UNIVERSITY OF LEEDS

Understanding and Validating Acxiom's Research Opinion Poll Data

Working Paper 10/06

Chris Thompson[1], John Stillwell[1], Martin Clarke[1] and Carl Bradbrook[2]

Version 1.0

September 2010

This Working Paper is an online publication and may be revised

[1]School of Geography
University of Leeds
Leeds
LS2 9JT
United Kingdom

[2]Acxiom Limited
Park House
Teddington
TW11 9AD
United Kingdom

Abstract

As a global leader in interactive multi-channel marketing services, Acxiom aims to collect data from a range of sources to aid organisations in customer profiling, retention and acquisition. One of the company's primary activities is the delivery of the annual Research Opinion Poll (ROP). The ROP is an optional household survey designed to capture detailed information about household consumption and expenditure across Great Britain (GB). Primarily used in the private sector, Acxiom's ROP has recently seen considerable growth of usage amongst organisations within the public sector. Consequently, this paper provides an independent review of the ROP as a means of validating the data for use in academia, primarily in the social sciences. The paper reveals that the ROP survey is delivered through a number of channels which enable Acxiom to generate over one million household responses a year. Subsequently, the ROP microdata are used in the construction of many of Acxiom's aggregate products such as the Acxiom Population Estimates, Acxiom's Aggregate Data and its behavioural geo-demographic classification system named PersonicX. Moreover, due to a number of estimation and modelling techniques which remain partially confidential, all three products are considered representative of the GB population at a range of geographic levels. The paper also offers a comparison to governement datasets including the 2001 Census, the Expenditure and Food Survey, the Labour Force Survey, the British Household Panel Survey, the General Household Survey and the Survey for English Housing. The ROP was found to compare favourably in areas such as sample size, geographic detail, consistency and the quality and accuracy of the data.

KEYWORDS: Acxiom; microdata; consumption; expenditure, review.

Acknowledgements

Glossary of Acronyms

ACORN A Classification Of Residential Neighbourhoods

AD Aggregated Data

APE Acxiom Population Estimates

BHPS British Household Panel Survey

EFS Expenditure and Food Survey

ESDS Economic and Social Data Service

FES Family Expenditure Survey

GB Great Britain

GHS General Household Survey

GOR Government Office Region

HRP Household Reference Person

ILU Information Lifestyle Universe

LAD Local Authority District

LFS Labour Force Survey

LSOA Lower Super Output Area

MAUP Modifiable Areal Unit Problem

NFS National Food Survey

NUTS 2 Nomenclature of Units for Territorial Statistics 2

OAC Output Area Classification

OMR Optical Mark Recognition

ONS Office for National Statistics

PAF Postcode Address File

PASW Predictive Analytics SoftWare

ROP Research Opinion Poll

SAR Sample of Anonymised Records

SEH Survey of English Households

SOA Super Output Area

List of Figures

List of Tables

Contents

1 Introduction

Academics tend to be sceptical about commercial survey datasets collected and processed by private sector organisations. They doubt the provenance of such data, worry about sampling bias and data quality issues, and prefer the comfort of using data from well-established surveys or censuses designed to capture details of every household. Yet there are ever growing volumes of unofficial data being captured through a number of different channels by different organisations which what with shrinking public sector funds, may become increasingly useful for social science research in the course of time. One particular company that openly advertises its data resources is Acxiom. Originally named Demographics, Acxiom was founded in 1969 by Charles D. Ward in Conway, Arkansas and is now recognised as one of the global leaders in interactive multi-channel marketing services. Boasting a worldwide annual turnover of over $1 billion, the company has over 35 years of experience using information to improve marketing and business results in a host of different countries across the globe (Acxiom, 2009). The mission of the company is to transform data collected from different sources (such as questionnaires or official registers) into actionable information which helps its clients understand their customer preferences, improve customer acquisition and retention, predict consumer behaviour and locate optimum retail sites (Blaszczyński *et al.,* 2006).

