | 21.0 | 9.8 | 0.4 | 9.0 | 8.3 | 3.5 | 11.6 | 0.7 | 11.0 |

Table 4.1 illustrates an extract from the results of this analysis. An indication of the presence of each household class in each region is given. This work is encouraging since it clearly indicates that others see the value of a means to investigate sub-unit social variation. However, since the microdata used already had a very fine spatial resolution means that the technique cannot be regarded of direct use for our purposes, since the SAR have a very coarse spatial resolution. The sub-regional analysis was only facilitated by the extraordinary nature of the available data. In order to attain a level of information similar to that offered by Openshaw et al.'s analysis for the UK,

an integration of *two* data sets was required, one providing locational specificity, the other providing detail on the population present. The literature describing research undertaken with foreign microdata illustrates many examples of their integration with other data and the use of microdata and aggregate data together. The following section explores what has been achieved.

### *4.3 - Integration*

Aggregate and disaggregate data have been used together in two ways. The first, which can be called 'ordered consultation', sees the identification of a feature of interest in data at one level and its subsequent exploration using data at another. A feature observed in microdata, for example can be subsequently explored at an aggregate level. Clark et al. (1989) used aggregate data at a city level to search for a relationship between city size and life quality following an exploration of variations in income and household characteristics from the PUMS.

The second type of integration might be called 'simultaneous consultation' and Navaratil and Doyle (1977) provide a good example. Focusing on migration, they sought to enhance aggregate approaches to the study of migration decision making. Two models of migration were constructed. The first used entirely aggregate data on migrant and destination characteristics. The second model retained aggregate destination descriptions, but utilised data from the PUMS to classify individuals in terms of migration behaviour. Navaratil and Doyle revealed that a model which simultaneously considered destination characteristics and *individual* migrant attributes, identified as significant factors which the aggregate model missed or misinterpreted. The individual / aggregate model performed better than its aggregate / aggregate competitor.

Census microdata have also been integrated with other microdata not stemming from a census. Akbari (1989) for example, combined Canadian PUST with a series of data sets about differences in the consumption of public services by immigrants. His work allowed the Canadian government to assess whether it was making a positive net gain in terms of tax income, by encouraging immigration. It was found that once established, many groups of immigrants made no extra demands on the welfare state, and contributed taxes to the extent that the Canadian government actually made a profit from encouraging immigration. There are a multitude of other examples in which

the excellent basic economic and socio-demographic information of census microdata has been used to enhance more specific survey data.

### 4.3.1 - Commercial Attitudes to Census Microdata

A difference between the academic and commercial attitudes to integration can also be noted. Given the potential advantages of having fine scale socio-spatial information for marketing purposes (which were identified in previous chapters), one might expect the funds and expertise of commerce to have been applied to an integration of aggregate and disaggregate census data.

American industry however, seems loathe to use the PUMS at all and appears to have a low opinion of the census as a whole. "I know many professionals who regard the decennial census as a fairly stale compilation of statistics," comments Sawicki (1993). There may be several reasons for this, but uncertainty over budget allocations in the USA has introduced doubt over the maintenance of a census programme in its current form. Removing publicly available PUMS is seen as a potential cost cutting measure. Industry may therefore be reluctant to build and market systems based on a data source whose future is uncertain and over which they have relatively little control. A second reason is that the census microdata in particular, do not provide the type of information that geodemographic analysts require. Information which the microdata provide can be more easily obtained for a small area from the aggregate census release without the problems introduced by a small sample size.

Geodemographic analysts however, do integrate *other* forms of microdata with aggregate census data to enhance the performance of their systems. Data available at postcode or household level are integrated with aggregate data about the area of interest. Through examining the first names of residents in a particular street for example, a system called MONICA (CACI, 1992) can produce a profile of their likely ages and inferred attitudes and aspirations. Other systems record court appearances or credit ratings and integrate them with an ED profile constructed from aggregate information (Spa Marketing 1992). Like the work of Openshaw et al. (1985), this kind of integration is interesting because it illustrates different approaches to solving the problems in hand. However, like the Openshaw work it does not offer solutions to the task identified in section 2.11. This is because the microdata in use are already spatially referenced at a fine scale, thus effectively removing any possible technical comparison between the two tasks. Geodemography is based on a spatial unit, and integration takes place through finding information which relates either to a location *inside* the unit (e.g. an address within a postcode)

or relates to a unit which *encloses* that of interest (e.g. an ED enclosing a postcode).

Abroad and in commerce then, microdata have been extensively used, often in some form of integration with other (sometimes aggregate) data. It would seem however, that no one working with foreign data or in geodemography has directly attempted the kind of integration presented in chapter 2, despite some obvious interest in related issues. In order to proceed, a UK focus will be restored and other ways in which attempts have been made to deal with the difficulties of aggregate data will be explored.

## *4.4 - Microsimulation Techniques*

Before the SAR data were made available to the public, a number of techniques were developed to tackle the problem of only having aggregate data for tasks which required disaggregate information. Collectively known as *microsimulation*, these techniques allowed the user to simulate a disaggregate population and information about it which could conform to known aggregate constraints. Given aggregate data at ward level for example, microsimulation was able to provide 'synthetic' households and individuals whose collective characteristics matched those already known. In their summary of the potential for microsimulation in the analysis and enhancement of census data, Birkin and Clarke (1995) allude to its potential for providing information about a population described by aggregate statistics. "The application of microsimulation methods may be thought of as a direct attack on this (the aggregation) problem... Because there is no prior aggregation, model outputs can be tailored to the requirements of the user and not constrained by the aggregation process" (p. 369)

In order to understand the potential which microsimulation offers, the essence of its mechanism must be clear; an exploration of one, typical microsimulation follows.

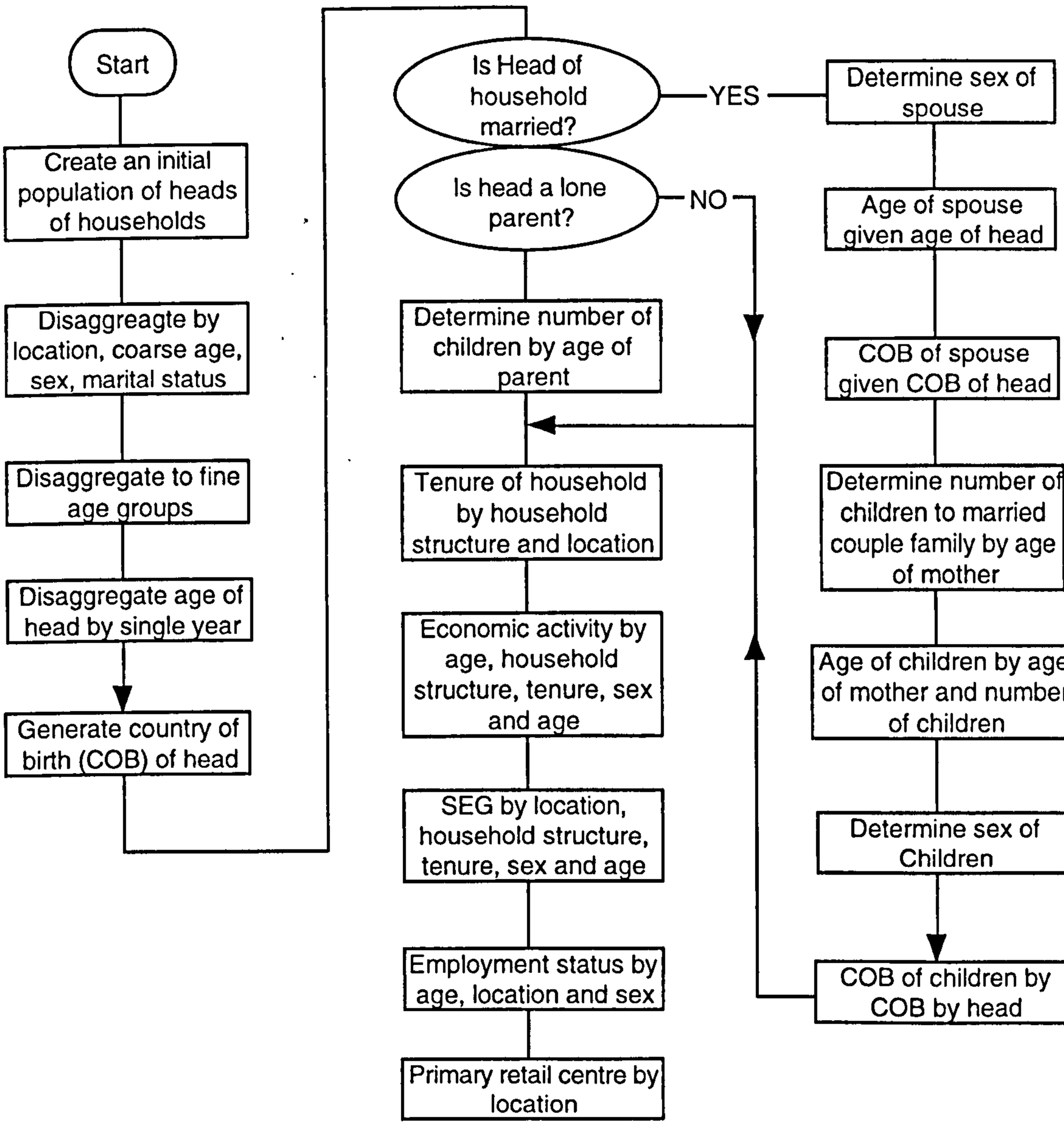
The work in question concerned the production of a microdata sample for the population of Leeds Metropolitan District (Birkin and Clark 1988). The synthetic population was to consist of individuals placed together in households, and was to conform as closely as possible to known aggregate parameters. The data to be simulated were effectively an individual, disaggregate set. The approach commenced with the generation of a synthetic sample of 50,000 heads of household in the form of simple arrays of values representing their characteristics and those of their family and household.

The first stage of the simulation was to assign each of these heads of household a location (within an ED), an age, a sex and a marital status. This was achieved through the employment of SAS table 02 (1981 tabulations) which cross classified these personal attributes for every ED. The allocation was carried out by constructing a very large cumulative frequency table of 1562 ED locations x 4 age groups x 3 marital status's.x 2 sexes. Samples were then drawn from this distribution using the Monte Carlo method to generate 50,000 heads of household distributed amongst the 1562 ED locations, each with their own basic attributes. Before proceeding with the generation of any further household features, the authors took the opportunity to assess the degree of sampling error involved in this process. Simulated age - sex distributions were aggregated for the entire city and compared to known distributions. They were found to be nearly identical and statistically significant well above the 99% confidence level.

The addition of other attributes to the household heads was carried out stage by stage. In each case, data were required that would yield a frequency distribution of the attribute to be added, based upon some or all of those attributes already assigned to the household. The data source was sometimes SAS tabulations, but more often other national survey data. For example, age of spouse (given the age of head of household) was estimated using a probability distribution from the 'household and family composition,' summaries published by ONS. Following the production of a probability distribution for the feature in question, Monte Carlo methods were employed to assign the developing households the attribute and iterative proportional fitting (Birkin and Clarke 1988, Johnston and Pattie 1993) constrained the generated population to fit the known parameters of the real population.

The sequence of attribute simulation is given in figure 4.1;

*Figure 4.1- Sequence of Attribute Simulation (Birkin & Clarke 1988)*




In some cases, the simulation of certain attributes was simplified by the presence of others. For example, Birkin and Clarke assumed only those heads of household which were married needed to be given partners. For those households to which a spouse had been allocated, children could be generated from a combination of national and small area survey data. National tables were employed to generate a probability distribution of family size in relation to mother's age. Generating family size in this way allowed a comparison of actual family sizes observed for an ED and synthesised family sizes. Expected family sizes could therefore be adjusted by weighting before allocation actually took place. This entire section of the simulation was, of course, only undertaken for those households which were allocated families.

The modelling of socio-economic group was carried out in a similar manner. The subsequent allocation of a specific occupation for each individual was regarded as more complex because of problems finding a suitable data source to generate the necessary probability distribution. (This work was pre-SAR data, which would have been an ideal source). After the simulation was completed, comparisons with known populations were undertaken to determine its quality. It was expected that the simulations would prove less accurate as the spatial scale at which testing took place became finer. This was because the characteristics were derived from a statistical model rather than directly from individuals. Initial sampling errors were thus likely to have been compounded and to become more significant at finer resolutions (Birkin and Clark 1988).

The output from this simulation process was a total number of households for each ED, together with characteristics of each household. These included the number of people it contained, the country of birth of the household head and the primary shopping location of the householder for non-food goods. In addition to this information, each household member was identified with a relationship to household head, an age, a sex, a marital status and an economic activity indicator.

Therefore, with some knowledge of the basic shape and size of a population in aggregate terms and some suitable data to inform an understanding of how household and individual attributes covary, microsimulation provides a means of synthesising a reliable model of population at the individual and household scale. A path is marked from aggregate and disaggregate data to an integrated product which offers plausible fine scale socio-spatial information. Since much of this work was carried out prior to the advent of the SAR, Birkin and Clark and others working in the same field, were faced with a number of difficulties which have subsequently been removed. As the SAR offer a complete, individual account of many of the characteristics which microsimulation was attempting to synthesise, it might have been used to generate the joint probability distributions necessary to create individuals. The quality of the SAR in terms of their coverage and their innate linkage to SAS and Local Base Statistics (LBS) would have greatly reduced the sample error propagation difficulties experienced with using smaller surveys as a data source. It would also have made validation of the simulated populations much easier by offering real, disaggregate households for comparison.

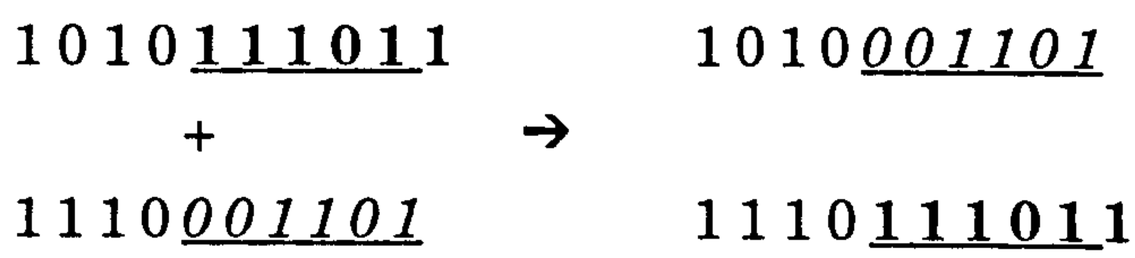
Although the advent of the SAR might have reduced the difficulty of a microsimulation approach to enhancing socio-spatial information, it has not removed the two main reasons why such a technique was not adopted for this research. Firstly, there is the difficulty of building a model which assigns a household new characteristics *based upon those with which it has already been assigned* (Williamson et al. 1993). Clearly this process has to have a beginning and an order of progress, but that order will profoundly affect the ultimate household form. Creating a relationship between two attributes is straightforward and a number of techniques from simple log linear models to more complex graphical Guassian models can be employed to do so. However, no comprehensive model of interdependencies can yet be built (either in mathematical or theoretical terms) when a large number of attributes are involved. Whilst Birkin and Clark are admirably open about the subjectiveness of their decision about attribute order, no account of its effect can be taken when evaluating the model. Secondly, the process is simply too complex to be efficient. It requires too great an investment in data and processing. Populations must be generated specifically for every ED under consideration and once started, the processing paths are themselves are specific because of the need to iterate until synthetic population parameters match known values. The remit of this research included a commitment to efficiency and a 'product' which might be applied to different areas with relative ease. Although the growth in computing power continues to reduce the importance of process efficiency, microsimulation cannot satisfy the remit of this research simply because it requires such large volumes of data from different sources.

Before moving on from microsimulation, there is one further piece of related research which is worthy of note. It is in fact an extension of the microsimulation principle, but rather than trying to *simulate* households and individuals it uses the SAR as a pool of households and individuals from which to re-populate small areas. The work was presented at a conference in 1993 and has now been accepted for publication (Williamson et. al. forthcoming).

### *4.5 - Genetic Algorithms*

Williamson et al. (1993) describe their attempts to select individuals and households from the SAR so that their aggregate characteristics match those given in the SAS for a particular ED. Central to their approach was the reduction of size of the task of finding the best 150 households with which to populate an ED from the 215,789 featured in the SAR data set as a whole.

Without a means to direct the search for suitable household combinations, the number of possible solutions to be tested exceeded a billion billion and thus any practical computational possibility. Williamson et. al. therefore tested an "intelligent" searching mechanism (p.7) known as a genetic algorithm (GA). GAs were originally developed in order to try and mimic evolution processes in nature (Holland 1975, Davis 1991). They utilise a list of values as a *chromosome* through which information is exchanged and passed from generation to generation just as DNA is re-combined and passed on during biological reproduction. By re-combining fragments of the 'fittest' chromosomes from each generation, the drive towards an optimum solution is maintained. When applied to the problem in hand each chromosome carried a list of household identifiers which reflected a possible combination of households to form an ED. Reproduction was carried out by combining two chromosomes so that the two 'parents' swapped parts of their sequences:




Other aspects of the evolutionary process were also mimicked, including mutation (in which a small possibility of incorrect transmission of information was introduced) and limiting parenthood to a small number of the fittest chromosomes for each generation. The authors acknowledged that by introducing these limitations a drive towards convergence was created (because with each generation the difference between parents and children decreased). There is an inherent risk that the solution offered might therefore not be the optimum, but they note that "the use of genetic algorithms implies a trade-off between solution optimality and search time" (p.8).

The fitness of a chromosome was established by aggregating the appropriate households by age, sex and marital status into a SAS look-alike format and then comparing the values with the real SAS (table 35). An index known as Total Absolute Error (TAE, Upton 1985) was utilised to asses the quality of the match. Had the two sets of figures matched exactly, TAE would have been 0 indicating that on age, sex and marital status the population drawn from the SAR was identical to the real ED population. If however, the SAR population was absolutely different from the real one, the maximum error would have been twice the total count for the SAS table in question (964 in the example presented).

Williamson et. al. report that their best result was a TAE of 151, however the results varied considerably from run to run. Each run took about 8 hours. Somewhat discouraged by the poor results, the team abandoned the GA system itself. They then attempted to mimic some of its features but used a more conventional 'hill climbing' approach as the basis for the new technique. This yielded a substantially reduced TAE of 28, in a much reduced processing time, however the technique does not appear to have been developed any further.

Although the research is intriguing, since it is the only work found which has attempted to achieve the aims of *this* research through an integration of SAR and SAS data, it does not provide a way forward. There are two main concerns; firstly, the technique attempts to select SAR records for an ED on the basis of just three attributes. Age, sex and marital status. Although this limited number was almost certainly selected for the purposes of technical development, the fact that such poor results were achieved with these variables does not give much hope for the extension of the technique to cover more attributes. Secondly, as with more conventional microsimulation techniques, the GA approach required a separate simulation for each ED analysed. As discussed in section 4.4, this research aimed to provide a more efficient solution which might avoid the need to undertake the entire simulation process for every ED of interest.

The Williamson et. al. approach was therefore not of direct use, but as with the Openshaw et. al. work (1985) the fact that other researchers had been interested in achieving the type of enhancement to socio-spatial data was encouraging. Their failure also confirmed the originality of this research and its potential value if success could be achieved.

## *4.6 - Data Fusion*

Given that microsimulation did not provide a means of integrating the SAR and SAS data which satisfied the particular research aims of this work, there was a need to explore other techniques through which the integration might be achieved. It must be recognised that in the reasonably narrow field of census data analysis, options left open for exploration were increasingly difficult to find. One technique offered some hope. Like geodemography, data fusion is a technique which bridges the gap between academia and commerce. It has found a niche in media research roles:

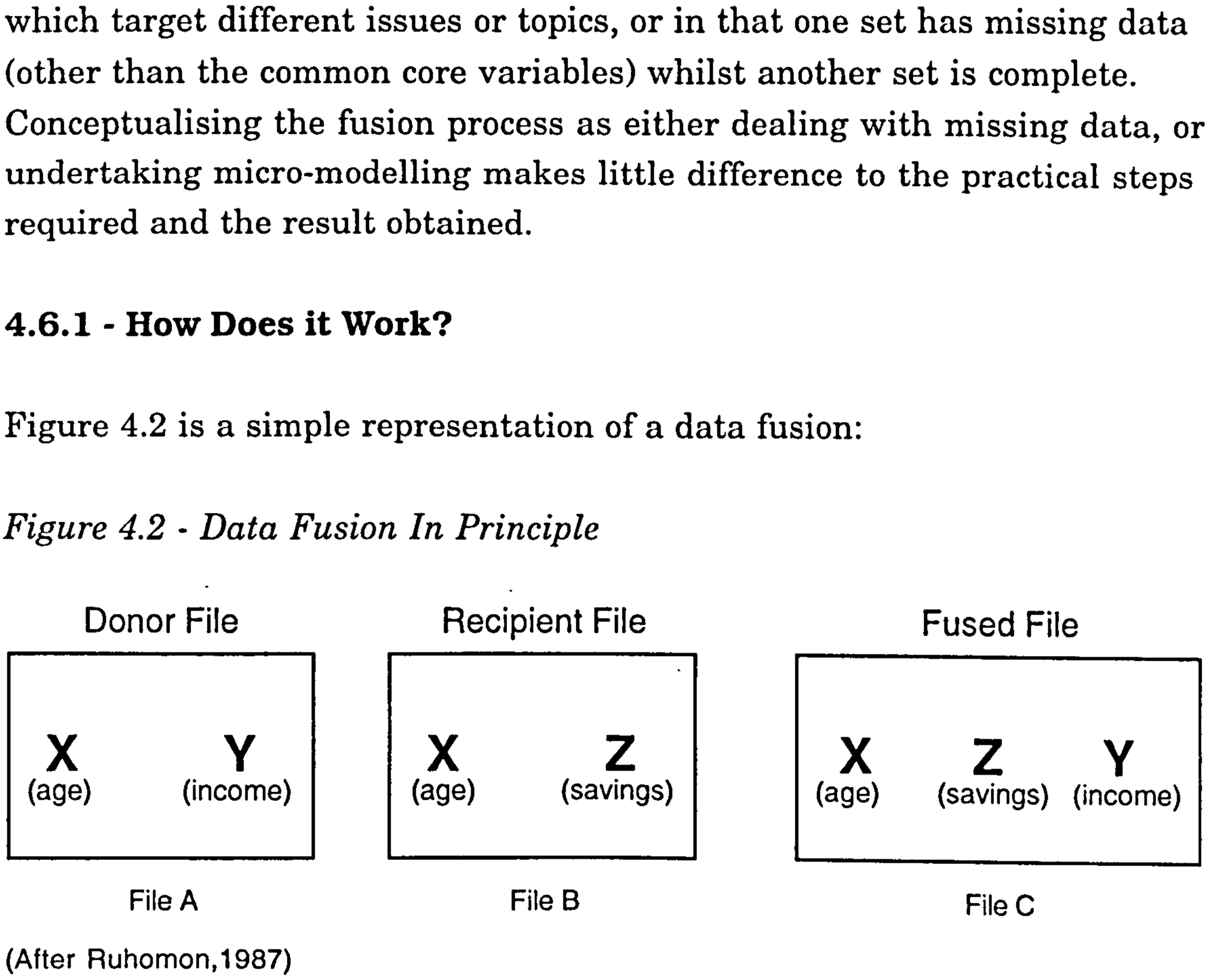
Data fusion appears in most of the literature covered (of which the small quantity emphasises its relative novelty) as a means to handle missing data (Baker et al. 1989). However, it is also described as a parameter-free micro modelling approach (Antoine and Santini 1987) because data fusion can be used in situations where there is a priori knowledge of the data that will not be available, (i.e. because that information was not collected in a survey). Data fusion is not the same as microsimulation because *entire sets of responses* are simply being donated in one step, from those for whom the desired information is held to those for whom it is not held. Response sets are not manufactured or suggested based on other sample distributions. Thus, the purpose of the fusion process is to take two sets of responses or attributes and somehow undertake a procedure which results in every individual having a complete set of information, regardless of whether they actually gave the complete set, or even whether the were given the opportunity to respond to the complete set. In simple terms, " the ideal world that fusion is trying to replicate is single source data where each respondent provides all that data under consideration." (Baker et. al. 1989)

The technique is suited to a situation in which two or more sets of survey response exist. A number of common *core* questions must be present in both data sets. The sets may differ however, either in that they contain questions which target different issues or topics, or in that one set has missing data (other than the common core variables) whilst another set is complete. Conceptualising the fusion process as either dealing with missing data, or undertaking micro-modelling makes little difference to the practical steps required and the result obtained.

### 4.6.1 - How Does it Work?

Figure 4.2 is a simple representation of a data fusion:

*Figure 4.2 - Data Fusion In Principle*




(After Ruhomon,1987)

Data fusion is structured by the relationship of one set of responses to the other (terms in *italics* are those taken from Baker et al. 1989). Fusion occurs between the *donor* survey which contains information (income in this case) to be transferred and the *recipient* survey (Baker et al. 1989). There must be some *common* variables between the donor and recipient data sets (age in this case). These will form the basis for deciding which respondents in the donor survey will pass data to which respondents in the recipient survey. The relationships between the common variables and those to be donated must thus be established and strong. The fusion process requires a test of similarity between the common variables of the donors and recipients in order to find the best pairings and it is the *matching algorithm* that assesses these similarities. The two individual respondents are then joined together if they are a suitable match, so that a response from the donating individual is acquired by the recipient individual. The degree of closeness between respondents can vary from exact duplication of all the common variables, to duplication of a defined sub-set of common variables (known as *critical variables*) with more approximate matching in others. The measure of similarity required in order to proceed with the matching process is conceptualised in terms of a multivariate distance; "Basically, the fusion process could be thought of as the global minimisation of the average distance between linked respondents." (Antoine and Santini 1987). The intricacies over whether donor respondents can be re-used or not, and what constitutes a minimum acceptable distance vary from fusion to fusion, however in all cases the principle is the same. In the data fusion literature, the matching of respondents is often discussed using the analogy of marriage.

Baker et al. (1989) describe the technical intricacies of an actual fusion, and the problems encountered. Two issues warrant attention. Firstly, in their attempts to define the algorithmic rules for the fusion process, they divided the critical variables into strata, within which 'marriage' would be allowed or not allowed. This division was undertaken without careful analysis of the data they were dealing with and the groupings of age and of class that they employed were ultimately shown to have limited the success of the fusion.

Secondly, in order to discourage a 'polygamy', when a marriage occurred, a weight was added to the donor involved which effectively increased its distance from other potential recipients. However, poor control of this weighting served only to increase the number inappropriate pairings, since donors with common features were being re-used and thus quickly taken out of the

marriage market. These two points are fine illustrations of the need to remain aware of how the technical procedure affects the end results.

### 4.6.2 - What Can it be Used For and How Good Is It?

In all the literature examined, data fusion has been employed to provide more comprehensive survey information. Despite Richardson's (1990) claims that "fusion does not work", (p. 472) (grounded in a rather a weak mathematical argument), success is claimed by most experiments. How success is actually defined however, is not always clear. In most exploratory research the success of a fusion is assessed by testing its power to recreate artificially removed data (i.e. fusion is carried out to attempt the replication of known values). In trying to quantify success of a fusion the results have been considered both at aggregate and individual levels. The results of Bronner (1983) are typical.

He employed the Dutch National Readership survey which consisted of a one hour face to face interview collecting information on demographics, reading habits, consumption behaviour etc. This was followed by an extensive psychographic questionnaire covering activities, interests, opinions, attitudes and self-concept. A total of 5,537 individuals completed the media/product interview *and* the psychographic questions, whilst 1,049 individuals only participated in the interview. The psychographic responses of the 1,049 respondents were thus to be simulated using data fusion. Following a stepwise regression to assess which of the media/product data were best suited to be employed as common variables, (there were 21), the fusion took place.

Comparison with the 'real,' data was carried out at aggregate level.

*Table 4.2 - Results From Bronner's (1983) Data Fusion Experiment*

| Proportion of Significant relations between the real and simulated scores for each item group in the psychographic questionnaire | |
|---|---|
| Item Group | Percentage of simulated responses with significant relation to truth |
| TOTAL | 85 % |
| Attitudes | 82 % |
| Interests | 93 % |
| Music, T.V. | 100 % |
| Behaviour | 89 % |
| Self-Concept | 77 % |

The table shows that fusion reproduced the distributions of responses in each item group relatively well. Results at an individual level were more revealing. with approximately two thirds of the responses correctly simulated. The distribution of the error was as shown.

*Table 4.3 - Distribution of Error in Bronner's (1983) Data Fusion Experiment*

| Number of items over which true and simulated scores differed | Number of Individuals receiving simulated scores |
|---|---|
| < 25 % | 13% |
| 25 - 37.5 % | 47% |
| 37.5 - 50 % | 32% |
| > 50 % | 8% |

Bronner concludes that these error margins are acceptable for his purposes. He wished to employ the fusion for aggregation and cross tabulation purposes, and thus the individual level variable - variable relations were of little importance. The overwhelming conclusions of all proponents of data fusion are that fusion is very good at simulating the marginal values of a data set, for predicting global aspects of the data but that it is not an acceptable method to project or predict specific individual behaviour, (Antoine and Santini 1987 p.) or specific individual attribute sets.

### 4.6.3 - Why is Data Fusion Not Appropriate For This Research?

Whilst data fusion is a fascinating concept, it is not a technique that provided a way forward for this research. This decision was made for several reasons.

Firstly, if the research task is considered from a fusionist's perspective it appears thus.

A SAS tabulation exists for each ED, representing a recipient data set. The SAR data represent the donor file, with those attributes that also appear in the SAS as the common and critical variables. Items of information about households, beyond the two or three crosstabulated in the SAS, represent the 'missing' data. It is clear that data fusion requires a number of discriminatory common variables, but the SAS only provide a maximum of three items for an individual. These three are simple demographic descriptors, and thus unlikely to be highly discriminatory. All the examples of data fusion examined have used at least 14 common variables, some have used almost double that number. It is certain that the SAS do not provide enough information on the populations within each ED to make fusion viable option. In addition to this difficulty lies the fact that each ED would have to be considered as a separate recipient file. In this respect the same accusation of inefficiency that was levelled at microsimulation can be levelled at data fusion. A more computationally efficient of dealing with multiple EDs is needed. Data fusion is thus rejected as a means to integrating the SAR and SAS data sets.

## *4.7 - New Ways Forward*

Given that data fusion does not provide the path to an integration of the SAR and SAS data it is certain that a new technique will be needed. Although the literature has not revealed a specific way forward towards the research aims, there are elements in the work covered which may prove useful. At least, in clarifying the shortcomings of these techniques, a more precise awareness of what must be achieved becomes available.

## *4.8 - Conclusions and Discussion*

This chapter has reviewed the research carried out using microdata in other countries. Many examples have been found of the integration of aggregate and disaggregate data, but none in which the data are themselves integrated in the manner discussed in chapter 2. The simultaneous, or ordered consultation approaches, (in which the two different types of data are used to provide contrasting or complementary insights to social phenomena) are interesting but entirely unsuited to the task of revealing more about social variation at fine spatial scales.

The Openshaw et. al. (1985) work provided an example of a highly desirable output. The proportional breakdown of household types for regions (table 4.1) fits precisely with second 'desirable enhancement' outlined in table 2.3 although the team utilised a unique data set to produced this analysis. Nonetheless, the format of the results provides an excellent model for this research to aim for, if an alternative route to its production can be found.

In terms of the techniques which have been reviewed, data fusion is not a suitable candidate for development. This is because it deals with data sets at the same level, (i.e. disaggregate survey data) rather than between levels and because the SAS and SAR data are not structured in a way that would make data fusion viable. Microsimulation certainly does provide information about social variation at fine spatial scales, satisfying the third 'desirable enhancement' (table 2.3). With the advent of the SAR, the volumes of data required for the simulation could be dramatically reduced. The SAR would yield probability distributions for many of the variables required. The advent of the SAR does not remove the intensity of the simulation process however. The disaggregate population for each ED must still be simulated separately. Ideally, a technique could be found (or created) that would establish relationships between the aggregate SAS and disaggregate microdata which were instantly applicable to all EDs. One conceivable means of achieving this would be to undertake a microsimulation for a large *sample* of EDs within an area and attempt to derive generalised relationships between the aggregate appearance of an ED and a 'typical' disaggregate population. Rather than provide a complete disaggregate population for every ED, such an analysis might present a table akin to that used by Openshaw et. al. (1985). The ED would be described as containing household types, rather than individual households. This would represent the achievement of enhancement 2 (table 2.3) but would not require the full microsimulation process for every ED. The mechanism for deriving these relationships is however, not clear.

Two conclusions of significance may be drawn from all that has been reviewed here. The first is that no existing technique will achieve all of the desired aims. This seems to be primarily because the task of generating disaggregate data for small areas is too intensive, and the data with which to attempt an Openshaw-like analysis do not exist. The second is that, an Openshaw-like analysis seems highly desirable. If a new technique has to be developed, it would be enormously useful to have an end product clearly defined. This would act to focus the theoretical and technical innovations which must be

undertaken. Adopting the Openshaw-like analysis provides the following information and raises the following questions:

- The new technique will seek to satisfy enhancement 2 (table 2.3) i.e. A measure of social mix within an area for which aggregate data are available, together with details on the nature of that social mix.

- The new technique may have to conceptualise the households in the SAR in terms of their 'class' or 'type'. The household type will become the base unit for the analysis, rather than the household or individual.

- The derivation of household type is certainly possible using the SAR data, however the relationship between the aggregate SAS and the presence of different household types is unclear. Determining this relationship now becomes the key task for the research.

The derivation of that relationship is explored, in theoretical terms, in chapter 5.

# Chapter 5

## Chapter Overview

This chapter outlines in principle, the path from raw data to a new model of social variation at a fine spatial scale. Social theory is applied to understand aspects of Small Area Statistics (SAS) data and a reconceptualisation of their nature is then presented. Re-thinking SAS data allows an analogy to be drawn between problems that Earth Observation (EO) researchers face in trying to classify land cover types from satellite images and the difficulty of relying on aggregate data sources such as the SAS for areal classification. A number of solutions to the EO problems are reviewed. A technique which models variation at a scale beneath that at which the data have been collected is highlighted as being potentially useful in the social geography context. The function of neural networks is explained, and their role in this technique is subsequently made clear. In order to attempt the technique, a classification of household types is required and the Samples of Anonymised Records (SAR) are proposed as an ideal data source with which to carry this out. The role of this classification in synthesising data required by the neural network is explained. Finally, the principles and techniques discussed are arranged in sequence, priming the reader for the following, technically based chapters.

## 5.0 - Introduction

Chapter 4 illustrated that other work had considered problems similar to those in this research, but that the approaches adopted had a number of shortcomings. It also introduced the kind of analysis which this research would ideally like to offer, but emphasised that the data required to provide it for the UK were not available. It was therefore necessary to develop an original approach to enhancing the level of socio-spatial information available from census data. Chapter 5 will describe the process through which an integration of SAR and SAS data was achieved, ultimately resulting in the model introduced in figure 2.8 and section 2.12. This account will commence with an explanation of how the social theory explored in chapter 3 was used to consider the nature of the data available and how they related to the phenomena of interest:

## 5.1 - Theory and Empirical Information

The reader will recall that the urban theory accepted in chapter 3, described the city as being composed of a number of socially homogenous

neighbourhoods, distinct from each other in both their spatial location and in their social character. Pause for a moment in order to consider a city and ask whether it really fits this description? Even attempting this ought to be difficult because the terminology used in the description raises several qualifying questions. How are neighbourhoods, social character and socially homogenous to be defined? A robust definition of neighbourhood, the social unit at which theory can freely operate, is in fact extremely difficult to provide. *Spatial* definitions are fraught with difficulty since individuals do not interact with each other in a manner limited by distance or an areal boundary (Bridge 1994, Schubert 1994). *Social* definitions are also difficult, since the social networks which physical neighbours might inhabit could differ completely in terms of both the individuals involved and the distances over which social interactions occur (Bell and Newby 1978). It is hard to be precise about the distance at which social interactions cease to become 'neighbourly'.

Attempting to apply the relevant social theory at the empirical level is therefore rather difficult. The most complete and detailed data set about city inhabitants (the SAS) is set in a rigid framework which is entirely unsuitable for dealing with abstract concepts such as 'the neighbourhood'. Consider the spatial aspects of the SAS data set; the data are structured within a spatial framework which divides the city into a number of small units (the EDs), the construction of which is essentially pragmatic. The boundaries are constructed so that it is not an unreasonable task for an enumerator to visit the properties for which they are responsible. The EDs must be large enough to contain a sufficient quantity of households, not be divided by any obstacles to collection, (such as a river, motorway or railway line) and must be bounded by recognisable features (Morphet 1993). In short, the construction of ED boundaries is not designed to be congruent with fluctuations in household type on the ground. Morphet's work (1993) shows that in a very limited number of cases the ED boundaries will coincide with social boundaries. Large tower blocks, estate edges or small suburban developments may all provide natural boundary lines for EDs, which might subsequently enclose reasonably homogenous and distinct household types. In other locations however, a convenient boundary placement might bisect a distinct and socially homogenous neighbourhood, dividing its households into separate EDs. Morphet's analysis of SAS data for EDs in Newcastle clearly showed that they did not enclose socially homogeneous areas.

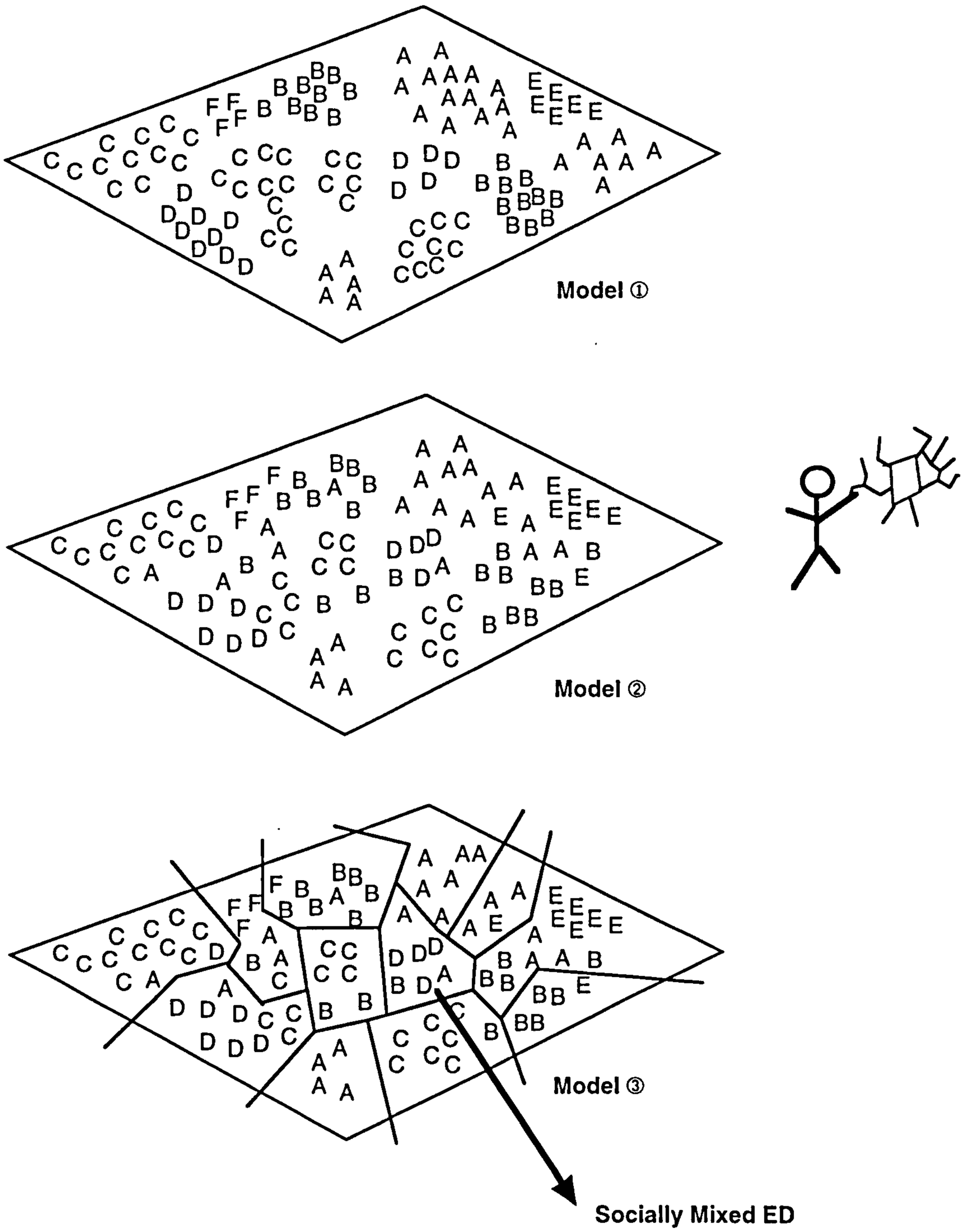
In addition to the mis-match between the rigid spatial framework of SAS data and the abstract neighbourhood concept of urban theorists, there is a

difference in terms of the social elements on which each is based. The SAS data consist of a limited set of categorised variables which describe physical and material characteristics. Urban theory freely considers a multitude of characteristics, feelings, opinions and achievements in its discussions of the 'neighbourhood'. There is no way in which the variables in the SAS can possibly provide an account of the differences between neighbourhoods and their social groups in these intangible terms. It is clear then, that there is one spatial unit and information system in the SAS, and a different social unit system and type of information on which the relevant theories are based. How can urban theory be used in order to understand the precise nature of SAS data and ED boundaries?

## *5.2 - Theory, Reality and a Way Forward*

According to the theoretical understanding gained in chapter 3, the city can be thought of as a social mosaic of distinct and homogenous 'neighbourhoods'. Figure 5.1, model ①, illustrates this situation by representing households of differing social type with letters of the alphabet, (for the purposes of this illustration, the nature of the social groups is not important). Model ② represents a more realistic situation in which some socially distinct groups remain, but the *spatial* pattern is more confused and less easily de-limited. This pattern results from the influence of fixed city entities around which the social processes leading to the residential location of different groups must flow. These entities include factors such as the availability of property, its type and the type tenancy offered (Knox 1982). In figure 5.1, a boundary system is applied to this urban mosaic. The results are seen in model ③. Some EDs have captured quite homogenous groups of households, whilst others (as illustrated) have captured a mixture.

