Case Study Examples to Demonstrate the use of Samples of Anonymised Records in Marketing Analysis

BARRY LEVENTHAL
BERRY CONSULTING

Occasional Paper 5

UNIVERSITY OF MANCHESTER
CENSUS MICRODATA UNIT

1994

Barry Leventhal
Berry Consulting
2 Charterhouse Mews
London, EC1M 6BB

ISBN 1 899005 06 4

Printed in Great Britain by:

Census Microdata Unit
Faculty of Economic & Social Studies
The University of Manchester
Oxford Road, Manchester M13 9PL.

Contents

Appendices

1. Introduction

In the Autumn of 1993 the Census Microdata Unit at Manchester University received two samples of Anonymised Records (SARs) from the 1991 Census. This was the first ever release of individual-level data from a population census in Great Britain, and was the outcome of lengthy discussions between the ESRC and the Census Offices.

The Census Microdata Unit (CMU) is responsible for disseminating information from the SARs to the community of census users, including commercial organisations under a special license agreement.

In order to demonstrate some of the potential applications in marketing, the CMU commissioned Berry Consulting (BCo) to produce two case study analyses using SAR data. This report describes the two case studies and presents the main results.

Both case studies were produced on a PC using standard statistical software.

2. Potential Applications of Census Microdata in Marketing

Marketing analysts and planners of all types need to know about the demographics of the population. Census microdata give us the most flexible possible source of demographic information and so have numerous potential applications.

First we shall consider the range of possibilities, before demonstrating these with two case studies.

Electronic Census Reports

Census microdata can be employed to provide an 'electronic alternative' to the printed census reports, and equally to obtain new reports that have never actually been published.

Conventional census results are provided in a long series of volumes which include topic reports, mainly at national or regional level, county monitors and county reports.

Despite this wealth of output, the potential analyst is liable to find these reports unwieldy to use. They may be inconvenient to access, may not provide exactly the required demographic relationships or level of detail, and results may need to be amalgamated across geographical areas.

The 'Electronic Census Report' (ECR) can be created simply by loading SAR data into a user-friendly software package - such as one of the many survey analysis systems or the 'USAR' program developed specifically for handling the SARs.

Armed with a desk-top ECR, the analyst could define and produce demographic tables either nationally or for required areas (as aggregations of districts or regions). The reports could then be passed onto other software for charting and presentation.

Market Planning

If the target market for a product can be defined demographically, census microdata may be employed to obtain the market size and understand its characteristics and regional dispersion.

Although conventional census output may be used in a similar way, it may not fit the required target market definition exactly, meaning that approximations and assumptions then have to be made. Since standard census reports are mainly two-way cross-tabulations, they will probably contain insufficient detail to look at the demographic characteristics of the target group.

This major application is demonstrated below, in Case Study 1.

Retail Site Planning

Knowledge of the population within a district is essential for planning locations of retail developments. Census microdata permits in-depth analysis of population structure and dynamics, and so could form a benchmark for site planning.

We must remember that the SARs do not identify small geographical areas, and so the Census Local Base Statistics or Small Area Statistics will probably still be required for catchment area analysis.

Market Share Evaluation

Organisations which hold information about their customers may use census microdata in order to benchmark their market against the population, and so obtain market share within each demographic group.

Customer Segment Evaluation

Many organisations are adopting customer segmentation as the basis for marketing to their customers. A set of customer segments could be 'mapped onto' the SARs, so as to understand the overall size and distribution of these segments in the population. From this, market shares within segments could then be obtained.

For example, Lifestage is often a key segmentor in financial services markets. The SAR user can employ the Lifestage groups used in the Census, or define a new set of segments to match those being applied to customers.

Demographic Modelling

Because the SARs consist of individual records for persons and households, they may be used as a testbed to develop models and segmentation systems. For example, a classification of households could be constructed, which should be a more accurate indicator of household behaviour than the area discriminators such as ACORN and MOSAIC. Area discriminators all suffer from the 'ecological fallacy' which means that they may not accurately describe the individuals living in each area. In principle, the household classification could be applied to the main census database held by OPCS/GRO(S) in order to obtain the breakdown of household types at small area level.

Similarly, models could be built to predict 'hard to collect' variables from other more readily available attributes. Case Study 2 demonstrates this type of analysis.

3. Case Study 1 - Quantifying a Target Market

Introduction

When marketing any product an understanding of the size and characteristics of the target market is obviously vital.

Often, the target market is defined in demographic terms and so census information should be the most exact data source for measuring market size and penetration. However, published census results, such as the topic reports and local statistics may not give the required count. For example, they generally provide population counts cross analysed by two demographics at a time and age is often banded into 5 or 10 year ranges. Therefore, in employing such sources, approximation and guesswork are often required.

Using census microdata, target markets can be defined exactly as the required combinations of census demographics.

This is demonstrated in Case Study 1 for a hypothetical target market:

"Heads of household aged 55+ who are income earners and own their homes outright."

This target market would be of obvious importance to companies marketing certain products and services, for example luxury goods and holidays.

In the Census, the Head of Household is taken as the first person on the schedule who is 16 or over and not a visitor. This procedure differs from the approach adopted in Market Research and we have not attempted to adjust for it in this case study. However, census microdata could be employed to identify an alternative Head in each household, and quantify the potential error in the census definition.

In Case Study 1, the target market is quantified and profiled by other attributes in order to help understand its characteristics.

Method

Using the Individuals SAR file, all records with age less than 16 were excluded.

The target market was defined as:

- Heads of household
- Aged 55+
- Earners (employed or self-employed)
- Owning home outright

The variables used from the SAR dataset were:

Relat (Relationship to household head)	=	1 (household head)
Age (Age)	ge	55
Ecposfhp (Economic position of family head)	=	1 (in employment)
Tenure (Tenure of household space)	=	1 (owner occ - outright)

The target records were flagged and cross-tabulated against the rest of the sample by a number of variables.

The cross-tabs gave the penetration of the target market within other attributes; these values were then indexed against the national average.