Despite being a global leader in interactive marketing services, Acxiom has been termed by John Meyer, the current company chief executive, as *"...the biggest company you have never heard of"* (Telegraph, 2009). This is not surprising considering Acxiom's unique selling point is built on the fact that it collects large volumes of sensitive consumer data on an annual basis. When the data collected through an array of sources are pooled together, the company's central database in Normanton houses information on over 60% of Great Britain's (GB) inhabitants including their age, income, address, spending habits and various lifestyle choices. Widespread proclamation of the existence of their sensitive data could be potentially damaging to the business as people would surely raise concerns over issues relating to 'big brother' and data protection. Nevertheless, for Acxiom's clients and partners who do know about the database and use it readily, the data represent a reliable source of information which no other company or organisation can provide. The high level of response to their annual data collection exercise, and the quick turnaround of the raw data into outputs, means that Acxiom can provide almost a mini-census of the national population each year.

This paper contains a detailed and critical appraisal of Acxiom's main source of information, its annual Research Opinion Poll (ROP) survey. The survey provides the microdata that are the foundation for most of Acxiom's aggregate data packages and, in essence, is the company's 'holy grail'. As there is no existing documentation which discusses the appropriateness of the ROP for use in social science research, the paper represents a completely original evaluation of the data (Sorenson *et al.,* 1996). To ensure an inclusive discussion, a number of the important questions and criteria set by Stewart (1984) for analysing secondary data sources will be utilised. Stewart argues that before using secondary data, the researcher must consider the purpose of collection, the methodology, the date of collection, the content and accuracy of the data, and its credibility. To ensure all these points are covered, every stage involved with the ROP survey will be analysed from start to finish. To begin, observations will be made on the questionnaire design, covering areas such as the placement and type of questions included each year. The way the survey has evolved between 2004 and 2009 will also be examined, thereby helping to understand the extent of consistency within the raw data and the impact that changes might have for time-series analysis. Thereafter, all steps in the data collection process will be outlined, paying specific attention to the distribution and compilation of each survey. This will cover issues about the response rate, the representativeness of the demographic profile and the extent of geographic penetration. The paper will then explain the manipulation and aggregation techniques used on the data collected to produce the various packages which Acxiom sells to clients. Finally, a section will be dedicated to the consistency, reliability and accuracy of the actual results produced from the ROP. Comparisons will be made on a number of core variables against alternative surveys which also offer information about household consumption and expenditure.

2. The Research Opinion Poll (ROP)

2.1 Survey Design

To ensure that Acxiom has the most accurate and up-to-date information about the GB population, the company collects primary data through a number of different channels. These include the paper-based ROP, the online ROP, other online-based questionnaires, product guarantee forms and holiday evaluation questionnaires. However, this paper will concentrate solely on the ROP survey which Acxiom designs, distributes and processes annually. The primary aim of the ROP is to gather detailed and up-to-date information on consumer

spending habits, preferences, socio-demographic information and the respondents' geographic locations. The combination of these different pieces of information allows for detailed insights into the spending patterns of different 'types' of people and geographic areas. Moreover, with the survey being distributed annually and including questions not asked on other public sector surveys, it also provides a unique source of time-series data on the demographic and socioeconomic characteristics of people and households across GB. The fact that the data are licensed by clients from both the public and private sectors demonstrates the value of the information collected by the ROP. Nevertheless, with Acxiom being a world leader in interactive marketing, the conventional use of the data is usually to help businesses in the private sector better understand and retain their existing customers, and locate new ones.