*Figure 5.1- Social Groups In the City and Their Relationship With ED Boundaries*




A useful conception of the SAS can be derived by using theoretical and technical understanding together. A grasp of how the spatial framework of SAS data interacts with the theoretical socio-spatial structures allows a reconceptualisation of EDs as enclosing households from a variety of social groups. Since the SAS are derived by the aggregation of census returns from an ED's population, it can be assumed that they are the result of carrying out a census on a specific mixture of social groups (N.B. although the term

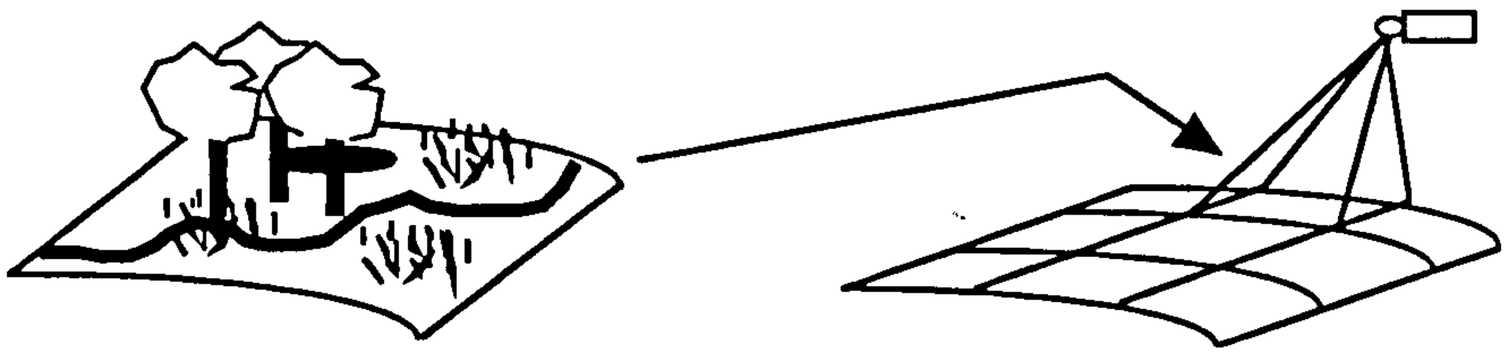
*mixture* is employed here, it is taken to include those situations in which the population is very 'pure'). If this conception holds firm, the SAS have information locked inside them about the proportional mix of differing social groups present within the ED. It is that information which this research has sought to exploit. The aggregate nature of the SAS masks specific household to household and individual to individual variation amongst the ED population. However if the typical household characteristics of differing social groups can be established and the relative presence or absence of those groups in the ED determined, an enhanced account of what kinds of people live in a specific ED (and the relative proportions in which they are present) can be produced. Thinking about the research task in this way alters its nature. If the SAS data are the result of a mixture of household types (which may be described in the SAR), obtaining information about that mixture becomes an issue of data *analysis* rather than one which tries to physically join two differing types of data together.

Having reconceptualised the data and problem in this way, it is possible to find an analogous situation in a related field of research. Earth Observation (EO) has a very similar set of difficulties with the classification of land cover from satellite images:

## *5.3 - The Earth Observation Analogy*

Satellites view the Earth and record the reflectance of electromagnetic radiation from its surface. This spectral reflectance can be used to determine a great deal about the type and quality of material covering the Earth's surface (Curran 1985). The satellites record their data via a grid or 'pixel' system in which the cell or pixel size is determined by the nature of the recording instrument. Usually, a single value of electromagnetic reflectance for each waveband (concentrating on different areas of the spectrum) is recorded for the pixel. As a result of the analysis of these waveband values and the application of knowledge about the spectral response of various kinds of surface, the pixel may be assigned to the most likely class of land cover. In some cases however, this single classification approach is far from ideal (see figure 5.2).

*Figure 5.2 - A Mixed Pixel*



Area of mixed ground cover types - trees, water and grass. This area is one pixel in size

Earth's Surface divided into pixels

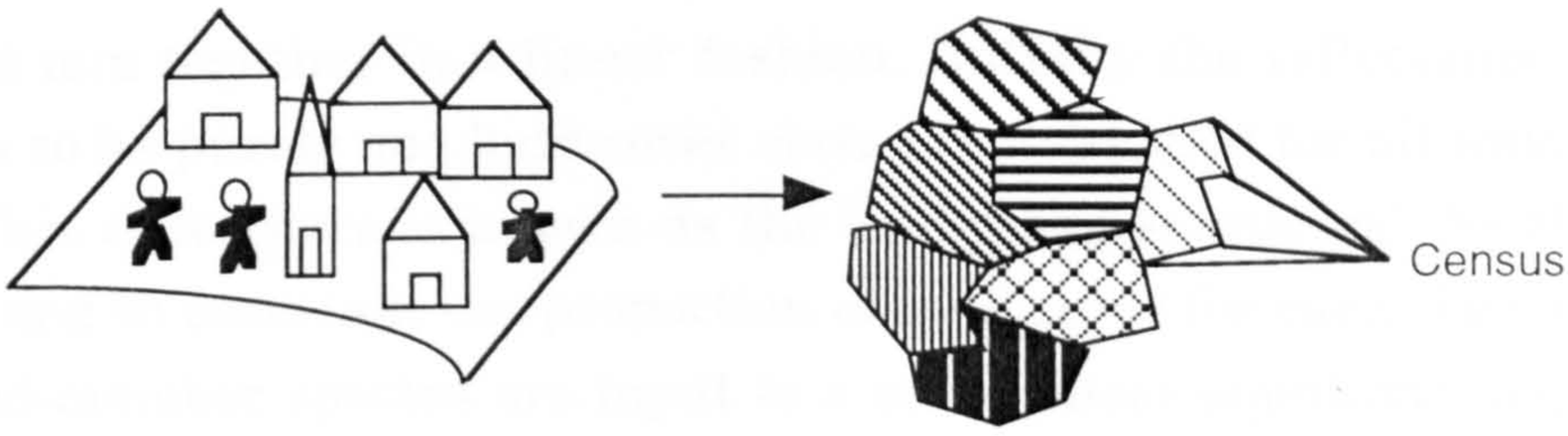
One numerical value per reflectance waveband for each pixel:

Band 1 - 25
Band 2 - 43
Band 3 - 56
Band 4 - 12
Band 5 - 69

In the diagram, a pixel which contains a mixture of land cover types is sensed by the satellite. Here, the scale of spatial variation on the ground is finer than the sampling scale imposed by the pixel structure. Thus, the variation in reflectance which exists within the pixel is lost when single numerical values for each waveband are recorded. These values are the result of the collective reflectance of the pixel. For some years, EO researchers classified each pixel according to its most likely or dominant land cover type, knowing that in doing so they were disregarding ground variation. However, technical and conceptual developments within EO have now moved pixel classification into a new realm of complexity. Today, the "objective is not to assign a pixel to a single class, but to a number of classes, where the sum of the proportions is one" (Atkinson and Cutler, 1995, p.13). Rather than offer an analysis which states that a pixel is 'predominantly' grassland, the aim is to offer an analysis which states the probable proportions of each land cover class present (e.g. sixty percent grassland, thirty percent forest and ten percent water). This kind of analysis will be referred to as *unmixing*, since its aim is to unmix the data into a model of its constituent parts.

The EO problem is directly analogous to the difficulty of dealing with social variation within EDs when limited to aggregate data sets. If a pixel is substituted with an ED and spectral response data with SAS from the census, the problem looks like this:

*Figure 5.3 - A Mixed ED*




An ED boundary encloses a socially mixed area

Small Area Statistics produced for each ED. Single values for a range of collective characteristics:

Variable1, 25
Variable2, 65
Variable3, 12
Variable 4, 13
Variable 5, 67

We have conceptualised an ED as composed of a variety of social groups, just as a pixel might be filled with a variety of land cover types. Like EO researchers, social scientists have classified EDs according to their most likely or dominant social group, knowing that they were disregarding variation within the ED. This process is the essence of geodemography (see 2.8). Unlike EO researchers, social scientists have not moved on to attempting a proportional classification of their small areas.

## *5.4 - Earth Observation Solutions*

Researchers in EO have a number of techniques which can provide potential solutions to the mixed pixel problem. Each solution is best suited to a slightly different set of circumstances. Given the fact that there are direct analogies between the EO situation and this research question, it was reasonable to review their solutions and determine if there was a method which had the potential to be adapted for use in a social geography context. The solutions can be categorised into three types of approach, linear mixture models, fuzzy classification and neural networks. In a recent study, Atkinson and Cutler (1995) compared these three techniques on a variety of axes to determine their performance characteristics:

### 5.4.1 - Mixture models

In mixture modelling, a set of simultaneous linear equations are solved to yield a proportional mixture of classes present in the pixel (Gillespie 1992, Foody and Cox 1994). The known values in the equations are the pixel

reflectance values for each waveband, the (initially) unknown values are the proportions of land cover for each class within the pixel. The assumption behind models of this kind is that the reflectance from different ground cover classes mix together in a linear fashion. Initially the reflectance from an area known to be purely one land cover class is established for all wavebands in use. This reflectance is known as the 'end-member spectra' (Settle and Drake 1993) and to ascertain the proportion of land cover for each class in a pixel, all the end-member spectra are input to a set of linear equations, together with the actual waveband values for the pixel in question. A solution is obtained using least squares approximation. The major limitation of this approach is that the number of wavebands must always be equal to, or exceed the number of ground cover classes of interest. This factor was unlikely to limit the adaptation of the approach for social geography since more variables (analogous to wavebands) were likely to be used than the number of groups desired. A second limitation is that an entire set of equations must be solved for each pixel in the study area, making the approach computationally intensive. In terms of accuracy, Atkinson and Cutler found this approach to be better than fuzzy classification, but not as good as a neural network. It was also shown to be highly sensitive to the quality of end-member spectra, or in simple terms, to the quality of the class definition in terms of its reflectance. This means that obtaining a very clear definition of the classes of interest becomes highly important and can be a major operational drawback for the EO researchers. Given the difficulty of producing a limited number of very strictly defined social groups, such a drawback might also affect an application of this technique to census data.

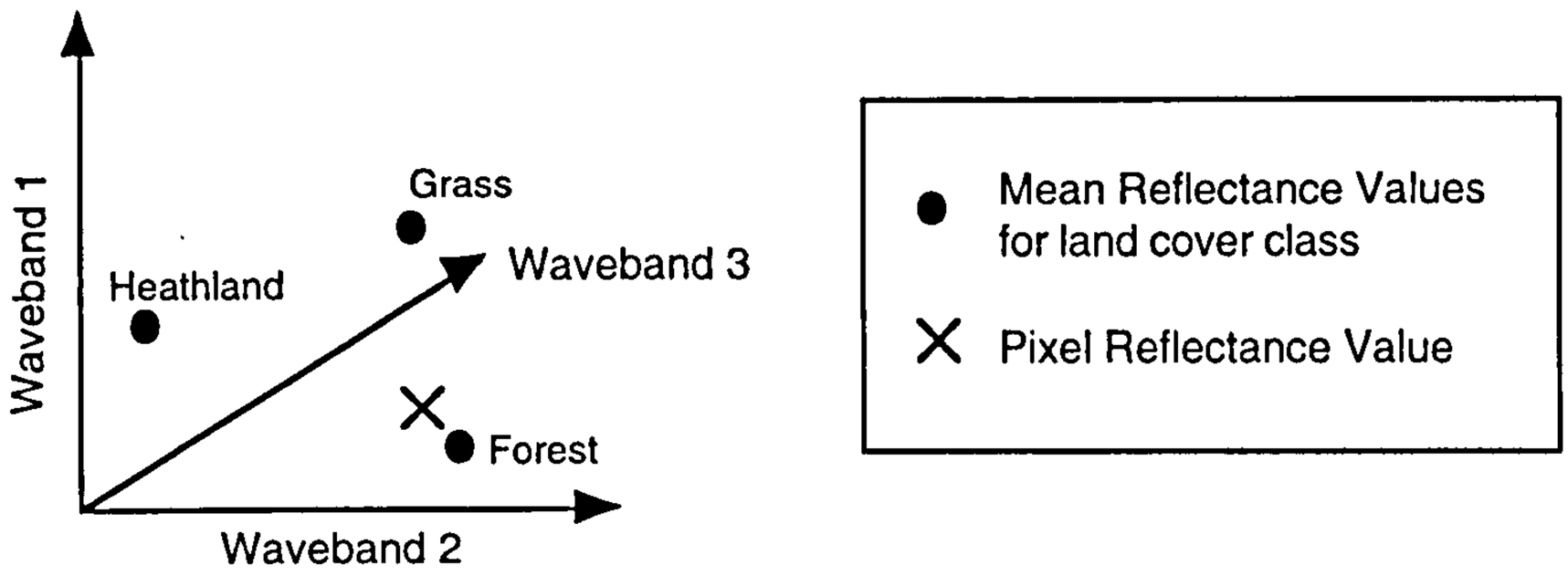
### 5.4.2 - Fuzzy classification

In fuzzy classification, a pixel is proportionally assigned to a number of ground cover classes (Kent and Mardia 1986, Gopal and Woodcock 1994). The *extent* to which the pixel is assigned to each class can be related to the presence of that ground cover class inside the pixel. Fuzzy classification is merely an extension of the more common 'hard' classification of pixels. Hard classification is achieved simply by plotting the pixel reflectance values in a multi-dimensional 'space' define by the wavebands being used. If mean waveband reflectance is known for each land cover class (effectively the end-member spectra), the pixel can be assigned to the cover class it has the highest probability of belonging to, thus achieving a pixel classification.

In fuzzy classification techniques, rather than finding the cover class to which the pixel is most likely to belong, the distance of the pixel values to a

variety of cover class means can be computed. These distances equate to the likelihood of pixel membership to each class. In simple terms, the closer a pixel reflectance value is to a class mean, the greater the proportion of the pixel land cover that is assigned to that class (Kent and Mardia 1986) (see figure 5.4). In testing, the fuzzy classification achieved the least accurate results, although it was relatively easy to carry out.

*Figure 5.4 - Fuzzy Classification*




### 5.4.3 - Neural network approaches

Here, a neural network is trained to asses the reflectance information from a pixel and offer a model of the proportional values of land cover classes present. The complexities of neural networks will be dealt with below, since this is the approach which seemed to offer the features required for an attempt at unmixing SAS data. The network approach utilises the same principle as the other two techniques – that the waveband information is the result of mixing of spectral responses from land cover types. However, the network learns to 'recognise' the mixture present from repeated exposure to examples of mixed spectra together with the actual proportional values which account for the spectra. The network has an ability to *generalise* its learning and thus accurately unmix pixel data which it has not previously seen.

Given a knowledge of social group schemes gained from investigations outlined in chapter 3, it seemed likely that an ability to deal with between ten and twenty social classes would be needed and Atkinson and Cutler suggest that the neural network approach would cope best with this demand. They do

note however, that a neural network solution was not the simplest or easiest means of unmixing. It was the potential (albeit an untried one) for a neural network approach to offer a high standard of unmixing for a large number of classes that confirmed it as the approach to experimentally apply to the problem in hand. The reader will therefore require a brief explanation of the concepts and technologies behind an artificial neural network (ANN), as it is more correctly known.

## *5.5 - Artificial Neural Networks*

Contemporary applications of ANNs are found in fields from image processing to bacterial classification, machinery control to forensic science. It is their ability to be artificially intelligent or perhaps more precisely to 'learn', that places them at the leading edge of computer development today. ANNs can make excellent pattern recognition systems, seeing pattern through noisy data (Gurney 1995). It is this ability to detect pattern which is exploited in their application to mixed data problems.

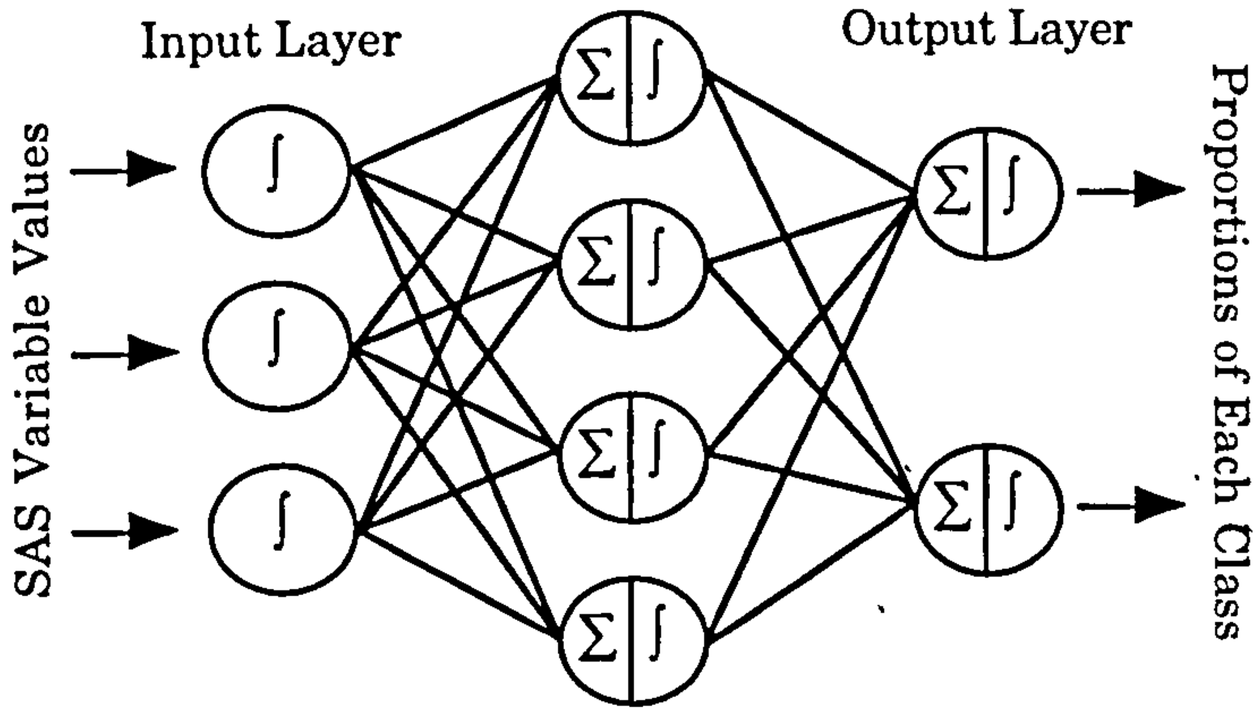
### 5.5.1 - Components and Mechanisms

The basic components and principles of an ANN are very simple. The network consists of an arrangement of *nodes* linked together in an *architecture*[3] (Aleksander and Morton 1990). Each node has a small amount of memory in order that it may 'remember' the value that last passed through it. The connections between nodes are modified by an adjustable *weight* and it is the ability of the network to manipulate the weights between nodes that makes learning possible. The way an ANN works is perhaps best illustrated by taking the reader through a simplified version of the *feed-forward* ANN which was utilised for this research:

---

[3]The reader should note that the architecture of an ANN has no physical component. An ANN is a piece of software which will run on any computer of sufficient specification.

*Figure 5.5 - ANN Architecture*




(i) There is an input node for *each category* of every SAS variable included in the analysis. Together these constitute the input layer. The role of this part of the net is simply to receive the incoming data and pass them into the ANN.

(ii) An input value is passed from the input node to every other node in the hidden layer along weighted connections. The initial weights of the connections are randomly determined (constrained to between +1 and -1) although in other types of ANN application these weights can be initialised in a number of alternative ways. The input values are adjusted by the weight and arrive at the hidden nodes as new values.

(iii) Each hidden node sums all the weighted signals that it receives and subjects that summed value to an activation function. The purpose of this function may be to constrain the value that leaves the node to within a certain limit, or transform the value in some way. The function which sums the inputs to the node is known as the propagation rule and can be defined thus:

$$net_j = \sum_i w_{ij} o_i$$

where $net_j$ is the value entering the node, $w_{ij}$ is the weight of the connection from node $i$ to node $j$, and $o_i$ is the output of unit $i$. Although the activation function can vary, the most common is the sigmoid function (Schalkoff 1992):

$$o_j = 1 / 1 + e^{-\lambda net_j}$$

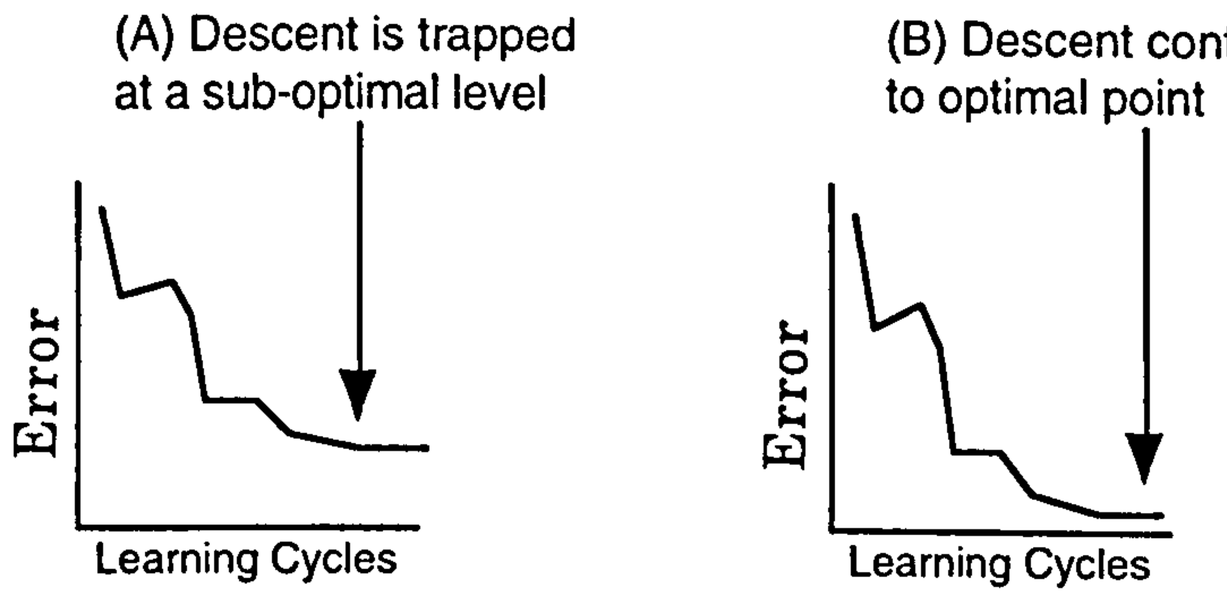
(iv) The adjusted signal leaves the hidden layer and is passed along another weighted connection to the output layer. Each output node sums the signals it receives and subjects them to another activation function before releasing a

single value as an output. In this case, if the input data has been SAS for a particular ED, the output would represent the proportion of a particular class present.

### 5.5.2 - Learning

Having seen how a value is passed through the network, there now follows a simple description of how this type of ANN learns to detect patterns. In the process known as training, the ANN is presented with a data set for which the desired output values are known. This set is called the *training data*. The desired outputs are presented to the net as *target values*. Following an initial random setting of link weights, the data are passed through the ANN and a comparison of the resulting value for each output node is made with the target value. If (as is likely) the values are not the same, a wave of weight adjustment is passed back through the network in order to try and reduce the difference between output and target values (here the node's capacity to remember what value they last dealt with is utilised). Once the weight adjustment is complete, the data are passed forward through the network again and the comparison is repeated, followed by any necessary adjustments as before. This 'backpropagation' of error, and subsequent re-try is repeated many times (Rumelhart et. al. 1986, Schalkoff 1992). Whilst it continues, the user is presented with numerical (and possibly graphic) information about the extent of the error which the ANN is currently making. The actual algorithm used to reduce the error generated on each pass of the data may vary, although it will usually employ *gradient descent* and the resultant learning curve is ideally an exponential decay (Beale and Jackson 1990). Some problems however, can occur when using gradient descent; the network can get stuck in a sub optimal solution (as shown in figure 5.6).

*Figure 5.6 - Sub-Optimal and Optimal Learning*




Graph (A) illustrates a situation in which the ANN has been unable to reduce error any further, but this is not an optimal solution because the error is not as low as it could be. The preferable situation, shown in (B), is achieved

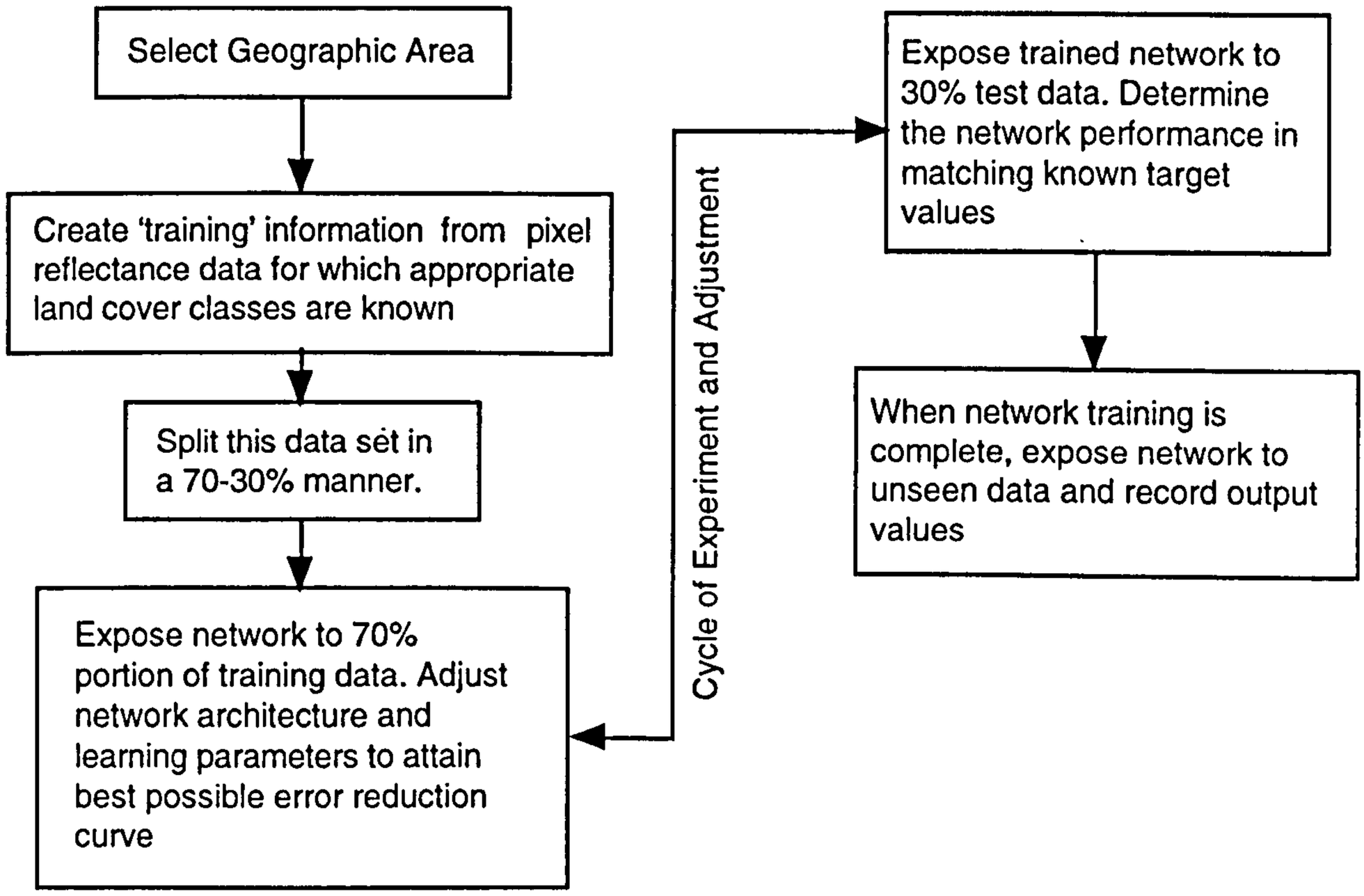
through utilising an algorithm which includes a momentum effect designed to carry the error descent over any flat spots.

The learning process has no fixed end point but if the user persists in training for too long there are likely to be two consequences. Firstly, the error from the ANN will begin to rise away from the optimal solution, and secondly the ANN may well have a reduced ability to generalise what it has learnt to unseen data. Too little training and the ANN will not be able to generalise at all. It is a delicate balance between network architecture, number of training cycles and algorithms that gives an ANN the ability to generalise its knowledge to unseen situations. In addition to this, the error with which the user is presented is only related to the target values in use. It is thus essential that the ANN be regularly tested against unseen data (but that for which the target values are known), to determine how accurately the trained network can perform.

## *5.6 - Unmixing Pixels in Practice*

The application of ANNs to an EO unmixing task generally proceeds like this:

*Figure 5.7: The Use of ANNs in Pixel Unmixing*




The EO researchers create their training data either through ground measurement of sub-pixel variations (an expensive and time consuming process) or through the use of information from a finer scale sensor. Perhaps,

for example, fine spatial scale information is available in a limited way (in terms of spatial or temporal coverage). This information can be used to train a network to unmix data at a larger spatial scale, which is may be more readily available.

## *5.7 - Unmixing Enumeration Districts ; Two Problems*

Each stage of the process outlined in figure 5.7 had to be replicated in some way in order to attempt an application of these techniques to SAS data. There were two obvious difficulties: (1) So far in this research, much attention has been paid to the lack of fine scale information about who lives where. Given that the ANN requires training data to learn how to unmix EDs, where would this data come from? (2) Secondly, the approach assumes that ground cover classes are obvious and quite easily defined. Woodland for example, is clearly not the same as scrubland and there is a corresponding difference in their spectral reflectance. In order to use 'social group' instead of land cover class, the groups must be clearly defined. Problem (2) will be considered first.

### 5.7.1 - Defining Social Groups; Desires and Constraints

The fact that social group schemes may vary in their content and basic themes has been established in chapter 3. For the *development* of this technique, the precise nature and number of the social groups into which an ED's population is unmixed will not matter a great deal, provided that a scheme is utilised which has a place for every household type. This is because the actual character of the groups plays no part in the unmixing process, and could therefore be defined in any number of ways. If the technique were to be used for a specific purpose, for example the detection of a particular type of household, the nature of the groups would be more significant. As Pahl comments, "there can be no true number of classes; the actual number put forward in any given analysis will depend on the problem that is being addressed" (1989, p.713).

A number of elements were required for the development of a suitable social grouping system. Firstly, an appropriate data source; the SAR. The SAR provide unprecedented detail in their description of the characteristics of households in the UK and offered an opportunity to accurately identify and group households with similar census attributes. Secondly, the technical demands which would be placed on the scheme had to be identified and understood. The unmixing method hinged on the fact that the SAS provide information about the number of various household characteristics present in an ED. The unmixing is achieved by deciding how these characteristics might

be accounted for by the presence of social groups which commonly display *those characteristics*. It is was therefore imperative that (1) the social groups were defined using variables which would be available in the SAS tables (albeit in an aggregate form) and that (2) the groups were defined through the possession of a *distinctive set of characteristics*. Households in the same social group needed to share a similar set of census variable attributes and possess different attributes from those in other groups. The social grouping scheme also needed to divide households into between ten and fifteen classes. This value is similar to the basic levels of geodemographic classification, or occupational classification, and also allowed the research to proceed without making the computational demands too great. Since it was expected that stages in the development process would be repeated a number of times, working with a highly complex classification system would have been prohibitively expensive in terms of processing time.

A further constraint was the need to work at a household level, rather than at individual level. Given that there are physically fewer households to cope with and potentially less variation amongst them, it was thought that working at this level would make achieving a workable classification more likely. Whilst this was a choice made through technical influence, households do offer a more stable proxy for social status than an individual's attributes, since the census does not deal well with the complexities of the individual. Although social theorists (Thompson, 1968 for example) may argue that the individual is the site of day to day class processes, the household was an acceptable medium through which to pursue the development of a socio-spatial model.

Against these technical constraints lay a number of methodological aims born from the desire to remain reflexive (see 3.7). The research aimed to reflect a number of axes in the distinction of different types of household which other social scientists suggested were important in the production of a classification system. A number of commentators have highlighted the good performance of classification schemes devised in this way (notably Sleight 1993, Blake and Openshaw 1994). These axes included occupation, lifestage (a combination of age and family status) and the more contemporary influence of consumption behaviour. Since the census does not contain any direct measures of consumption, the possibility of surrogate measures was explored.

The same principle lay at the heart of every classification system considered or tested; when using census characteristics to classify households, difference between the characteristics displayed by households was interpreted as *social*

*distance*. Households which 'looked' different, were treated as such. This was the point at which any mis-match between class theory and the data available (see chapter 3) was internalised by the methodology. Differences between households could only be conceptualised in terms of their appearance in the census. Differences in terms which were not related by the census could not be taken into account by the technique developed. It is important to recognise this because it helps to maintain an understanding of the potential sources of error in the final model and to evaluate how the method actually treated real people in real households. The actual *technique* utilised to divide the SAR data set into a number of social groups was also the subject of a negotiation between methodological desires and technological constraints. An understanding of exactly how each household would be treated by the technique was considered very important. The following chapter will explore all these ideas more fully, together with the precise techniques used.

Following this negotiated method of variable and technique selection, the social group scheme was created. With the groups into which the EDs were to be unmixed now defined, the difficulty of providing training data for the ANN remained. The ANN needed SAS data and the appropriate proportional value for each social group present in the ED. The only means of providing the ANN with this information was to synthetically generate it, given the absence of any suitable data. However, the process of generating census-like data was not without precedent (see 4.4).

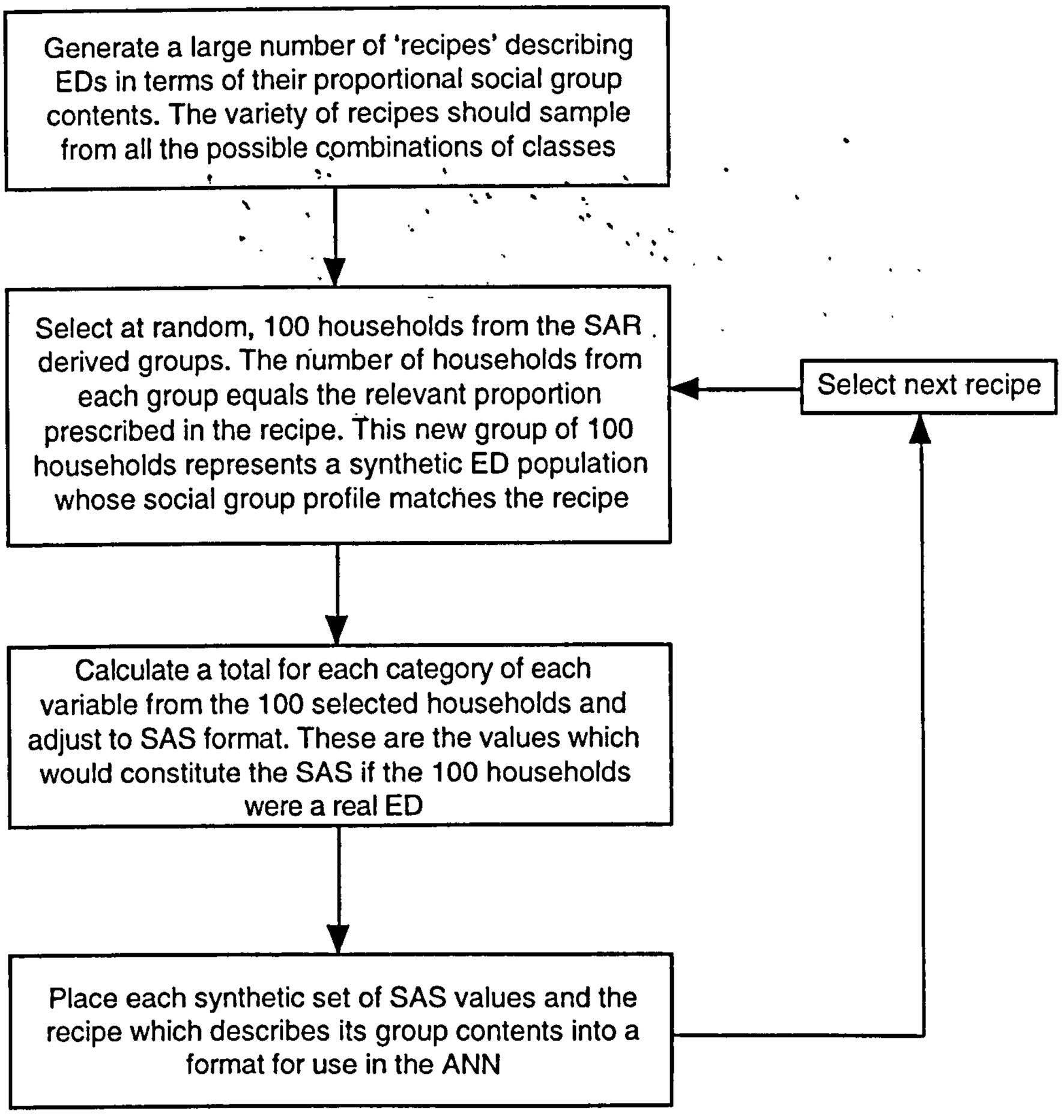
### 5.7.2 - Generating Training Data

After the classification process, the SAR data set could be split into groups of households. Each group was homogenous in terms of its census characteristics and yet distinct from the other groups. Households were drawn from each of the groups in known proportions and combined together to create an artificial ED population. By aggregating the census characteristics of the selected households and carrying out some re-coding of the data synthetic SAS were created, just as though the households lived together in real life and had been subject to a census. Figure 5.8 expands the principle and practice by which a large number of synthetic EDs and accompanying SAS-like data were produced.

With these pieces in place, the adaptation of the EO technique was complete. The synthetic census data, (with known group composition), were divided and network training took place on the larger portion in accordance with common practice (see figure 5.7). Initially, the ANN architecture and learning

parameters were set as though this was a normal image processing task. Subsequent experiments fine-tuned these parameters (see chapter 6).

*Figure 5.8 - Training Data Generation*

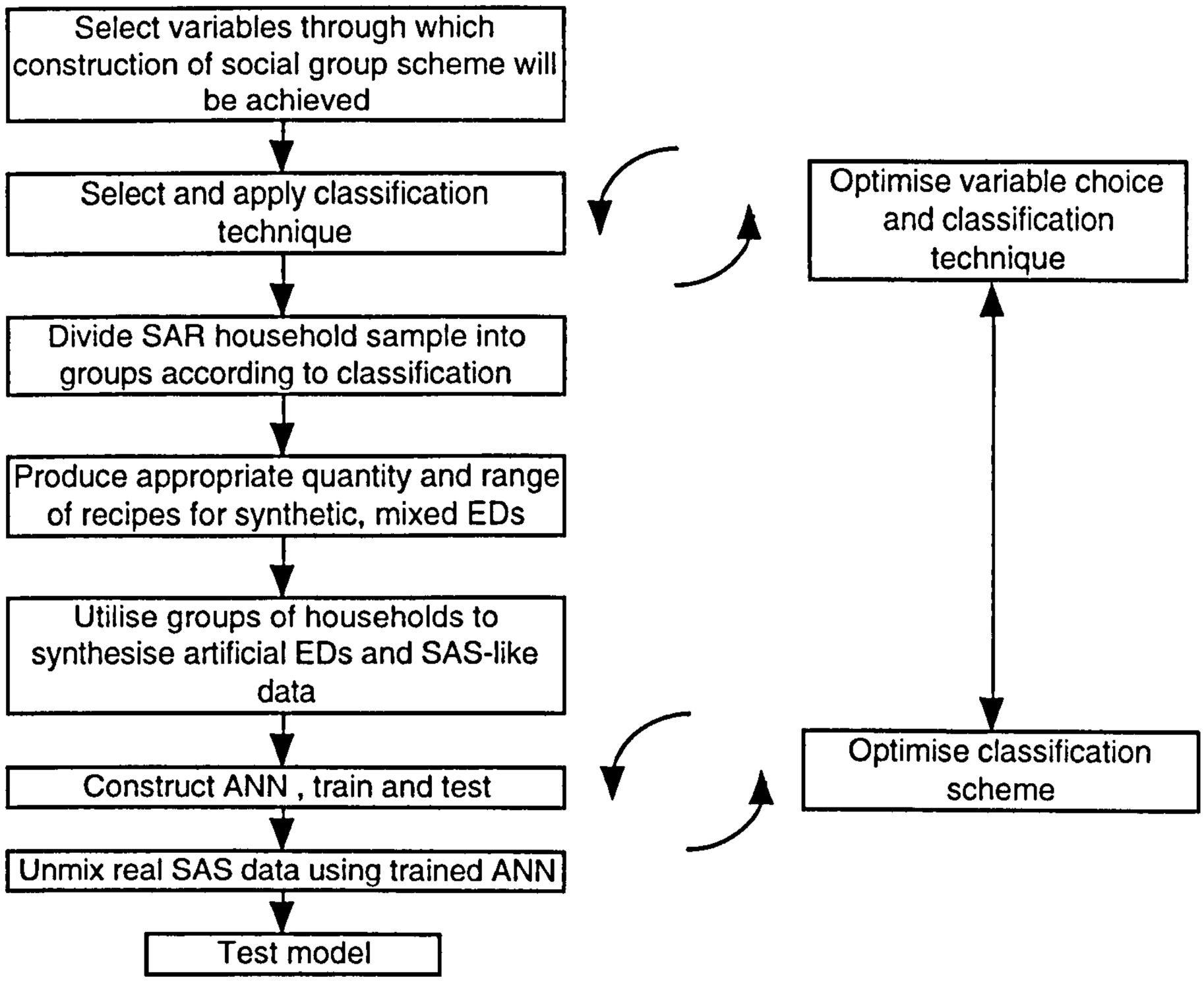



Once the ANN had been trained sufficiently, a model of social mix at fine spatial scale for the Southampton study area was produced. The relevant SAS data for all EDs in the study region were used as the input values and the model was produced very quickly since the ANN had already acquired the ability to unmix every ED in a single step. This was the great advantage of an unmixing approach over a microsimulation. Once trained, the ANN could unmix any ED extremely quickly. Its output values constituted an account of the proportional contents of each ED, in terms of the pre-defined social groups. Thus the steps were in place from two census data sets to a model providing detail of the extent and nature of social mixing in small areas .

## *5.8 - From Data to Model: A Process Structure*

These were the principles and stages through which the integration has been achieved. This presentation of the principles and stages in the unmixing process provides the 'bones' upon which the 'flesh' of operational technique will be built in the following chapters. It will be useful to arrange these bones in a structure to help order and present an account of what has been achieved:

*Figure 5.9 - Integration in Principle*




## *5.9 - Summary and Conclusions*

The pathway from SAS and SAR data to a model of socio-spatial variation has been outlined in this chapter. The key points from the chapter are these:

- The mis-match between empirical information about social and spatial phenomena and the theoretical accounts of those phenomena can be highlighted through a knowledge of social theory and the nature of the data used.

- In highlighting this disparity, a useful reconceptualisation of the spatial data in question has been achieved.

- As a result of this reconceptualisation, EDs are seen as units within which differing social groups live together. The relative proportional presence of these groups has been internalised by the census.

- Conceptualising EDs in this way allows the adoption of principles and practice developed by EO researchers to model the proportional presence of these different groups within small areas.

- Through innovative data manipulation, a method for applying the EO techniques to this social geography problem has been designed.

The principles and stages of the unmixing process have now been presented. Making the principles operational was a two stage process and is presented in chapters 6 and 7. Chapter 6 describes the design and implementation of the techniques needed to complete each stage of the unmixing process. Chapter 7 describes how these techniques were combined and refined through experiments, until the best performance was obtained.