Results

A total of 13,088 individual records were identified as belonging to the target market, out of 894,115 persons aged 16+ present in the Individuals SAR. Therefore the penetration of the target market was 1.46%, amongst adults aged 16+.

The identified records were then profiled by regions and key demographics indicating that the target market is:

- most concentrated in East Midlands and East Anglia, and lowest in Inner London
- mainly Married or Remarried
- weak among non-white ethnic groups
- skewed towards professional self employed occupations and farmers
- likely to possess amenities such as central heating and cars.

4. Case Study 2 - Demographic Modelling

Introduction

The purpose of the second case study is to demonstrate the ability to use census microdata for individual-level analysis, such as demographic modelling. This type of analysis cannot be undertaken with conventional census output which aggregates together individual results.

The objective of Case Study 2 is to model the propensity for households to belong to a particular target group in terms of other attributes or "predictors". Having developed such a model, it could then be employed to impute propensities on a separate dataset, for example a customer file or a research survey.

The nominated target for this exercise was membership of Social Class I, a group of obvious marketing importance. The predictor attributes were a set of simpler census variables coded at "100% level" in the 1991 Census.

A simple tree segmentation model was developed using the CHAID package for Chi-square automatic interaction detection. This method was chosen due to its clear visual presentation of the analysis results. However a number of other techniques could have alternatively been employed, including logistic regression and log-linear analysis.

Method

The Household SAR file contains data both at household level and at person level.

For the household level file the following variables were included:

cars	(Number of cars)
tenure	(Tenure type)
roomsnum	(Number of rooms)
cenheat	(Availability of central heating)
hhsptype	(Household space type)

For the person level file, all records where relat (Relationship to household head) = 1 (household head) were included and the following variables were retained:

age	(Age)
soclass	(Social class based on occupation)

These two files were then merged and each record was flagged as either being Social Class I or not (ie. head of household in Social Class I).

A similar result could be achieved using the individual-level SAR file, by selecting RELAT=0 to obtain heads of household with housing information attached.

A 1 in 20 sample was drawn for those not in the target group and combined with all households in the target group giving a sample of 20,626 records.

A tree model using CHAID was then developed on one half of the sample, validated on the other half and then re-run on the whole sample.

Results

The target group comprised 10,357 households in Social Class I. Taking a 1 in 20 sample of other households gave 10,269 non-target households for comparison. Therefore the target households accounted for 50.2% of the analysis sample, but their true penetration across all households was 4.8%.

The tree segmentation analysis identified that the most important predictors of Social Class I (amongst 100% coded variables) are Number of cars, Tenure, Household space type and Age of household head.

Two versions of the tree segmentation are presented in Appendix A:

a) the unweighted analysis, as obtained using CHAID

b) the corresponding results after weighting non-target households by a factor of 20, in order to produce realistic estimates for the penetration of Social Class I among all households

Therefore, from the weighted results we see that the tree analysis identified a series of sub-groups with Social Class I penetration ranging from 15.1% down to 0.1%.

5. Main Conclusions

By producing these case studies we may conclude that:

a) The SARs are a rich and flexible source of data for analysing target markets or customer segments.

b) SAR data enables the 1991 Census to be analysed in ways that could not be achieved from the conventional output products. Similar techniques may be employed to those commonly adopted elsewhere in market research and marketing.

c) Users will be able to manipulate SAR data with their own computers and software, therefore census analysis should be easier and more affordable than ever before.

APPENDICES

Appendix A : Presentation of Case Study Results

Case Study 1

Target Market

- Head of household
- Aged 55+
- Earners (employed or self employed)
- Owning home outright