It is not surprising that Acxiom chooses to use a paper-based survey because many argue that when dealing with such a large sample, the questionnaire is an indispensable tool when primary data are required about people, their behaviour, attitudes and opinions (Hay, 2005). The acclaimed accuracy and robustness of Acxiom's products for very small geographic areas are in part due to the quality and volume of data that the ROP survey delivers. Prior to sending out the survey, a large amount of effort goes into the design and structure of the questionnaire each year. As with any survey, it is crucial to develop a well formulated blend of different questions to elicit relevant information from participants. The Data Acquisition team within Acxiom have a remit to check the design and layout of all surveys, allowing Acxiom to test the responsiveness of particular factors on an annual basis, guaranteeing the various components of the survey perform in an optimum manner. The various factors which are tested include:

- the months in which people are most responsive;
- the type of people that are most responsive;
- individual question placement and wording to maximise response;
- the questions most suitable to place upfront (to encourage survey completion);
- Data Protection Act (1998) issues such as sensitive questions, questions that cannot be asked and additional Data Protection Act wording (e.g. ethnicity);
- return address (a regional postal return address is more responsive);
- prize draw offers and survey incentives; and
- survey size, style, font and paper used.

This type of work is crucial to the whole process as the final questionnaire must be one which will maximise the response rate and generate the most accurate results. As a result, the survey

predominantly contains closed questions because these are easy for respondents to answer and for Acxiom to code and standardise. Moreover, data gathered from closed-ended questions lend themselves easily to statistical analysis (Fink, 1995). Open-ended questions are avoided because their responses are more difficult to code and interpret. The wording of questions and potential answers are also kept relatively formal. This is because formal responses are believed to trigger a respondent to focus on the task of formulating precise answers (Morse, 1994; Ongena and Dijkstra, 2009). When survey questions resemble expressions commonly used in ordinary conversations, respondents will not be focused on the task of giving precisely formatted answers, yielding a high number of misleading responses. In addition, as many of the questions ask for quite sensitive information, Acxiom has traditionally adopted a funneling technique which *"...follows a gradual movement towards personal matters"* (Dunn, 2005, p.85). This ensures that personal information such as health, age, income and marital status are left to the end of the questionnaire. The positioning and wording of certain questions can also provide a form of quality assurance by helping to identify errors and false entries created by random ticking. For example, if a respondent ticks the 'no internet connection' box, checks would be made to indentify whether or not they have ticked any questions relating to their online shopping habits from home.

In a bid to entice more responses, financial incentives are also used; participants are offered the chance to receive both financial rewards and free information on completion of the questionnaire. Figure 1 displays various incentives which were offered on completion of the ROP survey distributed in January 2009. Every ROP delivered also includes a small pen to encourage the respondent to answer the survey straight away. Furthermore, Acxiom makes every attempt to ensure that the questionnaire 'caters for' each 'local area' within which it is distributed. For example, the first page of the survey has various statistics from the previous year drawn from the answers given by residents in the same locality. This may encourage participants to respond as they can see that other people's views on their neighbourhood are being taken seriously and put to use. Moreover, if they disagree with the previous consensus, they may be more inclined to answer the questionnaire themselves and provide a different opinion. Other techniques to make the survey appear more personalised include having a regional address for return of the questionnaire and placing familiar default place names in parts of the questions. Respondents are also reassured that they do not have to provide answers in the more sensitive sections of the questionnaire, such as family health, unless they want to do so.
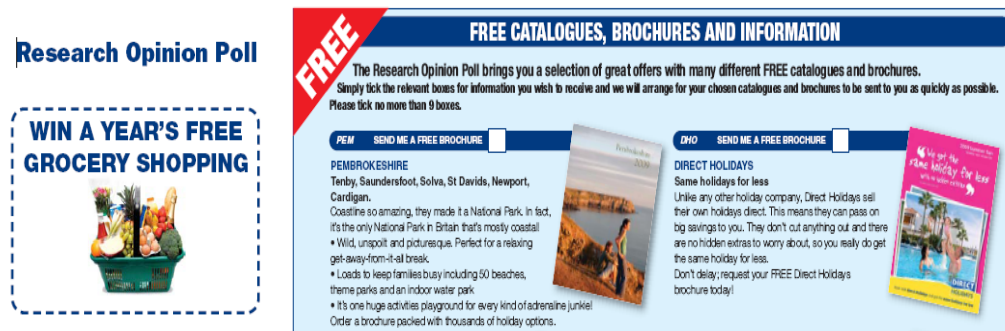
Figure 1. Incentives given on the Research Opinion Poll, January 2009