# Chapter 6

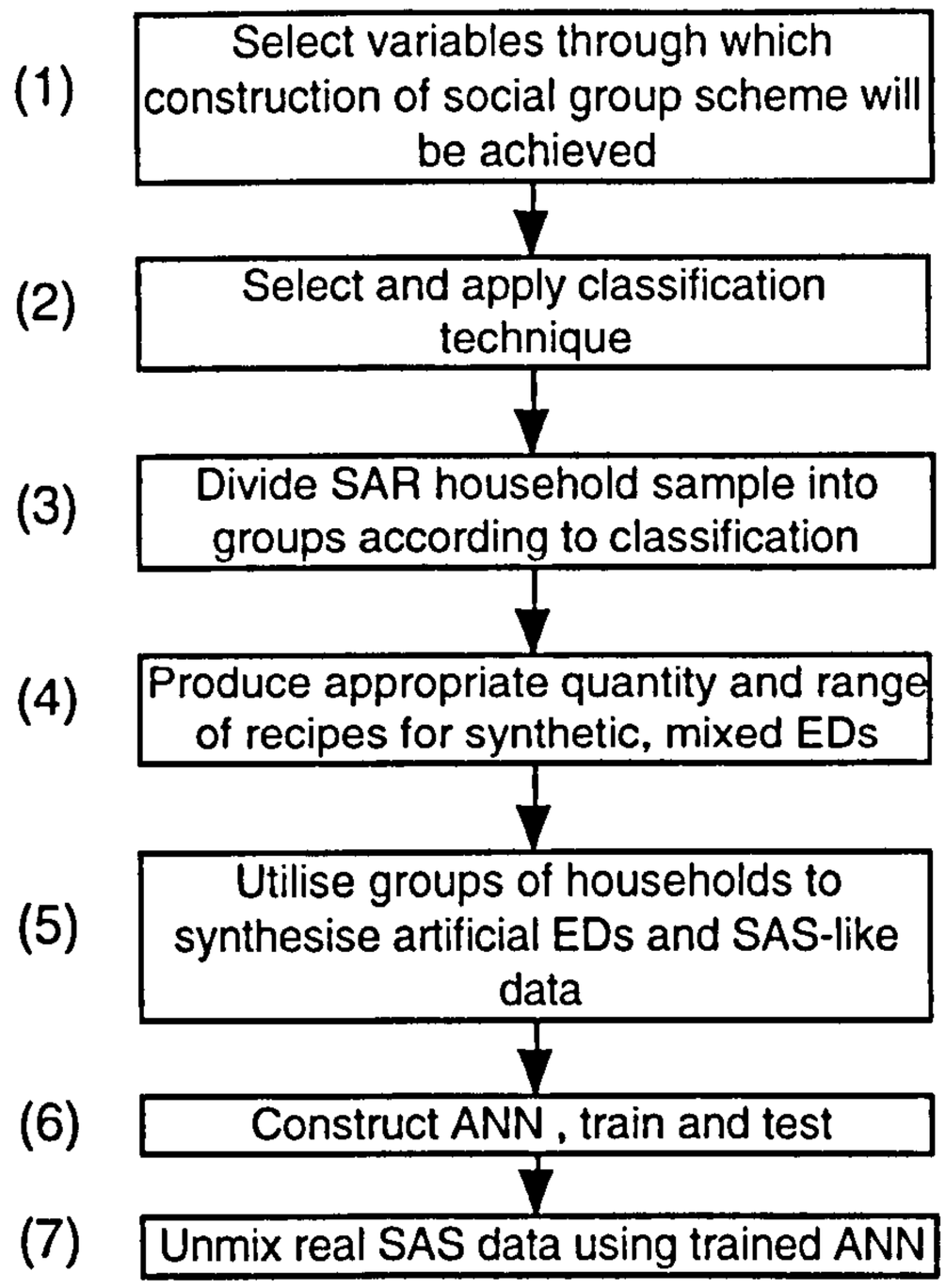
## Chapter Overview

In this chapter the technical and practical steps which produced a socio-spatial model from raw data are described. Census variables and software for the creation of a social group scheme are selected and tested. A method for producing synthetic enumeration district (ED) populations from the grouping scheme is developed and the synthesis of Small Area Statistics-like (SAS) data from these EDs is explained. The preparation and use of these data in training and testing an artificial neural network (ANN) to unmix EDs is then presented. Means of monitoring and testing the ANN learning process are explored followed by explanation of how the ANN was used to unmix real SAS data for Southampton. The chapter includes details of the data sources, extraction procedures and custom designed programs which facilitated the unmixing process.

## 6.0 - Introduction

The purpose of this chapter is to translate the principles outlined in chapter 5, into a series of practical steps, techniques and procedures. Chapter 7 will then show results actually obtained at each stage of development and how they influenced the refinement and use of these techniques. The flow chart in figure 6.1 (based heavily on figure 5.9) will act as a guide to the reader, so that they may determine precisely how the practical steps described here relate to the process as a whole.

*Figure 6.1 - The Unmixing Process: Steps Explored in Chapter 6*




The first task in the sequence is the selection of variables.

## *6.1 - Data Selection*

Getting the choice of data right was extremely important. The unmixed model was expected to be highly dependent on the grouping scheme used and the grouping scheme would be highly dependent on its raw data. The spatial origin of records in the (household - see section 5.7.1) Sample of Anonymised Records (SAR) is provided at UK standard region level (see section 2.10). The development of the unmixing technique was to be targeted on Southampton which offered the benefits of local knowledge and proximity for model testing. The first step in building a data set for the development process was to select records from the SAR's most appropriate geographical segment. Numerous studies have identified a regional variation in household type (see for example Dorling, 1994) and it was appropriate to construct a household classification using data most relevant to the area in which that scheme will be applied. The SAR area which encloses Southampton is The Rest of the South East (RoSE), and therefore only households from the RoSE region were used. This gave a sample size of 41043 households from which to construct the social grouping scheme.

Having determined a geographical sample from the SAR, the specific variables with which to work had to be selected. Chapter 5 has already

explored the 'negotiation' between methodological desires and constraints to which variable selection was subject (see 5.7.2). In practice this negotiation continued throughout the research and development. This continual process of optimisation resulted in the testing of several different variable sets. However, an initial set was required and it included these variables:

*Table 6.1 - Initial Variable Set for Classification*

| Theme | Variable |
|---|---|
| Consumption | Tenure of Household<br>Number of Cars Per Household<br>Presence of Central Heating<br>Number of Rooms In House |
| Demographic | Family Type<br>Number of Persons In Household<br>Age of Household Head<br>Sex of Household Head<br>Ethnic Group of Household Head |
| Economic | Socio-Economic Group of Household Head<br>Economic Activity of Household Head<br>Educational Attainment of Household Head |

The variables shown in table 6.1 fell broadly into three groups; consumption, demographic and economic oriented. The exploration of academic theory and commercial practice suggested that these themes would lead to the creation of an effective household classification scheme. The success of schemes which employ a similar set of variables is testament to its validity (Webber 1995).

### 6.1.1 - Variable Choice In Practice

All variable selections were subject to an intensive scrutiny. This scrutiny was intended to clarify the reasons for the inclusion of the variable and to determine the precise format of the information it provided. The selection of an occupation oriented variable for example, was clearly warranted. Many theoretical and commercial classification schemes include occupation as a key aspect of social difference (see section 3.8). It was intended to reflect this by including an occupation based variable in the classification scheme, however there were a number of different options available. A choice from the occupation oriented census variables available was required.

The census variable 'social class based on occupation' was the most hierarchical available. It has an implicit order, placing an individual in one of seven groups according to the perceived rank of their occupational title. Despite the aims of OPCS to "bring together, as far as possible, people with similar levels of occupational skill" (1992, p. 40) the structure has been widely criticised for its assumptions about and reproductions of, out-dated social divides (see for example Rex, 1970 and Saunders, 1987). The validity of these criticisms had to be considered and judged against whether this particular form of occupational information was vital to achieving the aims of the grouping process. Such a hierarchical construction did not seem to offer any particular advantages. In fact, the aim of the classifier was not to generate a hierarchy but to group households which appeared to be similar. Therefore, rather than utilise a processed version, raw occupational information was explored. The SAR data set offers over 350 different occupational classifications through the standard occupational classification system. These would be easily collapsed into more general groups, without a hierarchy making this variable ideal from a theoretical viewpoint. However, unfortunately the SAS do not use the same occupational classification at household level. The reader will recall the necessity of selecting variables common to the SAR and SAS (see section 5.7.1).

Another version of occupation information was available as 'socio-economic group'. This variable is designed to group together "people with jobs of similar social and economic status,"(OPCS 1992, p.41). The position of an individual within this categorical variable is a result both of their employment status and occupation. It offered a more detailed picture than the 'social class' variable whilst avoiding hierarchical overtones. Although 'socio-economic group' is a derived variable and thus further from the information that the individual had given on the census questionnaire, some separation of the categories which cover a variety of positions (such as manager) was offered. This might allow the separation of the large middle class groupings whose census characteristics might otherwise appear similar. Finally then, the variable socio-economic group was chosen. Its selection was a balance between the technical demands of the methodology and the desire to select the most suitable information available.

This detailed example is an indication of the process which lay behind every variable choice. Is the variable cited by others as being useful? Is it available in the correct form in both data sets? Does its use contravene the desire to utilise data closest to an individual's response? Ultimately, the technicalities of using

census variables to create a generalised scheme capable of differentiating between households, proved to be the most powerful element in their choice. Chapter 7 will reveal more about this.

Once the selection of variables had been completed, the initial data set was obtained (N.B. this was still stage (1), figure 6.1).

## *6.2 - Data Extraction*

The means of obtaining this variable set and all its subsequent mutations, were very straightforward. The SAR are held in an on-line repository, by the Census Microdata Unit (CMU) at Manchester University. The data required were extracted through a user defined SPSS syntax file and subsequently downloaded to a local UNIX platform over the internet. One syntax file extracted the variables required, for the specified households and specified regional area (RoSE in this case). Figure 6.2 illustrates an example extraction job:

*Figure 6.2 - SAR data Extraction Syntax*

```
TITLE ' DATA FOR ANALYSIS'.
GET FILE = 'db/census91/sars/spss/hhold/hhold.sav'/KEEP AREAHH RELATE
 TENURE SOCLASS ECONPRIM CENHEAT ROOMSNUM FAMTYPE ETHGROUP SEX QUALNUM
 PERSINHH AGE.(opens correct data file and specifies variables of interest)
SELECT IF (AREAHH EQ 07).(select RoSE data only)
SELECT IF (RELAT EQ 00).(specify head of household only)
WRITE OUTFILE = 'input.dat'/ TENURE (F3.0) SOCLASS (F3.0) ECONPRIM (F3.0)
 FAMTYPE (F3.0) QUALNUM (F3.0) CENHEAT (F3.0) CARS (F3.0) SEX (F3.0) AGE
 (F3.0) PERSINHH (F3.0) ETHGROUP (F3.0).(write data to file)
EXECUTE.
FINISH.
```

Some re-coding took place to ensure that all the variables had the same basic structure, following the extraction. After each download the data were checked for error which might have resulted from their transfer, both from Manchester to the local platform and from one manipulation package to another.

### 6.2.1 - Dealing With Error in the SAR

The need to check data sets for error when using them in a modelling environment is paramount (Unwin 1995). Unwin draws a distinction between

different kinds of error; issues of data *quality*, (i.e. their appropriateness for a task) to which we have already devoted considerable attention (see section 3.6 - 3.14 and 6.1.1), and issues of data *error*, regarding the accuracy of the numerical information. The CMU, which administers the SAR, publishes quite extensive documentation on the data's reliability. Their advice includes measures to account for known sampling error and known factors which might reduce the SAR accuracy (CMU 1993). The approach to dealing with error in the SAR depends heavily on the use to which they are being put. If the purpose of this research had been to provide estimates of the sizes of particular social groups, either regionally or nationally, the issue of sample bias would have needed to be addressed. However, a review of possible sources of non-sampling error was sufficient given the research tasks.

Non-sampling error may have been introduced during any of the operations to collect and process the data. This may affect the data in two ways; error introduced in a random way will increase variability in the data, error introduced in a systematic way will introduce bias (CMU 1993). Sources of errors include (i) Quality of Response - Incorrect form completion by respondent either through misinterpretation of the question or a deliberately incorrect response, (ii) Incomplete coverage - The census often misses those who are hardest to enumerate and (iii) Transcription and coding errors - These are introduced during processing.

Every effort is made by ONS and the CMU to ensure that error sources are detected and either corrected or quantified. As the SAR was employed here as a source of information about household variation, a reasonably relaxed attitude to data error was adopted. The types of error which might have affected this use of the SAR include mistakes in coding and the absence of particular household types. There were no specific actions which could be taken to counter these difficulties, other than an awareness of them and the possible influence on model accuracy which they might have had. Error was more likely to have stemmed from mistakes in local data manipulation than from the original source.

## *6.3 - Exploring Univariate Schemes*

With the data prepared, classification could proceed (stage (2), figure 6.1). The first experiments in this process were born both from curiosity and the need to determine whether complex multivariate classification of households would prove absolutely necessary. Although their utility was strongly doubted, a brief examination of two or three schemes was undertaken to determine their

quality and to start exploring the nature of the sample population. A simple method was employed for examining these classifications. A group was defined using SPSS, in which households shared a common characteristic (e.g. a household head in a certain occupation). The rest of the group's characteristics were then visualised in Excel as shown in figure 6.3. The graph shows a comparison of the characteristics of the group with those of the entire sample population. It shows which attributes were more (or less) common amongst households in the same group by using a Z-score. If a particular attribute occurred amongst the selected group in the same proportion as the wider sample population, the Z-score was zero. If the group displayed a higher proportion of a particular characteristic, the Z-score was positive (increasingly so with increasing proportion) and the Z-score was correspondingly negative when the feature occurred in a reduced proportion.

*Figure 6.3 - Univariate Classification*

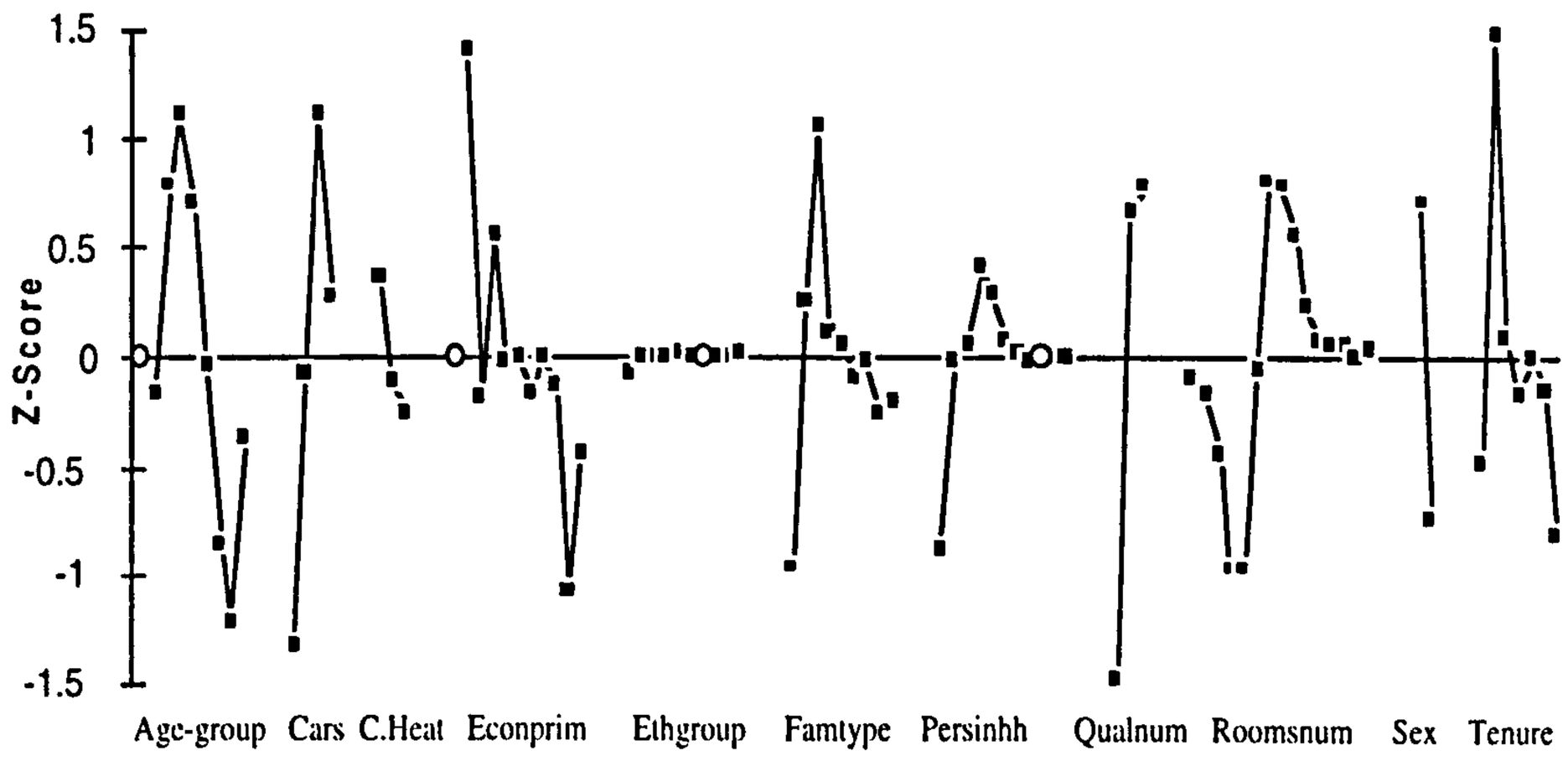



Figure 6.3 shows the Z-Scores for all households in which the household head was classified as a professional. The black squares represent the possible categories within each of the labelled variables. The white circles represent a category of an attribute for which there were no households in this class. For some variables, the categorical divisions are easily interpreted. Age group for example, increases from left to right and thus illustrates that this group contained a higher than average number of young - middle aged heads of household, and a lower number of elderly heads of household. Other variables are categorical. Tenure for example is divided as follows (left to right):

Owner occupied outright
Owner occupied with a mortgage
Rented Privately Furnished
Rented Privately Unfurnished

Rented With A Job or Business
Rented From a Housing Association
Rented From a Local Authority

Like many of the categorical variables used in the research, tenure has an implicit order, from outright ownership 'down' to council tenancy. Figure 6.3 shows that for this group, mortgaged ownership was well above average, with council tenancy significantly below average. Furthermore, the number of rooms in household (Roomsnum) illustrates that there is a high proportion of larger houses present. These features were as expected for a professional group.

The univariate class schemes explored included tenure, economic activity and occupation. However, the resultant structures were problematic. A significant problem with census based univariate classifiers is that they only separated the usual from the unusual. Consider for example, the variable 'economic activity'. There are a number of categories within this variable which relate to those in employment. A grouping scheme based on economic activity therefore gave a number of 'employed' groups with many similar characteristics, but in stark contrast to the one category for the unemployed. The variables were of fine enough resolution to distinguish between the extremes of the economic scale, but not between the middle or upper grades. Whilst this kind of distinction may be adequate for those focused on the divisions of a highly polarised society, it was not sufficiently detailed for the task outlined in section 5.7.1. It appeared certain that a multivariate scheme would be required.

## *6.4 - Multivariate Classification*

The use of multiple variables necessitated the adoption of a multivariate classification technique. A number of technical and methodological desires had to be balanced against each other and as with the variable selection, the classification technique adopted changed during the course of the model's development. Initially however, a number of options were examined:

### 6.4.1 - Conventional Approaches

There is a wide range of multivariate classification techniques available to the census data user, the majority of which have been developed using SAS. This section reviews both conventional and neural approaches:

Until recently, the k-means nearest neighbour classifier was the state of the art technique for classifying spatial data (Openshaw et al 1995). The k-means approach is based on nearest centroid sorting (Anderberg 1973). A case is

assigned to the cluster with the smallest distance between the case and the cluster centre (SPSS 1991). The process of assigning cases and re-computing cluster centres (from the average of cluster members) continues until no further changes occur in cluster centres or until a maximum number of iterations is completed SPSS (1991). This approach has provided successful classifications of census data, however it can be subject to a number of potential problems. The k-means technique has the tendency to become 'stuck' at a sub-optimal classification and can have difficulty with non-normal distributions (Evans & Webber 1994).

Other criticisms are more philosophical and include the fact that the classifier has the potential to impose an arbitrary structure on the population undergoing classification. There is also, apart from the selection of variables and manipulation of classification parameters, no means of introducing prior knowledge of the population into the classification (Openshaw et al 1995). The absence of an opportunity to influence the outcome of a classification was of considerable significance given the theories that lie behind issues of social and socio-spatial structure and the stated aim to use this knowledge (see section 3.13). The k-means method does however, have a proven history of effective classification and has been refined to the point of considerable efficiency. The technique was also readily available within the SPSS package. K-means was thus a candidate for use if no better option could be found, providing an effective means of objectively clustering households together.

### 6.4.2 - Neural Approaches

The advent of neural computing has offered researchers the chance to address a number of the criticisms of conventional multivariate classifiers. The basic principles of neural classification are similar to neural learning and the reader will require no further explanation of the technique. Following their extensive research, Openshaw et al. argue that "the use of a neuroclassifier provides a much more flexible and potentially superior means of generating census classifications," (section 7, 1995). It is argued that neural classification holds a number of advantages over more traditional approaches including the fact that the classification structure emerges from the data, rather than being imposed on them. Neural approaches allow the classification to move beyond the bounds of an 'either / or' classification system because they are inherently fuzzy (see section 2.13). This characteristic introduces the possibility of recognising different levels of group homogeneity. Neural approaches also facilitate the introduction of prior knowledge to the classification. This, claim Openshaw et al (1995), makes the classification 'more intelligent'.

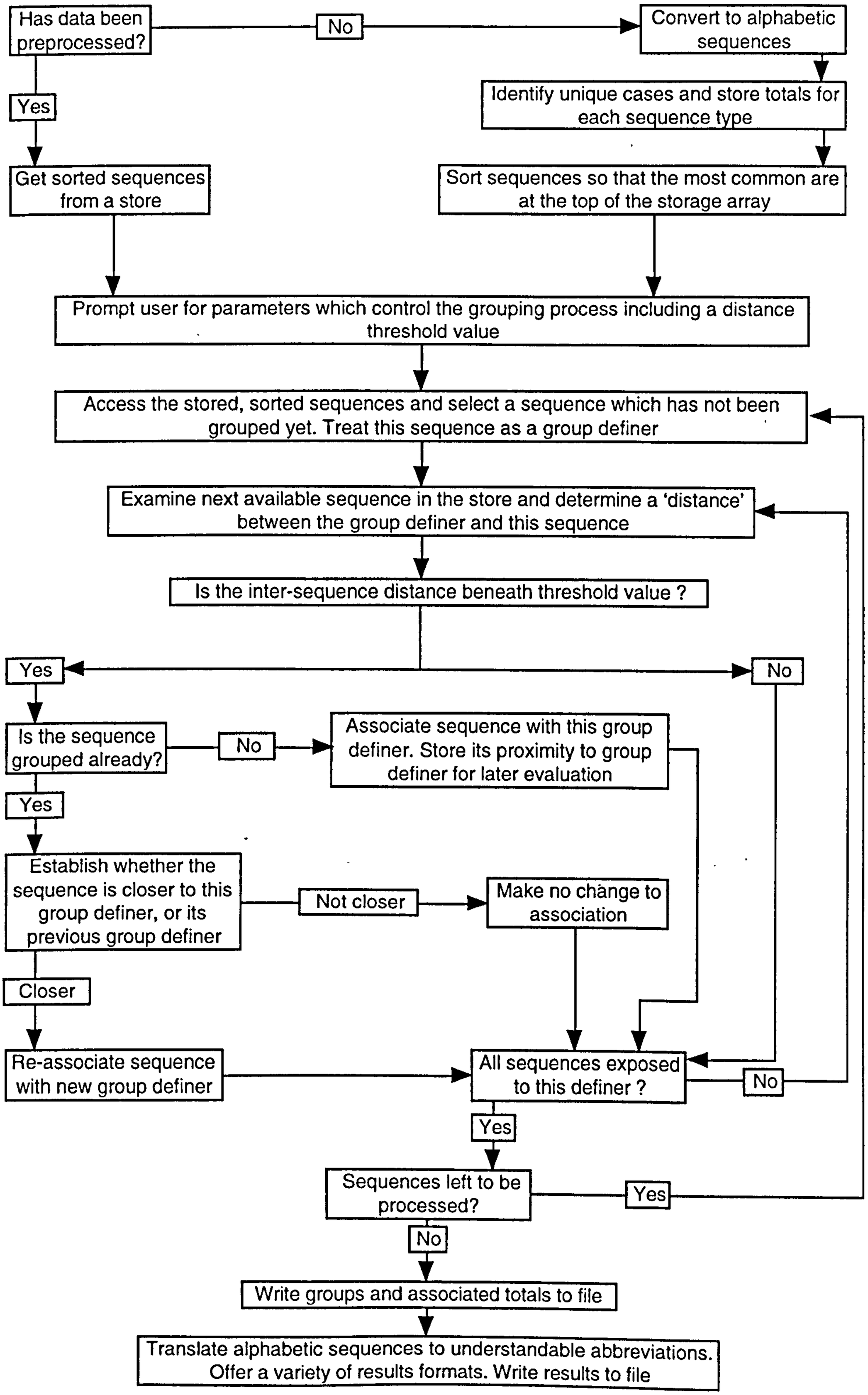
There are however, a number of potential disadvantages in the adoption of neural classification, perhaps primarily the lack of experience with the concepts and technologies. This leads to an uncertainty over the effects of variable choice, net architecture and the limitless potential for minute adjustment of the system's governing parameters. In additional to these, neural classification can be highly demanding in terms of computer power and time and with the use of an ANN to unmix the SAS, those demands were already building steadily.

Other ideas contributed to the rejection of both conventional classifiers and neural classifiers, in favour of a home-grown technique. These concerned the nature of the data in use, a desire to remain aware of the impact of technique on model output and a desire to involve relevant social theory at all appropriate opportunities (see section 3.7). Most of the reviewed research which used a census based classification, employed aggregate data. Given the relatively recent advent of the SAR this is not surprising. However there is a significant difference between the demands placed on an areal classifier and those for the classification of individual or household level data. One criticism of quantitative social geography highlighted in section 3.6.1 is that it can be remote from its subjects and therefore often fails to recognise that the data used represent real people. The application of a black box classifier to the SAR was initially felt to be undesirable, since there would be uncertainty over how each household had been treated. *(N.B. Extensive experiments eventually showed that the k-means classifier was needed, since it was capable of producing a much higher standard of classification (see 7.9.2). Before this was established however, development of a custom written classifier was undertaken.)*

### 6.4.3 - GroupMaker: A Custom Written Classifier

A classification scheme which allowed the structure to emerge from within the data, but which also retained a facility for the application of certain constraints would have been the ideal choice. A close dialogue between theories of social structure and the data themselves could be achieved using this kind facility. The best of both worlds was needed; a classifier that allowed the data to determine their own structure but that also allowed user intervention and control. Constraints on computational power, time available to access and learn systems and the technical competence of the researcher also had a role to play in the decision making process. Together, these issues

*Figure 6.4 - Flow Chart For The Custom Classifier GroupMaker*



made a powerful argument for a classifier that was custom designed and programmed. (See Appendix V for GroupMaker's code listing).

The classifier, called GroupMaker, was written in C and run on a UNIX platform. It was designed to run with either a command line interface or with a command file which allowed overnight running. Figure 6.4 illustrates the sequence of processing. As figure 6.4 illustrates, GroupMaker initially reduced the SAR data to a list of unique household character sets together with the total number of occurrences for each character set. This step was designed to significantly speed processing. A sorting function arranged the array of sequences with the more common ones at the top. This allowed for the fact that GroupMaker created groups whose nature depended on the order in which the sequences were encountered, to some extent. The most common household types in the population were given the greatest chance to become cluster centres. This condensed and sorted data set could be stored and re-used, saving time when different grouping parameters were being tested.

GroupMaker then prompted the user for the parameters with which the classification process was controlled. Once instructed, the program worked through the array of households forming groups by identifying those which were significantly different and those with similar characteristics. It is important to understand how grouping decisions were made and the range controls over that process.

### 6.4.4 - Principles of Distance and Controlling Classification

The classifier worked through the calculation of a *distance* between two variable sequences as a measure of the *difference* between them (see 5.7.1). A decision over whether two households were sufficiently similar to warrant being grouped together was based on this derived distance. An inter-variable distance was calculated which reflected how 'different' two households were on a specific variable. For example;

Household 1, Variable 1: Category A or Household 2, Variable 1: Category E

Distance = 4 units

The inter-household distance was simply the absolute sum of all the inter-variable distances. A number of parameters controlled the calculation and impact of the distance. Control over the classification process was achieved through the

manipulation of these parameters. There were two types of parameter; the distance limit (or threshold) and the controls over individual variables. The distance threshold represented the value beneath which an inter-household distance had to be in order for the two households to appear in the same class.

A number of options were available for determining the contribution that each variable made to the derived distance; (i) A variable could be switched off and take no part in the derivation of a distance value. This allowed the user to determine the significance of a particular variable in constructing a group structure. (ii) A variable could be assigned a weight through which the distance could be increased or decreased. This allowed the user to increase or decrease the influence of a variable in the grouping process. It also allowed the user to account for the fact that some variables have a higher number of categories than others and therefore, the potential to contribute a much greater value to the inter-household distance. (iii) A variable could be declared a 'key'. In order for households to be grouped together, they had to display *the same category of the key variable*. The key variable system allowed the user to try imposing class structures which were heavily based on one or two variables. (iv) A variable could be declared categorical. In those situations, any difference in category between households, whether it is an A-B (i.e. 1 unit) difference or an A-L (i.e. 11 units difference) contributed the same amount to the overall distance value. This reflected the fact that not all the variables in use had an implicit order in their categorical structure.

The controlling parameters provided the ability to dissect the SAR data in an vast number of ways from strict, subjective divisions to allowing the populations to settle into groups without any external influence. The features that were lacking in other classification techniques had been achieved, together with an understanding of how the classification worked that could only have been obtained through personal construction of the program.

## *6.5 - Analysing Group Schemes*

GroupMaker provided a variety of report formats and a file containing the households separated into groups according to their classification (stage (3), figure 6.1). Early experiments with the program illustrated the need for a means of assessing the groups' characteristics, differences between them and within group variation. Visualisation rather than statistical description seemed appropriate, given the complexity of the information and the frequency with which assessments of a group system would have to be made.

Data visualisation is a field of growing significance, reflecting a period of dissatisfaction with statistical description or testing (Dorling 1995). As Dorling comments "most of those [statistical] techniques require the researcher to know beforehand precisely what he or she is looking for" (p.354).

### 6.5.1 - Visualisation

The stored group details were used to generate a visualisation. The total number of households in each category, of each variable was established. The counts were then arranged in a grid where the rows were formed by the census variables and the columns were formed by the categories within those variables. The program adjusted the 'sizes' of the category columns within the variable rows so that despite the census variables having different numbers of categories, the resultant grid was always almost square.

This matrix was then passed into the PV-Wave data visualisation package. The package assigned a colour to each category within each variable by scaling the variation in the matrix to within a 0 - 255 scale. Hence, the colours that were seen represented a scaled representation of the relative presence of each category in the group. The fact that in many cases the variables used had an inherent order to them meant that rapid interpretation of the image could be learnt.

Images like that in figure 6.5 could also be used to assess within-group variation. The intensity of shading, and the number of cells not shaded at all, gave an indication of the homogeneity of the group. When viewed in sequence, the images showed the variety of group character. Where similar patterns occurred repeatedly, the groups could be judged to be too similar.

### 6.5.2 - Numerical Analysis

In addition to the diagrams, custom written software called Mode determined the modal household features of the group and gave a numerical indication of the degree of within group variation. An example of the output from Mode shown in figure 6.6. The modal sequence is shown in abbreviated format at the top of the output. These short abbreviations were used to speed the analysis of output from GroupMaker and Mode. NoIF for example, refers to the 'not in a family' category of the family type variable. The presence of each category is expressed as a percentage in this particular version of the output, (49% of households in Group 2 are outright home owners). Mode also provided information on the absolute numbers of household features present in a group.

*Figure 6.5 An Example of PV Wave Visualisation*

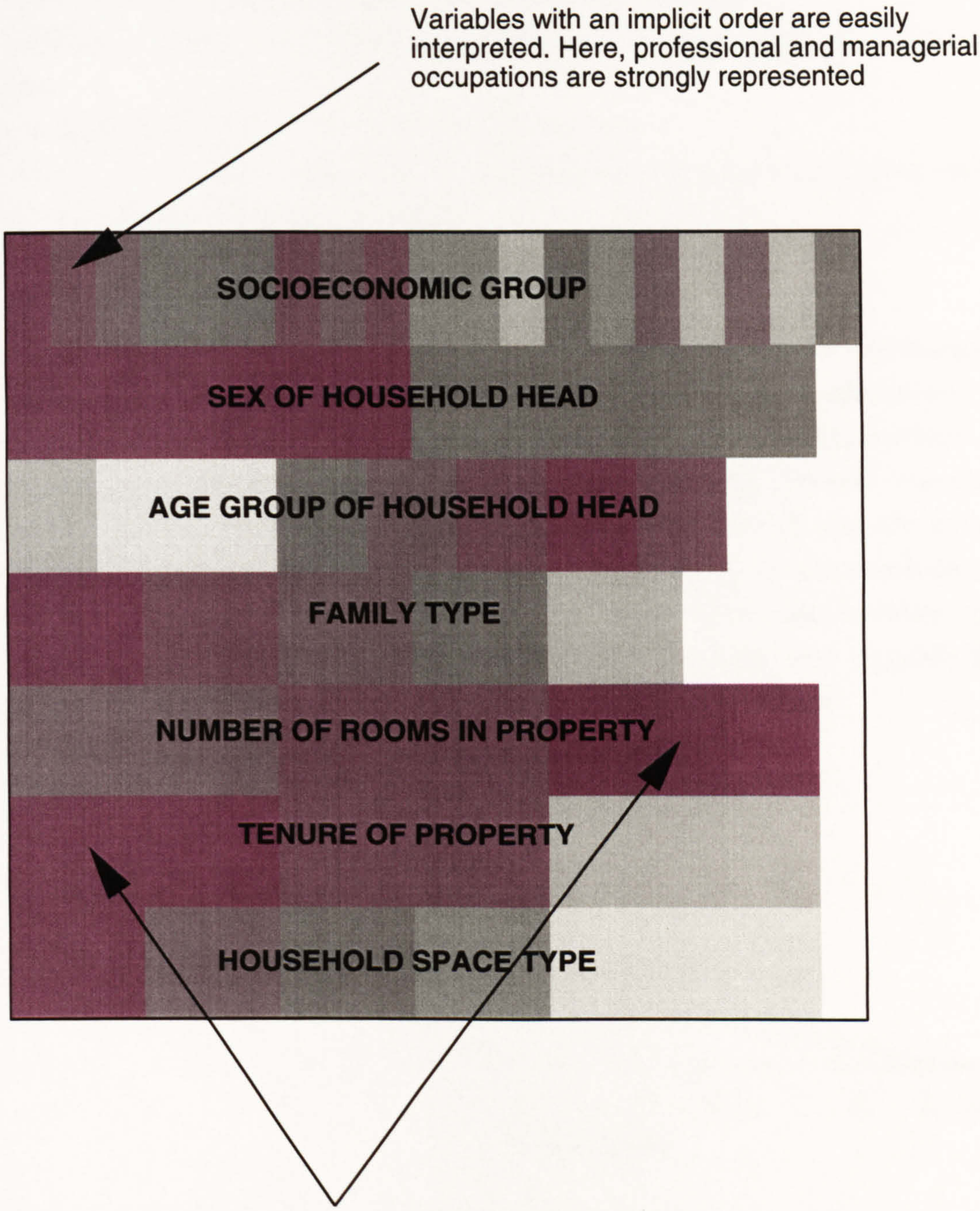


*Figure 6.6 - Output from Numerical Analysis of Groups*

Group Number :2

NotA Fema NoIF 4Rom OocO age6 Noac

V. 0> NotA (97.41) Inad (2.08) Armd (0.31) Agri (0.18) FaOw (0.01)

V. 1> Fema (59.14) Male (40.86)

V. 2> NoIF (59.56) CoNc (25.98) LoPs (9.13) CoDC (2.72) CNDc (2.61)

V. 3> 4Rom (26.38) 5Rom (24.82) 6Rom (16.75) 3Rom (14.57) 7Rm+ (8.22) 2Rom (6.51) 1Rom (2.76)

V. 4> OocO (48.69) RLoA (27.27) RPJH (14.90) OocB (9.14)

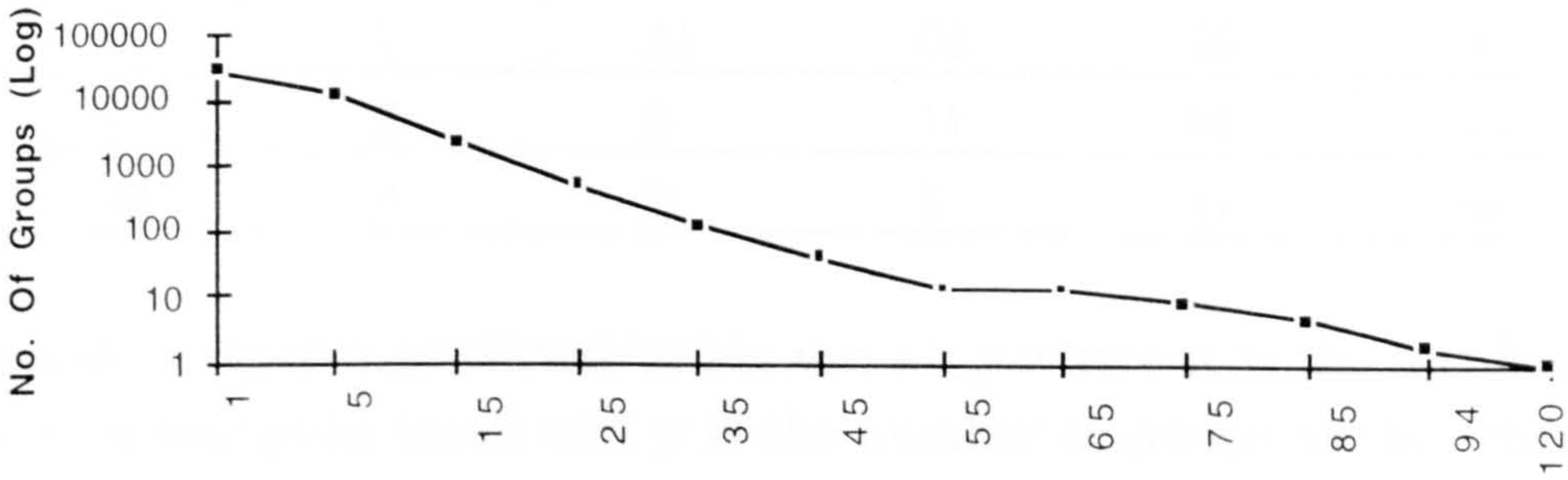
V. 5> age6 (41.33) age5 (29.25) age7 (10.85) age3 (5.78) age4 (5.26) age2 (4.77) age1 (2.76)

V. 6> Noac (83.51) 1acr (12.68) 2+ac (3.81)

## *6.6 - Initial Exploration*

Some calibration work was necessary to determine the numerical ranges within which distance thresholds and weight values worked effectively. This was achieved by repeatedly running GroupMaker and varying just one parameter of interest. To determine the effect of varying distance thresholds for example, the influence of all the variables was equalised and the threshold value varied from one at which no groups were formed, to one at which every households formed 'a group'. The derivation of a relationship between the number of groups and distance threshold value was then used to guide the choice of threshold value when other parameters were varied.

*Figure 6.7 - Number of Groups against Distance Threshold*




A combination of GroupMaker, visualisation software and numerical analysis thus provided a potent means of exploring potential grouping schemes and the resultant division of households.

## *6.7 - Synthetic Small Area Statistics Generation*

The practical means to produce a household grouping structure was now established. The groups of households were moved into separate files and after a conversion from alphabetic to numeric information, the data were ready to be used in the generation of the synthetic EDs and SAS.

### 6.7.1 - Recipes for Enumeration Districts

The first task was the production of a large number of proportional recipes which would prescribe the mixture of groups used for ED synthesis (stage (4), figure 6.1). The production of these recipes was a two stage process. The first stage used a relational spreadsheet (built in Excel) to generate a sequence of random numbers whose sum was precisely 100. Each number would represent the proportional presence of one group. The spreadsheet created a variety of distributions, rather than always having one or two values providing the majority of the 100. Once the recipes were created, they were exported to a UNIX platform to become the input for stage two: A program increased the number of recipes by reading the original sequence into an array, copying and reordering it so that every group would 'experience' every value in the recipe:

*Table 6.2 - Ensuring Variation In The Training Data*

(assume 6 groups)

| Sequence | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 |
|---|---|---|---|---|---|---|
| Original | 64 | 18 | 0 | 2 | 5 | 11 |
| Seq 2 | 11 | 64 | 18 | 0 | 2 | 5 |
| Seq 3 | 5 | 11 | 64 | 18 | 0 | 2 |
| Seq 4 | 2 | 5 | 11 | 64 | 18 | 0 |
| Seq 5 | 0 | 2 | 5 | 11 | 64 | 18 |
| Seq 6 | 18 | 0 | 2 | 5 | 11 | 64 |

The number of recipes produced ($n$) is $n = x \times y$ where $x$ is the number of sequences produced in Excel and $y$ is the number of groups in the scheme. This ensured variety in the training data generated from the recipes.

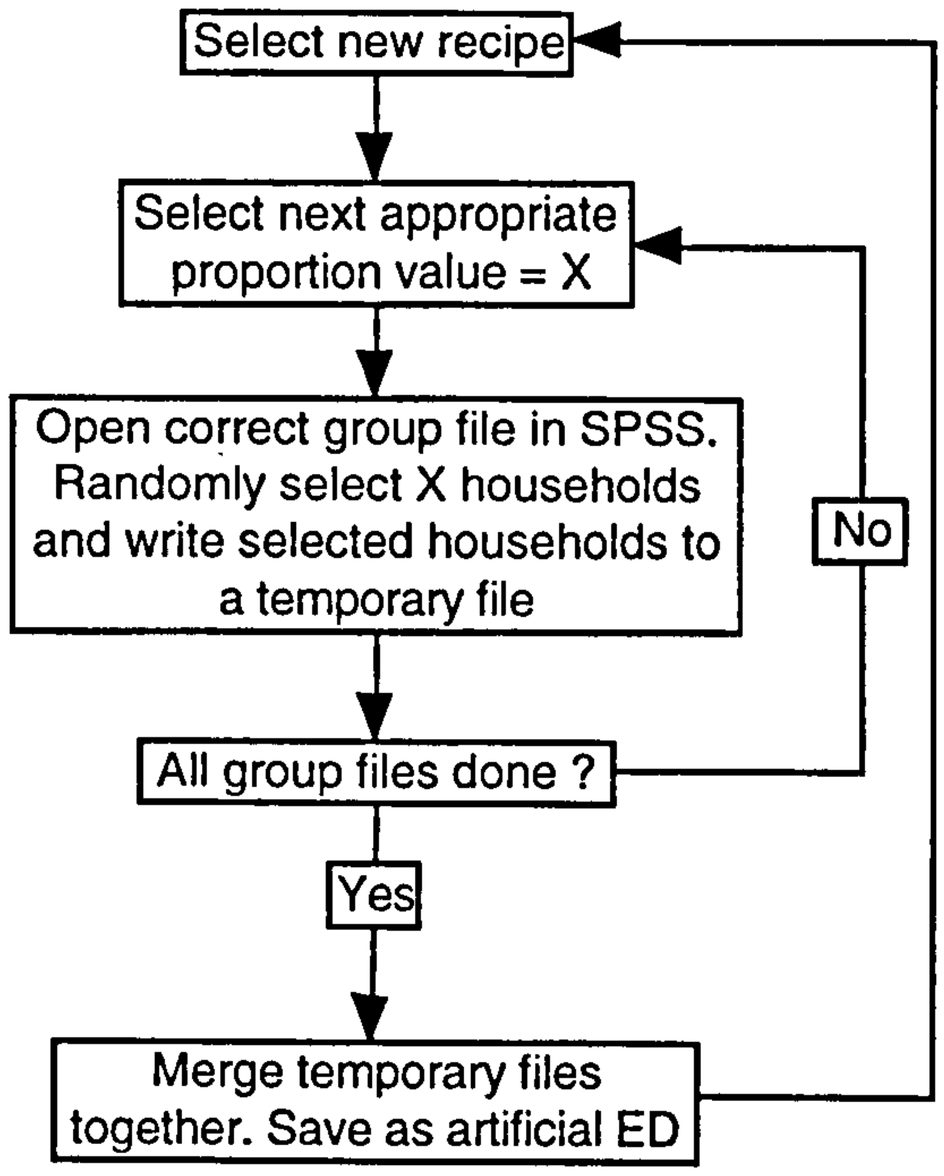
### 6.7.2 - Synthetic Enumeration District Generation

(Stage 5, figure 6.1) Using these recipes required some manipulation of the household data. The artificial neural network (ANN) required an input for every possible category of every variable, whether present or not (see section 5.5.1). Data from the census are coded so that if for example, the household was owner occupied with a mortgage, its tenure code was 2. To be read by the

ANN the code had to be changed to 0 1 0 0 0 0 0, since mortgaged ownership represented one tenureship from a possible seven. This translation was carried out and the data were read into SPSS. Each group was then saved as a separate file.