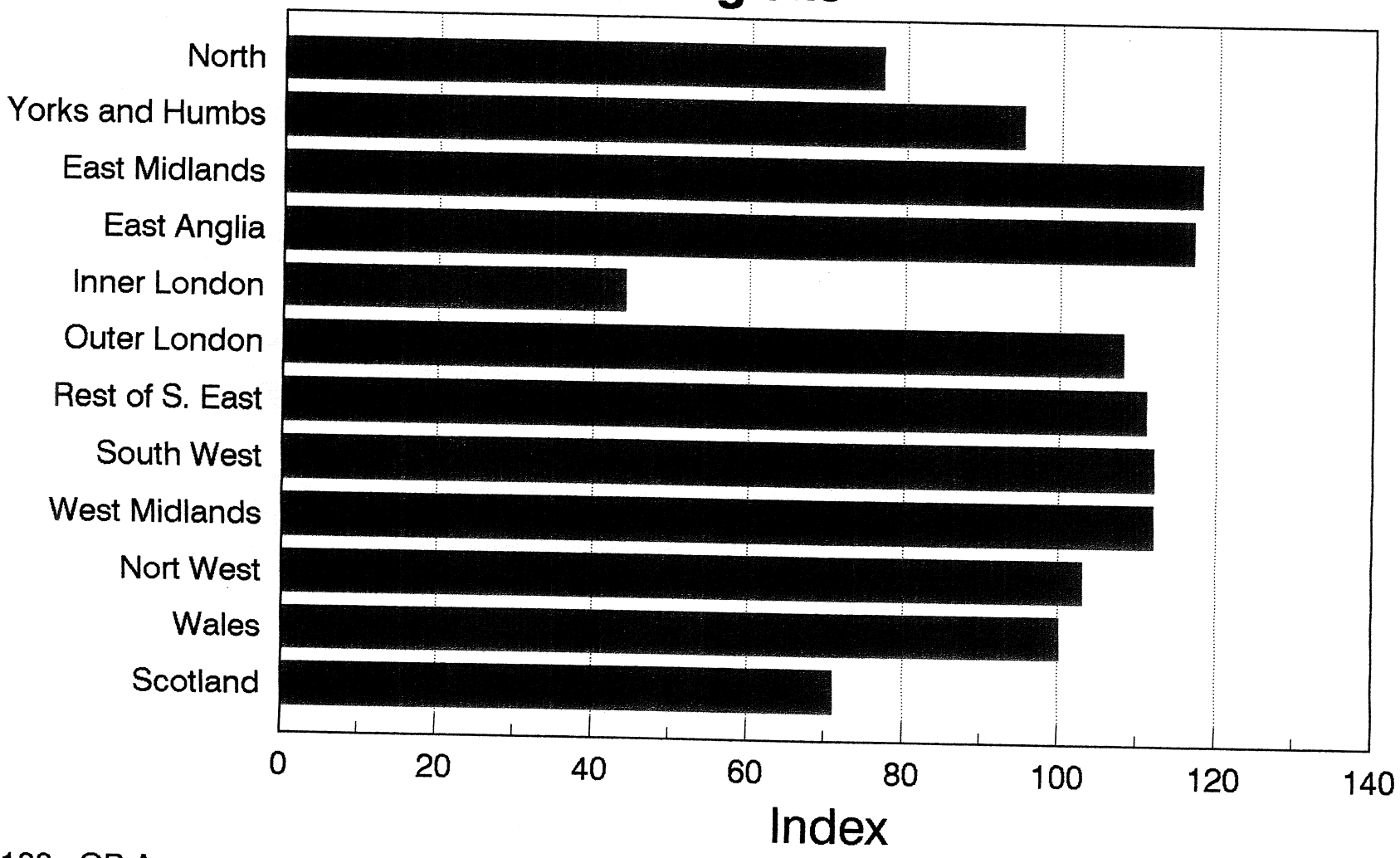
Base

- Individuals SAR aged 16+

Sample Sizes

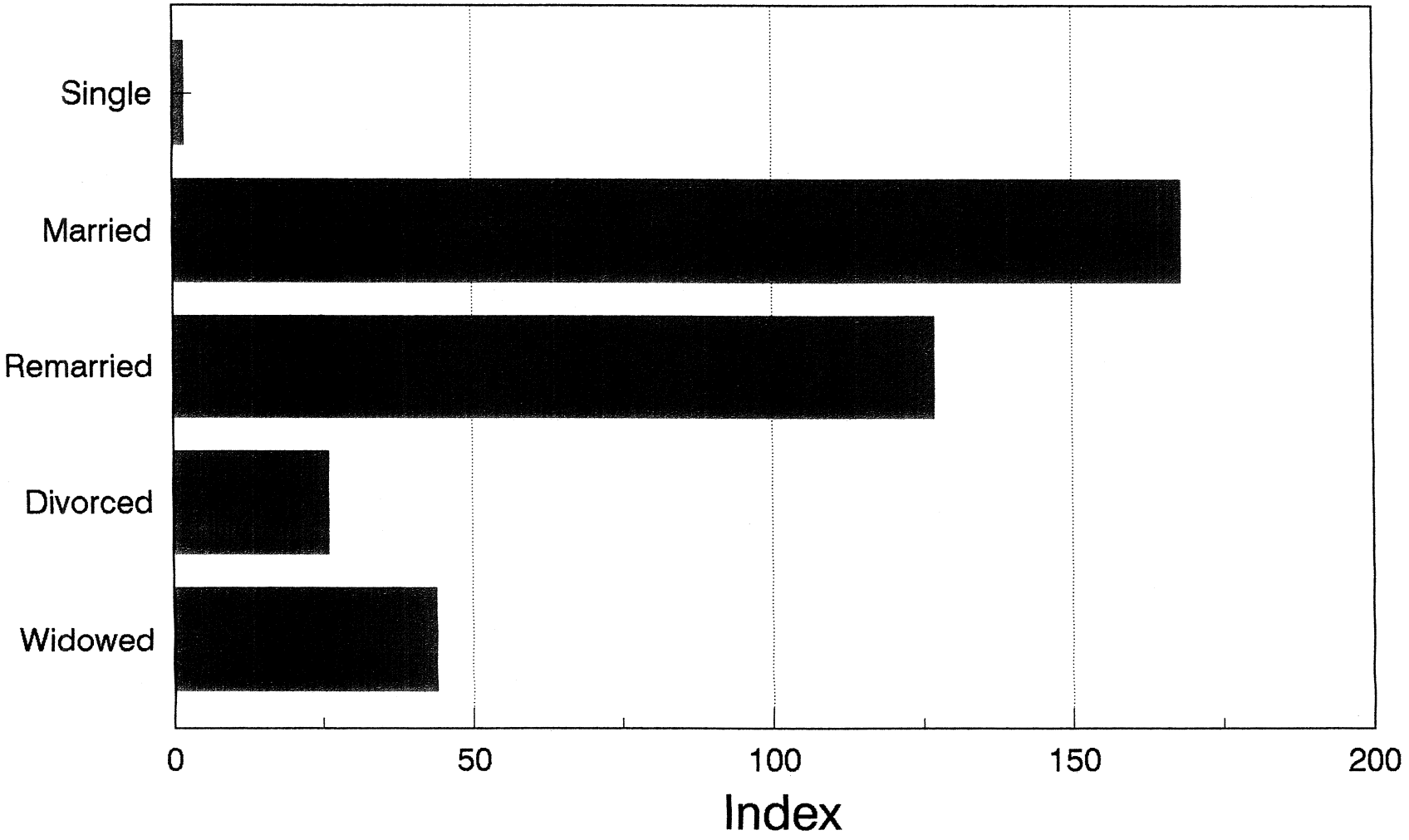
Target Market:	13,088
Base:	894,115
GB Average Penetration:	1.46%