2.2 Questionnaire Structure

As previously mentioned, the ROP provides one of the main sources of information for the company and is the basis of their unique selling point. The survey has evolved dramatically since the early 1990s as its popularity has increased. Since 2004, the survey has been made much more consistent to allow for more accurate time-series analysis and consequently this review considers only 2004 and more recent ROPs. However, there have still been some significant alterations within this time period. Table 1 demonstrates the number of questions and sections in each survey between 2004 and 2009. The sections are listed in the order in which they appeared on the survey for each year.

Table 1. ROP questionnaire structure, 2004-2009

Year	Questions	Sections	Section Contents
2004	147	8	*Hobbies & Activities; Shopping; Personal Care; About Your Home; Computer/Internet; Smoking; Motoring; You and Your Family.*
2005	163	14	*Hobbies & Interests; Shopping; Drinks; Smoking; Pets; You & Your Family; Motoring; Charities; Family Health; TV & Telephone; Computing & Internet; About Your Home; Financial Planning; Information Guides.*
2006	148	22	*Groceries; Hobbies; Shopping; Your Interests; Drinks; Your Home; Outgoings; Your Occupation; Charities; You & Your Family; Pets; Family Health; Motoring; Financial Products; TV & Telephone; Computing & Internet; Local Area; Tobacco; Financial Planning; Planning Your Future; Information Guides.*
2007	136	25	*Groceries; Shopping; Newspapers; Hobbies; Books; Home; Home Improvements; Your Local Area; Occupation; Outgoings; Financial Products; You & Your Family; Motoring; Cars; Charities; Family Health; Telephone & Internet; Shopping Channels; Leisure; Entertainment; Pets; Tobacco; Financial Planning; Retirement; Education.*
2008	133	27	*Groceries; Shopping; Newspapers; Hobbies; Entertainment; Environment; Home; Home Improvements; Your Local Area; Charities; Occupation; Business Owner; You & Your Family; Family Health;*

Year			
			Health Concerns; Outgoings; Internet; Telephone & TV; Financial Products; Financial Planning; Holidays; Pets; Education; Tobacco; Leisure; Motoring; Cars; TV Viewing.
2009	*130*	*26*	Groceries; Shopping; Your Local Area; Hobbies; Newspapers; Coffee; Insurance; Environment; Internet & TV; You & Your Family; Occupation; Outgoings; Home; Leisure; Financial Products; Charities; Telephone; Credit Crunch; Financial Planning; Family Health; Technology; Education; Cars; Pets; Tobacco; Shopping Vouchers.

Table 1 indicates that the ROP offers a large number of questions across a range of different areas. For example, the 2009 survey boasts 130 questions spread across 26 different sections. The survey covers topics such as consumption and expenditure (Groceries, Shopping, Newspapers and Outgoings), preferences and opinions (Environment, Charities and Local Area), health and education (Family Health, Education and You & Your Family), demographics and geography (You & Your Family and Home), and the economy (Occupation, Financial Products, Financial Planning and Credit Crunch). Figure 2 exemplifies how the various sections were divided up and presented on the survey in 2009. Although the name and order of the sections has changed slightly from year-to-year, the style of the survey has remained consistent since 2004.