Combining households together in the proportions prescribed by a recipe sequence was carried out by SPSS and controlled by a syntax file. The syntax file performed a series of looped actions:

*Figure 6.8 - SPSS and Synthetic ED generation*




The SPSS job was run in batch mode. The result was one file for every recipe, containing 100 households from the household groups in the prescribed proportions.

### 6.7.3 - Small Area Statistics from Enumeration Districts

There was one further stage in this process. At this point each of the files described above represented a synthetic ED population. The information held in those files had to be converted to a format to which real SAS could be matched. A program converted the 100 household sequences to a single set of SAS values by simply totalling the number of households falling into each category of each variable. The conversion also accounted for that fact that some of the variables in use appeared as a 10% sample for the real EDs. The program also placed all the synthetic SAS data together in a file with the

recipes which accounted for each sequence. The training data and target values were now ready to use in ANN training.

The reader will recall that in order to effectively train and evaluate the performance of an ANN, some test data is required (see 5.7). The synthetic SAS set was thus split, with approximately one third retained to use in ANN testing. Both testing and training data sets were converted to a format which the ANN software could read.

## *6.8 - Neural Network Training*

The ANN package used was the Stuttgart Neural Network Simulator (SNNS) Version 4.0. It originated in the University of Stuttgart (SNNS 1995). The package was downloaded across the internet in code form and complied and configured to run on the UNIX platform that was available. The package utilises an X Windows graphical user interface and control over its architecture and parameters is achieved through a number of dialogue boxes. The ANN can also be run in batch mode through the use of command files. This feature enabled much of the processor intensive training and testing to be undertaken at night.

### 6.8.1 - Network Architecture and Parameters

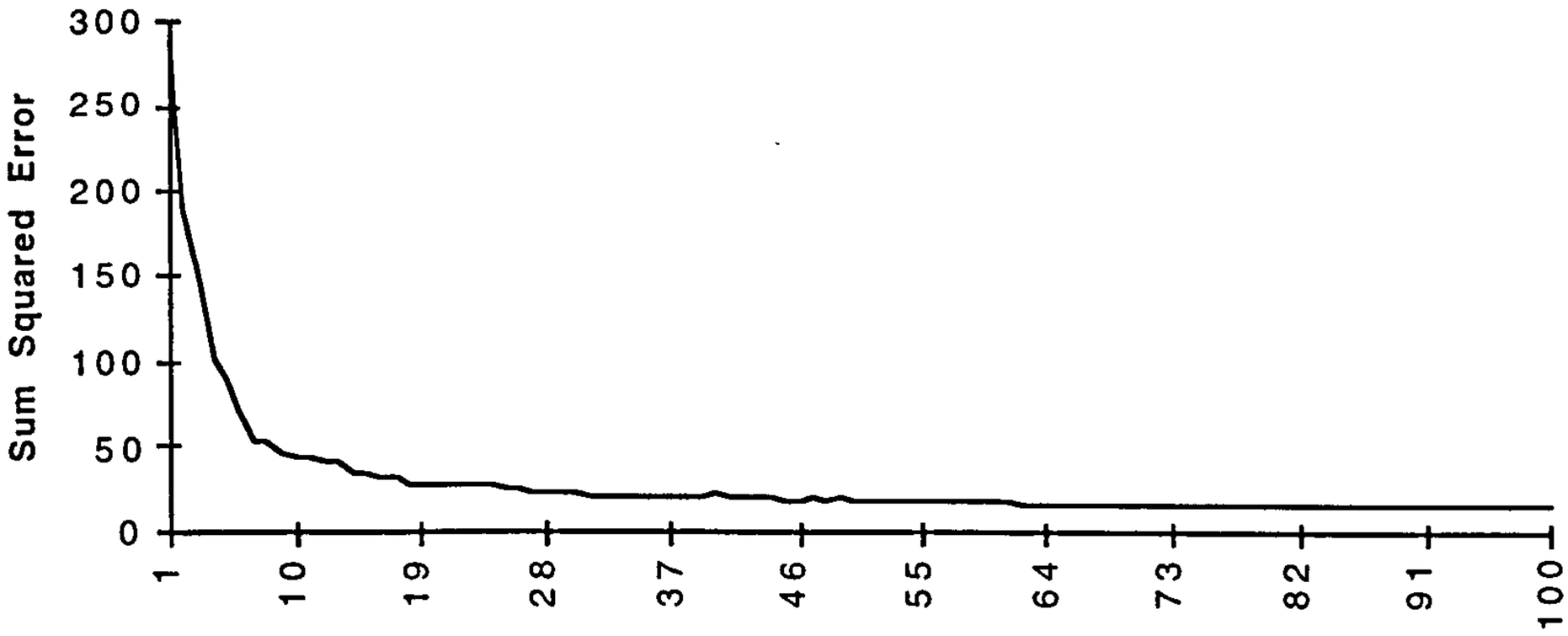
It was recognised that the network architecture and learning parameters would be adjusted as aspects of the ANN performance were revealed. Therefore, as with the selection of an initial set of variables for classification, a network architecture and set of parameters were adopted with which to begin training (stage (6), figure 6.1). These choices relied on advice from Earth Observation researchers, used to designing networks to unmix pixels. They suggested the following design (Foody, pers. com.) (see section 5.5.1 for an explanation of the ANN terminology)

*Table 6.3 - Initial ANN Architecture and Formulae*

| ANN Feature | Initial Setting |
|---|---|
| Input Layer | number of nodes (n) = number of variables x number of categories within variables |
| Hidden Layer | number of nodes = 2n +1 |
| Output Layer | number of nodes = number of social groups |
| Activation Function | Act_Logistic (constrains to between -1 and +1) |
| Learning Algorithm | BackPropagation Momentum |

Appendix III provides an explanation of the back propagation formula used by SNNS. With the ANN set up and training data ready to be utilised, training began. Initially the ANN was trained on 100 epochs (cycles through the entire training set) and a close watch was kept on the error that the ANN was making. SNNS provide both detailed numerical accounts of the error reduction and a graphical account (a typical version of which is shown in figure 6.9).

*Figure 6.9 - Typical ANN Training Curve*




The ANN records its error in three numerical ways. There is a measure of the total sum of squared errors for the learning (SSE), the mean squared error (MSE), which is the SSE divided by the number of patterns (EDs) that the ANN encounters in one epoch and finally the SSE divided by the number of output nodes (SSE/O). The SSE/O was the most appropriate indicator of the level of error the ANN was making. The closer to 0 this value became, the better the

UNIVERSITY OF SOUTHAMPTON LIBRARY

ANN had learnt to predict target values from the training data. However, a very small SSE/O did not necessarily indicate that the ANN had acquired the desired knowledge. The ANN performance with real SAS data depended on its ability to *generalise*, which was assessed at the end of each training run by using the test data.

## *6.9 - Neural Network Testing*

Once trained, the ANN and its configuration could be saved for use in unmixing. To test its performance, the trained ANN was presented with the test data but not with the target values. The test data were passed through the network and the output values were saved, together with the target values. The two sets of figures were then compared in SPSS.

Two versions of the comparison were tried. The first calculated a correlation coefficient between the predicted and target values. This provided a measure of how well the network was modelling the target values and assumed that when summed, the predicted values for each ED would equal 1.0 (representing 100% of the population accounted for). The second version of the test constrained the predicted values to sum to 1.0, since the nature of the ANN activation function meant that the predicted values were unlikely to do so precisely. This adjustment was found to slightly raise the quality of the predictions. In the early development stages, this test was used to guide the operation and adjustment of the ANN. Details of these adjustments and further tests will be found in chapter 7.

## *6.10 - Using The Trained Network to Produce a Model*

Assume for the moment, that the ANN had now been optimised and trained. The last stage in the sequence from raw data to model is the application of the trained ANN to real ED data for the case study area (stage (7), figure 6.1). The appropriate SAS data for the 418 EDs in Southampton were acquired:

### 6.10.1 - Extraction of Real Small Area Statistics

The SAS data, like the SAR, were accessed on-line from their repository. Again, a syntax file specified the variables required and their format. The SAS data were obtained from Manchester University's data sets service using a dedicated package called SASPAC91.

*Figure 6.10 - A sample SASPAC91 Syntax File*

```
INPUT SYSTEM FILE NAME = hantsesh (looks at right data file)
INCLUDE EDLEVEL JP (selects EDs for Southampton only - there are 418)
SAVE VARIABLES S580008 S580015 S580029 S580036 (specify data)
OUTPUT DELIMITED FILE NAME = soton.dat (output file name)
END
```

The ANN training took place using synthetic SAS for EDs with 100 households in them. In order to make the real SAS compatible with the trained ANN, the SAS counts were scaled as if every ED in Southampton also contained 100 households. Once extracted and coded appropriately, the scaled SAS data were converted into a format compatible with SNNS using the same code which generated the training data files.

### 6.10.2 - Model Creation

Actually creating the model of socio-spatial structure from the trained ANN and SAS data was straightforward. The SAS data were passed through the ANN once, and the predicted proportional values for each social group, for each ED, were recorded. The ease with which the model was produced for so many EDs emphasised the benefits of using an ANN approach as opposed to repeated micro-simulation techniques (see section 4.4). It took about 3 minutes to produce the model for the 418 EDs which comprised the study area. The predicted values were extracted from the SNNS output file and converted to a text matrix. This matrix became the input for a number of visualisations and interpretations which will be explored in the following chapters. Figure 6.11 illustrates a small segment of it.

*Figure 6.11- Output Matrix from Unmixing*

| ED Code | %Grp 1 | %Grp 2 | %Grp 3 | %Grp 4 | %Grp 5 | ... |
|---|---|---|---|---|---|---|
| 25JPFA01 | 23 | 0 | 0 | 11 | 32 | ... |
| 25JPFA02 | 74 | 21 | 0 | 0 | 1 | ... |
| 25JPFA03 | 0 | 0 | 2 | 3 | 88 | ... |
| ... | ... | ... | ... | ... | ... | ... |

This final stage completed the path from variable selection to model, shown in figure 6.1.

## *6.11 - Summary and Conclusions*

The path from raw data to a working model has been described in terms of the techniques needed to complete each stage. The reader should now be familiar with the stages from data selection, through classification and separation of the household sample into groups, to production of training data, through ANN training and testing to model production. The following are the main points from this chapter:

- A set of variables was selected with which to build a household classification

- A classifier was custom written, because none of the available classifiers satisfied the demands of the research process as a whole

- The household groups resulting from the classification process were used to generate training and testing data for an ANN.

- An ANN was configured and trained to model the contents of synthetic SAS data in terms of the household groups present.

- The trained ANN was used to unmix real SAS data for the study area.

Experiments with techniques and data have led to improvements in this process and challenged some of the principles and practices which were initially adopted. This chapter has outlined the starting point for this development process. Chapter 7 will expand upon it.

# Chapter 7

## Chapter Overview

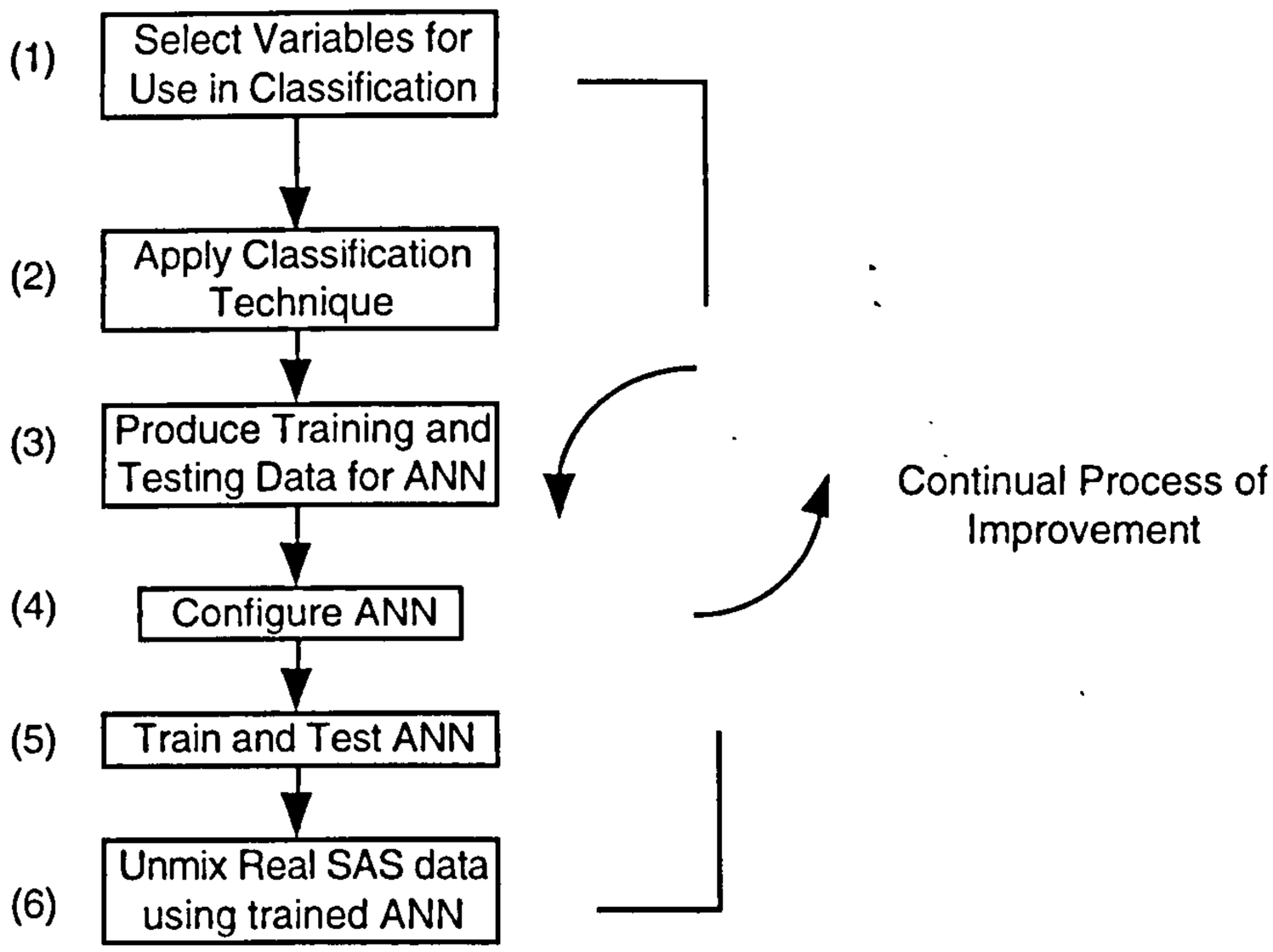
This chapter provides an account of the development of the unmixing technique which was introduced in chapter 5 and operationalised in chapter 6. A chronological structure is adopted and the reasons for this are made clear. Evaluation and adjustments to the initial variable set are made, before the production of a six group household classification. These six classes are used in the first attempt to unmix real Small Area Statistics (SAS) data. This step was intended to confirm that the unmixing approach was plausible, before a more complex classification scheme was developed. The six class model is shown to have been successful and the development of a more appropriate, twelve class classification is pursued. However, tests on the resulting model suggest that the classification is not performing as well as expected. Enhancements to the classification are proposed, but GroupMaker is shown to be incapable of providing the improvements. K-means cluster analysis is tested as an alternative means of group production and found to be successful. Unmixing proceeds with a new, fifteen class structure which results in the final model of Southampton (see figure 2.8).

## 7.0 - Introduction

This chapter employs a chronological structure to convey the real nature of the development process, rather than a rationalisation of it. In addition, the chronological approach assists in describing the cyclical nature of the development. The quality of the model produced by the unmixing technique is dependent on the quality of the classification scheme used and the training process. Thus, if a model is produced and testing suggests potential improvements, the entire process from variable selection to ANN training must be repeated. A chronological structure emphasises this and helps to present the stages of development in a coherent way.

The chapter is therefore divided into three sections, defined by the number of groups in the classification scheme under development ( six, twelve and fifteen). Details of experiments and the results from each stage are given. Reference will be made throughout the chapter to the flow chart shown below . This is intended to provide the reader with a context for each development or experiment described:

*Figure 7.1 - Integration in Principle - An Outline*




## *7.1 - The Sequence of development*

Pressure to determine whether the unmixing approach would work was clearly evident in the development strategy. The construction of a classification scheme with which to test the approach was a priority. This first classification scheme needed only to exhibit the characteristics which were thought to be necessary for the technique to work, i.e. to group households with similar census attributes together and separate them from households with different sets of attributes (see section 5.7.1). Once constructed, the scheme would be used to train an ANN in the manner outlined in previous chapters (5 and 6) and tests on a small number of Southampton's enumeration districts (EDs) would determine the plausibility of the model produced.

The development of a six class model was intended to confirm that the ANN approach offered a chance of success, or to reveal any fundamental flaws. In the latter situation, there would still have been time to explore other possible solutions whilst in the former, the approach could be fine tuned.

## *7.2 - Obtaining a Working Set of Variables*

An appropriate pool of census variables with which to build the classification had already been established (see section 6.1 and stage 1, figure 7.1). Exploratory work with GroupMaker and the visualisation system (see section 6.5.1) began. It was possible to explore the relationships between these census variables by initially allowing every variable to exert an equal influence on the grouping process. GroupMaker's flexibility was then utilised in controlling which variables took an active part in the clustering process, and which were carried along without influence. This allowed an understanding to be built of the significance of each variable within the clustering process. Two notable changes were made to the raw data as a result of these initial experiments.

### 7.2.1 - Variables with little specific influence in the classification

The variables 'educational attainment of household head', 'ethnic group of household head', 'presence of central heating' and 'number of cars in household' were all excluded from the classification process because experiments showed that they contributed little to it:

The household Sample of Anonymised Records (SAR) divides the educational attainment variable into 3 categories; no higher qualifications, one higher and two or more higher qualifications. Since the vast majority (87.9%) of the household heads in the SAR sample used fell into the first category, the variable was not exerting a strong influence on the classification. It simply contributed to the identification of those at one extreme end of the social spectrum. Combinations of other variables, usually including socio-economic group, were capable of separating the same households just as effectively.

The variable ethnic group of household head was also shown to have little influence on the classification because of the relatively small numbers of non-white or non-British persons in the SAR sample. This reflects the demographic character of The Rest Of The South East (RoSE) region. The relative affluence of the RoSE region might also explain the poor performance of the central heating and car ownership variables. Whilst they were effective at highlighting the small number of households at the extremes of the economic spectrum, they had little to contribute to the separation of the numerically dominant, 'middle class' groups.

### 7.2.2 - Variables with Redundant Categories

The discriminatory value of some *categories* within the remaining variables was also questionable. Experiments with household tenure for example,

suggested that the various categories of private renting had little independent discriminatory power. However, these categories were shown to be distorting the classifications obtained because of the manner in which GroupMaker calculated distance between households and therefore cluster membership (see 6.4.4). They were thus collapsed together and replaced by a single 'private and association renting' category.

*Table 7.1 - Variable Adjustments following Initial Experiments*

| Tenure As Recorded By the Census | Tenure Variable as used in the classification |
|---|---|
| Outright Ownership | Outright Ownership |
| Owner Occupier, Buying | Owner Occupier, Buying |
| Rented Privately Furnished | Rented or Association tenancy |
| Rented Privately Unfurnished | |
| Rented with a Job or Business | |
| Housing Association Tenancy | |
| Local Authority Tenancy | Local Authority Tenancy |

| Family Type as Recorded By the Census | Family Type variable as used in the classification |
|---|---|
| Married, No Children | Couples, No Children |
| Cohabiting, No Children | |
| Married, Non dependent Children | Couples, Non Dependent Children |
| Cohabiting, Non Dependent Children | |
| Cohabiting, Dependent Children | Couples, Dependent Children |
| Married, Dependent Children | |
| Lone parent, Dependent Children | Lone Parents |
| Lone parent, Non Dependent Children | |
| Not In Family (Single Person) | Not In Family |

A similar reduction was applied to the family type variable, as table 7.1 shows. The variable 'number of rooms' was also adjusted. It was top coded at seven rooms, (a value selected by determining the distribution of high numbers of rooms within the data set).

This process of rejection and recoding resulted in a new set of variables with which to classify the household sample:

*Table 7.2 - Variables for Six Class Scheme*

| |
|---|
| Socio-economic group of household head |
| Sex of household head |
| Family type of household head |
| Number of rooms in house |
| Tenure of household |
| Age of household |
| Number of economic active residents in household |

The representation of consumption themes appeared rather weak within this set of variables. This development was due to the performance of the consumption related variables in the classification process. This was itself a topic of interest since it suggested that the consumption based variables were not closely related to the socio-economic variables upon which class structures have traditionally been based (see section 3.10). Here lay an example of a direct conflict between the aim to adhere to theoretical constructs and not be constrained by technical issues, and the need to obtain a high quality classification. In this case the classification process would have been hampered by the theory.

It seemed probable that the constraints of the census questionnaire limited the utility of the consumption theme. Contemporary notions of consumption cleavages do not rely on the primitive, surrogate measures which the census provides (Pahl 1989). However, the experiences of Earth Observation (EO) researchers suggested that for the unmixing methodology to succeed, the quality of classification needed to be as high as possible. The decision to proceed with a smaller variable set was made knowing that the absence of those variables would represent a direct loss of information about the ED populations in the final model.

## *7.3 - Production of a Six Class Test Classification*

Using this set of variables, GroupMaker generated a six class classification scheme (stage 2 - fig. 7.1). The six classes were characterised as follows:

*Table 7.3 - Six Class Scheme: Class Characteristics*

| No. | Description |
|---|---|
| 1 | Younger families with mortgages, professional / managerial heads of household |
| 2 | Single, predominantly female, elderly |
| 3 | Families with mortgages, manual working heads of household |
| 4 | Young, single manual workers, council tenants |
| 5 | Families with manual working heads of household, council tenants |
| 6 | Young single women in employment, council tenants and mortgaged owners |

This was a crude classification of the sample population. However, sufficient differences existed between the groups to offer the ANN a chance of unmixing an ED, as proposed in chapter 5. The next section describes this first attempt at unmixing.

## *7.4 - The First ANN Unmixing*

The most important aims for the first phase of ANN testing were to establish (i) whether the ANN could be trained so that a significant correlation between target values and network predictions for class presence could be obtained and (ii) whether plausible proportional values were obtained when using real ED data for Southampton. The training and testing data were prepared as described in section 6.7 (stage 3, fig 7.1) and following the advice of EO researchers (Foody 1995 pers. comm.) the ANN was configured as follows (stage 4, fig. 7.1).

*Table 7.4 - ANN Configuration for Six Class Training*

| | |
|---|---|
| No. of Input Units (Census Variables) | 47 |
| No. Of Hidden Units | 15 |
| No. Of Output Units (Classes to be identified) | 6 |
| Weight Initialisation Function: | Random |
| Update Function: | Topological Order |
| Learning Function: | Back Prop. Momentum |

Different combinations of parameters were tested to determine their effect on learning performance, as shown in the table below. $\eta$ refers to the step width

of gradient descent, μ to the momentum term and c to the flat spot elimination value (see appendix III). The ANN was trained, and subsequently tested as described in 6.6. The next section describes the results obtained:

## *7.5 - Results*

*Table 7.5 - Results from ANN training using a Six Class Scheme*

| Value of | Test 1 | Test 2 | Test 3 | Test 4 |
|---|---|---|---|---|
| η | 0.2 | 0.8 | 0.8 | 0.8 |
| μ | 0.1 | 0.2 | 0.2 | 0.2 |
| c | 0.1 | 0.1 | 0.1 | 0.01 |
| No. Of Epochs | 100 | 100 | 200 | 200 |
| **Correlation** | **0.890** | **0.895** | **0.902** | **0.916** |
| Error Per Node | 0.96 | 0.878 | 0.788 | 1.4 |

The quality of ANN learning was difficult to ascertain. For each test, the first indicator of learning quality was provided by the amount of 'error per output node'. This value reflected the difference between the known class proportions in the training data set and ANN estimates for those values. The low error per output node indicated that the ANN had established a relationship between the input values and the target values. This statistic was *not* an effective indicator of the trained ANN's ability to generalise its learning to all situations, or of whether each of the classes had been modelled equally well. A correlation coefficient between the trained ANN estimates for each class and the known class composition of an *unseen test data set* provided a better indicator of the learning quality. This value was related to the ANN's ability to generalise its learning and the very high correlation values obtained in the tests suggested that the ANN was capable of unmixing test data sets with a reasonable degree of accuracy.

A more stringent test was the application of the trained ANN to real SAS data for a selection of EDs in Southampton (stage 6, fig. 7.1). Testing the ANN in this way achieved, for the first time, an increase in the available information about the residential population within an ED. It was recognised however, that the quality of the model was dependent not only on the ability of the ANN to unmix the SAS, but the applicability of the six class classification to the population living within those EDs. Given the crude nature of the classification scheme, some error in the model was expected. Tests took place on five EDs which were well known to the researcher:

### 7.5.1 - Testing the six class model on EDs in Southampton

*Table 7.6- Proportional Presence of Each Class in EDs (expressed as percentage)*

| Class | Northam | Bassett | Flowers | S. University | Lordswood |
|---|---|---|---|---|---|
| 1 | 1.07 | 75.75 | 17.87 | 78.96 | 44.62 |
| 2 | 87.5 | 9.54 | 15.76 | 5.69 | 23.33 |
| 3 | 4.16 | 4.02 | 12.05 | 4.10 | 6.24 |
| 4 | 2.80 | 2.80 | 21.79 | 4.58 | 8.29 |
| 5 | 1.30 | 3.81 | 17.38 | 2.45 | 10.94 |
| 6 | 3.17 | 4.09 | 15.14 | 4.23 | 6.57 |
| Total | 100% | 100% | 100% | 100% | 100% |

Table 7.6 shows the proportional presence of the six classes for five selected EDs in Southampton. Northam is an area of high rise council flats, homogenous housing and a high proportion of single person households. The description offered by the model was in keeping with the area's social character, although it is unlikely to be completely dominated by 'single, female pensioners'. The presence, albeit in small numbers, of groups 3, 4 and 6 may suggest that the social mix is more complex than the model indicated. Nevertheless, this was considered a reasonable account of the ED's population given the crude nature of the classification scheme employed. Bassett was presented by the model as being rather more socially mixed than it would appear to be. This is an affluent area on the northern outskirts of the city. The dominance of group 1 (managerial and professional families) was certainly accurate, and the presence of a retired population plausible too. The presence of groups characterised by council tenancy was unlikely. Again, this kind of model error was possibly a feature of the crude nature of the classification used.

The Flowers ED was correctly highlighted as the most socially heterogeneous. This is primarily an area of council housing, but significant levels of right to buy purchases have distorted the tenancy profile of the area. Whilst the ED is certainly not affluent, it is demographically and economically mixed. Given the width of class definitions used, an accurate judgement about the area's social mix was not possible. However, the technique clearly indicated that this is a socially mixed residential zone. A similar analysis for the remaining EDs was possible. In both cases the model offered a plausible account of the

residential population. It was concluded that the model had performed well given the relative width of the class definitions and therefore the difficulties in understanding the resident profile provided.

The unmixing technique certainly justified further development and testing. In particular, the experiment had shown that the quality of the classification would be paramount in determining the utility of the model produced. The next section explores how this first attempt at unmixing was developed through the construction of a more suitable classification scheme:

### *7.6 - Production of a 12 Class Scheme*

The next stage in developing the model was to determine an approximate number of classes to generate. Although work with six classes had provided ample evidence for the need to use a larger number, there were also several factors which suggested that the number of classes should be kept down. There was a need to balance the accuracy which a high number of classes might bring to model interpretation, with practical concerns. These concerns included the implications of greater number of classes for the production of training data and ANN training time.

*Figure 7.2 - Factors influencing the number of classes to be used*

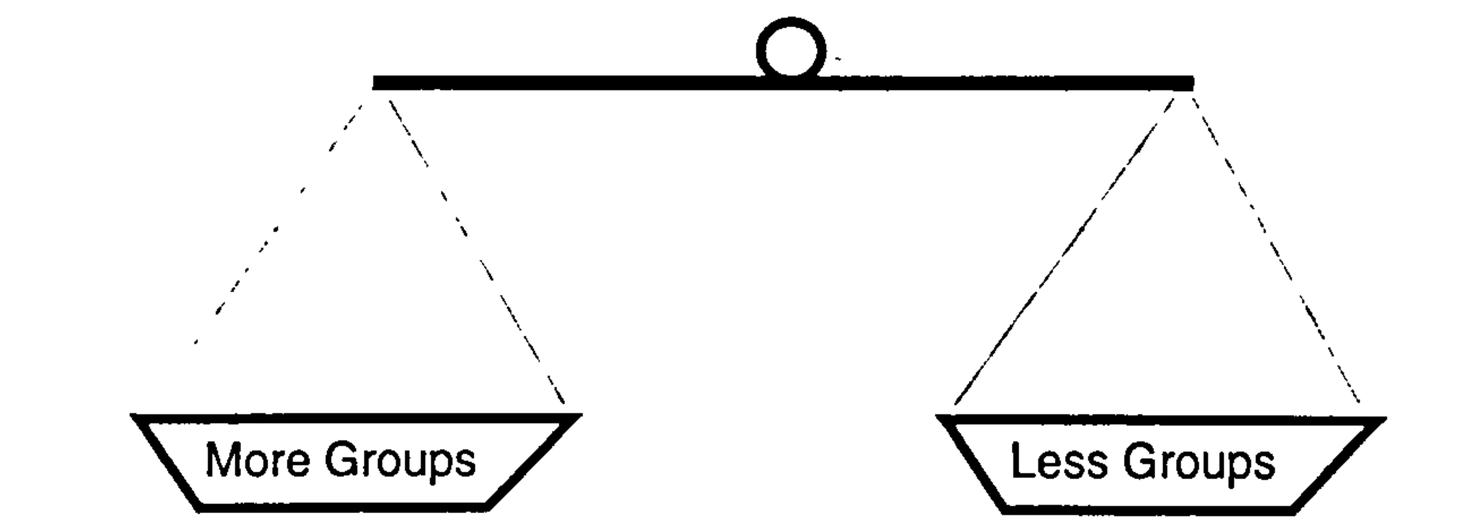



Figure 7.2 summarises the arguments for and against increasing the number of classes. The number of classes used related directly to the identity a household would have within the classification because this number would affect the precision with which a household's characteristics could be

described. A greater number of classes would assign the households a more specific identity and the desire to relate the identity of a household as clearly as possible has already been stated in sections 3.6 and 3.7. From a practical point of view however, the length of time taken to produce of training data for the ANN (see 6.7.2) was directly related to the number of classes involved and little was known about the ability of an ANN to handle large numbers of similar classes.

To determine a target number, other class schemes were reviewed, (the UK census and ACORN's most basic classification of households, for example). These suggested a number anywhere between 10 and 40 classes. Practical limits to the production of training data were also explored. A number of classes greater than 20 would have required more disk space and processing time than was available. A range of 10 - 15 classes was chosen, with the precise number selected to be determined by the quality of the scheme produced. This range was the product of a negotiation between technical constraints and the influence of current practice and theory. The next stage was to generate schemes with a greater number of classes within the specified range and determine their quality:

### 7.6.1 - Production of a Basic Scheme

Early experiments with GroupMaker had helped develop a successful approach to classification. This was to initially allow all the variables equal weight in the classification process and determine the distance threshold at which the required number of groups would be formed. From this starting point (at which the nature of the groups had been determined by the data themselves, rather than any subjective influence), the weights on each variable were adjusted to alter the nature of the classes produced. It was during this process that the influence of theory could be brought into the classification.

Table 7.7 details the initial weight configurations used (stage 2, fig 7.1). They corresponded to an equal contribution from all variables. The weights were derived by standardising the largest inter-household distance value which each individual variable could provide. This ensured that a variable with 19 categories (and thus a largest possible distance contribution of 18 units) would contribute the same maximum distance as a variable with just 5 categories.

*Table 7.7 - Initial Variable Weights in the Classification Process*

| Variable | Weight Value |
|---|---|
| Socio-economic group | 0.526 |
| Sex of Household Head | -2.0 (treat as categorical variable) |
| Family Type | -2.0 (treat as categorical variable) |
| Number of Rooms in Property | 1.428 |
| Tenure of Household | -2.0 (treat as categorical variable) |
| Age of Household Head | 1.428 |
| Number of Econ. Active Residents | 3.333 |
| Distance Limit | 36 |

These weights produced an eleven class structure which had several shortcomings. Too many groups highlighted elderly households, female headed households and those not in work. Previous analysis of the SAR data (see section 6.3) had shown that these groups were not numerically dominant in the sample. The classification scheme had not effectively highlighted or dissected the middle classes. Whilst the scheme had detected some of the polarisation in society, (the distinction between the home owning elderly and those reliant on council property for example), further influence was needed to make a more even division of households. This influence was brought to bear through manipulation of the variable weights.

### 7.6.2 - Manipulating the Basic Scheme

By reducing the influence of age group (i.e. lowering its weight so that two households with heads of differing age are less likely to be separated) and of sex, it was hoped that groups containing older household heads would collapse together. (Since a higher proportion of older household heads were female, reducing the influence of sex was necessary to help achieve this collapse). Separation and sub-division of the middle classes was attempted by raising the influence of property size and socio-economic group. These adjustments were reasonably successful although after several experiments the 'mortgaged families' still remained grouped together. By further raising the influence of family type, number of rooms in property and socio-economic group, whilst allowing the distance threshold to rise slightly (to take account of the increased weights contributing to the calculation) a 13 group scheme was produced, as shown in table 7.9. Table 7.8 provides the weights used.

The tabular arrangement in table 7.9 is intended to help the reader assess the nature of the groups. Where no mention is made of a specific variable, it may be assumed that this variable was not notably homogenous within a class.

*Table 7.8 - Variable Weights Used In An Interim Classification*

| Variable | Weight Value |
|---|---|
| Socio-economic group | 5.0 |
| Sex of Household Head | 2.0 |
| Family Type | 2.5 |
| Number of Rooms in Property | 3.428 |
| Tenure of Household | -2.0 (treat as categorical variable) |
| Age of Household Head | 1.0 |
| Number of Econ. Active Residents | 3.333 |
| Distance Limit | 43 |

This scheme was close to satisfying the requirements of the methodology. It reflected the ranges of household types, highlighting those at the extremes of the social scale whilst still providing a reasonable separation of the middle classes. It was thought that by raising the influence of tenure, the mixing of households exemplified by groups two and eight could be reduced. Given the significance attached to tenure by some class theorists (Saunders, 1987, for example), this was felt to be desirable. Group 12 needed to be dissolved because it only contained just one household.

Visualisation had also suggested that there was scope for sharpening the differences in the family types dominating each group. Accordingly, weights for family type and tenure were raised and the distance threshold adjusted to prevent greater numbers of classes being produced. The weights shown in table 7.10 were the result.

*Table 7.9 - An Interim Classification: Characteristics of Classes*

| Class number | Household | Economic | Demographic |
|---|---|---|---|
| 1 | Larger, mortgaged | Managerial / professional | Middle aged couple with children |
| 2 | Either avg. Outright owned, or avg. Council rented | Not applicable | Elderly |
| 3 | Average mortgaged | Manual / unskilled, 2+ ec. Ac. Residents | Young / middle aged couple with children |
| 4 | Larger mortgaged | Own account workers, 1 ec. Ac. Resident | Young / middle aged couple with children |
| 5 | Average outright owned | Skilled labour / non manual | Older / elderly couples without children |
| 6 | Small rented | Junior non-manual | Young single women |
| 7 | Larger mortgaged | Non-manual | Middle aged single women |
| 8 | Mixture of mortgaged and council rented | Skilled / semi skilled, 2 + ec. Ac. Residents | Middle aged couples, non dependent children |
| 9 | Council rented | Skilled / semi skilled | Middle aged couples with children |
| 10 | Rented and mortgaged | Managerial / professional | Young single |
| 11 | Mortgaged and council rented | Skilled / unskilled | Young / middle aged |
| 12 | Single member | | |
| 13 | Rented | Armed forces / own account workers | Young couples |

*Table 7.10 - Variable Weights Used to Produce 12 Class Scheme*

| Variable | Weight Value |
|---|---|
| Socio-economic group | 5.0 |
| Sex of Household Head | -2.0 (treat as categorical variable) |
| Family Type | 3 |
| Number of Rooms in Property | 3.428 |
| Tenure of Household | 2.0 |
| Age of Household Head | 1.0 |
| Number of Econ. Active Residents | 3.333 |
| Distance Limit | 65 |

### 7.6.3 - A Working Structure

Table 7.11 characterises the resultant classes. This structure seemed to satisfy the requirements for a classification. It provided a reasonable separation of the middle classes, whilst still identifying those at the extreme ends of the social scale. The single household group had been removed and distinctions between the groups were more clear. With a working classification now developed, the next stage was to refine an appropriate ANN configuration (stage 4, fig. 7.1) for unmixing. For this purpose, the classification was used to produce training and testing data as described in section 6.4. The next section explores their use in the development of the ANN configuration.

*Table 7.11 - Twelve Class Scheme: Class Characteristics*

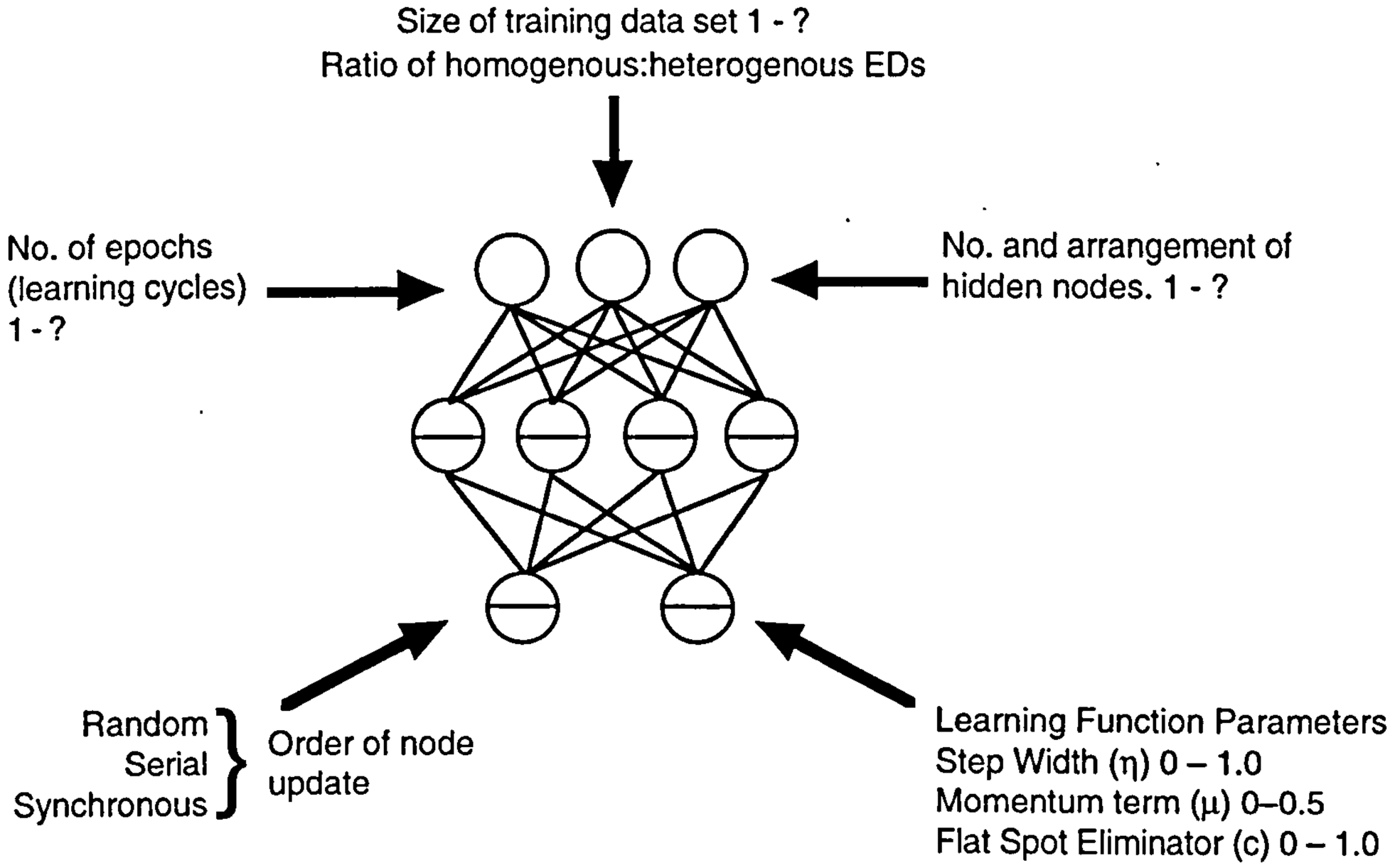
| Class No. | Household | Economic | Demographic |
|---|---|---|---|
| 1 | Larger, mortgaged | Managerial / professional, 2+ ec. ac. residents | Middle aged couples with children |
| 2 | Average, owned outright, some council tenancy | Not Applicable | Single, elderly females |
| 3 | Mortgaged, some council tenancy | Self-employed, or manual workers.2+ ec. ac. residents | Young couples with dependent children |
| 4 | Average, owned outright | Skilled manual workers, junior non-manual | Older couples without children |
| 5 | Larger, owned outright | Junior Non-manual | Older single women |
| 6 | Mostly average, mortgaged | Ancillary workers | Single people |
| 7 | Council tenants | Skilled / semi-skilled manual workers, 2+ ec. ac. residents | Couples with dependent children |
| 8 | Council tenants, some mortgaged owners | Non-manual workers | Lone female parents, some single women |
| 9 | Smaller mortgaged | Skilled / semi-skilled workers | Younger, single men |
| 10 | Privately rented | Skilled / Own Account workers, 2+ ec. ac. residents | Couples |
| 11 | Smaller, council tenants | Personal service / non manual | Young single women |
| 12 | Large, owned outright | Farmers / Own Account workers, 2+ ec. ac. residents | Couples without children |

## *7.7 - ANN Development*

Although the ANN had functioned well when trained on six classes, little ordered experimentation had been undertaken to establish what its optimum configuration might be. To achieve this, a methodical sequence of experimentation with the ANN was undertaken using the twelve class structure.

The number of adjustable ANN parameters and the potential range of those adjustments, made searching for the optimum network configuration a sizeable task. The diagram shows the elements of the ANN configuration and training process which could vary against each other.