Penetration of Target Market
SAR Regions



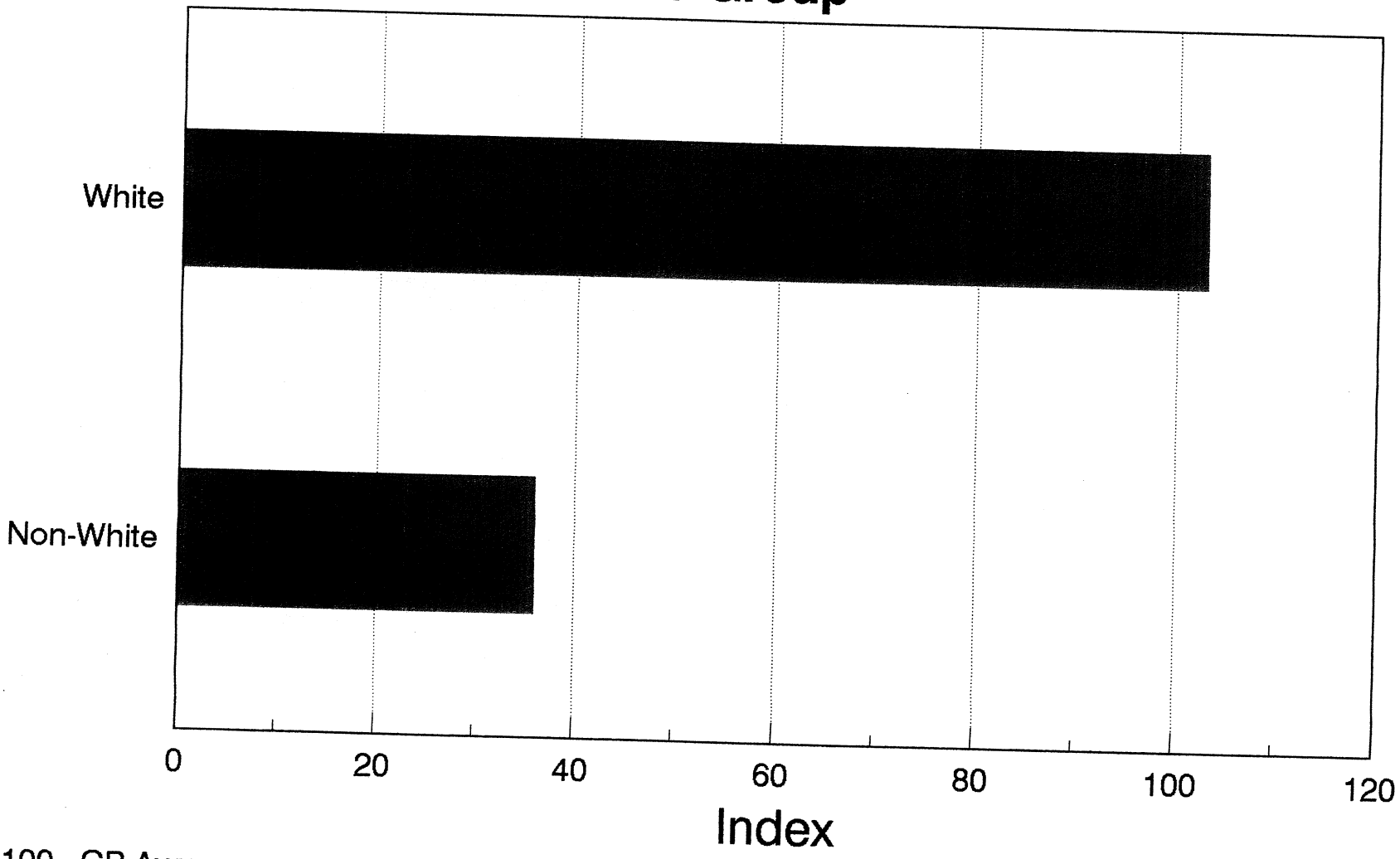
100=GB Average

Penetration of Target Market
Marital Status



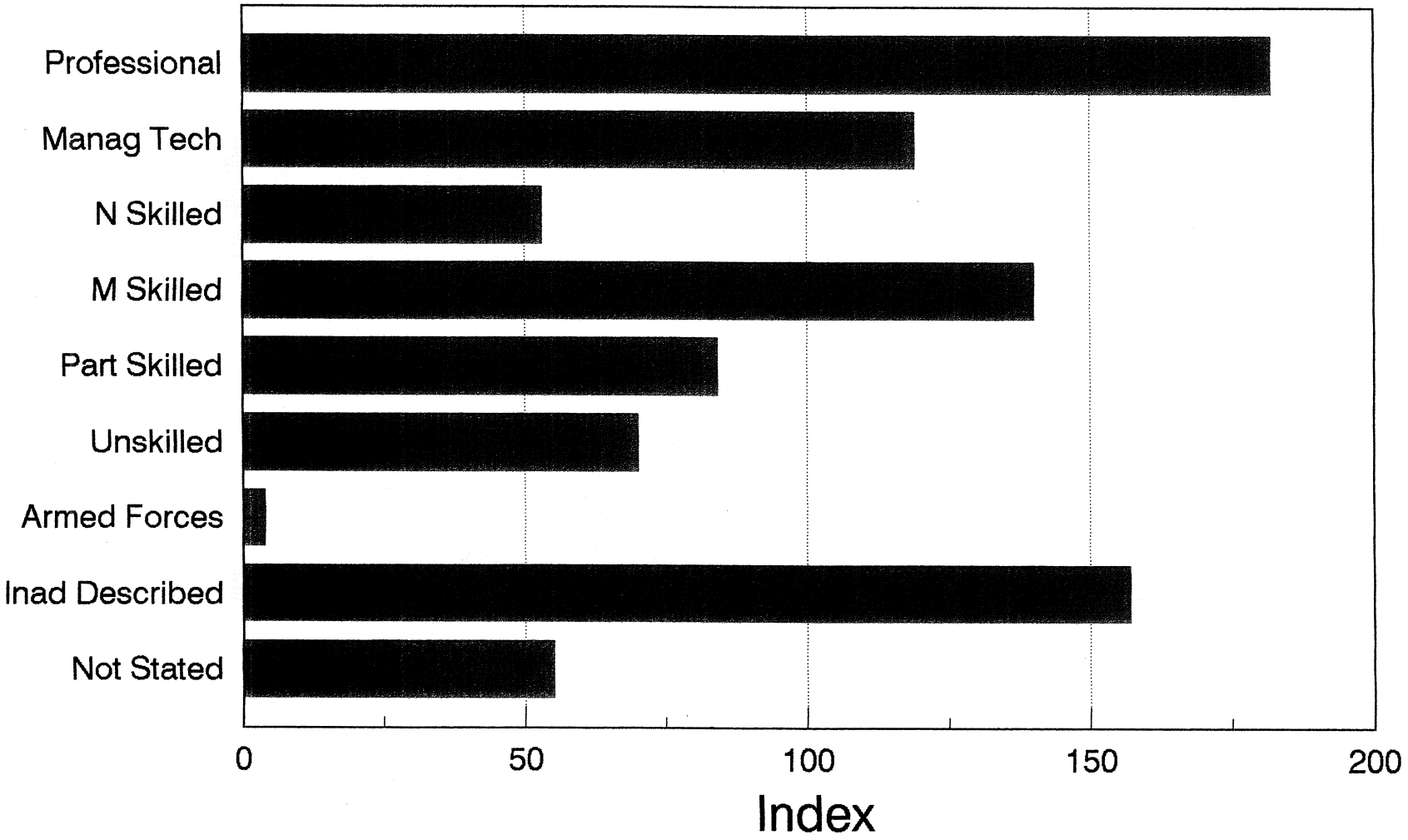
100=GB Average

Penetration of Target Market
Ethnic Group



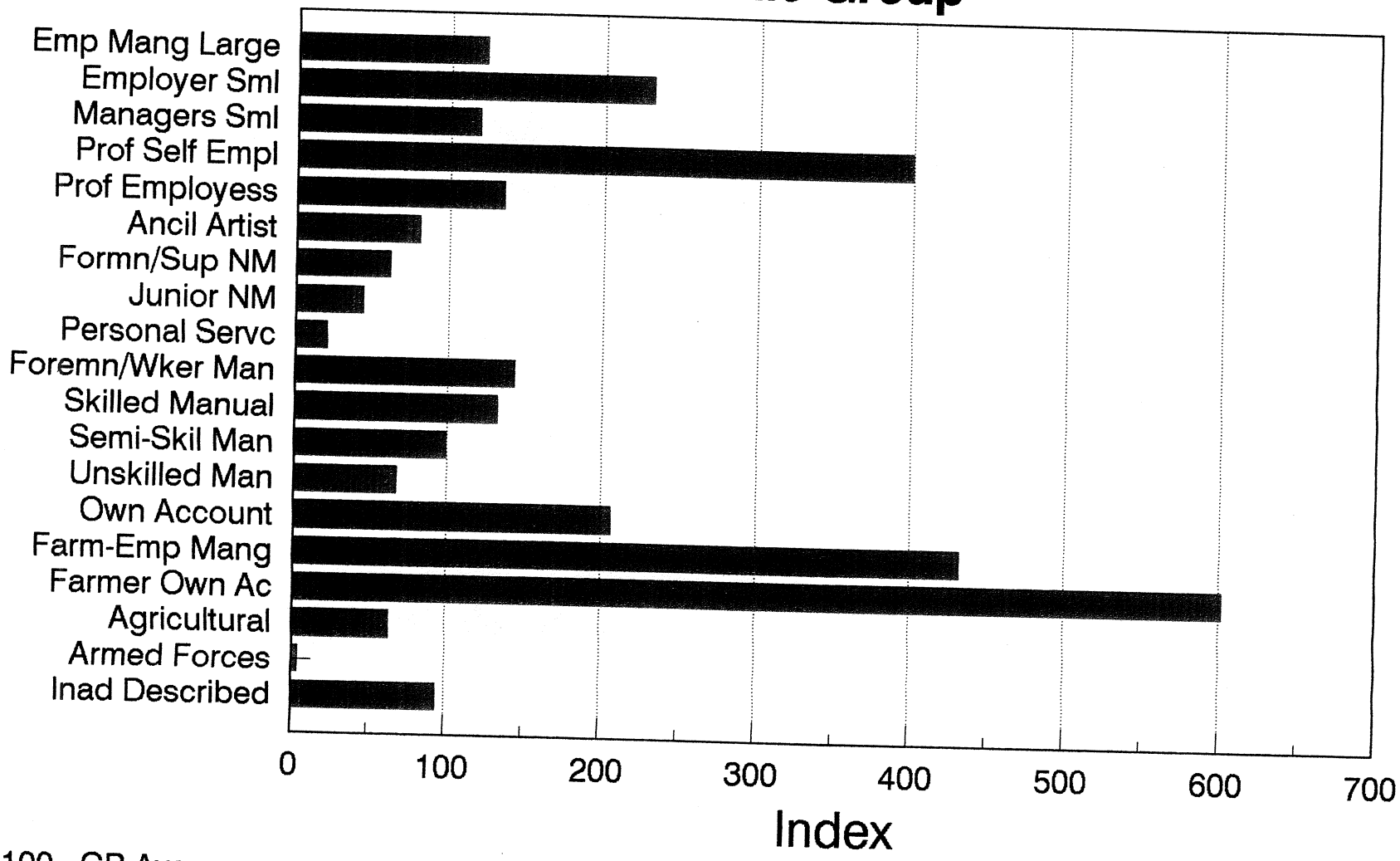
100=GB Average

Penetration of Target Market
Social Class (Based on Occupation)