*Figure 7.3 - Range of ANN Parameters With Influence over Performance*

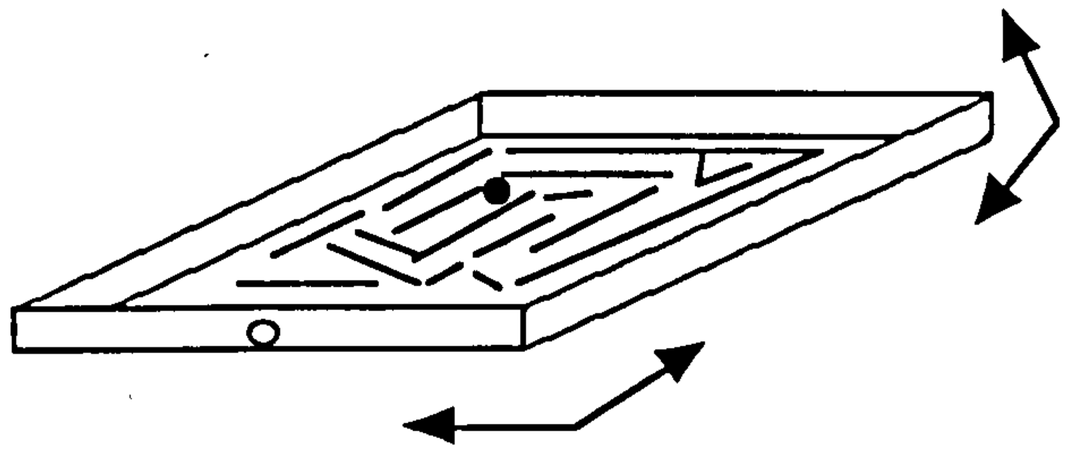



The strategy for exploring this potentially vast range of network configurations was i) to try and minimise the number of configurations to be tested and ii) to carry out experiments at night, using the ANN's batch processing system.

### 7.7.1 - Order of Experimentation

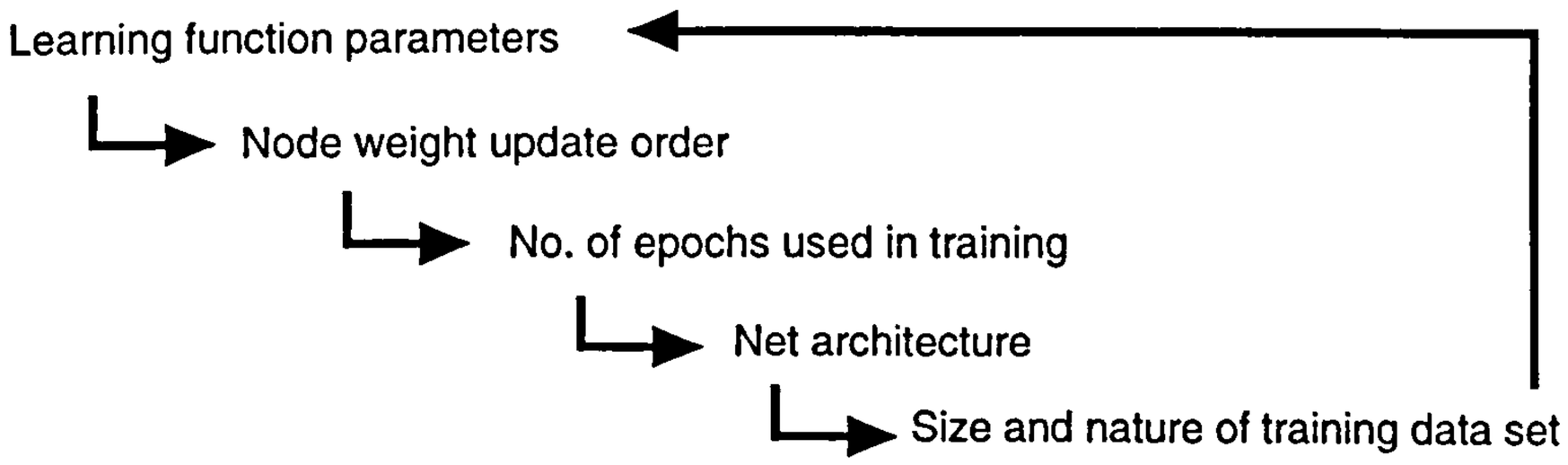
Training the ANN (stage 5, fig. 7.1) was analogous to playing the game in which a steel ball must be guided through a maze by tilting the maze and trying to influence the way the ball will roll. The progress of the ball towards the finish was akin to the quality of ANN learning achieved, the tilting of the maze akin to the adjustment of network parameters.

*Figure 7.4 - The ANN Training Maze*



The difference between the game and training an ANN is that in the game, the route through the maze is visible to the player. With the ANN, the parameters could be changed with only limited prior knowledge of the effect. Experiments focused on exploring one group of parameters at a time. An apparently optimum setting for that group was determined and then retained whilst altering values for the next group of parameters. Figure 7.5 illustrates the order and cyclical nature of this process:

*Figure 7.5 - Order of adjustment in search for Optimum ANN Parameters*




Appendix IV contains details of the tests carried out. It provides a description of the various combinations of parameters tested, together with correlation value (between know class proportions in a test data set, and the ANN estimates) for selected combinations. The following section discusses the key findings from these experiments:

### 7.7.2 - Key Findings From Experiments with ANN Parameters

A number of characteristics of the ANN training process became clear during the experiments (see 5.5.1 for a reminder of ANN terminology). Firstly, the performance of different combinations of ANN parameters was polarised. Either the combination virtually prevented learning taking place, or the combination appeared to achieve excellent results. There was no steady ascent of a learning quality curve during the sequence of experiments. Experiments showed that some parameters and algorithms were definitely not suited to

this purpose and that the order in which node weights were updated was significant. Both random and synchronous update order appeared to dramatically reduce ANN performance, as did the arrangement of the hidden nodes into two connected layers, rather than a single one.

The most significant enhancements to ANN performance were brought by adjusting the number of hidden nodes in the architecture and the number of epochs. The optimum balance between these two parameters gave the ANN maximum learning experience of the training data, whilst retaining its ability to generalise that learning for use with unseen data. Experiments using very small numbers of hidden nodes to very large numbers and 50 epochs to 1000, showed that the ANN appeared to perform best with 95 hidden nodes and 200 epochs.

Training used a data set of 538 synthetic EDs (see section 6.7) and a test data set of 174 synthetic EDs. Consistently using the same training and testing datasets would have provided a reliable basis for comparing the performance of the ANN when trained in different ways. However, having used the same training and testing datasets to optimise the ANN parameters, performance was found to be lower than expected when tested using different data sets. Varying the training and testing data and ensuring a spread of homogenous and heterogeneous EDs was found to be important. These experiments led to a convergence on the following set of parameters:

*Table 7.12 - Optimal ANN Parameters for a Twelve Class Model*

| Network Parameter | Value |
|---|---|
| $\eta$ | 0.8 |
| $\mu$ | 0.2 |
| c | 0.01 |
| No. of Epochs | 200 |
| No. of Hidden Nodes | 95 |
| Weight Update order | Topological |
| Initialisation Function | Random |

A typical correlation between the known class values in the test data set and the ANN's estimates was 0.93. However, the performance of individual classes typically remained variable as shown in table 7.13.

*Table 7.13 - Typical Correlation Values ($R^2$)between ANN Output and Known Target Values, by class*

| Class Number | Value | Class Number | Value |
|---|---|---|---|
| 1 | 0.896 | 7 | 0.887 |
| 2 | 0.939 | 8 | 0.931 |
| 3 | 0.798 | 9 | 0.698 |
| 4 | 0.838 | 10 | 0.899 |
| 5 | 0.817 | 11 | 0.909 |
| 6 | 0.947 | 12 | 0.784 |

The table shows quite significant variation in model performance, between classes. This performance varied from test set to test set and despite further experimentation, consistently high performance remained illusive.

A decision was made to carry out an analysis of several real EDs in Southampton using the trained ANN. An appraisal of the model would then determine whether further work would be carried out (stage 6, fig. 7.1).

## *7.8 - Testing on Southampton EDs*

SAS data were prepared for the five EDs used in the previous test (section 7.4.2) and three additional EDs, whose social composition was thought to be more heterogeneous. Appropriate SAS data were prepared for the ANN and passed through it. The resultant unmixed model was then assessed using local knowledge to determine if the model had produced a plausible account of the EDs population and a number of model parameters which indicated its quality in less subjective manner.

The twelve class model performed reasonably on the EDs selected for testing. It gave a plausible account of the contents of each ED which was in broad agreement with the researcher's subjective assessment of the kinds of households resident in each area. However, two aspects of the model's performance gave rise to concern.

### 7.8.1 - Failure to Account for All Variation

Firstly, the values output from the ANN were not consistently summing close to 1.0. Had they done so, it would have indicated that the ANN could account for all the variation within the SAS for the EDs under test. Failure to consistently reach this target suggested that a degree of uncertainty remained

in the model and that it was unable to offer a complete explanation for the EDs' populations. There were two possible explanations for this difficulty.

1) There was no guarantee that all the households in Southampton possessed a combination of characteristics also present in the SAR sample, from which the class scheme had been derived. The ANN is trained on the basis that the ED data with which it is presented will only describe households that fit into the classification scheme. When it encounters aggregate data which contain households of an unfamiliar nature, error may occur. 2) ANN training may have been insufficient to allow generalisation of learning to all situations. Correcting this deficiency was more likely to require a larger training data set, or a change in network architecture.

### 7.8.2 - Difficulty of Model Interpretation

The second aspect of concern was over difficulties in interpreting the model. A clear indication of the extent of social mixing inside an ED could indeed be obtained from the model. However, where some classes contained a mixture of household types themselves, precise identification of the characteristics of an ED's population was hampered which might lead to mis-classification of the ED's population. Given that one intended advantage of the unmixing method over the wholesale classification of an ED was to be the identification of small pockets of 'hidden households' (see section 2.11) it was necessary to insure that the classes could be clearly interpreted.

It was clear that the model needed further refinements before an analysis of the whole study area could be attempted. The next section describes how those refinements were achieved.

## *7.9 - Refinements*

Refinements to the model were intended to solve the problems outlined above, as far as possible. However, differences between the range of households covered in the SAR sample and those present in Southampton was a problem without a direct solution. There could be no change to the SAR sample itself and it was felt that utilising a wider sample (by including records for outside the Rest of the South East (RoSE)) might equally well introduce household types not appropriate for Southampton (see section 6.1). The RoSE sample offered the most appropriate base from which to build a household classification scheme.

Exhaustive experiments with the ANN training process also suggested that any enhancements to model performance would be unlikely to come from altering learning parameters. Refining the classification scheme was the only option which remained. By attempting to give classes a more precise definition, the interpretation of the model might be made easier and ANN performance might also be enhanced as each class was made easier to distinguish. One means of making the classification scheme more precise was to increase the number of classes within it, thus reducing the need for less similar households to be grouped together.

### 7.9.1 - Further Classification

The balance between the number and quality of classes in the classification scheme was now confirmed as a key aspect of the methodology explored in this research. Experiments with GroupMaker had clearly demonstrated the relationships between variable structure and the resultant classification. To significantly change the nature of the classification, the variables used in the grouping process needed to be changed. This, 'back to the beginning' approach to the refinements took the development back to stage 1 in figure 7.1.

Three changes were made to the variables used. In two instances the number of categories within a variable was reduced. The variable 'number of rooms in house' was collapsed from seven categories (1,2,3....7+) into three categories (1-3 rooms, 4-6 rooms, and 7+ rooms). This action was intended to make the groups more homogenous by reducing the range through which house size could vary. This was clearly a reduction in the detail of information available about the ED population, but did not represent a loss of discriminatory power of the variable. Work with the SAR had revealed that households *do not naturally cluster into a convenient number of classes*, within which there is little variation, but between which distinct boundaries may be drawn (see sections 3.8, 6.3, 7.3 and 7.5). In order to construct a classification scheme which provided this kind of division, it was necessary to reduce the number of ways in which household characteristics could vary. The variable 'household tenure' was also further collapsed from four categories, to three. The outright owner occupied category was merged with the owner occupied buying category, creating a single 'private ownership' category. Again, this collapse represents a significant loss of detail within the model, and it might be argued (by Saunders 1987 for example) the removal of a significant distinction between the economic circumstances of different households. The loss of detail was offset by the chance to enhance the quality of the classification and interpret the model with more certainty, if less specificity. With this decision

and the arguments associated with it, came a clear indication of the difficulty of trying to build a general purpose classification scheme. Through the reduction of the number of categories in a variable, it would be possible to interpret the unmixed model with greater certainty, because the variation within a specific class has been reduced. However, if tenure is of specific interest, the reduction in the information available for a specific class might render the model less useful.

Decisions about which variables to collapse and which to preserve, should be purpose led. The key question must be the purpose for which the model is intended. If, for example, the aim of an analysis was to identify pockets of housing owned outright by a specific socio-economic and age group, the ability to identify those elements in a class would be paramount. The classification need not be so complex for a more general query concerned with the relationship between social and private housing and other demographic characteristics. For the purpose of developing the technique, a balance has been sought between model complexity and the quality of interpretation.

The third change to the variable set was the inclusion of a variable to distinguish between different property types. This was prompted by walking in residential areas around the University in Southampton and noting that quite distinct property types housed people whose demographic or economic characteristics might appear similar from the census, but whose lifestyle or lifestage were clearly different. In a sense, it returned an element of the consumption concept (see section 3.10) to the classification. This variable replaced 'number of economically active residents', which was felt to be the weakest of those used previously in terms of its discriminatory power.

*Table 7.14 - The Variable 'Household Space Type'*

| Household Space Type |
|---|
| 1 - Unshared, purpose built; detached, semi-detached or terraced |
| 2 - Unshared, purpose built flats |
| 3 - Unshared converted |
| 4 - Unshared, not self contained |
| 5 - Other spaces, not self contained |
| 6 - Non permanent accommodation |

### 7.9.2 - Experiments and Failure

Experiments with this new variable set proceeded in a similar manner to that previously outlined in 7.5, (stage 2, figure 7.1).

*Table 7.15 - Variable Set Used for Further Classification Experiments*

| Socio-economic group |
|---|
| Sex |
| Age |
| Family Type |
| Roomsnum |
| Tenure |
| Household Space Type |

The variables were initially allowed an equal influence on the classification process (see section 7.6.1). At a distance threshold value of thirty, a fourteen group classification was produced. Whilst this classification covered the social scale quite effectively, only four of the groups were particularly homogenous. The tenure and socio-economic group variables were particularly mixed in a number of classes. A process of adjustment, as described previously (see section 7.5) was undertaken, but with unsatisfactory results. It seemed possible to either (i) produce a classification in which a sensible social range was reflected in the groups, but in which the groups themselves remained more heterogeneous than was desired, or (ii) to produce a classification in which many of the groups were extremely homogenous, but which did not deal effectively with the range of household types.

The twelve class scheme (see table 7.11) had provided a division of the SAR sample which efficiently reflected the range of household types present within it. It was intended that by controlling class formation in the same way as before, but allowing higher numbers to form, the same social range would be illuminated with more homogenous classes. Extensive tests with GroupMaker suggested that this was not going to be the case. The households would not cluster themselves, or be forced to cluster, in the desired way. Further changes to the variable set would have been possible using the understanding of variable co-variance gained from using GroupMaker. However, the changes already described represented the limit beyond which variable adjustment would make the class definitions meaningless. The reason for undertaking this work was to provide *more* information about

households inside EDs (see chapter 2). Further reductions in the specificity of class definitions would have compromised this aim.

In seeking a possible explanation for this difficulty and thus a possible remedy, attention was turned to how GroupMaker actually formed its classes (see section 6.4.3). GroupMaker used the most common sets of household characteristics as the basis for its clustering. Once a household was utilised as a 'cluster centre', or 'group definer' the centre did not change. Every other household in the sample (barring the group definers) was compared with this one to determine if it should join the group or not. The classification process did not adjust the centre against which classification decisions were made as the group grew. So, although GroupMaker did provide excellent control of how each variable contributed to the classification process, it was not capable of adjusting the set of characteristics to which a specific household would be compared, based on the whole character of a growing group. This might explain why certain household types were being separated into several small but similar groups. If a number of households with distinct features formed group definers early in GroupMaker's processing, those groups would be maintained and attract other households. A technique which proceeded through an iteration process, re-examining the distances between cluster centres and between centres and cluster member households, might well have collapsed those households together and made more meaningful distinctions. To test this idea, it was decided to run the SAR sample through a k-means classifier. Whilst this kind of classifier did not offer the same flexibility as GroupMaker (see discussion in section 6.4), it did utilise the kind of iteration discussed above.

## *7.10 - K-Means Cluster Analysis*

The classifier is available in the SPSS software package (see section 6.4.1 for details). The k-means algorithm was instructed to produce fifteen groups from the SAR data set (which remained the same as that shown in table 7.15). SPSS produced the clusters in about 8 minutes.

### 7.10.1- Cluster Results

Visualisation of the groups (as described in section 6.5.1), showed that a classification scheme which satisfied the requirements had been produced. Not all of the groups were entirely homogenous (see group 12, table 7.16, for example) but most were. The classification gave a good account of the range of social positions described within the data. However, it should be noted that the basic nature of the groups differed little from that produced by GroupMaker.

The difference lay in the quality of the clustering, with the k-means approach able to adjust the groups until they reached a level of homogeneity that GroupMaker could not have achieved. The resulting classes can be characterised by the descriptions in table 7.16, overleaf. Table 7.17 illustrates the number of households from the SAR sample, in each group.

*Table 7.17 - Numbers of Cases in Each of the Fifteen Groups*

| Group No. | No. Of Cases | Group No. | No. Of Cases |
|---|---|---|---|
| 1 | 5453 | 9 | 1968 |
| 2 | 2506 | 10 | 2530 |
| 3 | 5242 | 11 | 1770 |
| 4 | 4372 | 12 | 1840 |
| 5 | 4157 | 13 | 1373 |
| 6 | 2985 | 14 | 386 |
| 7 | 3129 | 15 | 199 |
| 8 | 3133 | | |

## *7.11 - ANN Training with Fifteen Classes*

With a different classification scheme now in place, stages 3 - 6, in figure 7.1, needed to be re-visited. Training and testing data were constructed for the new scheme, as described in section 6.7 (stage 3, figure 7.1). Using the extensive experience of ANN training gained in earlier stages of the research, an optimal set of ANN parameters and architecture was quickly established (stage 4, figure 7.1). This differed little from that used with twelve classes:

*Table 7.18 - Optimum Parameters for Use with a Fifteen Class Model*

| Network Parameter | Value |
|---|---|
| $\eta$ | 0.8 |
| $\mu$ | 0.2 |
| c | 0.01 |
| No. of Epochs | 100 |
| No. of Hidden Nodes | 91 |
| Weight Update Order | Serial |
| Initialisation Function | Random |

*Table 7.16 - Descriptions Of The Fifteen Classes Produced By Cluster Analysis*

| No. | Housing | Economics | Demographics |
|---|---|---|---|
| 1 | Avg. size, privately owned houses, some flats | Retired | Single elderly females |
| 2 | Avg. size, privately owned houses | Retired | Elderly childless couples |
| 3 | Avg. size, privately owned houses | Skilled / semi-skilled employment | 30's couples with dependent children |
| 4 | Avg. - large, privately owned houses | Managers | 30's couples with dependent children |
| 5 | Avg. - large, privately owned houses | Professionals / ancillary workers | 30's couples, some with dependent children |
| 6 | Avg. size, privately owned houses (some council tenancy) | Skilled / semi-skilled employment | 50-60's couples without children |
| 7 | Avg. size, privately owned houses / flats (some council tenancy) | Manual labour | 40-50's single men |
| 8 | Avg. size, privately owned houses | Junior non-manual employment | 30's couples, without children |
| 9 | Avg. - large privately owned houses | Managers and professionals | 50-60's couples without children |
| 10 | Avg. size, privately owned houses / flats (some council tenancy) | Junior non-manual employment | 20-40's single females, some lone parents |
| 11 | Avg. size, privately owned houses | Professionals and ancillary workers | 20-30's single people |
| 12 | Avg. size, privately owned houses / flats, some private renting | Either no socio-economic group or service workers | 30-40's couples with dependent children or lone parents |
| 13 | Avg. size, privately owned houses / flats, some private renting | Mangers | 20's-40's single people |
| 14 | Small rented, shared property | Skilled / semi-skilled manual or junior non-manual workers | 20-30's single people |
| 15 | Small, privately owned / rented non permanent accommodation | Retired | Single elderly |

This configuration was used with a training data set of 750 synthetic EDs. Following testing on a further 300 unseen EDs, the ANN performance was assessed: The correlation between ANN predictions and known class values was 0.963, with network output adjusted to sum to 1.0. Analysis of how close to 1.0 the summed ANN outputs were revealed a range of 0.725 to 1.325, with an average of 0.984. The excellent performance of *every* class is show by table 7.19:

*Table 7.19 - Correlations between Target Values and ANN Outputs, by Class*

| Class Number | Value | Class Number | Value |
|---|---|---|---|
| 1 | 0.97 | 9 | 0.96 |
| 2 | 0.97 | 10 | 0.96 |
| 3 | 0.96 | 11 | 0.97 |
| 4 | 0.95 | 12 | 0.96 |
| 5 | 0.97 | 13 | 0.97 |
| 6 | 0.95 | 14 | 0.96 |
| 7 | 0.95 | 15 | 0.96 |
| 8 | 0.96 | | |

These results suggested that the ANN was accounting for a very high proportion of the variation within the EDs, and that it was accurately reproducing the class compositions of EDs with which it was presented. It seemed likely that this enhanced performance was achieved through the use of a classification scheme consisting of more homogenous groups as a result of the iterative nature of the k-means classifier and the reduced variance within the initial data set.

## *7.12 - Forward to Testing*

Given the quality of the results obtained with this new ANN a decision was made to proceed to testing the resulting model on the City of Southampton as a whole. A model for the entire city was produced by preparing real SAS data for every ED within the city boundary as described in 6.10.2. Each set of SAS were passed through the ANN and a set of output values recorded. The visualisation and analysis of that model is presented in the following chapter. At this point, the technical development within the research was assumed to have been completed. Only the discovery of significant shortcomings or model error through a city-wide analysis would prompt a re-development.

## *7.13 - Conclusions and Summary*

The research presented in this chapter has built a path from aggregate and disaggregate census data sets, to a model of social structure within an ED. No insurmountable technical obstacles were encountered in the translation of the concept, outlined in chapter 5, to a practical reality. There was little difficulty in adapting the ANN to carry out this task. The concept of unmixing an area into its constituent classes has, in technical terms at least, been successfully translated from a different academic field, into social geography. Chapter 8 will reveal if the success has gone beyond the ability to operationalise the concept.

The significance of the classification scheme in this success, and in the utility of the final model cannot be underestimated. There was no prior expectation that the SAR sample would cleave easily into homogenous groups of distinct character. However, the difficulties experienced in producing a classification which offered the ANN a chance of successful learning were greater than prior estimations had allowed for. This difficulty may be explored from two points of view. Firstly it suggests that the concept of society as divided into many distinct groups, with mutually exclusive membership, may be outmoded. This work did not find strong evidence for a systematic structure which is reflected across multiple characteristics. Secondly, it suggests that the means of measuring these divisions may need to be re-calibrated to explore other variables along which these divisions may be more apparent. The attributes recorded in the census did not provide a good basis upon which to explore society's structures.

It was stated in section 3.13 that technical decisions in this research would be the product of a negotiation between what was conceptually or theoretically desirable and practically possible. The work described in this chapter, more than in any other chapter, was the theatre for those negotiations. What has become clear is that those negotiations have often been extremely brief and one sided. In order to develop this technique, the classification scheme had to conform to a particular specification. Attaining that specification required an almost entirely technocentric approach to the classification. Manipulation and adjustment was undertaken because it was known or hoped that this would bring the classification closer to the required specification. Prior to those adjustments, their impact on the nature of household representation was considered but at no point did this consideration alter the technical process subsequently undertaken.

These considerations led the researcher to the conclusion that whilst the negotiation process was valuable and revealing, the impact of its lessons needed to be elsewhere. They need to have an impact at the model interpretation and employment stages where an understanding of the limitations of the model and how it represented people and households could be most effectively utilised. They could also bring tremendous benefit if utilised when planning primary data construction, as for example with the UK census. Class theory and the possibilities for the empirical exploration of class are two aspects of an important field which remain very distant from each other. Clarification of the impact of that distance would greatly enhance the negotiations which give rise to data sets. These points will be expanded upon in chapter 9.

In summary then, the key points from chapter 7 are these:

- The path from raw data to a model of socio-spatial structure has been completed in principle and in practice.

- Classification of the SAR proved difficult. GroupMaker was enormously valuable in the exploration of household structures and in gaining an understanding of the significance of different household characteristics for class structures.

- A classifier which was able to adjust its cluster centres proved more effective at building a structure that satisfied the rigorous demands of the unmixing methodology. The *nature* of the groups it produced did not differ from GroupMaker but the classification was of better *quality*. The quality of the classification was significant for both the ANN learning and the ease of model interpretation.

- A set of optimal parameters for the use of a feed-forward ANN to unmix aggregate census data has been established. The ANN learning process has been assessed as far as possible and found to have been successful.

From here, the research progressed to an evaluation of the model for Southampton as a whole. This is described in chapter 8.

# Chapter 8

## Chapter Overview

In this chapter the model of Southampton's socio-spatial structure is tested and explored. The difficulties of testing a new kind of information are discussed and a three-fold approach to overcoming those problems is proposed. Comparison with an existing model is described, followed by details of field research which sought to determine if the model was congruent with the character and appearance of a sample of areas. These subjective assessments are supplemented by an attempt to re-construct the original small area statistics for each enumeration district in Southampton, using the unmixed model for guidance. The quality of the reconstruction serves as an indicator for the quality of the model. Having established the model's success, an exploration of what it reveals about Southampton and its residents follows. A number of new types of analysis and information are presented. Finally, the chapter speculates on the potential applications and developments of the unmixing technique. This review is based only upon the possible applications and developments in which the aggregate unit would be spatial. The removal of bounded spatial units from the analysis and the application of the unmixing technique to aspatial contexts is discussed, together with the wider conclusions of this research, in chapter 9.

## 8.0 - Introduction

Two questions remain unanswered about the model of Southampton's socio-spatial structure: how accurately has the city's population been represented, and what new information does the model reveal about that population? This chapter seeks an answer to those questions. Since an answer to the first question will determine the reliability of an answer to the second, the two will be tackled in order. Once the technique's credibility has been established, the final section of the chapter suggests a range of possible developments for the technique and a number of scenarios in which it might find application.

## 8.1 - How Accurate is the Unmixed Model?

In section 2.11 it was suggested that the model represented a *new* level of information about an ED's residential population. This new and unique nature made evaluation problematic. When Earth Observation (EO) researchers wish to evaluate the accuracy with which pixels have been unmixed, they can usually turn to imagery with a finer spatial resolution or

rely on field observations (see section 5.6) to provide data for a valid comparison with their model. In the application of their techniques to census data, there were no other data against which to test the resultant model, making a robust evaluation difficult. A threefold approach to assessment was adopted as an alternative to comparison with finer resolution data. Each approach will be introduced briefly before a detailed account of the results.

Firstly, single area classifications for each enumeration district (ED) were derived from the unmixed model (UM model). These were then compared with a geodemographic classification of Southampton's EDs (see section 2.8). This comparison was intended to determine broadly whether the UM model's account of Southampton's socio-spatial structure was similar to that provided by a conventional technique. Secondly, a number of EDs were visited by the researcher and the UM model's account of each area was subjectively assessed. This approach was intended to determine if the model made intuitive sense 'on the ground'. Thirdly, an attempt was made to recreate the original Small Area Statistics (SAS) values for each ED by sampling from the household level Sample of Anonymised Records (SAR). Estimates of class presence from the UM model were used as a constraint in the sampling process. The proximity of the recreated values to the originals served as a measure of the model's quality. The following sections describe each approach in detail.

### 8.1.1 - Comparison With A Single Area Classification

Geodemography (as described in 2.8), provides a single classification for each ED, based on appropriate aggregate statistics. To compare a geodemographic classification of Southampton with information provided by the UM model, a geodemographics package was used. GBprofiler is one such package available for use by the academic community (Blake and Openshaw 1994) and it offers a range of class numbers with which to work, from ten to one hundred. The ten class scheme was chosen for this analysis since its classes had characteristics closest to that used by the UM model. A comparison of the general social morphology which the models describe was thought to be a valuable guide to the plausibility of the UM model.

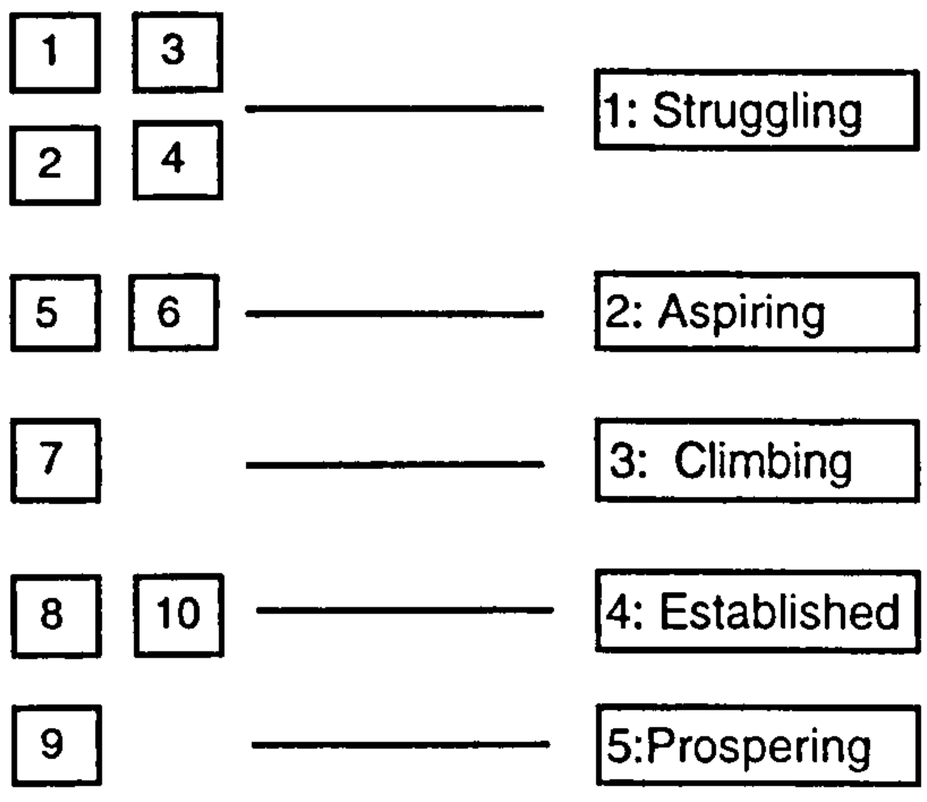
A single classification value for each ED was required from the UM model for comparison with GBprofiler's model. Each ED was therefore assigned a classification in accordance with its *proportionally dominant class*. For example, the ED shown in table 8.1 would be assigned to class 14.

*Table 8.1 - An Illustration Of The Proportionally Dominant Class*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ED | 14.8 | 9.2 | 7.7 | 12.6 | 3.4 | 0.7 | 1.2 | 0.15 | 0.5 | 0.8 | 4.7 | 3.4 | 11.8 | 22.4 | 6.0 |

This means of classification was certainly crude, however since the comparison with GBprofiler could not be made with any real precision, it was regarded as sufficient. This lack of precision in the comparison was a result of the difficulty in trying to match the classifications used by the two models. It was not possible to directly equate any class or group of classes used by GBprofiler with those used by the UM model because the classifications were derived from different variable sets and through different techniques. Without a direct relationship between the two classification schemes, any numerical analysis of the differences between GBprofiler and a single classification derived from the UM model would have been meaningless. Although the classes used by GBprofiler and the UM model were not directly comparable, both systems attempted to account for the range of household types in society and *some* similarity was evident. To facilitate this crude comparison, the two classification schemes needed to be matched as closely as possible and a process of aggregation was employed to achieve that match. The ten class classification from GBprofiler was collapsed to a five class system using five key-words drawn from the class descriptions; struggling, aspiring, climbing, established and prospering (see table 8.2 for a review of the class characteristics of GBprofiler's scheme).

*Figure 8.1 - Key to GBprofiler's Ten to Five Class Collapse*



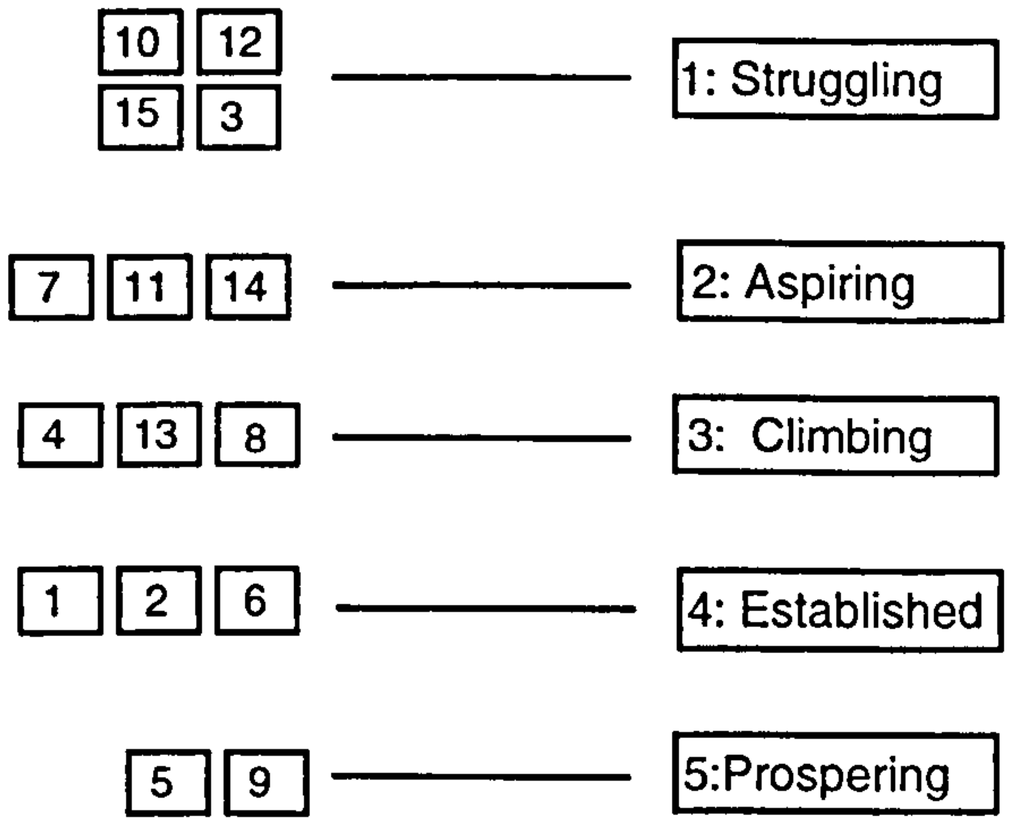

The fifteen class scheme used by the UM model was also collapsed subjectively to a five class scheme. The collapse attempted to match the GBprofiler five class scheme as closely as possible with the key words again used as a guide.

*Table 8.2 - Class descriptions for GBprofiler (copy of table 2.1 for convenience)*

| Class | Description |
|---|---|
| 1 | Struggling; multi-ethnic areas - pensioners & single parents - high unemployment - LA rented flats |
| 2 | Struggling; Council tenants - blue collar families & single parents - LA rented terraces |
| 3 | Struggling; Less prosperous pensioner areas - retired blue collar residents - LA rented semi's |
| 4 | Struggling; Multi ethnic areas - less prosperous private renters - young blue collar families with children - privately renting terraces and bedsits |
| 5 | Aspiring; Academic centres and student areas - young educated white collar singles and couples - privately rented bedsits and flats |
| 6 | Aspiring; Young married suburbia - young well-off blue collar couples and families - mixed tenure terraces |
| 7 | Climbing; Well-off suburban areas - young white collar couples and families buying semi's and detached houses |
| 8 | Established; Rural farming communities - Mature well-off self employed couples and pensioners - owning or privately renting large detached houses |
| 9 | Prospering; Affluent achievers - mature educated professional families - owning and buying large detached houses |
| 10 | Established; Comfortable middle agers - mature white collar couples and families - owing and buying semi-detached houses |

An ED thus assumed a new classification based on its dominant class *and* the new grouping into which that dominant class had been placed.

*Figure 8.2 - Key to The Unmixing Classification, Fifteen To Five Class Collapse*




(see table 8.3 for a review of the UM model's class scheme. This table is a copy of 7.16 provided for convenience).

The process of collapse resulted in two loosely comparable classifications. When mapped, the two versions of Southampton's social morphology could be compared (see figure 8.3).

Visual comparison of these two images revealed a number of interesting features. Firstly, the UM model showed greater local variability in its classification, (i.e. more differences between neighbouring EDs). The GBprofiler model divided the city into broad segments, with local variation more marked on the city fringes. The pattern described by GBprofiler was in keeping with the researcher's knowledge of the area in general terms; higher status areas are found in the north and north east of the city, lower status areas in the south and south central and in outlying areas to the east and west.

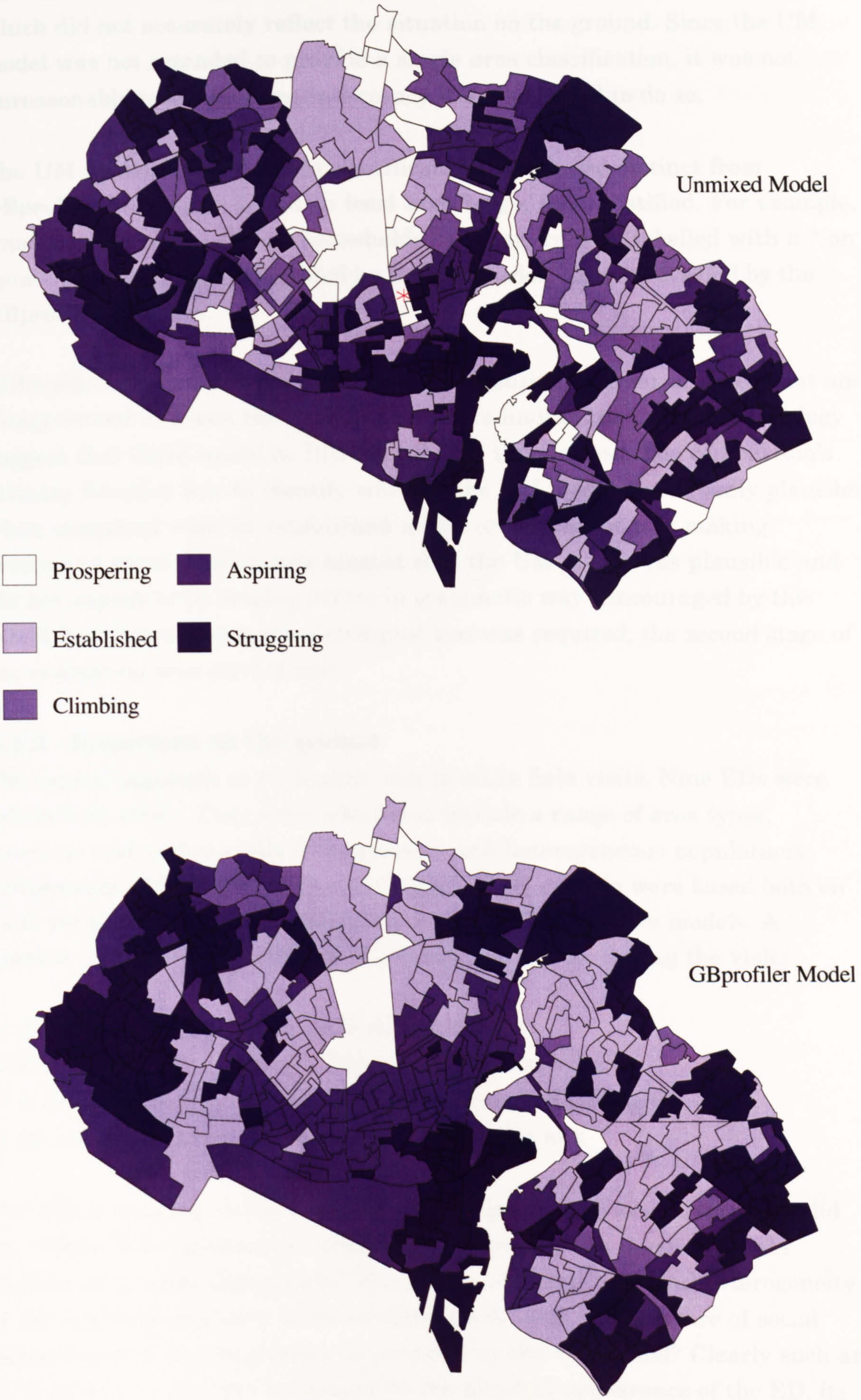
The more variable pattern described by the UM model might be partly a function of its production. For example, suppose an area was socially mixed and contained forty households in class 9 (labelled as prospering (see figure 8.2)). Suppose it also had thirty households in class 10 (labelled as struggling) and a further thirty households in class 12 (also labelled as struggling). Since class 9 was individually dominant in the area when a single classification

was derived, that ED was assigned to 'prospering'. This was despite the fact that the majority of its households were described as 'struggling'. This kind of

*Table 8.3 - Class Descriptions for the Fifteen Class Scheme (see also 7.16)*

| No. | Housing | Economics | Demographics |
|---|---|---|---|
| 1 | Avg. size, privately owned houses, some flats | Retired | Single elderly females |
| 2 | Avg. size, privately owned houses | Retired | Elderly childless couples |
| 3 | Avg. size, privately owned houses | Skilled / semi-skilled employment | 30's couples with dependent children |
| 4 | Avg. - large, privately owned houses | Managers | 30's couples with dependent children |
| 5 | Avg. - large, privately owned houses | Professionals / ancillary workers | 30's couples, some with dependent children |
| 6 | Avg. size, privately owned houses (some council tenancy) | Skilled / semi-skilled employment | 50-60's couples without children |
| 7 | Avg. size, privately owned houses / flats (some council tenancy) | Manual labour | 40-50's single men |
| 8 | Avg. size, privately owned houses | Junior non-manual employment | 30's couples, without children |
| 9 | Avg. - large privately owned houses | Managers and professionals | 50-60's couples without children |
| 10 | Avg. size, privately owned houses / flats (some council tenancy) | Junior non-manual employment | 20-40's single females, some lone parents |
| 11 | Avg. size, privately owned houses | Professionals and ancillary workers | 20-30's single people |
| 12 | Avg. size, privately owned houses / flats, some private renting | Either no socio-economic group or service workers | 30-40's couples with dependent children or lone parents |
| 13 | Avg. size, privately owned houses / flats, some private renting | Mangers | 20's-40's single people |
| 14 | Small rented, shared property | Skilled / semi-skilled manual or junior non-manual workers | 20-30's single people |
| 15 | Small, privately owned / rented non permanent accommodation | Retired | Single elderly |

*Figure 8.3 - Five Class Images of The City*



mis-classification may have been enough to introduce variation into the image which did not accurately reflect the situation on the ground. Since the UM model was not intended to provide a single area classification, it was not unreasonable to expect some inaccuracy if it was forced to do so.

The UM model has also made classifications which are distinct from GBprofiler but which, based on local knowledge, seem justified. For example, small pockets of prospering households in Inner Avenue (labelled with a * on figure 8.3) have been highlighted by the UM model, but not revealed by the GBprofiler analysis.

Although considerable time could be spent identifying areas of agreement and disagreement between the models, the severe limitations in the methodology suggest that there would be little of value in this process. The comparison's primary function was to identify whether the UM model was broadly plausible when compared with an established model or whether it was making systematic error. The images suggest that the UM model was plausible and did not appear to be making errors in systematic way. Encouraged by this result, but aware that a more stringent test was required, the second stage of the evaluation was carried out:

### 8.1.2 - Inspection on the ground

The second approach to evaluation was to make field visits. Nine EDs were selected for study. They were selected to include a range of area types, locations and both socially homogeneous and heterogeneous populations. Judgements about which EDs could satisfy these criteria were based both on local knowledge and the information available from the two models. A number of items of information were prepared for use during the visits.