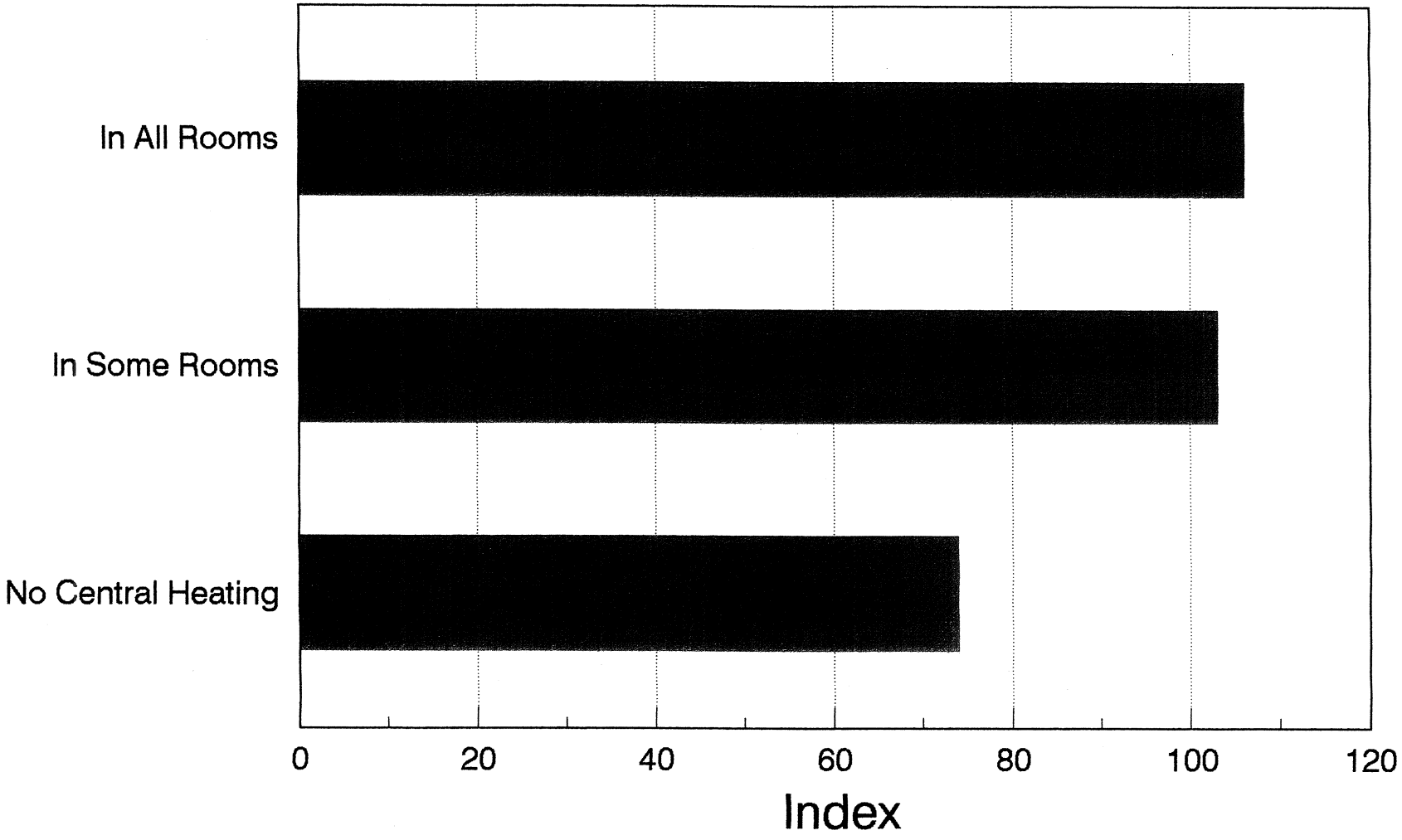
100=GB Average

Penetration of Target Market
Socio-Economic Group

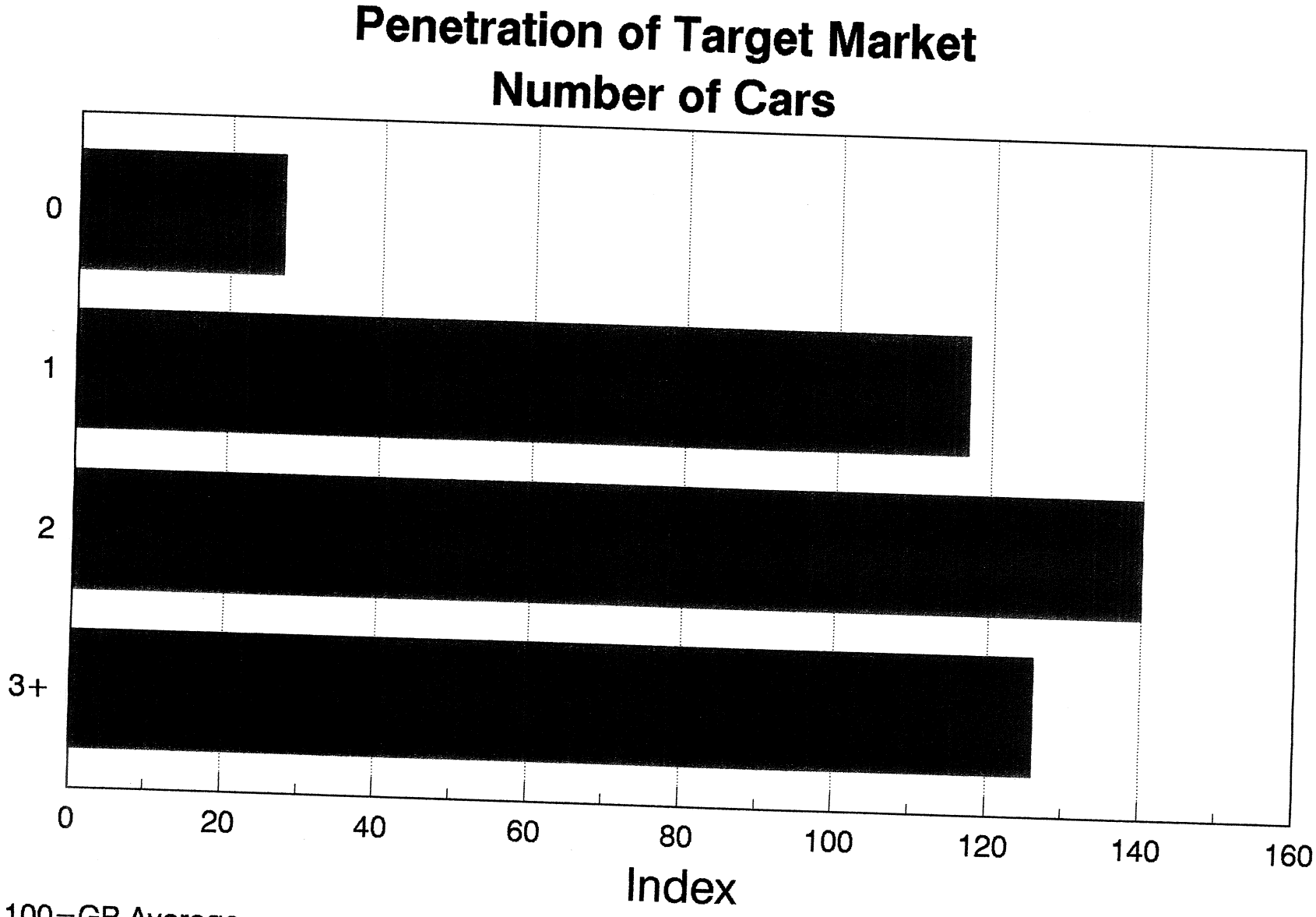


100=GB Average

Penetration of Target Market
Availability of Central Heating



100=GB Average


Penetration of Target Market
Number of Cars
0
1
2
3+
0
20
40
60
80
100
120
140
160
Index
100=GB Average

Case Study 2

Target Group

- Heads of households in Social Class I

Base

- Household SAR

Analysis Technique

- Tree Segmentation (CHAID)

Candidate Predictors

- Number of cars
- Tenure type
- Number of rooms
- Availability of central heating
- Household space type
- Age of household head

Sample Sizes

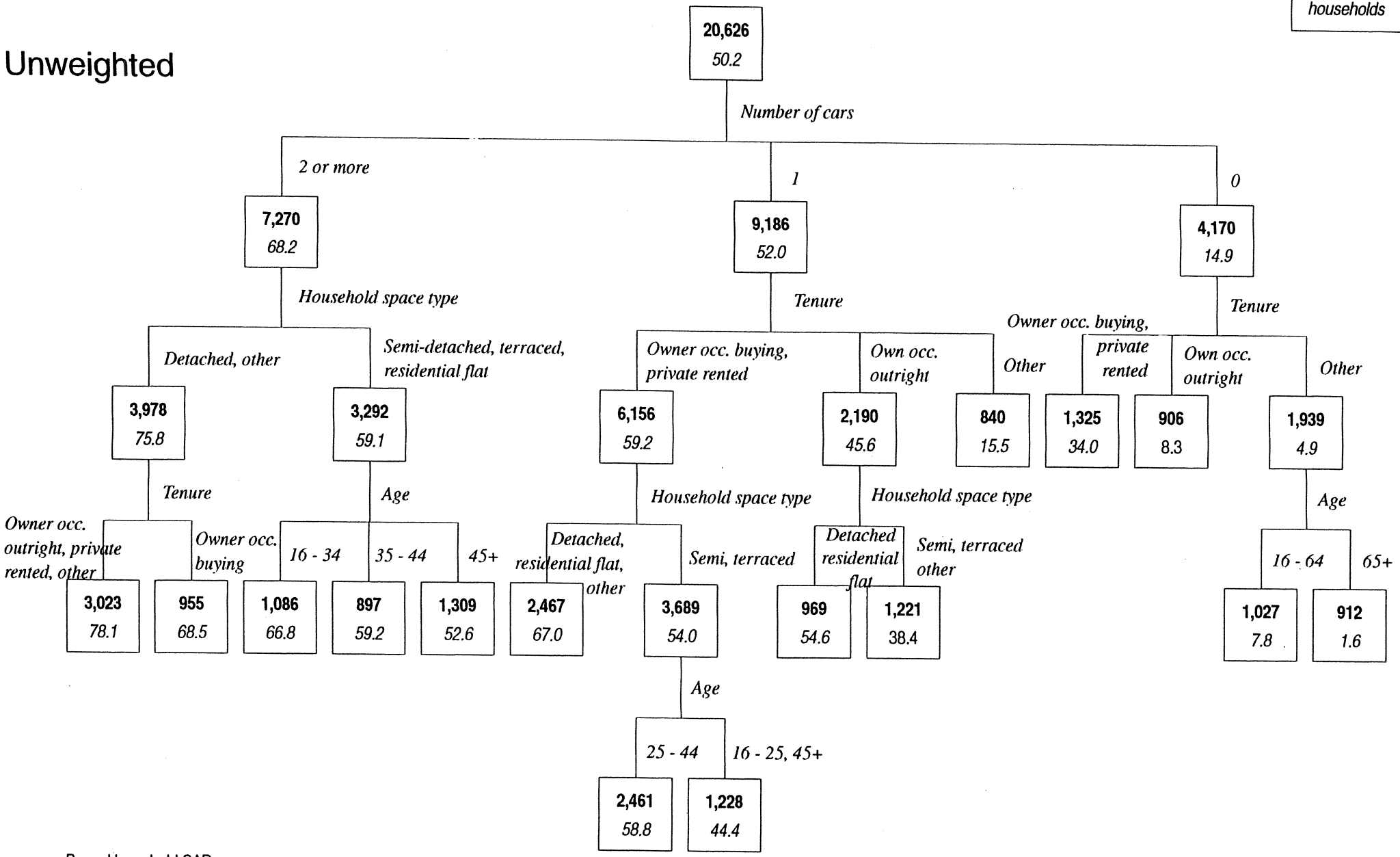
Target Group:	10,357
1 in 20 sample of other households:	10,269
Penetration of Target Group within sample:	50.21%