1) A GBprofiler classification (10 class scheme)
2) A dominant class from the UM model (15 class scheme)
3) A measure of social mixedness derived from the UM model
4) An indication of the major components of that mix

The EDs were then visited and a number of questions posed at each site; did the GBprofiler classification seem appropriate? Did the dominant class classification seem appropriate? Were there any signs of social heterogeneity or homogeneity and were these consistent both with the measure of social mixedness and its components as provided by the UM model? Clearly such an evaluation was strongly influenced by the physical appearance of the ED, its

housing stock and the preconceptions of the researcher. However, observing and assessing the relationship between the model and the real appearance of the areas was felt to be essential and valid, despite the reliance on qualitative judgements. The visits represented an opportunity to assess the model's representation of the researcher's own reality.

*Table 8.4 - Information for ED Visits*

| ED Code | UM Class | GB Class | Mixture | Next Class | Next Class |
|---|---|---|---|---|---|
| JPFD07 | 8 | 10 | quite mixed | 3 | 2 |
| JPFH14 | 12 | 1 | pure | 1 | 3 |
| JPFJ16 | 12 | 10 | middle | 3 | 7 |
| JPFP04 | 1 | 10 | quite pure | 2 | 3 |
| JPFH13 | 12 | 2 | v. mixed | 9 | 14 |
| JPFL04 | 3 | 3 | middle | 8 | 2 |
| JPFE25 | 1 | 3 | v. pure | 2 | 14 |
| JPFN05 | 4 | 8 | v. pure | 7 | 12 |
| JPFB07 | 1 | 1 | quite pure | 3 | 2 |

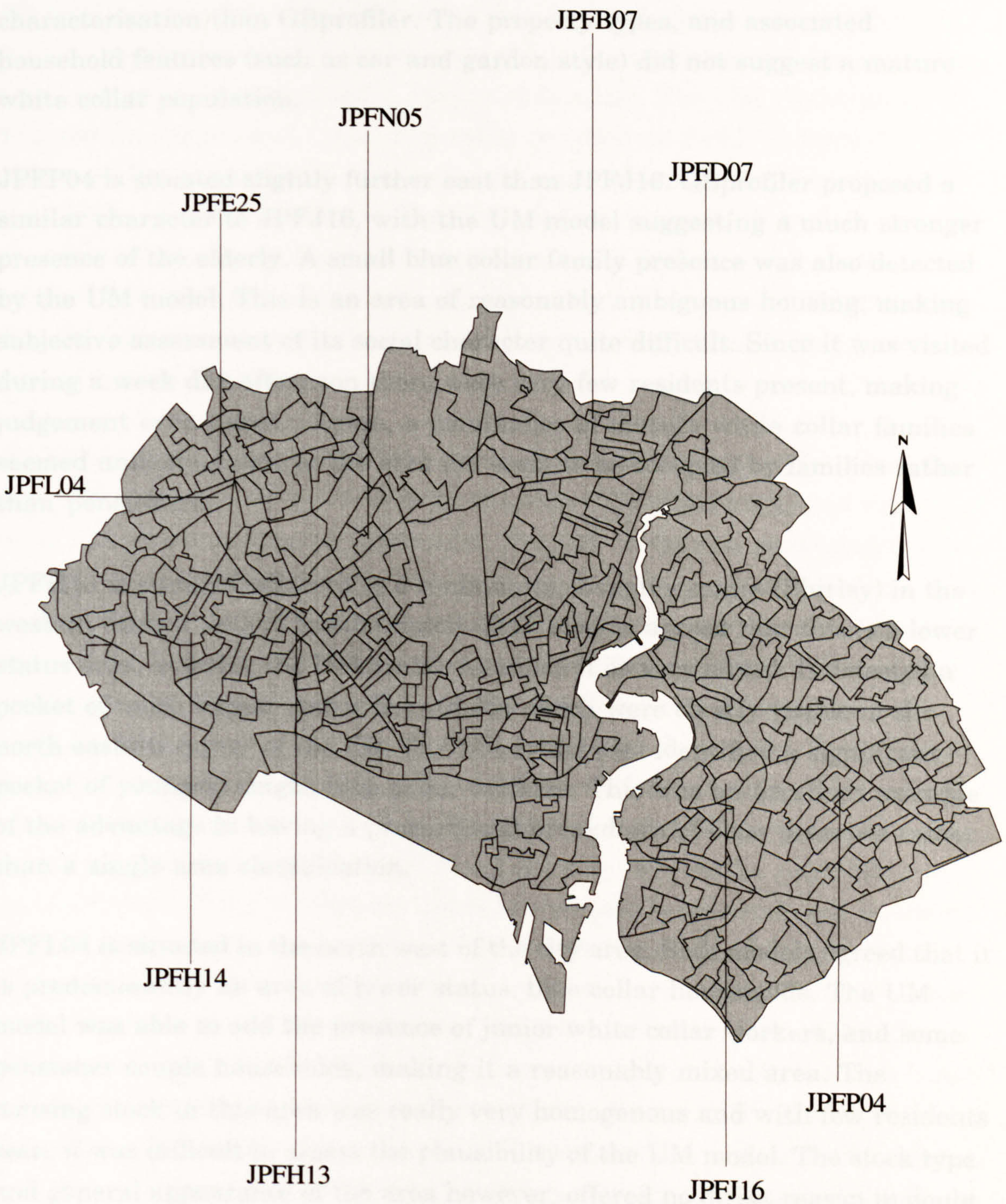
Table 8.4 shows the single area classification derived from the UM model, the GBprofiler classification, the degree of social mixing detected and the two next biggest classes after the dominant one, for each ED. Figure 8.4 shows the locations of these EDs. They are explored below:

JPFD07 is situated close to Southampton's north east fringe. Both models identified it as a white collar family, owner occupier area. The UM model suggested that the area was reasonably homogenous, with a pensioner element and some blue collar family presence too. The area visit confirmed that these assertions seemed entirely plausible. A reasonable number of smaller bungalow type properties suggested that the presence of retired households was likely.

JPFH14 is situated in the mid south west of the city area. Both models identified this area as one of lower status, with high levels of unemployment and council tenancy. The UM model suggested that the area was highly homogenous with perhaps a lower status pensioner element and some blue collar families too. The site visit confirmed this assessment as plausible.

JPFJ16 is situated close to the eastern bank of the Itchen, in the southern central area of the city. The models were not in agreement in their

*Figure 8.4 - The Location of EDs Visited During Model Testing*



assessment of this area. GBprofiler suggested that it contained mature white collar couples and owner occupying families. The UM model suggested a much lower status mixed area containing many unemployed couples, some blue collar singles and families and a mixture of tenancies. The visual appearance of the area suggested that the UM model had achieved a better characterisation than GBprofiler. The property types, and associated household features (such as car and garden style) did not suggest a mature white collar population.

JPFP04 is situated slightly further east than JPFJ16. GBprofiler proposed a similar character to JPFJ16, with the UM model suggesting a much stronger presence of the elderly. A small blue collar family presence was also detected by the UM model. This is an area of reasonably ambiguous housing, making subjective assessment of its social character quite difficult. Since it was visited during a week day afternoon there were very few residents present, making judgement even harder. Again, a dominance of mature white collar families seemed unlikely, however the area did seem to be occupied by families rather than pensioners.

JPFH13 is situated adjacent to a community shopping centre (Shirley) in the western central area of Southampton. Both models agreed that this is a lower status area, however the UM model described it as very mixed. It detected a pocket of much higher status households which were clearly visible in the north eastern corner of the ED. The UM model also identified a significant pocket of younger, single, blue collar workers. This was an excellent example of the advantage in having a proportional breakdown of class presence rather than a single area classification.

JPFL04 is situated in the north west of the city area. Both models agreed that it is predominantly an area of lower status, blue collar households. The UM model was able to add the presence of junior white collar workers, and some pensioner couple households, making it a reasonably mixed area. The housing stock in this area was really very homogenous and with few residents seen, it was difficult to assess the plausibility of the UM model. The stock type and general appearance of the area however, offered no direct reason to doubt it.

JPFE25 is a development of high rise flats, close to JPFL04 (above). Both models were in broad agreement that this is predominantly a lower status, pensioner area. However, the UM model, using its more complex

classification system, was able to illustrate the relative proportion of couples and single person households, as well as detecting young, single, blue collar households. It was not possible to directly access the estate, however, local knowledge suggested that the UM model was entirely plausible.

JPFN05 is just to the west of Southampton common. Both models agreed that this is an affluent area, comprising large semi-detached and detached houses with mature professional and managerial families. The UM model also detected the presence of single blue collar people households in flats. It suggested there may be an underlying presence of lower status, unemployed households, possibly in bedsits. The visit confirmed that both models seemed plausible. Some houses had clearly been converted to flats and bedsits, suggesting a good performance by the UM model. Of particular interest was that the UM model had pinpointed an ED containing groups from opposite ends of the social scale.

JPFB07 is situated on the northern fringe of the city. Both models agreed that this area is dominated by the elderly although GBprofiler suggested a group of lower social status than the UM model. The UM model also suggested a blue collar family presence, which was in keeping with the appearance of the area. In fact, neither classification seemed particularly appropriate when the area was visited. This ED contains a number of properties built by Collins, a local philanthropist. The properties have a distinct character and despite their small size remain extremely popular. Today Collins houses are desirable and tend to be expensive, although price varies according to location. The census does not contain information about the relative market value of specific properties and both classifications would have focused on the physical characteristics of the properties, which in this case were small and terraced. The UM model was equipped with a better understanding of how household characteristics can cluster together and has correspondingly provided a better characterisation of the area. However, it has no understanding of the specific nature of the properties in the area. Both models rely on a source which is not sufficiently complex to relate the unique nature of some of the properties in the city. This is an excellent example of the shortcomings of quantitative models of social features, demonstrating the mis-match between the complexities of the empirical world and the simplicity of the data. The marginally better performance of the UM model illustrated some advantages over the GBprofiler model but also demonstrated its own shortcomings.

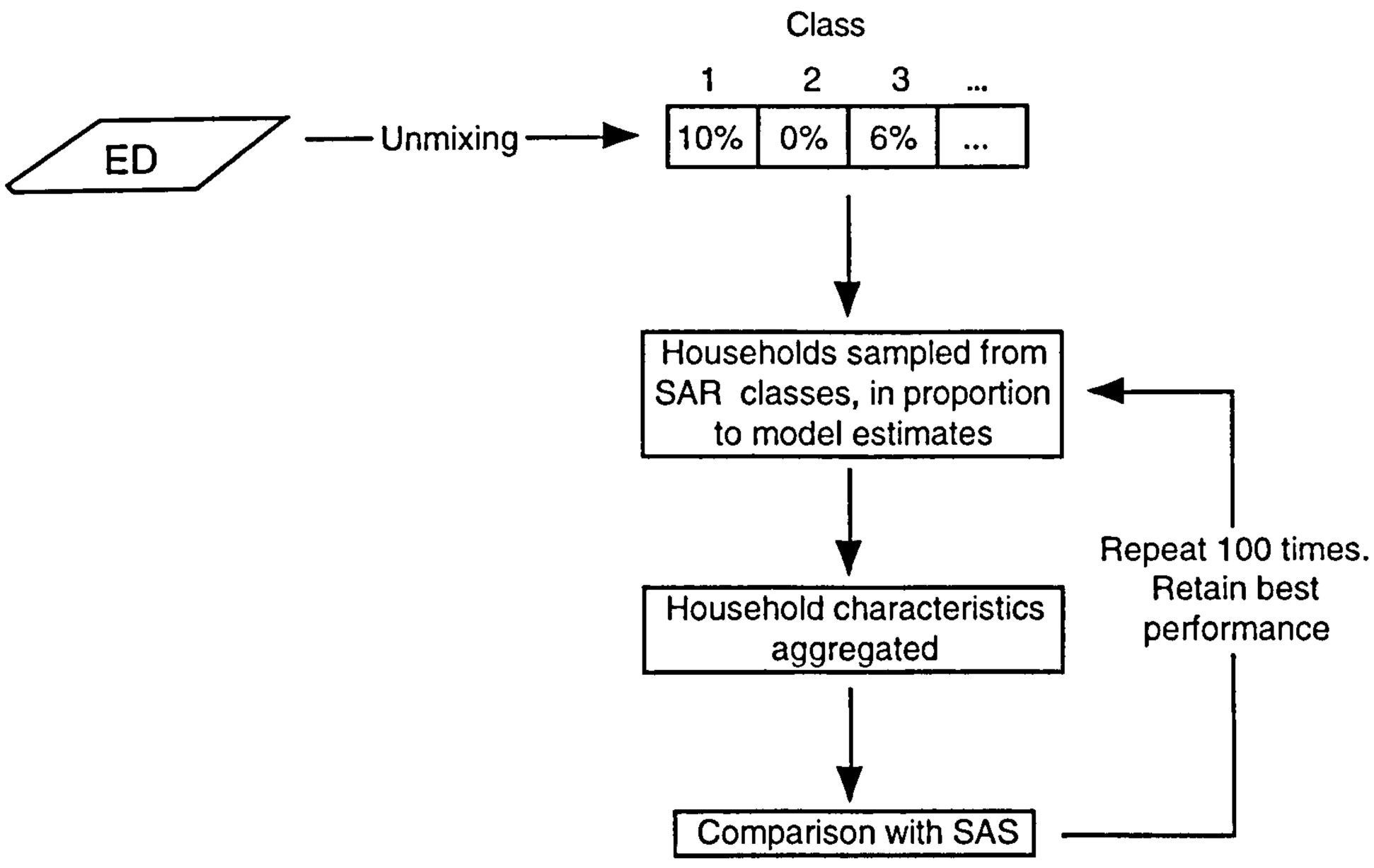
When these results were viewed as a whole, it was apparent that the UM model did appear to be unmixing EDs in a realistic way. The visits were especially useful for revealing the utility of the complete proportional breakdown of class presence in the ED. The detection of pockets of households which were visibly distinct from the rest of the area was a graphic illustration of both the dangers of relying on single area classifications and of the accuracy of the new model. Whilst these visibly distinct households were quite easy to relate to the model findings, the presence of households whose distinctive characteristics were demographic or occupational was much harder to confirm. The site visits thus compounded the case for a measure of model quality that did not rely on visual or subjective checks.

### 8.1.3 - Reconstruction of Small Area Statistics from Model Outputs

Although the comparisons with GBprofiler and the field visits had provided firm (if subjective) evidence for the success of the unmixing technique, a more objective measure was needed. A third test, which could evaluate the model's performance in every ED, would help to reveal any systematic error:

Since there were no data against which to make a suitable test, an alternative means of evaluation was devised. The intention was to answer this question; "if the ED were populated with households matching these classifications, how closely would the original SAS values be reproduced?" The UM model had estimated the proportional presence of households from different classes in the ED. The test would re-sample an appropriate number of households from each SAR-derived class and determine how closely their aggregate characteristics matched those of the real ED population. This was akin to an unconstrained micro-simulation (see section 4.4) since the test would involve the synthesis of a disaggregate population for each ED. The characteristics of that population would then be aggregated for comparison with the real SAS values. Figure 8.5 illustrates each stage of the process.

*Figure 8.5 - Rebuilding SAS Data From The Unmixed Model*

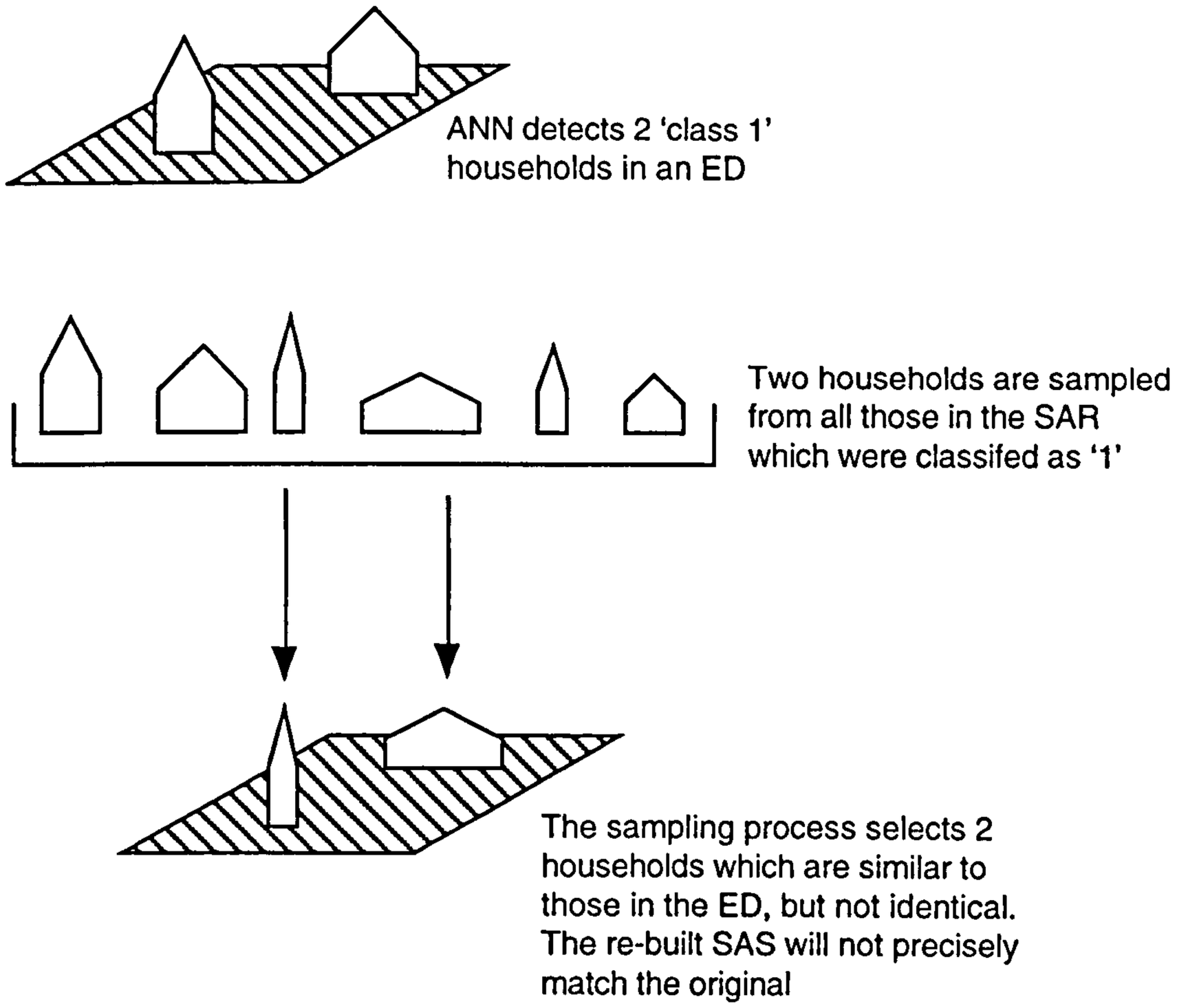



The flaws in this re-building technique were apparent from its conception. It was clear that precise reconstruction of the original SAS values would be extremely unlikely for two reasons. Firstly, the classes derived from the SAR data were not homogenous, despite their k-means classifier origin. Therefore, even if the unmixing technique had identified the presence of households from a particular class, the households re-sampled from that class's defining group would not necessarily match those present in the particular ED. This is demonstrated by figure 8.6.

The second potential source of error was related to the nature of the SAR sample itself. The SAR classes were based on a large sample of both urban and rural households. There was therefore no reason to expect the full range of household types found in the SAR to be present in Southampton and it seemed likely that certain household types would be more common in Southampton than in the SAR. For example, it was expected that the housing stock profile of the space-pressured city would be skewed towards smaller houses and flats when compared with the whole SAR sample. Whilst the unmixing technique would have detected this skew and reflected it in the proportional class model, there was no way of constraining the re-sampling process to reflect it. In fact, it was hoped that recurring differences between the rebuilt SAS values and their originals might illuminate the applicability of the SAR classification to Southampton. This would then assist in evaluating the consistency of class structures between different kinds of places. The

applicability of class structures to different locations was an important factor to consider when making an assessment of the model's potential uses.

*Figure 8.6 - A Potential Source of Error in the SAS Re-building Process*




These difficulties ensured that the true value of the exercise lay in detecting any large discrepancies between the two sets of aggregate data or any systematic variation. It was not a highly precise measure of the UM model's accuracy.

### 8.1.4 - Results

The most straightforward measure of error from this test was the *absolute difference between a rebuilt SAS count and the original*. The average of this 'per variable error' served as an indicator of the model plausibility for an ED. The average error for a variable across the 418 EDs in the study area was 9.78, with a standard deviation of 3.84, a minimum of 2 and a maximum of 38. Since the total number of households varies between EDs, this figure may be expressed more meaningfully as a percentage of that total. Expressed as a percentage of the total number of households in each ED, the mean error was 5.11% with a standard deviation of 1.79%, a minimum of 2.29% and a maximum of 13.89%. Thus, a SAS value reconstructed from the UM model

differed by an average of 5.11% from the original. This value seemed acceptably low, given the potential for error inherent in the reconstruction process. It was, however, the variations in this error rate which were of more interest (see table 8.5 for details of the error associated with each variable):

Figure 8.7 uses a measure known as an *entropy score* to illustrate the relative 'social purity' of each ED[4]. The entropy score is higher when the ED is more 'mixed', or heterogeneous. Figure 8.7 illustrates that the greatest error levels were consistently found in those areas which identified as socially homogenous and also in which the population is at the lower end of the social scale. There are a number of possible explanations for this. It may have been that the population inside those EDs was *not as homogenous* as the model suggested. Directed to sample from just one or two classes, the SAS rebuilding process did not yield a population with enough variation in household type and differences between the original and rebuilt SAS values were therefore introduced. Secondly, it may have been that the population in those areas was *more homogenous* than the model suggested and in sampling from SAR derived classes which do contain some variation in household type, the reconstructed SAS contained more variation than the originals.

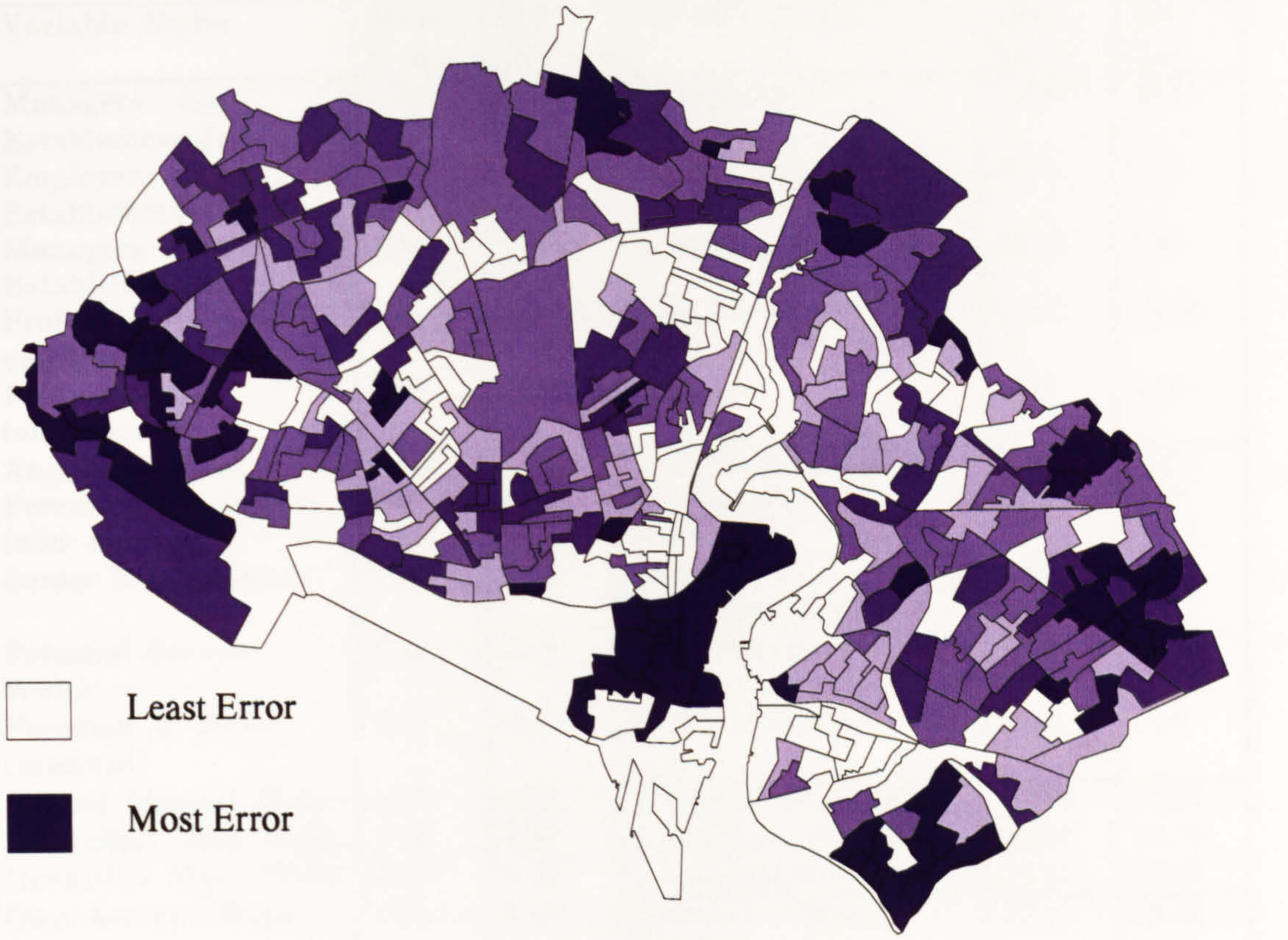
The EDs in which model performance has been notably poor do contain very homogenous *housing stock*. It is possible that this affected the model's performance since, whilst economic and demographic characteristics still varied within the area, housing characteristics were constrained. The classification scheme utilised may not have been flexible enough to accommodate these situations. Indeed, if the performance of individual variables within the reconstruction process is examined, those relating to property features (tenure, numbers of rooms in house and household space type) are associated with the higher error rates (see table 8.5). This suggests that the relationship between property and other household characteristics, which from the classification process appeared to be quite strong, may be subject to a locational influence.

The process of residential location in a city is often described in competitive terms, with families striving to attain location in a particular place (see

---

[4]The proportional presence of each class was treated as a probability in order to calculate the entropy score. It was calculated such that $H = -\Sigma\, p(x)\log p(x)$ (where $p(x)$ is the proportion in which a particular class was present in an ED, and H was the entropy measure (Klir and Folger 1988). This measure is used in Earth Observation research.

*Figure 8.7 - Mapped Error and Entropy*



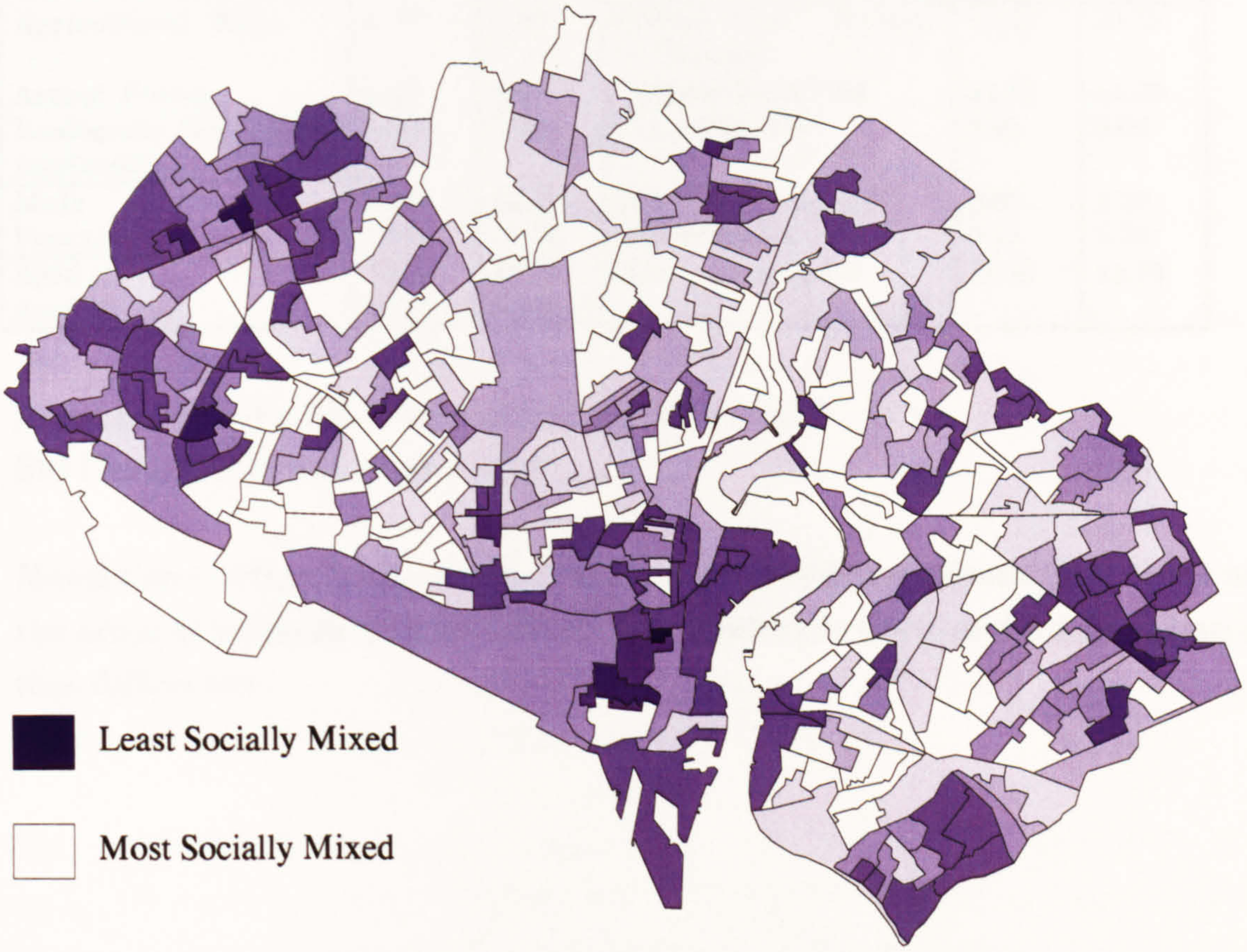



*Table 8.5 - SAS Re-building; A Summary of Results*

| Variable Name | Mean Error | Std. Dev. | Variable Name | Mean Error | Std. Dev. |
|---|---|---|---|---|---|
| Managers - Lge Establishments | 0.56 | 0.71 | Aged 45 - 59 | 14.84 | 10.55 |
| Employers - Smll Establishments | 0.99 | 1.02 | Aged 60 - 64 | 7.10 | 5.06 |
| Managers - Smll Establishments | 1.04 | 1.07 | Aged 65 - 74 | 10.29 | 7.86 |
| Professionals (self employed) | 0.68 | 0.94 | Aged 75 - 84 | 11.77 | 10.39 |
| Professionals (employees) | 0.93 | 1.01 | Aged >85 | 4.18 | 4.10 |
| Ancillary Wrks | 0.93 | 1.09 | Couples, No Kids | 1.56 | 1.27 |
| Foreman & Supervrs (non man) | 1.42 | 1.27 | Couples, Dependent Kids | 1.27 | 1.07 |
| Junior Non Manual | 0.95 | 0.96 | Couples, Non Dependent Kids | 0.98 | 0.87 |
| Personal Service Wrks | 0.44 | 0.61 | Lone Parents | 1.12 | 1.12 |
| Foreman & Wrks (manual) | 1.87 | 1.45 | Not In a Family | 1.46 | 1.27 |
| Skilled Manual Wrks | 1.05 | 0.95 | 1 - 3 Rooms in House | 22.78 | 14.57 |
| Semi-Skill Man Wrks | 1.03 | 0.89 | 4 - 6 Rooms in House | 24.29 | 18.78 |
| Unskilled Man. Wrks | 0.95 | 0.88 | 7 + Rooms in House | 34.21 | 21.36 |
| Own Account Wrks | 1.54 | 1.16 | Privately Owned Property | 40.27 | 29.34 |
| Farmers (emps / mgs) | 0.01 | 0.10 | Privately Rented Property | 23.65 | 17.49 |
| Farmers (own account) | 0.03 | 0.16 | Council Tenants | 44.89 | 38.27 |
| Agricultural Wrks | 0.17 | 0.39 | D'ched, Semi - D'ched or Terrace | 37.33 | 33.58 |
| Armed Forces | 0.38 | 0.67 | Purpose Built Flat | 44.35 | 44.29 |
| Inadequate Desc / Not Applicable | 1.56 | 1.30 | Converted Flat | 9.40 | 9.69 |
| Male | 19.51 | 11.27 | Not Self Contained | 0.65 | 1.18 |
| Female | 19.64 | 11.28 | Other Spaces | 8.45 | 9.73 |
| Aged < 29 | 15.76 | 12.19 | Non - Permanent | 10.41 | 11.79 |
| Aged 30 - 44 | 13.58 | 10.00 | | | |

Average error for all counts across all EDs = 9.78

Std Deviation of this error = 3.84

Mean Error refers to the average difference between a rebuilt SAS count and the original value in all 418 EDs. Std. Dev refers to the standard deviation of this difference.

Herbert and Johnston 1978, for example). In this competitive market place housing stock choice may be constrained for many residents. It is therefore reasonable to expect local differences in the relationship between the physical aspects of stock and the social characteristics of its occupants. This is a significant finding for those who might wish to apply the technique to large areas. Either the classification scheme employed must be sufficiently complex to deal with different relationships between stock type and occupant characteristics, or different classifications should be applied to areas with different stock profiles.

The performance of other variables however, was good. Socio-economic group, which was clearly a significant variable in the classification, was re-constructed almost precisely. Least error occurred in those categories with few cases in the city (e.g. agricultural employment), since these are the more distinctive household types, with a lower probability of mis-classification.

It is difficult to make an assessment of the UM model's performance through the SAS reconstruction technique without being able to quantify the sources of error in the reconstruction. Possible sources have been identified and include sampling households from within heterogeneous classes, the SAR classification not being suited to Southampton's population or model inaccuracy. It seems most likely that error will have been drawn from all three sources. However despite this, the model seems to have performed extremely well. Improvement in the reconstruction, and indeed in model performance, might have been achieved through adjustments to the original SAR sample. An aggregate analysis of Southampton's population for example, would have revealed broad discrepancies between their characteristics and those of the SAR sample, highlighting the over or under representation of certain features. By increasing the numbers of households in the SAR sample with characteristics common to Southampton's population and possibly removing those not found, the data set would have yielded a more classification more appropriate for Southampton. However, this location specific enhancement would have reduced one advantage that the UM approach has over microsimulation – speed. Future work may shed light on the possibilities for using a more task specific classification system with the technique. This will be discussed later in this chapter.

The combined evidence from comparisons with GBprofiler, site visits and reconstruction of SAS values suggests that the UM technique has produced a plausible and realistic characterisation of the ED populations. The model has

both qualitative and quantitative accuracy, i.e. it unmixes EDs into the correct types of household and in the correct proportions. It is therefore appropriate to analyse the new model and review what it revealed about Southampton's socio-spatial structure that was previously unknown.

## *8.2 - New Information From This Research*

The new model directly altered or increased the available information about Southampton's socio-spatial structure in five ways. This section will review each of them in turn:

### 8.2.1 - Analysis in Aggregate

The UM model provided a new means of determining Southampton's aggregate social characteristics (i.e. the relative proportions of different household types for the entire city). This kind analysis was directly akin to that carried out by Openshaw et. al. (1985), see section 4.2.2 and is typical of the kind of assessment upon which strategic planning decisions are informed or location decisions may be based (for new stores or leisure facilities for example). Prior to this work, such an analysis might have been based upon the numbers of EDs classified into different geodemographic categories or upon Local Base Statistics from the census. The UM model facilitated a different approach:

The proportional presence of each class in the whole city was derived from the model by aggregating its output. By recording the presence of *all* classes in every area rather than assigning the entire population of an ED to a single identity, each class was represented more realistically. *Every* household in *every* class had a chance to contribute to the aggregate analysis.

*Table 8.6 - Aggregate Analysis of Southampton from Unmixing*

| Class | % Of Southampton's Households |
|---|---|
| 1 | 14.9 |
| 3 | 13.9 |
| 12 | 10.8 |
| 2 | 9.7 |
| 7 | 9.3 |
| 4 | 7.6 |
| 14 | 7.2 |
| 5 | 6.1 |
| 8 | 5.0 |
| 15 | 4.2 |
| 6 | 3.0 |
| 13 | 2.6 |
| 11 | 2.5 |
| 9 | 1.9 |
| 10 | 1.4 |

*Table 8.7 - Aggregate Analysis of Southampton from GBprofiler*

| Class | % Of Southampton's Households |
|---|---|
| 10 | 17.0 |
| 5 | 16.3 |
| 3 | 12.9 |
| 8 | 12.2 |
| 1 | 12.0 |
| 6 | 11.8 |
| 2 | 10.3 |
| 7 | 3.4 |
| 4 | 2.9 |
| 9 | 1.2 |

This analysis is directly in keeping with the aims of this research, as described in chapters 2 and 3; the true characteristics of a residential population are more directly represented by the model because every household has a chance to contribute its own identity to the aggregate figures. With an analysis based on single area classification, a household contributes

only as another count for the class to which it is assigned as part of its area of origin.

Again, the fact that the two classification schemes do not match makes it difficult to offer a meaningful comparison between the two aggregate pictures. As the two tables show, the higher number of categories employed by the UM model classification has allowed the dissection of Southampton's 'middle classes' into a higher number of smaller groups. To attempt a more meaningful comparison between the two aggregate pictures, the class collapse system can again be used. See figures 8.1 and 8.2 for a key to the collapse:

*Table 8.8 - Comparison of the Aggregate Analyses of Southampton*

| Class | % of Households in GB Profiler Scheme | % Of Households in UM Model scheme |
|---|---|---|
| Struggling | 38.1 | 30.3 |
| Aspiring | 28.1 | 19.0 |
| Climbing | 3.4 | 15.2 |
| Prospering | 1.2 | 8.0 |
| Established | 29.2 | 27.6 |

The major difference between the two versions of the city's social structure is that an analysis based on the unmixed model suggests a greater number of Southampton's households are in the middle classes. Whilst both models show that 'struggling' is the largest category, the UM model paints a less bleak picture, with notably higher numbers in the 'climbing' category and a very much higher number of 'prospering' households. This might suggest that the prospering neighbourhoods in Southampton are particularly incongruous with the ED boundaries - either relatively small areas within large EDs or straddling a number of different EDs. Either of these situations would reduce their impact on a single area classifier such as GBprofiler because their presence would be diluted or obscured by households around them. The UM model however, is capable of detecting their presence.

The differences between the aspiring and climbing classes may also be a result of the enhanced household detection capability which the UM model has. However, given that the distinction between a climbing household and an aspiring household is not particularly clear, these differences might well be

largely accounted for by the crude nature of the collapse process which facilitated this comparison.

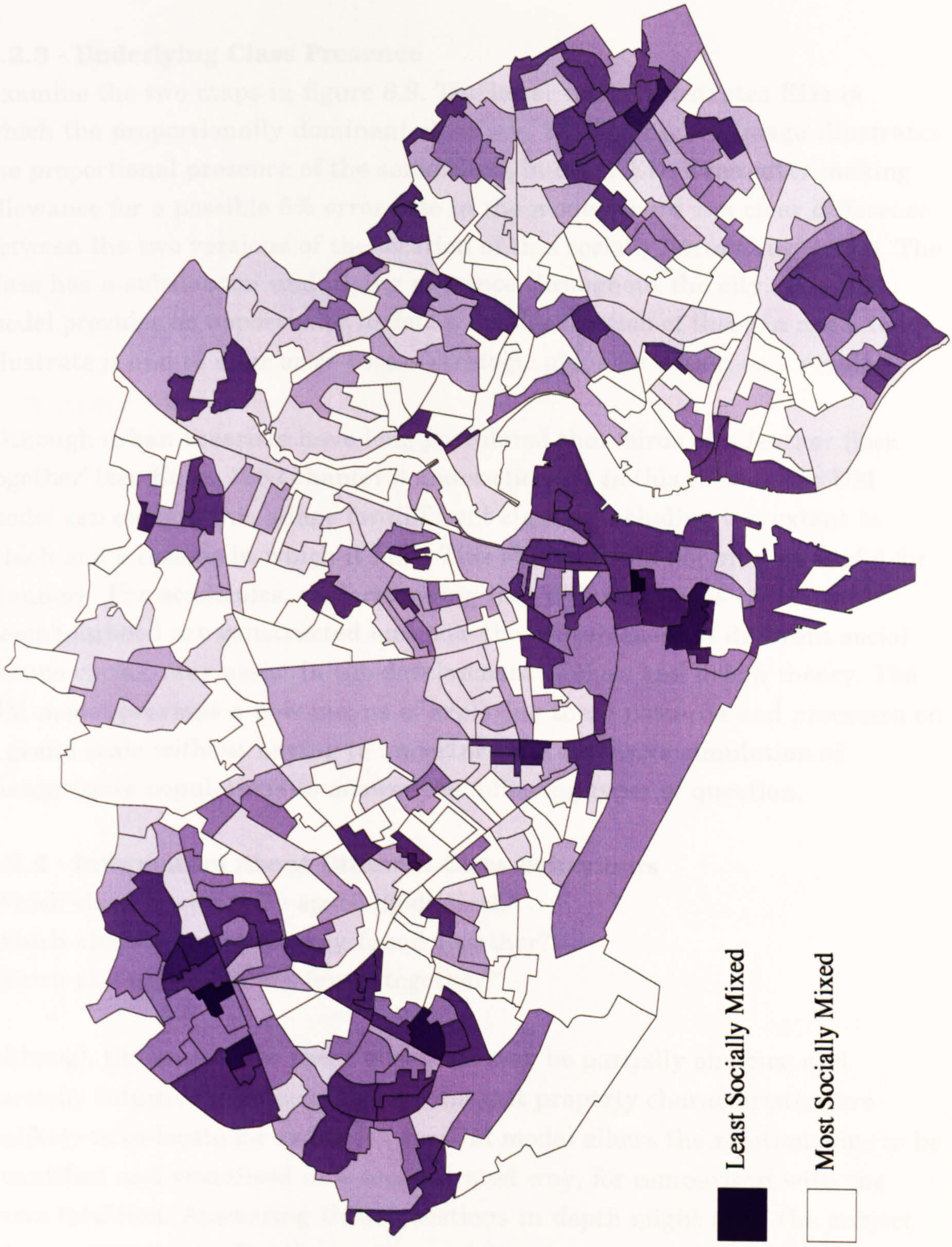
A planner might well be cheered by the UM model's account of Southampton's social structure which is rather more optimistic than the GBprofiler version. However, nearly one third of the households remain in the struggling category whichever way the city is modelled. The single area classification can be a powerful persuader in the planning process. If resources are to be effectively allocated to this sizeable, needy population, a planner might wish to know how reliable the single area classification is, and how this needy population relate (in spatial terms) to those who are apparently rising through the social scale. The UM model can assist in answering these questions:

### 8.2.2 - Social Mixing and the Reliability of Single Area Classification

Figure 8.8 was produced by mapping a measure of the 'purity' of the class mixture in each ED using an 'entropy' measure (see section 8.1.4). The map illustrates a concentration of socially homogenous areas in the central southern area of the city and in north western and eastern city fringes. This entropy score would be of interest to those concerned with the social morphology of the city and as measure of the applicability of, or confidence in, single area classification. The lighter the tone in the image, the less appropriate a single area classification would be for the ED, since it would mask variation in character amongst the resident households. The quite extensive spread of the lighter shades is a powerful illustration of how important it is to replace single area classification with a measure which reflects what is actually present in small areas.