TREE SEGMENTATION ANALYSIS

Target Group: Households in Social Class I

Unweighted

Key: **Sample Size** / *% of Class I households*



Base: Household SAR

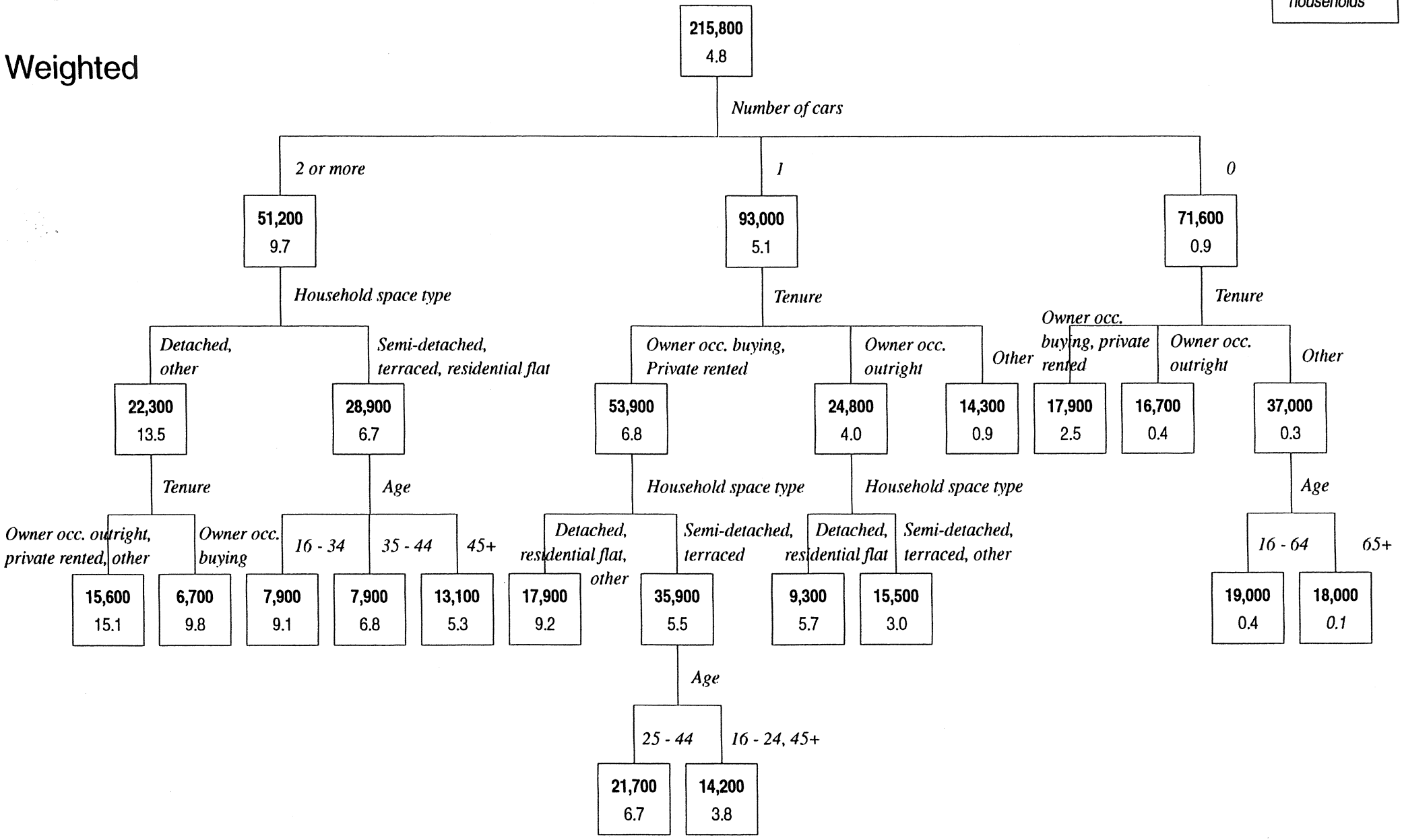
Analysis based on all Social Class I Households and 1 in 20 sample of non-Social Class I

TREE SEGMENTATION ANALYSIS

Target Group: Households in Social Class I

Key: **Sample Size** / *% of Class I households*

Weighted



Base: Household SAR

Analysis grossed up to represent all SAR households

Appendix B : Tree Segmentation

Two main packages are available for producing tree segmentation analysis.

Automatic Interaction Detector (AID)

AID is a powerful, statistically appropriate tool useful in identifying population groups (segments) that differ in their probability of an outcome such as response to mailing or purchase of a product.

AID is an explanatory tool. It gives some clear insights into the underlying factors associated with differences in value of the outcome variable.

AID identifies interactions between the predictor variables ie. situations where the outcome depends on a particular combination of predictors.

AID works by checking each predictor variable in turn and selecting both the predictor and cut-off point which segment the file most effectively for the dependent variable. This splits the file into two subgroups. Each of these subgroups is then tested, in turn, in order to find the predictor which best segments it further, giving four subgroups. The process continues until there are no further groups that can be split meaningfully.

AID provides a visual display of results in tree diagram form, where each mode represents a subgroup of the file and is accompanied by the sample size and average score on the dependent variable.

Chi-Squared Automatic Interaction Detection (CHAID)

The techniques described above were designed for situations where the dependent and predictor values are quantitative. They may be extended to handle 0/1 variables, such as response or ownership, but are less suitable for categorial data.

CHAID is a variation upon the AID approach, which requires all variables to be categorical. CHAID also segments the sample in a stepwise manner, producing a tree analysis as the result. However, CHAID selects subgroups using a different criterion (a chi-squared measure as described above) and can generate multi-way splits, rather than only two-way.

Other titles available from the Census Microdata Unit:

Census Microdata Unit Occasional Papers

No. 1 *Problems of Imputation in the 1991 Census*, (1 899005 02 1), Amarjit Sandhu

No. 2 *Bias, Sampling Error and Coverage: the preliminary validation of the Samples of Anonymised Records from the 1991 Census*, (1 899005 03 X), Steve Simpson, Ed Fieldhouse & Amarjit Sandhu

No.3 *An Introductory Guide to Analysing the SARs*, (1 899005 04 8), Liz Middleton

No.4 *Resource Allocation using measures of relative social needs in geographical areas: the relevance of the signed chi-squared, the percentage, and the raw count*, (1 899005 05 6), Steve Simpson

SARs User Guide (1 899005 01 3)

Manchester Census Group

The Ethnic Dimensions of the 1991 Census: A Preliminary Report, (1 89900500 00 5), Roger Ballard & Virander Singh Kalra

£3.00

ISBN 1 899005 06 4

Further copies of this paper may be obtained from:

Census Microdata Unit
Faculty of Economic & Social Studies
The University of Manchester
Oxford Road, Manchester M13 9PL