If this confidence indicator is coupled with a more complex analysis of what actually comprises the mixture, the potency of the new level of information now available is made clear. Consider an ED whose entropy score lies in the middle of a pure - mixed range. Suppose also, that a geodemographic analysis of that ED has placed it into a category which is associated with housing need. The entropy score suggests that the area may be socially mixed and therefore that not *all* the households in the area might actually be in housing need. On the other hand, the entropy score might be reflecting the presence of a number classes *all* of which *are* associated with housing need. The advantage of the UM model is that it allows the user to explore precisely what kinds of households are actually present in an ED and in what numbers. If the ED is identified as a mixture of households in need and not in need, resource allocations can be adjusted accordingly.

*Figure 8.8 - Entropy As A Measure of Social Mixing*



In this way, an increased level of information has enhanced the potential for social justice. The following section expands on this concept, using a slightly different type of analysis.

### 8.2.3 - Underlying Class Presence

Examine the two maps in figure 8.9. The lower image illustrates EDs in which the proportionally dominant class was 12. The higher image illustrates the proportional presence of the same class, in every ED. Even after making allowance for a possible 5% error rate in the model, there is a clear difference between the two versions of the location of this socially 'struggling' group. The class has a substantial underlying presence throughout the city. The UM model provides an opportunity to detect the distribution of those in need, to illustrate it and to alter or re-assess strategic plans for action as a result.

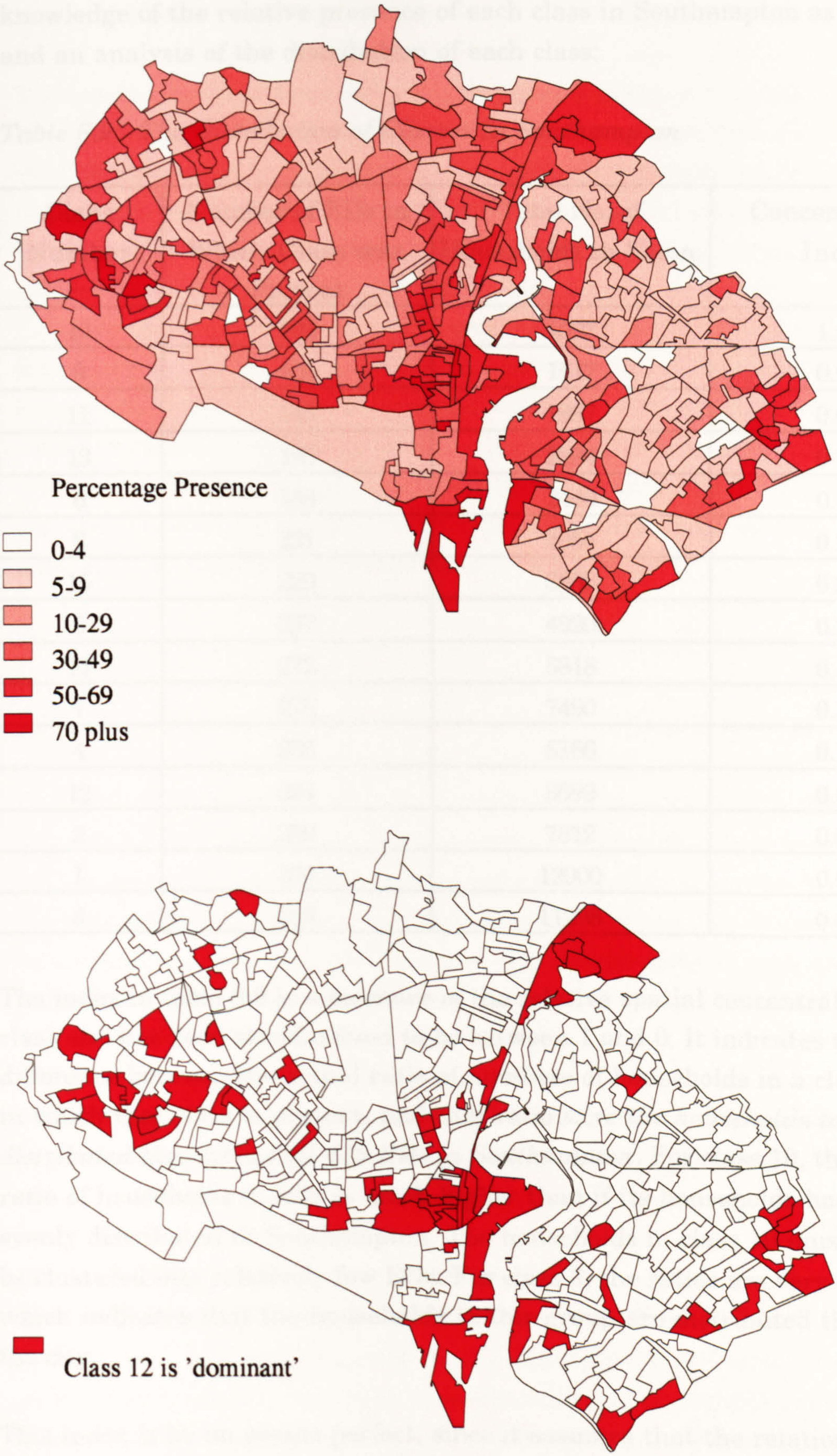
Although urban theorists have long postulated that 'birds of a feather flock together' (see Knox, 1982 chapter 3 and section 2.1 in this thesis), the UM model can explore this adage for different classes, including the extent to which and locations in which it fails. This ability would not only be useful for planners. For academics, understanding how different types of neighbourhood are constructed and how the cohesiveness of different social groups varies, may assist in the development of class and urban theory. The UM model provides a new means of exploring these patterns and processes on a grand scale without having to undertake the extensive simulation of disaggregate populations to answer the following types of question.

### 8.2.4 - Information About Different Class Behaviours

'Which classes tend to be spatially distinct?'
'Which classes are commonly found together?'
'Which classes are rarely found together?'

Although the answers to these questions may be partially obvious, and partially intuitive (classes with very distinct property characteristics are unlikely to co-locate for example), the UM model allows the relationships to be quantified and visualised in a sophisticated way, for comparison with the users intuition. Answering these questions in depth might form the subject for an entire thesis. For the purposes of this chapter, a crude indicator has been devised to briefly explore the first question and, more significantly, to illustrate the potential for the UM model to assist in answering it.

*Figure 8.9 - Class 12 Proportional Presence vs. Dominant Presence*



Determining which classes tend to be most spatially distinct requires a knowledge of the relative presence of each class in Southampton as a whole, and an analysis of the distribution of each class:

*Table 8.9 - The Distribution of Classes In Southampton*

| Class Number | Number of EDs in Which Class was Present | Total No. of Households in Soton. | Concentration Index |
|---|---|---|---|
| 10 | 102 | 1097 | 1.00 |
| 9 | 106 | 1517 | 0.95 |
| 11 | 145 | 1987 | 0.60 |
| 13 | 165 | 2094 | 0.48 |
| 6 | 168 | 2440 | 0.47 |
| 8 | 221 | 4054 | 0.27 |
| 15 | 229 | 3356 | 0.25 |
| 5 | 257 | 4920 | 0.18 |
| 14 | 272 | 5818 | 0.15 |
| 7 | 278 | 7490 | 0.14 |
| 4 | 305 | 6166 | 0.10 |
| 12 | 324 | 8699 | 0.07 |
| 2 | 359 | 7812 | 0.03 |
| 1 | 386 | 12000 | 0.00 |
| 3 | 389 | 11298 | 0.00 |

The index in table 8.9 is a measure of the relative spatial concentration of each class and has been standardised to be between 1 and 0. It indicates the difference between the actual ratio of numbers of households in a class to EDs in which that class is present, and that ratio *were the households to be evenly distributed throughout all 418 EDs in Southampton*. For class 10, the actual ratio of households to EDs is much higher than if its households had been evenly distributed in Southampton. The households in class 10 must therefore be clustered into relatively few EDs. For class 3, the ratios are very similar which indicates that the households in this group are distributed throughout the city.

This index is by no means perfect, since it assumes that the relationship between household numbers and presence in EDs is linear, however it does

effectively illustrate that classes exhibit different degrees of spatial concentration.

The relationship between the index value and the class character is not straightforward, but it does make some intuitive sense. The classes which exhibit greatest spatial concentration are those which might be expected. For example, class 10 is amongst the lowest on the social scale, containing a high number of lone parents and single flat dwellers. The social and accommodation characteristics of this class suggest that they are likely to found in quite specific areas of the city, possibly in high rise developments or flat-converted housing. Class 9, at the opposite end of the social scale, contains some of the most successful residents with few or no dependants. They are likely to reside in 'exclusive' areas, which are few in number in Southampton and therefore also suggest a higher spatial concentration for the class. Both classes are amongst the smallest which, due to the crude nature of the index's construction, will have increased their score.

The index illustrates the tendency of a class to cluster *within* EDs, but does not reveal anything about the tendency of the class to cluster in *physical space*. It takes no account of whether the EDs in which a class is present were adjacent to each other, or clustered strongly in groups and more complex spatial analysis would be required to investigate such precise spatial morphology. Were this analysis to be undertaken however, the 'neighbourhood' as defined by spaces containing similar household type, would be revealed. A 'neighbourhood map' would greatly assist in meeting the challenges of investigating urban theories based on the 'neighbourhood' in a quantitative way. This example has illustrated the potential of the UM model information to take a role in that investigation.

Not all the analyses which are facilitated by the new technique need to be at the city level. A considerable volume of information now exists for every small area.

### 8.2.5 - Detailed Analyses of Small Areas

The most detailed information which the UM model offers is the proportional breakdown of class presence for an ED (see figure 2.8). Conversion of the proportional information into estimates of actual household numbers is easily achieved. Although this volume of information is not easily visualised, it is ideal for manipulation in a spreadsheet or database package. Summary statistics could be derived from it, with anomalies or interesting features in

those summaries guiding the user into the full detail of the model itself. For use in this research, the information was accessed and manipulated in a geographical information system (GIS). The GIS drew a map of the study area with each ED shaded according to a particular summary statistic. Where an ED or group of EDs was of particular interest (perhaps highlighted by the summary variable in some way), it could be selected and the detailed class breakdown for that area displayed or further analysed. The GIS provided a convenient means of managing the model, allowing the user to see as much or as little detail as desired.

### 8.2.6 - How good is the model?

Thus far, the chapter has been dedicated to illustrating the results obtained from the model, to relating details of the tests carried out upon it and to illustrating how the information which the model provides has been or might be used. Given the results from the threefold testing process and the illustrations of analysis presented above, the conclusions must be that the model is accurate in a qualitative and quantitative way, and a potentially rich and flexible source of information, both for the academic and the public or commercial researcher. It appears that the model is ideally suited for basic management and manipulation within a GIS.

## *8.3 - Future Developments?*

The flexibility and detail of the UM model has been illustrated and its accuracy has, as far as possible, been assessed. However, during the course of this analysis, it has become clear that the *real* value of this research has been in the development of a *methodology* for the integration of aggregate and disaggregate data. The classification scheme utilised in this particular example divides the residential population into generalised classes with which, as has been shown, useful analysis of the city may be undertaken. However, it is the potential to employ this technique with different, perhaps more specific, classification schemes (or even different data), that seems most significant.

There are four ways in which the unmixing approach might be customised for use in a specific role; (i) the use of a different SAR based classification, (ii) the association of variables of interest with a SAR based classification, (iii) the integration of disaggregate data not from the census, with SAS and (iv) the integration of aggregate and disaggregate data sets, neither of which are drawn from the census. These will be explored in turn.

### 8.3.1 - Different SAR-based Classifications

The SAR classification used to develop the unmixing method was designed to be a generalised social structure, reflecting broad systematic differences between households on economic, demographic and housing axes. It is highly likely that different kinds of behaviours or resource needs will be associated with the different groups, but they were not designed to reflect any of these specifically. However, it may be that a user is aware of a systematic variation in a particular behaviour or need which is directly related to a combination of three or more variables featured in the census. The classification system could in fact be far more simplistic than has been used here, even simply dichotomous: households are either 'of interest' or 'not of interest'. For a planner, frustrated at the limited range of cross tabulations provided by the SAS, the unmixing approach might be used to model how many households of interest there are in an area.

The analysis could also be more complex; for example, suppose news is received of an impending ship yard closure. Following the closure, the pressure exerted on social services by different households would vary according to their characteristics and circumstances. If the vulnerability to the closure (i.e. employment in the ship construction industry) and details of other household circumstances were known, an unmixing approach would be able to determine which areas of the town might be worst affected. The classification on which the unmixing would be based should separate those households at risk (i.e. those in which a ship worker resides) from the others, and also distinguish between households with potentially different resource demands, based on their demographic and economic circumstances. This kind of analysis would not be possible using aggregate census data, since the precise combinations of household characteristics required are not available in the SAS tabulations.

### 8.3.2 - Variables of Interest Associated With SAR- based Classes

It also seems likely that users may be interested in an aspect of household circumstances which was not actually a variable in the census, but which may be strongly associated with combinations of variables which were featured in the census. For example, differences in morbidity are associated with social class, household circumstances, occupation and gender (Curtis and Taket, 1996, Jones and Moon 1987). Using a SAR- based classification which separated various combinations of household characteristics and the subsequent association of likely morbidity rates with those classes, a model of the probability of morbidity in small area could be produced. Again, this kind

of analysis might be useful in the allocation of resources, but also might find a role in the continued investigation of 'area effects' in medical sociology and medical geography. The investigation of area effects concerns determining whether people's location has an impact on their health over and above their own personal characteristics. An unmixed model which was able to provide projected morbidity rates by small area, could then be compared with survey data which described actual rates to determine the spatial patterning of any discrepancies. For this analysis to be successful, the association between the household type and the specific behaviour or feature of interest would need to be very strong. The user would also need to be aware of the difficulties inherent in establishing the quality of the model before drawing solid conclusions from it.

### 8.3.3 - Non Census Based Disaggregate Data

The third possibility is the use of a non-census based disaggregate data source. Provided that this source contained a reasonable number of variables common to the SAS, so that network training could be carried out as described in section 6.7, there is no reason why the unmixing technique could not be attempted. As with the suggestion above, the association between the variables of interest and those common to the SAS data would have to be very strong for the resultant model to be meaningful. If this were the case, it raises the intriguing possibility of being able to assign the results of a survey to households in EDs throughout the country. The user would however, need to be certain that the applicability of the survey data would not vary according to geographic location. This idea is in fact a hybrid of the unmixing concept and data fusion (see section 4.6). The unmixing technique would model sets of household characteristics from aggregate data, and survey responses would then be donated to those 'households'. This 'two stage' modelling process might not yield particularly good results because of the large number of potential error sources, but is nonetheless an interesting idea.

### 8.3.4 - Non Census Based Aggregate and Disaggregate Data

The fourth, and final suggestion is that the unmixing takes place using aggregate and disaggregate data which are all non-census based. The data need not have even come from the same source. One possible area in which this version of unmixing technique might be applied is in retrospective analysis of surveys. Suppose for example, a county council has undertaken an extensive survey of housing need for its five year strategy review. The results of the survey have been published, on an aggregate basis, in the council's planning documentation. Subsequently, a researcher who is interested in how

the various dimensions of housing need are related to household circumstances and to location, wishes to dissect the results of the council's survey with these questions in mind. By carrying out a survey which explores the interaction of household characteristics and housing need, and which uses the same core variables as the council's original survey, the researcher can take advantage of the council's survey work. Their aggregate results could be unmixed to show how what kinds of households comprise the need in each area.

All of the proposals above are speculative, none has been tried in practice. However it should be clear that once a researcher starts to think in an 'unmixing way', the possibilities rapidly expand. What will be certain is that the ability to produce a classification scheme which is strongly related to the phenomena of interest will be paramount in determining whether the unmixing approach is successful. This research has developed the concept and technique. Clearly there is scope for its refinement and experimental application in different situations.

## *8.4 - Conclusions*

The unmixed model of Southampton has been tested in a quantitative and qualitative manner. Whilst the nature of the new information makes it difficult to assess with precision, it would appear that the unmixing technique is capable of producing an accurate and plausible model of socio-spatial structure. The complex nature of the model's output does make for some difficulties in its visualisation and analysis, however the unique nature of the information now available means that more meaningful summary statistics for the EDs can be utilised. Where these summary statistics show something of interest, the full version of the model can be brought to bear in the search for an explanation or some explanatory context. The model is also suited to analysis area by area, or class by class.

Only the most basic analysis of Southampton's socio-spatial structure has been undertaken here, however it has been enough to suggest that the theories of residential differentiation may require some adjustment in the light of what the model shows about how different kinds of household co-locate. Hopefully, this basic analysis has also shown the model's potential to enhance social justice in the allocation of resources.

The speculative extensions to the technique in the chapter's final section have introduced a bewildering range of possibilities for the application of this

technique. It may be that the nature of the data sets available to users are not as conducive to unmixing as one would hope. However, once the technique is tested and developed, it may influence the way in which data sets are collected and used. One interesting possibility lies in the ever present threat of a reduction in the decennial census funds. If the UK census programme was to follow the American model, in which not everybody fills in a census form with the same level of detail, the potential for unmixing to synthesise the 'missing' data is intriguing. Small area statistics might be based on a 100% coverage of the population in core variables and use a disaggregate sample and unmixing to provide further, synthetic tabulations. Much development work will be required in the meantime. The following, final, chapter will draw together these and other significant points from the research. It will also explore the potential for the application of the technique without spatial boundaries and to non spatial data.

- The unmixed model of Southampton's socio-spatial structure has been shown to be plausible.

- New information about the city and its inhabitants has been demonstrated.

- Difficulties in dealing with the large amount of information which the model provides, have been experienced. GIS may provide a useful platform for managing the model.

- Useful summary statistics, in conjunction with the full version of the model may provide the best way of working.

- The potential for the development of the unmixing technique, beyond a reliance on census data and into a diverse range of applications has been illustrated.

# Chapter 9

## Chapter Overview

This chapter presents the concluding comments and analysis for the research as a whole. Since each chapter has been concluded at its close, there is no chapter by chapter review. Instead, the most significant achievements and findings from the research are brought together and their implications are discussed. The chapter is divided into two sections. Firstly the technical achievements of the research and its potential for further development are explored. Secondly, the relationships between theory, data and method within the research are reviewed.

## 9.0 - Introduction

"A conclusion should not be permitted to introduce new material, but it should be permitted to re-constitute the old" Harvey (1973) p.286

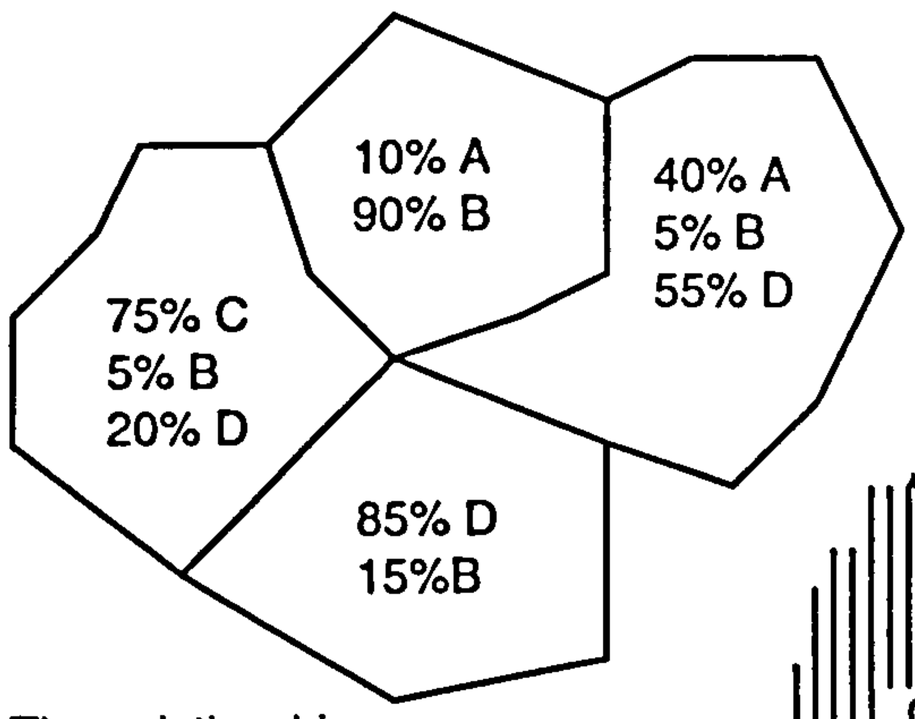
Harvey is not entirely correct in this assertion. The role of a conclusion is to draw together the findings of the research (the old) but also, through this synthesis, to look beyond its immediate confines towards the potential for its development and its implications (the new). This final chapter will do both of these things. It is divided into two sections. The first will review the achievements of the research in a technical sense and explore questions this raises. The second will examine the conduct of the research and review its interaction with, and contribution to, methodology and theory. The chapter will conclude with a short series of summary statements.

## 9.1 - Technical Achievements

This research has produced a new model of socio-spatial structure for the City of Southampton (see figure 2.8) and more importantly, a method which will facilitate a similar analysis for any part of the United Kingdom. The tasks outlined in chapter 2 (see table 2.3) have been completed. Through the unmixing technique, two enhancements to information available about residential populations have been created. Model users now know the extent of social-variation within an ED and the nature of that variation. The model has been tested as far as possible and found to be accurate in both a qualitative and quantitative sense (see chapter 8). The model's potential to be useful to the academy and the commercial world has been demonstrated (see section 8.3).

The research concentrated on establishing a technique to integrate disaggregate and aggregate data but it is appropriate to consider how the technique might be further refined or developed. Perhaps the most exciting prospect for developing the model would be to explore the possibility of removing its dependence on the bounded spatial unit. The transfer from a unit classification (which recognises the variety of households present within that unit) to some kind of surface analysis, presents both an intriguing challenge and a potentially rich source of information. Where 'neighbourhoods' have been artificially divided by the ED boundary construction process, an unmixed surface offers the possibility of their reconstruction through an evaluation of the 'contents' of adjacent EDs.

*Figure 9.1 - The Potential for Generating a Neighbourhood Surface*




(1) The relationship between adjacent EDs is assessed in terms of their class content.




(2) Using some kind of surface generation algorithm, the classes are 'joined' in space, to produce neighbourhood boundaries which are based on household type rather than administrative units

The spatial reconstruction of neighbourhoods might be assisted through the use of other reference data on house type and location, drawn from council data bases. It is likely that the success of neighbourhood reconstruction would depend on the classification scheme employed in the initial unmixing process.

In many ways, a surface based model would greatly assist quantitative exploration of theories of residential influence (for example in the construction of class identity or in the assessment of area effects on mortality

or morbidity rates). One of the greatest difficulties facing researchers interested in studying the effects of neighbourhood is that statutory or census boundaries do not conform to ecological patterns in residential areas (Morphet 1993). However, even if it were possible to reconstruct neighbourhood boundaries based on the household types present on the ground, this would still constrain the concept of neighbourhood to a spatial definition, rather than a social one. Nonetheless, it is an intriguing possibility.

In addition to potential for the development of the model's spatial context, there is also potential for development in an aspatial context. Although this research was inspired by situations in which the aggregate unit of interest is a spatial one, numerous situations exist in which this is not the case. In education for example, units may be formed in many different ways. The cohort, school, subject, pupil's background or ability might all form units of interest. The unmixing technique provides the potential for analysis 'within' these aggregate units. Provided suitable aggregate and disaggregate data exist, there is no reason why the technique could not be tried in a wide variety of situations.

## *9.2 - Problems with Innovation*

Although the technique has been successfully developed and it would appear that many possible applications remain to be explored, like all new concepts, the unmixing technique has raised some potential problems. These concern determining the quality of the information which has been produced and managing its quantity and nature.

### 9.2.1 - Limitations on Model Testing

As noted in section 8.1, the research has produced a new level information. There is no comparable data source with which to evaluate the unmixed model of Southampton, or that for any other location to which the technique is applied. Whilst it has been possible to apply tests which suggest that the technique performs well, it has not been possible to accurately quantify its inaccuracies. The tests have merely confirmed the model's *plausibility*. One possible solution to this problem would be to negotiate access to much more detailed crosstabulations of census data at enumeration district (ED) level. The Office of National Statistics will provide 'custom' designed crosstabulations, although only within their remit to protect the identities of individual census respondents. It might be possible to negotiate access to crosstabulations which provide enough information about the combinations of household characteristics within an ED to offer a better indication of the

model's quality. Until some means of rigorously evaluating the new model is found, users adopting the unmixing technique must remain aware that it is, to an extent, of unknown reliability.

### 9.2.2 - Quantity and Nature of The New Information

The unmixing of an ED into 15 constituent classes has massively increased the information available about areas. With a single number area classifier, the socio-spatial structure of the City of Southampton can be described with 418 items of information. With the 15 class unmixed model, this number rises to 6270 items. Additional complexity is provided by the fact that variance also exists within the classes themselves (see section 7.10.1). The management and analysis of this quantity and quality of data is a sizeable task. Discussion has already been presented (see chapter 8) of the utility of summary measures which describe the 'dominant class' in an area or provide a measure of its social cohesiveness. These have been applied with the proviso that a much greater depth of information lies behind these indicator variables, should the user wish to access it. However, given the difficulty of visualising and managing the full form of the model and the tendency to build aggregate, summary statistics from it, a question arises over whether this kind of model marks any real progress. Is the blanket application of an entropy score to a spatial unit actually any enhancement over the application of a single number area classification? Until visualisation techniques are developed which can cope with fuzzy classifications of the kind developed here, there is clearly a balance to be sought between the complexity and the utility of information.

This is an extremely important issue if the technique is to be applied elsewhere and by others. The author has spent time considering the 'within class' variation and the 'within ED' variation and how the two are related. Other users would need to do the same, but there can be no guarantee that they will do so. Just as many researchers from many backgrounds fail to recognise the significance of the ecological fallacy (which this research was designed to tackle) and as a result misuse or misunderstand data, so the new model is prone to misuse or misunderstanding. The research may have successfully produced a technical solution to problems with aggregate data, but it cannot address the (potential) 'human' problems with its use. Whatever the technique employed, control of the quality and applicability of the data analysis remains with the researchers in question. This thesis has provided the potential for significantly enhanced analyses of aggregate data and an example of good practice in quantitative research, but it cannot force people to take heed.

## *9.3 - Theory / Data / Method*

In this research, the relationship between social theory and quantitative technique has been a close and fruitful one. From the development of a strategy for integration to variable selection and manipulation, relevant theory and technical understanding have merged to assist in and clarify decision making (see chapters 3 and 6). The research has highlighted the gulf which currently exists between the concepts and constructs which census data should ideally utilise and illustrate, and those which actually exist. The rigid nature of the data in question and the desire to produce a working method led to a dominance of 'technical interests' in the decision making process. The research experience suggested that enhancements and adjustments to the census data should be made at the point of their construction (see section 7.13). Theoretical insights into the shortcomings of the data make for a comprehensive and honest research practice, but do not actually solve the problems which inadequate data may create.

Theory was not only used to decode and understand the variables and concepts in the data sets used. The drawing of an analogy between earth observation's analysis of remotely sensed imagery and the 'image' of the city that census data produce, was a direct result of understanding the ways in which residential processes are thought to work (see section 5.2). If the acid test of adopting a clear theory-method partnership is to determine how different the outcome of research would be without that partnership, the development of that analogy and the subsequent success of this research showed clearly that such an approach is beneficial. Indeed, the drawing of analogies between the research problems faced by those in other disciplines and those in geography may provide a useful tool for research development. However, these analogies can only be drawn, when the nature of the data is understood and the mechanisms through which people are represented within those data are clear. Quantitative and qualitative science as a whole needs the approach successfully demonstrated here. The theory-data-technique connection must be made explicit in research. When this occurs, each element in the chain will be enhanced.

The research has produced an analysis which may have something to contribute to social theory, as well as having drawn from it. Just as the theory of residential differentiation was used to create an analogy between earth observation and urban geography, the fuzzy classification model which resulted has provided a means for a quantitative exploration of that theory.

The residential co-location of different social groups can now be explored on a large scale at a new level of detail. Systematic analysis of the structure of districts and neighbourhoods may enhance theories which explore their creation and maintenance.

Beyond the direct implications for urban social structure research, this work has also made interesting inroads to the problem of balancing ecological and individual analysis. When taken alone, each approach to social phenomena can be misleading. The ecological approach can miss the diversity of individual lives, lost in the averaging or totalling process. The individual analysis may reveal the processes of life in detail, but may not be easily generalised to a wider population, or show how that individual differs or is similar to 'others' in another situation or place. The research described in this thesis illustrates how theory and technique can be utilised to commence the construction of a 'middle ground'. The unmixing process draws information from the individual and the area, making it easier to see how the relationship between individual and area actually works. The potential for using primary disaggregate data to enhance secondary aggregate data offers a bridge between qualitative understanding drawn from survey and interview work, and larger scale quantitative illustration. Provided that the primary data are constructed in with a sufficient number of variables to match the secondary data, the unmixing technique may be able to connect the two different scales of analysis.

It should now be clear that by making the theory-data-method connections explicit and giving them a high profile in the research process, this research has made an original and important contribution to social and quantitative geography. For the reader interested in pursuing the methods and themes explored this research, its key components are detailed in Mitchell, Martin and Foody (forthcoming).

In addressing a technical difficulty in the use of aggregate data to study or make inferences about residential populations, the work has yielded a new approach to handling census data. This new approach has been derived from, and contributed to, theories of urban process but the concepts and techniques used and developed here are applicable in a much wider arena. It may be that in taking the first steps in unmixing, the work has begun to forge much stronger bonds between different kinds of data and different kinds of science.

## *9.4 - Key Messages*

- A new kind of information about who lives where has been produced

- The concept and technique used to do this has many other potential applications

- An understanding of the relationship between theory, data and technique has been paramount in this success.

*"All this time there was just the sun and the silence, with the sound of the little spring and the three notes." Camus - The Outsider, 1942*

## *Appendix I: The Small Area Statistics*

- The Small Area Statistics (SAS) represent the output from the 1991 Census Of Population with the finest spatial resolution.

- The SAS are arranged into a tabular format. Each table is available for every enumeration district (ED), and may thus be aggregates to every higher spatial unit in the census geography (see Martin 1993).

- The country is divided into approximately 130,000 EDs.

- There are approximately 9,000 SAS values for each ED.

- Each table within the SAS provides a cross-tabulation for between two and four variables derived from the census questionnaire. Each cell within a SAS table will contain the total count of individuals or households with that combination of characteristics, within the ED which the SAS are describing, i.e. the data are aggregate.

- To preserve the anonymity of enumerated residents (and to save money), not all the information recorded by the census questionnaire is used in the SAS. For some questions, only a 10% sample is processed:

<u>Questions Processed for 100% of Respondents:</u>
sex and date of birth
marital status
whereabouts on Census night
usual address
term-time address of students
usual address one year before census
country of birth
ethnic group
long-term illness
Welsh/Gaelic language
economic activity in the week before the Census
type of accommodation and extent of sharing
number of rooms
tenure of household
household amenities
availability of cars

lowest floor level of accommodation (Scotland only)

Questions Processed for 10% of Respondents:
relationship in household
hours worked
occupation
name and business of employer
workplace
journey to work
higher qualifications

- Users may identify a particular combination of characteristics in the SAS using a referencing system. For each ED, every available cell can be specified by an alphanumeric code.

- A very small number of EDs have populations too small to release SAS without the risk of disclosing the identity of respondents. SAS are not released for areas with fewer than 50 usually resident people and 16 resident households. Data for these *suppressed* EDs are amalgamated with those of a neighbouring ED. Those EDs which have received extra counts in this way, are identified.

- In order to further protect the anonymity of enumerated individuals and households, a random adjustment of all 100% SAS counts (either by -1, 0 or +1) takes place. The exceptions to this adjustment are tables 1, 27 and 71 and the establishment count in table 3.

- Data for households which were absent from their usual address on census night, and which did *not* subsequently respond to the questionnaire as requested, were imputed using data collected from forms *completed* by absent households.

Information in this appendix was compiled from Martin, 1993 and OPCS 1991.

## *Appendix II - The Samples of Anonymised Records*

- The Samples of Anonymised Records (SAR) consist of two disaggregate data sets drawn from the 1991 Census of Population

- The sample at individual level contains 1,116,181 records, (1 in 50 of the enumerated population). The sample at household level contains 215, 789 households (with 541, 894 persons within those households). This is a 1% sample of the enumerated population.

- The spatial origin of a record in the individual sample is restricted to the equivalent of large local authority areas with a population size of at least 120,000. For the household sample, the spatial origin is restricted to within one of the Registrar General's Standard regions. The exception is the South East, which has been split into inner London, outer London and the rest of the south east.

- The sampling procedure which generated the SAR was as follows:
  The household level SAR was drawn from the 10% sample of records which are fully coded for the small area statistics (see appendix I). Households were ordered according to their geographical location and then placed into groups of ten. One household was selected at random from each group. Families were allocated a family number to enable their identification as units within households. The individual level SAR was selected from the remaining households to insure that there were no cases appearing in both samples. The individuals in these households were ordered into groups of nine. Two individuals were selected at random from these groups.

- A guide to sampling error in the SAR is given in the 'User Guide to The SARs' published by the CMU at Manchester.

- Table AII.1 provides a detailed breakdown of the variables available in the SAR. (source: User Guide to the SAR, CMU 1993)

# *Table AII.1 - Variables Available in the Samples of Anonymised Records*

Notes: The maximum number of categories as available on the full census database

| Item | Household (1%) sample | | Individual (2%) sample | |
|---|---|---|---|---|
| | No. of categories (maximum) | Other details | No. of categories (maximum) | Other details |
| Geographical area of enumeration | 12 | Standard regions of England (with split of S.East into Inner London, Outer London & Rest), Wales & Scotland | 278 | Local authority districts over 120,000 population. Others amalgamated to form areas over 120,000 |
| Housing/household information | | | | |
| Accommodation type | 14 (14) | Detached, semi-detached or terraced house; purpose-built flat in a commercial or residential building; converted or not self-contained accommodation in a shared house or flat | As household (hhld) sample | |
| Availability of amenities | | | | |
| - bath/shower | 3 (3) | Exclusive, shared or no use | As hhld sample | |
| - inside WC | 3 (3) | Exclusive, shared or no use | As hhld sample | |
| - central heating | 3 (3) | Full, part or none | As hhld sample | |
| Cars (number of) | 4 (4) | Top coded: 3 or more | As hhld sample | |
| Floor level (lowest) of of accommodation (Scotland) | 7 (101) | Basement, Ground, 1st/2nd, 3rd/ 4th, 5th/6th, 7th to 9th, 10t or higher | As hhld sample | |
| Number of household (accommodation) spaces in dwelling | 4 (35) | Top coded: 4 or more | Not included | |
| Number of persons (enumerated) in household | 12 (99) | Top coded: 12 or more | Not included | |
| Number of residents in household | | Derivable | 4 (99) | 0, 1, 2 to 5, 6 or more |
| Number of dependent children in household | | Derivable | 2 (99) | 0, 1 or more |
| Number of pensioners in household | | Derivable | 2 (99) | 0, 1 or more |
| Number of persons with long-term illness in household | | Derivable | 2 (99) | 0, 1 or more |
| Number of persons in employment in household | | Derivable | 3 (99) | Top-coded: 2 or more |
| Number of rooms | 15 (19) | Top-coded: 15 or more | Not included | |
| Number of persons per room | | Derivable | 5 | Ranging from less than 0.5 to greater than 1.5 |
| Tenure | 10 (10) | Owner-occupier or Rented (Public sector or Private) | As hhld sample | |
| Wholly moving household indicator | 2 (2) | Yes (all resident household members are migrants from the the same address) or No | Not included | |

*Table AII.1 - Continued*

| Item | Household (1%) sample | | Individual (2%) sample | |
|---|---|---|---|---|
| | No. of categories (maximum) | Other details | No. of categories (maximum) | Other details |
| Individual information | | | | |
| Age | 94 (111) | Single years 0 to 90, 91/92, 93/94, 95 and over | As hhld sample | |
| Status in Communal establishment | Not applicable | | 3 (7) | Visitor, resident staff or resident non-staff |
| Type of Communal establishment | Not applicable | | 15 (35) | Hotel,hospital,nursing home, etc |
| Country of birth | 42 (102) | | As hhld sample | |
| Migrants - distance of move (km) | 13 | 5, 10, 20 and 50 km bands; top-coded above 200km. | As hhld sample | |
| Distance to work (km) | 8 | 10km bands; top-coded above 40km; 0-9km band split 0-2, 3-4 and 5-9 | As hhld sample | |
| Economic position - primary | 10 (12) | Employee, Self-employed, Unemployed, Student, Retired etc. | As hhld sample | |
| - secondary | 7 (10) | | As hhld sample | |
| Economic position of family head | | Derivable | 3 (12) | Employed, unemployed or inactive |
| Employment status | 5 | Employee, manager,self-employed etc | Not included | |
| Ethnic group | 10 (10) | | As hhld sample | |
| Family head indicator | 2 (2) | Yes or No | Not included | |
| Family number | 5 (5) | Used to identify individual's family | Not included | |
| Family type | 8 (8) | Married or Cohabiting couple family with or without children or Lone parent family | As hhld sample | |
| Gaelic language (Scotland only) | 5 (8) | Ability to speak, read or write Gaelic | As hhld sample | |
| Hours worked weekly | 72 (99) | Single hours 0-70, 71 to 80, 81+ and over | As hhld sample | |
| Industry of employees and self-employed | 185 (334) | Mainly third digit (groups) of 1980 SIC | 60 (334) | Mainly second digit (classes) of 1980 SIC |
| Limiting long-term illness | 2 (2) | Yes (individual has illness) or No | As hhld sample | |
| Marital status | 5 (5) | | As hhld sample | |
| Migrant - geographical area of former residence | 13 | Standard regions of England (with split of South East), Wales, Scotland, outside GB. | As hhld sample | |
| Occupation | 358 (371) | Mainly unit groups of 1990 SOC | 73 (371) | Mainly minor groups of 1990 SOC |
| Number of higher educational qualifications | 3 (7) | 0, 1, 2 or more | As hhld sample | |

*Table AII.1 - Continued*

| Item | Household (1%) sample | | Individual (2%) sample | |
|---|---|---|---|---|
| | No. of categories (maximum) | Other details | No. of categories (maximum) | Other details |
| Level of highest qualification | 3 (3) | Higher degree, First degree, Above GCE 'A' Level | As hhld sample | |
| Subject of highest qualification | 88 (108) | Mainly third digit of Standard Subject Classification | 35 (108) | Mainly second digit of S.S.C |
| Relationship to household head | 17 (17) | | 8 (17) | |
| Resident status | 3 (3) | Present resident, Absent resident, Visitor | As hhld sample | |
| Sex | 2 (2) | | As hhld sample | |
| Sex of family head | | Derivable | 2 (2) | |
| Social Class | 8 (8) | | As hhld sample | |
| Social Class of family head | | Derivable | 8 (8) | |
| Socio-economic group | 19 (20) | | As hhld sample | |
| Term-time address of students and school-children | 4 | Inside or outside region of usual residence | As hhld sample | |
| Transport to work (mode) | 10 (10) | | As hhld sample | |
| Visitor - geographical area of residence | 13 | Standard regions of England (with split of South East), Wales, Scotland, outside GB | As hhld sample | |
| Welsh language (Wales only) | 5 (8) | Active use of (speak, read or write) | As hhld sample | |
| Workplace | 5 | Inside or outside region of usual residence | 5 | Inside or outside SAR area of usual residence |

## *Appendix III: Backpropagation in SNNS*

The formula for backpropagation used by SNNS is

$$\Delta w_{ij}(t+1) = \eta \delta_j o_i + \mu \Delta w_{ij}(t)$$

where

$$\delta_j = (f_j'(net_j) + c)(t_j - o_j) \text{ if unit } j \text{ is an output node}$$

$$\delta_j = (f_j'(net_j) + c)\sum_k \delta_k w_{jk} \text{ if unit } j \text{ is a hidden node (Mache 1995)}$$

$\delta_j$ is the difference between a teaching value $t_j$ and output value $o_j$

$net_j$ is the net input at unit $j$ at step $t$

$\eta$ is a learning parameter which specifies the step width of the gradient descent and thus, to some extent, the speed of learning. For our purposes this value was set initially at 0.8.

$\mu$ is the momentum term. This value specifies the amount of 'old' weight change (i.e. adjustment on the previous wave of adjustment) to 'new' change. By retaining some of the old change, the drive to optimum learning is maintained. This is because the reduction in error which would have been achieved if weight change had proceeded without interference is instead only partially carried out. This has the effect of raising the total error that the ANN is making and maintaining its impetus to reduce error further. The momentum term value was initially set at 0.2. By convention, the momentum term and the learning parameter sum to 1.0.

$c$ is the flat spot elimination value. This is a constant value which is added to the output of the activation function in an attempt to 'nudge' learning over any flat spots on the error curve (as shown in figure 5.6). This value was initially set at 0.1.

## *Appendix IV - Experiments to Determine Optimum ANN Parameters*

As explained in section 7.7.1, experiments to determine optimum parameters for the ANN, took place in groups. In each group, a set of parameters was held constant whilst one other was varied. The experiments are presented in their groups, with the parameters which were held constant as a *group heading* (shown in italics) and the parameter which varied described beneath:

The following abbreviations have been used to keep the presentation concise:

1a. SA1 = Standard ANN architecture number 1. This consisted of 47 input nodes, 15 hidden nodes and 12 output nodes.
1b. SA2 = Standard ANN architecture number 2. This consisted of 47 input nodes, 95 hidden nodes and 12 output nodes.
2. LP = Learning parameters, controlling the operation of the back propagation algorithm. There are three parameters (see appendix III) $\eta$, $\mu$, and $c$ and they are presented in that order in this appendix. For example, LP; 0.8, 0.2, 0.01 refers to a $\eta$ value or 0.8, a $\mu$ value of 0.2 and a $c$ value of 0.01.
3. UP = Update order. SNNS allows the user to chose the order in which weight updates are applied to nodes.
4. EP = Epochs. The number of cycles in which the ANN is exposed to the training data.
5. IN = Initialisation. The initialisation of the ANN weights was always random, unless stated.

<u>Experiments Carried Out:</u>

*SA1, UP = Topological Order, EP = 100*

1. LP = 0.8, 0.2, 0.01
2. LP = 0.7, 0.3, 0.01
3. LP = 0.2, 0.2, 0.01
4. LP = 0.5, 0.5, 0.01

*SA1, LP = 0.8, 0.2, 0.01, EP = 100*

5. UP = Random Order
6. UP = Serial Order
7. UP = Synchronous Order

*SA1, UP = Topological, LP = 0.8, 0.2, 0.01, IN = RM_Random_Weights*

8. EP = 100

9. EP = 50

10. EP = 500

11. EP = 1000

*UP = Topological, LP = 0.8,0.2, 0.01, EP = 100*

12. Hidden Layer = 95 units

13. Hidden Layer = 10 units

14. Hidden Layer = 2 rows or 7 units

*SA1, UP = topological, LP = 0.8, 0.2, 0.01, EP = 100*

15. Training Data not presented in a random order

*UP = Topological, LP = 0.8,0.2, 0.01, EP = 100*

16. Net architecture; 19 + 2 layers of 14 input nodes, 3 layers of 5 hidden nodes, 2 layers of 6 output nodes; Training data presented in random order

17. Training data not presented in random order

*SA2, UP = Serial, LP = 0.8, 0.2, 0.01*

18. EP = 50, IN = RM_Random_Weights

19. EP = 100

*SA2, IN = RM_Random_Weights, LP = 0.5, 0.5, 0.01, UP = Serial*

20. EP = 50

21. EP = 80

22. EP = 100

*SA2, IN = RM_Random_Weights, LP = 0.8, 0.2, 0.01, UP = Serial*

23. EP = 50

24. EP = 80

25. EP = 100

*Net architecture = 47 input nodes, 2 layers of 47 hidden nodes and 12 output nodes, IN = RM_Random_Weights, LP = 0.5, 0.5, 0.01, UP = Serial*

26. EP = 50

27. EP = 80

28. EP = 100

*Net architecture = 47 input nodes, 2 layers of 47 hidden nodes and 12 output nodes, IN = RM_Random_Weights, LP = 0.8, 0.2, 0.01, UP = Serial*

29. EP = 50
30. EP = 80
31. EP = 100

*Net architecture = 47 input nodes, 2 connected layers of 47 hidden nodes and 12 output nodes, IN = RM_Random_Weights, LP = 0.5, 0.5, 0.01, UP = Serial*

32. EP = 50
33. EP = 80
34. EP = 100

*Net architecture = 47 input nodes, 2 connected layers of 47 hidden nodes and 12 output nodes, IN = RM_Random_Weights, LP = 0.8, 0.2, 0.01, UP = Serial*

35. EP = 50
36. EP = 80
37. EP = 100

*SA2, UP = Serial, LP = 0.8, 0.2, 0.01*

38. EP = 100
39. EP = 500

N.B. With the exception of experiments 26 - 37, each bundle was repeated with different training and testing data sets. This brought the total number of experiments to 72.

The results table, AIV.1, provides a correlation between target values and ANN output. Where the experiment was undertaken using more than one training or testing data set, an average result is shown.

*Table AIV.1 - Correlations From Experiments With ANN Parameters*

| Test No. | Correlation | Test No. | Correlation | Test No. | Correlation |
|---|---|---|---|---|---|
| 1 | 0.87 | 14 | 0.80 | 27 | 0.88 |
| 2 | 0.90 | 15 | 0.92 | 28 | 0.85 |
| 3 | 0.90 | 16 | 0.90 | 29 | 0.92 |
| 4 | 0.92 | 17 | 0.91 | 30 | 0.91 |
| 5 | 0.02 | 18 | 0.87 | 31 | 0.92 |
| 6 | 0.91 | 19 | 0.88 | 32 | 0.90 |
| 7 | -0.09 | 20 | 0.90 | 33 | 0.92 |
| 8 | 0.91 | 21 | 0.92 | 34 | 0.91 |
| 9 | 0.89 | 22 | 0.89 | 35 | 0.91 |
| 10 | 0.90 | 23 | 0.90 | 36 | 0.91 |
| 11 | 0.86 | 24 | 0.90 | 37 | 0.92 |
| 12 | 0.96 | 25 | 0.89 | | |
| 13 | 0.88 | 26 | 0.87 | | |

There was considerable variation in the results achieved class by class, when testing took place on different training data sets. The following tables illustrate this variation for test 19.

*Table AIV.2 - Class By Class Correlations for Test 19; Test Data Set 1*

| Class | Correlation | Class | Correlation |
|---|---|---|---|
| 1 | 0.95 | 7 | 0.97 |
| 2 | 0.98 | 8 | 0.94 |
| 3 | 0.96 | 9 | 0.96 |
| 4 | 0.97 | 10 | 0.96 |
| 5 | 0.94 | 11 | 0.97 |
| 6 | 0.97 | 12 | 0.97 |

*Table AIV.3 - Class By Class Correlations for Test 19; Test Data Set 2*

| Class | Correlation | Class | Correlation |
|---|---|---|---|
| 1 | 0.90 | 7 | 0.89 |
| 2 | 0.93 | 8 | 0.88 |
| 3 | 0.80 | 9 | 0.86 |
| 4 | 0.81 | 10 | 0.89 |
| 5 | 0.90 | 11 | 0.90 |
| 6 | 0.88 | 12 | 0.78 |

## *Appendix V - GroupMaker Code*

```
/* GroupMaker reads in a set of SAR cases from a given file. It, re-codes the
 * data into an alphabetic sequence, examines the set of sequences and retains
 * a copy of each unique code, together with the total occurrences of that code.
 * This code & total information is sorted so that the most common codes are at
 * the top of the data set, and this sorted version is passed to a grouping
 * function, which allows the user to group codes together based on their
 *  relative similarity or difference. Difference is a controllable notion,
 * including the ability to state that 'key variables,' must be identical if
 * two or more codes are to be grouped */

#include<stdio.h>
#include<string.h>
#define STR_LEN 16
#define EXPECTED 41000
#define NAME_LEN 30
#define ENG_LEN 100
#define STD 10

        /* this function assigns each numerical portion of the input sequence an
         *  alphabetic equivalent */

int
alpherise(int code)


{
        char letter;

        char alpha [26] = {'A','B','C','D','E','F','G','H','I','J','K','L',
                           'M','N','O','P','Q','R','S','T','U','V','W','X','Y',
                           'Z'};

        code = code-1;

        letter = alpha[code];

        return(letter);


}


        /* this function acquires the parameters for the conversion */
        /* and then collects each variable in turn, sends it off to be */
        /* converted and then writes the new letter-code in a file */

        void
        coder(int *numcasesp, int *numvarp)

        {
                int i, n, code, numline, nomvar;
                char letter;

                char in_name [30];

                FILE *infilep, *outfilep;

                printf("Please enter the name of the file that you want to convert\n");
                scanf("%s", in_name);
```

```
infilep = fopen(in_name,"r");
outfilep = fopen("data.dat","w");

printf("Please enter the number of lines to be read\n");
scanf("%d", &numline);


printf("How many variables are there in the file?\n");
scanf("%d", &nomvar);

printf("\nConverting to alphabetic sequence\n\n");

/* the next section reads from the file requested above,
 * gets one variable value, sends it to alpherise to be
 * converted and then writes it to an other file called
 * data.dat */

for(n=0;n<numline;++n){          /* this loop will start a new line when
                                    one case is complete */


        for(i=0;i<nomvar;++i){   /* this loop will collect each variable
                                  * in turn */

                fscanf(infilep,"%d",&code);

                letter = (alpherise(code));

                fprintf(outfilep,"%c",letter);
                }

           fprintf(outfilep,"\n");   /* this starts the new case line when
                                      *  all variables are converted */

        }

        *numcasesp = numline;
        *numvarp = nomvar;


        fclose(infilep);   /* closes all the files */
        fclose(outfilep);


}

                /* the next bit writes a report to a file based on what's
                 * in the store and the totals found for each stored sequence */

char *
report( char store [] [STR_LEN],
        int totals [],
        int numinstore)

{

        int n, numexclude;
        int filter;
        char outpt [NAME_LEN];
        char fname[30];
        char junk [NAME_LEN];
```

```
FILE *dataout;
FILE *totsout;

dataout = fopen("data/store", "w");

totsout = fopen("data/totals", "w");

for(n=0;n<numinstore;++n){


                fprintf(dataout,"%s\n", store[n]);
                fprintf(totsout,"%d\n",totals [n]);
        }

        fclose(dataout);
        fclose(totsout);

}

        /* The next two functions are a recursive sorting mechanism */


void
place_largest(int totals[],  char store [][STR_LEN], int n)


{
        int temp,j,max_index;
        char worker[STR_LEN];
max_index = n-1;
for(j=n-2;j>=0; --j)

        if(totals[j] < totals[max_index])
                max_index = j;


if(max_index != n-1){

temp = totals[n-1];
strcpy(worker, store[n-1]);

totals[n-1] = totals[max_index];
strcpy(store[n-1],store[max_index]);

totals[max_index] = temp;
strcpy(store[max_index], worker);

        }
}


void
select_sort(int totals[],
            char store [][STR_LEN],
            int n)

{


        if(n>1){

                place_largest(totals,store,n);
                select_sort(totals,store,n-1);
```

```
        }

}

        /* this function will write a variety of reports to a specified file */

char*
writer(  char engsto[][ENG_LEN],
         char enggrp[][ENG_LEN],
         int totals[],
         int grouping[],
         int grptotal[],
         int numgroups,
         int numinstore,
         double cont[],
         int limit,
         int numvar)
{

         FILE *outpt;
         FILE *recout;

         int i, count, choice,p,totnum;

         char rec,junk;

         char outname[30];
         char recname[30];

         printf("\n\nPlease enter a name for the report file>   ");
         scanf("%s", outname);

         outpt = fopen(outname, "w");

         printf("\n\n***Please enter your choice of output***\n\n");
         printf("Number of groups and the total number of components [1]\n");
         printf("Group Definers and totals  with components and sub-totals [2]\n");
         printf("Just the group definers with group totals [3]\n");

         totnum = 0;

         for(p=0;p<numgroups;++p){

                  totnum = totnum + grptotal[p];
         }

         printf("Total number of households grouped is: %d\n", totnum);

         choice = 0;

         scanf("%d", &choice);


         /*printf("Record group numbers and totals for sorting; y/n ?....");

         scanf("%c%c",&junk, &rec);

         if(rec == 'y'){

                  printf("Filename for record? > ");
                  scanf("%s",recname);
```

```
        recout = fopen(recname, "w");
        for(i=0;i<numgroups;++i)
                fprintf(recout,"%d, %d\n",(i+1),grptotal[i]);

        fclose(recout);
}*/

fprintf(outpt,"\n\n**** Results of Grouping Using %d distance limit****\n\n",limit);

for(i=0;i<numvar;++i)
        fprintf(outpt,"Variable %d, had the control %.1lf\n", i, cont[i]);

fprintf(outpt,"\n\n\n");

if(choice == 1){

        fprintf(outpt,"Number of groups is %d\n", numgroups);
        fprintf(outpt,"Group Number                    Totals\n");

        for(i=0;i<numgroups;++i){

        fprintf(outpt,"%d %30d\n", (i+1), grptotal[i]);
        }

}

if(choice == 2 || choice == 3){
        fprintf(outpt,"Number of groups is %d\n\n", numgroups);
        fprintf(outpt,"Group Defined by                              Total Components\n\n");

        for(i=0;i<numgroups;++i){

                fprintf(outpt,"%s          %d\n", enggrp[i],grptotal[i]);
        }


        if(choice == 2){


fprintf(outpt,"\n\n\n**********************************************************************\n\n\n\n");

                for(i=0;i<numgroups;++i){
                fprintf(outpt,"GROUP NUMBER %d\n\n",(i+1));
                fprintf(outpt,"Group defined by %s\n", enggrp[i]);
                fprintf(outpt,"Group Total is %d\n\n", grptotal[i]);

                for(count = 0;count<numinstore;++count){

                        if(grouping[count] == i)

                                fprintf(outpt,"%s          %d\n", engsto[count],
                                        totals[count]);
                }

        fprintf(outpt,"\n\n\n\n");
        }

        }
}
```

```
        fclose(outpt);

}
                        /*constructs an understandable code from the letters.
                         *Uses a  reference file called 'text. */

char*
english(char in[][STR_LEN],
        char out[][ENG_LEN],
        int numtodo,
        int numvar)

{
        int i,v, temp, pain;
        char v1,v2,v3,v4,v5, junk, voc, move;
        int adrs,n;
        FILE *txtin;

        int varnum, z, start, stop;
        int right, numrets;

        right = 0;
        numrets = 0;
        adrs = 0;
        start = 0;
        temp = 0;
        /* do for all in the store */
        for(i=0;i<numtodo;++i){

        txtin = fopen("text", "r");

        for(v=0;v<numvar;++v){ /* do for each variable */



                voc = in[i][v];
                temp = ((int)voc - 65);

        /* move to right line */


        if(v!=0){               /* unless it's the first variable*/

                while(numrets != v){

                        fscanf(txtin,"%c", &move);
                        if(move == '\n')
                                ++numrets;

                }

        }

        /* look in the right place */

        adrs = temp * 5;


        for(n=0;n<adrs;++n){ /* go to correct abrev. for the letter code */

                fscanf(txtin,"%c", &junk);
        }
```

```
            /* get the the abrev. */

fscanf(txtin,"%c%c%c%c", &v1,&v2,&v3,&v4);

z=start;

out[i][z] = v1;
out[i][++z] = v2;

out[i][++z] = v3;
out[i][++z] = v4;
out[i][++z] = ' ';

start = ++z;

            adrs = 0;

  } /* close do for all variables */


numrets = 0;
fclose(txtin);
start = 0;



} /* close do for all loop */



} /* end of function */


            /* this function will write the groups to files */

char*
scribe(char store[] [STR_LEN],
            char grpdef[] [STR_LEN],
            int totals [],
            int grptotal[],
            int numingroup[],
            int grouping[],
            int numgroups,
            int numinstore)

{

            int i,grpno, z; /* loop counters */
            char fname[10];



            int deftot,value;

            FILE *grpout;
            FILE *totout;


            i=0;
```

```
printf("Enter name for GROUPS file > ");
scanf("%s", fname);
grpout = fopen(fname, "w");

printf("Enter name for TOTALS file > ");
scanf("%s", fname);

totout = fopen(fname, "w");



while(i<numgroups){

        fprintf(grpout,"%d\n", i);

fprintf(grpout,"%d\n", grptotal[i]);

fprintf(grpout,"%d\n", numingroup[i]);

for(z=0;z<numinstore;++z){

        if(grouping[z]==i){

                fprintf(grpout,"%s\n", store[z]);
                fprintf(totout," %d\n", totals[z]);
                deftot = deftot + totals[z];
        }

}

fprintf(grpout,"%s\n", grpdef[i]);

                value = grptotal[i]-deftot;
                fprintf(totout," %d\n", value);
                deftot = 0;
                ++i;

        }

        fclose(grpout);
        fclose(totout);
}


        /* This function will group the sorted codes together according to the
        * manner specified */


char*
compact(char store[] [STR_LEN],
        int totals[],
        int numinstore,
        int numvar)


{


        char grpdef[EXPECTED][STR_LEN];
        char enggrp[EXPECTED][ENG_LEN];
        char engsto[EXPECTED][ENG_LEN];
        double cont[STR_LEN];
        int grouping[EXPECTED];
        int grptotal[EXPECTED];
        int numingroup[EXPECTED];
        int grouped[EXPECTED];
```

```
int diffrec[EXPECTED];
int junk,j,i,v,z, totgrp;
double diff,temp,fdiff;
int limit, groupnumber, numgroups, found, new;
int subtot,tot, count, a, atot;

printf("\n\n***Now ready to group sequences***\n\n");
printf("Please enter the distance limit that you want to use\n");
scanf("%d", &limit);


/* Initialise lots of things */

diff = 0;
found = 0;
z=0;
totgrp = 0;
numgroups = 0;

printf("\n\n*** Please input the control parameters for each variable *** \n\n");
printf("To declare a variable as a key enter -1, or enter a weight\n");
printf("Enter 0 if that variable is to have no influence.\n");
printf("Enter -2 if the variable is truely categorical\n");
/* get the control parameters */

for(i=0;i<numvar;++i){

        printf("What is the control for variable number  %d?\n",i);
        scanf("%lf", &cont[i]);
}


printf("Now grouping cases\n");


while(!totgrp){ /* while not every case is grouped */


/* move to the next available, non-grouped sequence and copy it to grpdef*/

        i=0;
        while(!new && i<numinstore){
                if(grouped[i]!=1 && grouped[i] !=2){
                new = 1;
                strcpy(grpdef[z], store[i]);
                grptotal[z] = totals[i];
                grouped[i] = 2;
                grouping[i] = -99;
                numingroup[z] = 1;
                ++numgroups;


        }else
                ++i;
}

if(!new  && i==numinstore){

        printf("All codes are now grouped\n");
```

```
                totgrp = 1;
        }
        new = 0;


/* search all in store for those that are within 'difference' limit*/

                for(i=1;i<numinstore;++i){


                        for(v=0;v<numvar;++v){ /*look at each variable*/
        if(store[i][v] != grpdef[z][v]){

        /* if the two letters are different */
                temp = 0;

                if(cont[v]<0){/* if this is a special variable*/

                        if(cont[v] == -1)
                        diff = limit+1;

                        if(cont[v] == -2){
                                temp = 10;
                                diff = diff + temp;
                        }

                }

                if(cont[v]>0){

                                temp = (abs((int)store[i][v] - (int)grpdef[z][v]));
                                /*temp = temp*temp;*/ /* square distance */
                                diff = diff + (temp*cont[v]);

                        }


                } /*ends the 'differing variables loop */


        } /* ends do for all variables loop */


        if(diff<limit){ /* if store[i] is within difference limit, flag it as GROUPED, and
                                * grouping[i] = [z] */

        found = 1; /* switch on found flag */

        if(grouped[i] == 1){/* if this sequence is already grouped */


                if(diffrec[i]>diff){ /* if the previous difference
                                        * is more than this one */

                        a = 0;
                        atot = 0;
                        a = grouping[i];


                        diffrec[i] = diff; /* set new 'previous distance */
                        grouping[i] = z;   /* change grouping to this one */
```

```
                    ++numingroup[z];


                    subtot = 0;
                    subtot = grptotal[z];
                    grptotal[z] = subtot + totals[i];


                    atot = grptotal[a];
                    grptotal[a] = atot - totals[i];
                    --numingroup[a];


                    diff = 0;       /* reset diff*/
                }
}

if(grouped[i]!=1 && grouped[i] !=2){/* if the sequence isn't grouped
                                        & isn't a grpdef */

        grouping[i] = z;
        grouped[i] = 1;
        ++numingroup[z];
        subtot = 0;
        subtot = grptotal[z];


        subtot = subtot + totals[i];
        grptotal[z] = subtot;

        diffrec[i] = diff;
                diff = 0;
        }

        }else{ /* if store[i] isn't within the difference limit */

        diff = 0; /* reset diff */
        }


}/* close the do for all loop */


if(!found){ /*if we've gone through all of them and not found a match */

                ++z;

}else{ /* if it has been found*/


                ++z;
                found = 0; /* reset found */
        }


}/* close the do while we haven't done them all loop */


scribe(store, grpdef, totals, grptotal, numingroup,grouping, numgroups, numinstore);


printf("\nNumber of groups is %d, hit ctrl C to exit at this point\n", numgroups);
```

```
        printf("Now translating sequences\n");

        english(store, engsto, numinstore, numvar);
        english(grpdef, enggrp, numgroups, numvar);


        writer(engsto, enggrp, totals, grouping, grptotal, numgroups, numinstore, cont,limit,
numvar);



}

char *
control(int numinstore,
        int numvar)


{

        /*read in the data */

        char store[EXPECTED][STR_LEN];
        int totals[EXPECTED];
        char newcode[STR_LEN];
        char ch,pro,junk;
        int i,j,z;
        FILE *datin;
        FILE *totin;
        datin = fopen("data/store", "r");
        totin = fopen("data/totals", "r");

        /* get the store */

        i=0;
                for(z=0;z<numinstore;++z){

                        for (ch = getc(datin);

                                        ch!= '\n' && ch!= EOF && i<STR_LEN - 1;

                                        ch = getc(datin))

                                newcode[i++] = ch;


                        newcode[i] = '\0';

                        /* Discards any characters that remain on input line*/


                        while (ch!= '\n' && ch != EOF)

                                ch = getc(datin);

                        strcpy(store[z],newcode);

                        i=0;

}
```

```
/* get the totals */

for(z=0;z<numinstore;++z){

                fscanf(totin,"%d", &totals[z]);
        }

/*keep sending it to compact*/

compact(store, totals, numinstore, numvar);

}

        /*  The main program runs the coder, sort functions, grouper
         *  & report writer whilst also comparing each  newcode with
         *  all known sequences in 'store.' */

int
main(void)

{
        int numinstore;
        int numcases, numvar;
        char newcode[STR_LEN];
        char store[EXPECTED] [STR_LEN];
        char prec, repsort, junk,cont;
        char filename[30];
        int totals [EXPECTED];
        int temp;
        char option;
        char worker[STR_LEN];
        int n,i,ch,found, t, pass, sorted,h,tot;
        FILE *inptp;


        numinstore = 0;

        /* here we run the coding function to convert the numeric sequences to
         * alphabetic ones */

        printf("Hello. Is there a file prepared for grouping, called 'data/store'?\n");
        scanf("%c", &option);

        if(option == 'n'){

                printf("Has the input already been converted into file data.dat y/n?\n");
                scanf("%c%c",&junk, &prec);


                if(prec == 'y'){

                        printf("Please enter the number of cases and variables>\n");
                        scanf("%d%d", &numcases, &numvar);

                }else

                        coder(&numcases, &numvar);

                inptp = fopen("data.dat", "r");

                printf("Now processing cases...\n");

                for(t=0; t<numcases; ++t){
```

```
/* gets one line of data. Returns an empty string
 * on end of file. If data line will not fit in allotted space,
 * stores portion that does fit and discards rest of input line */


                i=0;

                for (ch = getc(inptp);

                                ch!= '\n' && ch!= EOF && i<STR_LEN - 1;
                                ch = getc(inptp))
                                newcode[i++] = ch;

                newcode[i] = '\0';

/* Discards any characters that remain on input line*/

                while (ch!= '\n' && ch != EOF)
                        ch = getc(inptp);

                i=0;

                found = 0;      /* sets the 'found,' flag to zero */


/* the next section assess the string, and either totals or stores */

                /*printf("Processing case number %d\n", t);*/

                while(!found && i < numinstore) {

                        if(strcmp(store[i], newcode) == 0)

                                found = 1;

                        else ++i;
}

                if (found){


                        totals[i] = totals[i] + 1;


                }else{

                        numinstore = numinstore + 1;
                        strcpy(store[i], newcode);
                        totals[i] = 1;

                }

printf("It's done now\n\n");

/* lets sort the two components by totals */

printf("Sorting the sequences by total found\n\n");

select_sort(totals,store,numinstore);

printf("There are %d sequences in store\n", numinstore);
printf("Do you want to write the results to a file?\n");
scanf("%c%c",&junk, &repsort);
```

```
            if(repsort == 'y'){

                    report(store, totals, numinstore);
            }

}

if(option != 'n'){

            printf("enter number of sequences in store, and the number of variables\n");
            scanf("%d%d", &numinstore, &numvar);

}
            control(numinstore,numvar);

            return(0);
}
```

## *Reference List*


Akbari, A. (1989) The benefits of immigrants to Canada - Evidence on Tax and Public Services. *Canadian Public Policy* 15 pp 424-435

Aleksander, I., Morton, H. (1990) *An Introduction to Neural Computing*, Chapman and Hall, London

Anderberg (1973) *Cluster Analysis for Applications,* Academic Press, New York

Antoine, J., Santini, G. (1987) Fusion Techniques: Alternative to Single Source Methods? *European Research*, August, pp178-187

Atkinson, P.M., Cutler, M.E. (1995) *Mapping Sub-Pixel Variation in Land Cover In the UK from AVHRR Imagery* DRA / University of Southampton, Southampton

Badcock, B. (1984) *Unfairly Structured Cities*, Blackwell, Oxford

Baker, K., Harris, P., OBrien, J. (1989) Data Fusion: an appraisal and experimental evaluation *Journal of the Market Research Society* Vol.31 No.2 pp 153-212

Barr, R., (1993) A Review, in Dale, A., Marsh, C. (eds) (1993) *The 1991 Census Users Guide*, HMSO, London

Beal, R., Jackson, T. (1990) *Neural Computing: An Introduction*, Institute of Physics Publishing, Bristol

Beatty, C., Fothergill, S. (1996) Labour Market Adjustment in Areas of Chronic Industrial Decline - The Case of the UK Coalfields *Regional Studies*, Vol. 30 No. 7 pp 627-640

Beaumont, J. (1991) *An Introduction to Market Analysis,* Concepts and Techniques in Modern Geography 53, GeoBooks, Norwich.

Beaumont, J., Inglis, K. (1989) Geodemographics in Practice - Developments in Britain and Europe, *Enviornment and Planning A* Vol. 21 pp. 587 - 604

Belanger, A., Rogers, A. (1992) The Internal Migration and Spatial Redistribution of the Foreign Born Population in the U.S. 1965-70, 1975-80 *International Migration Review* 26(4):1 pp 342-69

Bell, C. Newby, H. (1978) Community, Communion, Class and Community Action: The Social Sources of New Urban Politics in Herbert, D.T. Johnston, R.J. (eds) *Social Areas in Cities: Processes, Patterns and Problems* John Wiley and Sons, Chicheter

Birkin, M. Clarke, M. (1988) SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and examples *Environment and Planning A* Vol. 20 pp 1645-1671

Birkin, M., Clarke, G. (1995) Using microsimulation methods to synthesize census data in Openshaw, S. (ed) (1995) *Census Users Handbook,* GeoInformation International, Cambridge

Blake, M., Openshaw, S. (1994i) Selecting Variables For Small Area Classifications of 1991 UK Census Data *School of Geography Working Paper,* University of Leeds

Blake, M., Openshaw, S. (1994ii) GBprofiler:

Blalock, M. (1964) *Causal inferences in Nonexperimental Research* , University of N.Carolina, Chapel Hill

Breton, R., Isajiw, W., Kalbach, W., Reitz, J. (1990) *Ethnic Identity and Equality: Varieties of Experiance in a Canadian City* University of Toronto Press, Toronto

Bridge, G. (1994) Gentrification, Class and Residence - A reappraisal, *Environment and Planning D - Society and Space,* Vol. 12, No. 1 pp 31 - 51

Bronner, A.E. (1983) A way to handle non-response: Simulation of missing values *European Research* July pp 87 - 92

Brown, P. (1993) Personal Communication

Bryant, B.E. (1993) Stages of the 1990 Census data Collection *American Demographics* March

Burgess, E., McKenzie, R. (1925) *The City,* University of Chicago Press, Chicago

Burnham, D. (1983) *The Rise of The Computer State,* Random House, New York

Burrough, P., Frank, A.(1995) Concepts and Paradigms in spatial information: are current geographical information systems truly generic? *International Journal of Geographical Information Systems*, Vol. 9 No.2 pp 101-116

CACI (1992) MONICA (Commercial Literature)

Castells, M. (1977) *The Urban Question,* Edward Arnold, London

Coates, D. (1989) Britain. in Bottomore, T., Brym, R. (eds) *The Capitalist Class: An International Study,* Harvester Wheatsheaf, London

CCN Marketing (1993) The MOSAIC Postcode Classification System, (Commercial Literature)

CDMS (1993) Geodemographic Products (Commercial Literature)

Census Microdata Unit (1993) *A User Guide to the SARs* , University of Manchester, Manchester

Clark, D., Kahn, J., Ofek, H. (1987) City Size, Quality of Life and the Urbanization Deflator of the GNP 1910-1984 *Southern Economic Journal* 54 pp 701-714

Crompton, R.(1993) *Class and Stratification: An Introduction to Current Debate.* Polity Press, Cambridge

Crompton, R. (1995) Womens Employment and the middle class in Butler, T., Savage, M. (eds) *Social Change and the Middle Classes*, UCL Press, London

Curran, P.J.(1985) *Principles of Remote Sensing* Longman, Harlow UK

Curry, M.R. (1994) Image, Practice and the hidden impacts of geographic information systems *Progress in Human Geography* Vol. 18, No.4 pp 441-459

Curtis, S., Taket., A (1996) *Health and Societies,* Arnold: London

Dale, A., Marsh, C. (eds) (1993) *The 1991 Census Users Guide*, HMSO, London

Da Vanzo, J. (1978) Does unemployment affect migration? - Evidence from microdata *Review of Economics and Statistics* 60: 504-514

Davies, L. (1991) *Handbook of Genetic Algorithms*, Van Nostrand Reinhold: New York

Davies, R. (1984) Introduction in Davies, R., Rogers, D. (eds) *Store Location and Store Assessment Research,* John Wiley and Sons, Chichester

Denham, C (1993) Census Geography - An Overview in Dale, A., Marsh, C. (eds) *The 1991 Census Users Guide*, HMSO, London

Dooley, M. (1986) The over educated Canadian - Changes in the relationship among earnings, education and age for Canadian Men *Canadian Journal of Economics* 19 pp 142-159

Dorling, D. (1994) The Negative Equity Map of Britain *Area,* Vol. 26, No. 4 pp 327-342

Dorling, D. (1995) Visualizing changing social structure from a census *Environment and Planning A*. Vol. 27 pp 353-378

Driedger, L. (1987) *Ethnic Canada: Identities and Inequalities*, Copp Clark Patman, Ontario

Drudy, S. (1991) The Classification of Social Class in Sociological Research *British Journal of Sociology* Vol 42, No 1 pp 21 - 41

Ellis, M., Barff, R., Renard, B. (1993) Title Unknown,*Growth and Change* 24(2) pp106-90

Euro-Direct (1992) *Neighbours and Prospects* (Commercial Literature)

Evans, G (1992) Is Britain a Class Divided Society? A Re-Analysis and Exetension of Marshall et als Study of Class Consciousness in Sociology Vol 26, No2 pp233-258

Evans, N., Webber, R. (1994) Advances in Geodemographic Classification Techniques for Target Marketing, *Journal of Targeting, Measurement and Analysis for Marketing*, 2, pp. 313-321

Flowerdew, R., Goldstein, W. (1989) Geodemographics in Prcatice - Developments in North America, *Environment and Planning A,* Vol 21, No. 5 pp 605-616

Foody, G., Cox, D. (1994) Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions *International Journal of Remote Sensing* Vol. 15, 619-631

Foody, G., Curran , P. (1995) Environmental Remote Sensing From Regional to Global Scales John Wiley

Gaarder, J. (1991) *Sophie's World,* Phoenix, London

Georgescu-Roegen (1971) *The Entropy Law and The Economic Process*

Giddens, A. (1974) *Positivism and Sociology*, Heinemann, London

Giddens, A. (1980) *The Class Structure of the Advanced Societies*, Hutchinson, London

Gillespie, A. (1992) Spectral mixture analysis of multi-spectral thermal infrared images *Remote Sensing of Environment* Vol, 34, 121-132

Goldthorpe, J., Hope, K. (1974) *The Social Grading of Occupations: A New Approach and Scale,* Clarendon Press, Oxford

Goodchild, M., Gopal, S. (eds) (1989) *The Accuracy of Spatial Databases*, Taylor and Francis, London

Goodchild, M. (1995) Geographic Information Systems and Geographic Research in Pickles, J.(ed) (1995) *Ground Truth: The Social Implications of Geographic Information Systems* , Guilford, NewYork

Gopal, S., Woodcock, C (1994) Theory and Methods for Accuracy Assessment of Thematic Maps Using Fuzzy Sets *Photogrammetric Engineering and Remote Sensing* Vol. 60, 181-188

Goss, J. (1995) Marketing the New Marketing: The Strategic Discourse of Geodemographic Information Systems, in Pickles, J.(ed)*Ground Truth: The Social Implications of Geographic Information Systems* , Guilford, NewYork

Gregory, D. (1995) Between the book and the lamp: imaginative geographies of Egypt, 1849-50 *Transactions of the Institute of British Geographers*, NS 20, pp 29 - 57

Gregory , J. (1995) The Southern Diaspora and The Urban Dispossessed - Demonstrating the Census Public Use Microdata Samples *Journal Of American History,* Vol 82, 111-134

Griffin, R., Phoneix, A. (1994) The Relationship between Qualitative and Quantitative Research: Lessons from Feminist Psychology *Journal of Community and Applied Psychology*, Vol. 4, pp 287 - 298

Grigsby, J., McGowan, J. (1986) Still in the Nest *Sociology and Social Research* Vol. 70 No. 2 pp146-148

Gurney, P (1995) Lecture Notes *http://www.shef.ac.uk/psychology/gurney/notes/index.html* Department of Psychology, University of Sheffield

Halsey, A. (ed) (1988) *British Social Trends since 1900,* Macmillan, London

Hamilton, M., Hirszowicz, M. (1993) *Class and Inequality: Comparative Perspectives*. Harvester, London

Harley, J.B. (1989) Deconstructing the Map *Cartographica* Vol 26, No2 pp 1 - 20

Harley, J.B (1990) Cartography, ethics and Social Theory *Cartographica*, Vol 27 No2 pp1-23

Harvey, D. (1973) *Social Justice and the City*, Edward Arnold: London

Harvey, D. (1989) *The Urban Experience* , Blackwell, Oxford

Herbert, D.T. Johnston, R.J. (eds) (1978) *Social Areas in Cities: Processes, Patterns and Problems* John Wiley and Sons, Chicheter

Hodges, K. (1991) Editor's Note on Senate Proposals for the Census in American Demographics, August 1991

Hoffman, L. (ed) (1980) *Computers and Privacy In The Next Decade,* John Wiley and Sons, Chicester.

Holland, J. (1975) *Adaptation in natural and artificial systems*, Univeristy of Michigan Press, Ann Arbor

Huxley, A. (1959) *The Human Situation* Flamingo, London

Jarman, B. (1983) The Identification of Underprivileged Areas, *British Medical Journal* Vol. 298, pp. 1705-9

Johnson, J., Roseman, C. (1990) Increasing Black Outmigration from LA: The role of household dynamics and Kinship systems *Annals of the Association of American Geographers* 80 (2) pp 205-222

Johnston, R.J. (1984) The Region in 20th Century British Geography, *History of Geography Newsletter*, 4, 26-35

Johnston, R.J. (1985) *The Future of Geography,* Methuen, London

Johnston, R.J. (1986) *On Human Geography* , Blackwell: Oxford

Johnston, R., Pattie, C. (1993) Entropy maximising and Iterative Proportional Fitting, *Professional Geographer* 45, No. 3, 317-322

Jones, K., Moon, G. (1987) *Health, Disease and Society,* Routledge: London

Kelsall, K.,Kelsall, H.,Chisholm, L. (1984) *Stratification* [2nd edition] Longman: Aspects of Modern Sociology

Kent, J., Mardia, K. (1986) Spectral Classification Using Fuzzy Membership Models *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 10 pp. 659-671

Klir, G., Folger, T.(1988) *Fuzzy Sets, Uncertainty and Information,* Prentice Hall, London

Knox, P.(1982) *Urban Social Geography* , Longman, Harlow

Ley, D. (1987) *A Social Geography Of The City*, Harper and Row, New York

Ley, D. (1989) Fragmentation, coherence and limits to theory in human geography in Kobayashi, A., Mackenzie, S. (eds) (1989) *Remaking Human Geography Unwin* Hymen, Boston

Lichter, D. (1989) Race, Employment hardship and Inequality in the American Non-Metropolitan South, *American Sociological Review* Vol. 54, No.3 pp 436 - 446

Maguire, D., Goodchil, M., Rhind., D (eds) (1991) *Geographic Information Systems: Principles and Applications*, Longman, London

Marsh, C., Wrigley, N. (1988) Report of the ESRC working group, *Environment and Planning A*, Vol 20 pp 851-889

Marsh, C. (1991) Microdata from the 1991 Census of Population in Britain: applications in marketing research, *Journal of the Market Research Society* Vol 33, No.4 pp 275-284

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., Walford, N. (1991) The Case for Samples of Anonymised Records from the 1991 Census,*Journal of the Royal Statistical Society* , 154 Part 2, pp305-340

Marshall, G. (1982) *In Search of The Spirit of Capitalism*: Hutchinson, London

Martin , D. (1991) *Geographic Information Systems and Their Socio-economic Applications,* Routledge, London

Martin, D. (1993) *The U.K. Census of Population* CATMOG, GeoBooks, Norwich

Mitchell, R., Martin, D., Foody, G. (forthcoming) Unmixing Aggregate Data; Estimating The Social Composition of Enumeration Districts *Environment and Planning A.*

McQuillan, K. (1988) Report on the Feasibility of Producing a Pulic Use Sample Tape from the 1986 Census *Federation of Canadian Demographers* April

Morphet, C. S. (1993) The mapping of small-area census data - A consideration of the role of enumeration district boundaries *Environment and Planning A*, Vol 25, pages 1267-1277

Myers, D. (1993) Local Census Data *American Demographics* ,June, p52

Navaratil, F., Doyle, J. (1977) The Socio-Economic determinants of migration and the level of aggregation *Southern Economic Journal*, 43 pp 547-559

Nichols, T. (1979) Social Class: official, sociological and Marxist. in Irvine, J., Miles, I., Evans, J (eds) *Demystifying Social Statistics*, Pluto, London

Offe, C. (1985) *Disorganized Capital*, Polity Press: Cambridge

OPCS (1991) *Census Definitions*, OPCS, Fareham

OPCS, (1992) Census News Letter

Openshaw, S.(1984) *The Modifiable Areal Unit Problem*, CATMOG 38 Geo Books, Norwich

Openshaw, S., Sforzi, F., Wymer, C.(1985) A Multivariate Classification of Individual Household Census Data for Italy *Papers of the Regional Science Association* No.56 pp113-125

Openshaw, S.(1991) A view on the GIS crisis in geography, or using GIS to put Humpty-Dumpty back together again *Environment and Planning A*, 23 pp 621-8

Openshaw, S. (1994) Computational human geography: towards a research agenda in *Environment and Planning A* vol 26, pp 499-508

Openshaw, S (1995) Human Systems modelling as a new grand challenge area in science *Environment and Planning A* Vol 27, pp 159 - 164

Openshaw, S., Blake, M., Wymer, C. (1995) Using neurocomputing methods to classify Britains residential areas School of Geography Working Paper, University of Leeds

Pahl, R. (1989) Is The Emperor Naked? Some questions on the adequacy of sociological theory in urban and regional research. *International Journal of Urban and Regional Research* 13, (4) 711, 20

Pahl, R. (1993) Does Class Analysis Without Class Theory Have a Promising Future? A reply to Goldthorpe and Marshall *Sociology* Vol 27, No2 pp253-258

Payne, J.N.,Coy, J., Milner, P.C., Patterson, S. (1993) Are Deprivation Indicators a Proxy for Morbidity? *Journal of Public Health Medicine*, Vol. 15 , No.2, pp. 161-170

Pickles, J.(ed) (1995) *Ground Truth: The Social Implications of Geographic Information Systems* , Guilford, NewYork

Pratt, G. (1989) Quantitative techniques and humnistic - historical materialist perspectives in Kobayashi, A., Mackenzie, S. (eds) (1989) *Remaking Human Geography* Unwin Hymen, Boston

Preston, V., McLafferty, S., Hamilton, E (1993) The Impact of Family Status On Black, White, and Hispanic Womens Commuting *Urban Geography* 14, 3 pp 228-250

Richardson, C. (1990) Data Fusion Does Not Work *Journal of The Market Research Society* Vol. 32, 472-473

Ruhomon, B.(1987) Data Fusion: Its role in media research and media planning *ADMAP* June, pp 57 - 61

Rumelhart, D., Hinton, G., Williams, R. (1986) Learning internal representation by error propagation. In Rumelhart, D., McClelland, J. (eds) *Parallel distributed processing: exploration in the microstructures of cognition,* MIT Press, Cambridge MA.

Sack, R. (1980) *Conceptions of space in social thought*, Macmillan, London

Sack, R. (1982) Realism and realistic geography *Transactions of the Institute of British Geographers*, Vol 7, pp 504 - 509

Saunders, P. (1987) *Social Theory and The Urban Question*. Unwin, Hyman: London

Savage, N., Edwards, C. (1984) *A guide to The Data Protection Act,* Financial Training, London

Sawicki, D (1993) Analysis with local census data - Portraits of Change, by Myers, D. *Journal of the American Planning Association* Vol 59, No1 pp120-121

Sayer, A. (1978) Mathematical Modelling in Regional Science and Political Economy: Some Comments *Antipode* Vol. 10, pp 79 - 86

Sayer, A. (1989) On the dialogue between humanism and historical materialism in geography in Kobayashi, A., Mackenzie, S. (eds) (1989) *Remaking Human Geography* ,Unwin Hymen, Boston

Sayer, R. (1982) Misconceptions of Space in Social Thought *Transacations of the Institute of British Geographers*, Vol 7, pp 494 - 503

Schalkoff, R. (1992) *Pattern Recognition: Statistical, structural and neural approaches*, Wiley, New York

Schubert, H. (1994) The Significance of spatial distance and social-relations for old age help *Geographische Zeitschrift* Vol 82, 4, 226-238.

Settle, J., Drake, N. (1993) Linear Mixing and the estimation of ground cover proportions. *International Journal of Remote Sensing,* Vol. 14 pp. 1159-1177

Shevky, E., Bell, W. (1995) *Social Area Analysis,* Stanford University Press, Stanford, California

Simpson, S., Dorling, D, (1994) Those Missing Millions: Implications for Social Statistics of Non-response to the 1991 Census *Journal of Social Policy*, Vol. 23 No. 4, pp 543-567

Sleight, P. (1993) The New Classifications *Market Research Society Working Paper,* Market Research Society, London.

Smith, N (1992) Real wars, theory wars *Progress in Human Geography* 16 (2), 257 - 271

SNNS (1995) Users Mannual

Soja, E.W. (1980) The Socio-Spatial Dialectic *Annals of the Association of American Geographers* Vol 70. No.2 June pp. 207 - 225

Spa Marketing (1992) Commercial Literature

SPSS (1991) *SPSS Professional Statistics Guide V 6.0,* SPSS, London

Stark, D. (1980) Class struggle and the Labour Process *Theory and Society*, 9 (1)

Sui, D.Z. (1994 )GIS And Urban Studies: Positivism, Post-Positivism and Beyond *Urban Geography*, Vol 15:3, pp258-278

Taylor, P., Johnston, R. (1995) Geographic Information Systems and Geography in Pickles, J.(ed) (1995) *Ground Truth: The Social Implications of Geographic Information Systems* , Guilford, NewYork

Thompson, E.P. (1968) *The Making of The English Working Class*, Penguin, Harmondsworth, Middlessex

Tomlinson, R., Calkins, H., Marble, D. (1976) CGIS: a mature, large geographical information system in *Computer Handling of Geographical Data,* UNESCO Press, Paris

Townsend, P., Davidson, N. (1982) *Inequalities in Health; The Black Report*, Penguin, Harmondsworth

Unwin, D. (1994) Cartography, ViSC and GIS *Progress in Human Geography* Vol. 18, No.4, pp 516-522

Upton, G. (1985) Modelling cross-tabulated regional data in Nijkamp. P., Leitner, H., Wrigley, N. (eds) *Measuring The Unmeasurable,* Martinus Nijhoff Publishers, Amsterdam

Ward, P., Mortonjones, A.J., Pringle, M., Chilvers, C. (1994) Generating Social Class Data, *Primary-Care Public Health* ,Vol. 108, No.4, pp.279-287

Webber, R. (1994) *Alternative Approaches to classification* CNN Commercial Literature

Weber, M. (1970) *Essays from Max Weber* In Gerth, H., Mills, C. (eds) Routledge, London.

Williamson, P., Birkin, M., Rees, P. (forthcoming) The estimation of population microdata using data from small area statistics and samples of anonymised records *Environment and Planning A*

Williamson, P., Birkin, M., Rees, P. (1993) *The Simulation of Whole Populations Using Data from SAS, SAR and national surveys*, Paper presented at the Research on the 1991 Census Conference, University of Newcastle-Upon-Tyne, 13-15 Sept 1993

Willis, P. (1977) *Learning to labour: How working class kids get working class jobs* , Saxon House, Farnborough

Wilson, D., Marta, T. (1989) Employment Returns to Migration *Urban Geography* 10 pp540-561

Wright, E. (1979) *Class Structure and Income Determination*, Academic Press, New York

Wrigley, N. (1988) *Store Location Analysis* , Routledge, London

Wrigley, N (1993) Revisiting the Modifiable Areal Unit Problem and Ecological Fallacy. Chapter For Festschrift for Peter Hagget.

Xu, J., Lathrop, R. (1995) Improving simulation accuracy of spread phenomena in a raster-based Geographic Information System *International Journal of Geographical Information Systems* Vol 9. No2. pp153-168