Figure 2. Section of the annual ROP, January 2009

The choices of questions on the survey are not solely down to Acxiom; there are two types of question which appear on the form. The first type includes the core questions which are asked consistently so that continuity over time for key variables can be maintained. Examples include the questions on address, age, income, sex, socioeconomic status, tenure and grocery spend. The core questions typically feed into the construction of Acxiom's core products as they allow for consistent time-series analysis. It is mandatory that specific core questions appear on each survey because Acxiom is committed to providing data which will support time-series analysis going forward. Therefore, all changes to the survey that may impact on time-series analysis are stringently reviewed. In addition to the core questions, Acxiom's survey programme provides a mechanism for clients to place their own questions on the survey. These are termed 'sponsored' questions because their existence on the survey is paid for by the client. For example, Yorkshire Forward sponsors a large number of questions on the survey in an attempt to find out people's opinions about certain issues relating to their local area. Despite Yorkshire Forward having a regional interest, all sponsored bespoke questions are delivered nationally across GB. Sponsored questions are not ideal for time-series analysis as once the company stop paying for their existence on the survey, the questions are removed. The process to get bespoke questions onto the ROP is a collaborative one. This is because Acxiom want to ensure they are asking suitable questions which will add value to their data. The course of action is shown in Figure 3.

Clients, with Acxiom's support, meet to discuss key topics of interest and data applications

Both organisations agree suitable questions to ask

The Sponsorship and the Analysis teams at Acxiom meet to discuss the final question wording and the anticipated use of the collected data

The proposed questions will be mocked up by Acxiom in a survey template

Clients review the questions wording before sign-off

Figure 3. The process of securing sponsored questions on the ROP

Any other questions on the survey are those which have been devised by Acxiom. However, to ensure Acxiom gets a return on all questions, they are integrated with a business mentality. For instance, any additional questions will be added because Acxiom feels that the data produced will be of value to potential customers. Acxiom aims to be as up-to-date and current as possible which means the ROP is constantly evolving. A prime example of this is the 'Credit Crunch' section added in 2008 to collect data specifically on the impact of the financial crisis which began in 2007 (Nesvetailova and Palan, 2008; Langley, 2008).

2.3 Multi-channel Data Collection

The ROP survey is rolled out twice a year, initially in September and then in the following January. The results from these two surveys are combined into one data file and are packaged up as the data for the latter year. For instance, the surveys distributed in September 2008 and January 2009 are combined to form the results for 2009. September and January are chosen specifically because extensive research by the Data Acquisition team has found that the greatest response rate occurs in these months. During this time of year, respondents are more likely to fill out the survey because bad weather and decreasing levels of daylight mean people are at home and indoors for more of the time than they would be during the spring or summer. The survey forms are delivered through direct mail which provides a controlled and reliable method to survey a large number of households (Bradburn, 2004). Acxiom uses a variety of sources of addresses to ensure all parts of the country are surveyed and all demographic groups are well represented. This meant that, in 2009 for example, only 0.4% of all Middle Super Output Areas (MSOAs) across GB did not return a response. As a result, this reduces the amount of estimation and modelling that Acxiom has to undertake to create data accurate at low levels of geography for their various data packages as described in Section 5.

On average, Acxiom receive over 1 million household responses from the ROP each year, which makes it the largest annual paper-based survey in GB and the largest population study outside of the Census of Population. Figure 4 shows the raw counts of household responses received in each Government Office Region (GOR) of GB in 2009. It is clear that the South East provided the most responses with over 140,000 households returning the survey, followed by the North West with 115,000 and the East of England with just over 100,000.

However, these are just the raw counts and are not adjusted by the number of households in each area or by the number of questionnaires sent out. Consequently, Figure 5 shows the number of responses from the 2009 survey divided by the total number of households present when the survey was distributed for each NUTS 2 region. It would have been more appropriate to divide by the number of questionnaires sent out, however this figure remains confidential. To find out the total number of households present at the time of the survey, the 'total households' variable from Acxiom's 2009 Aggregate Dataset (AD) was used. This is essentially calculated by aggregating the total number of households within the 2009 Royal Mail Postcode Address File (PAF). Figure 5 paints a different picture than Figure 4 as the North East is shown to have the highest response rate (4.7%) despite contributing the lowest number of counts. Wales, the East Midlands, and the East of England also have high response rates well over 4%. The lowest and hardest to reach respondants reside in London, Scotland and the South East. London in particular has a very low response rate with 2.8% of households returning the survey.
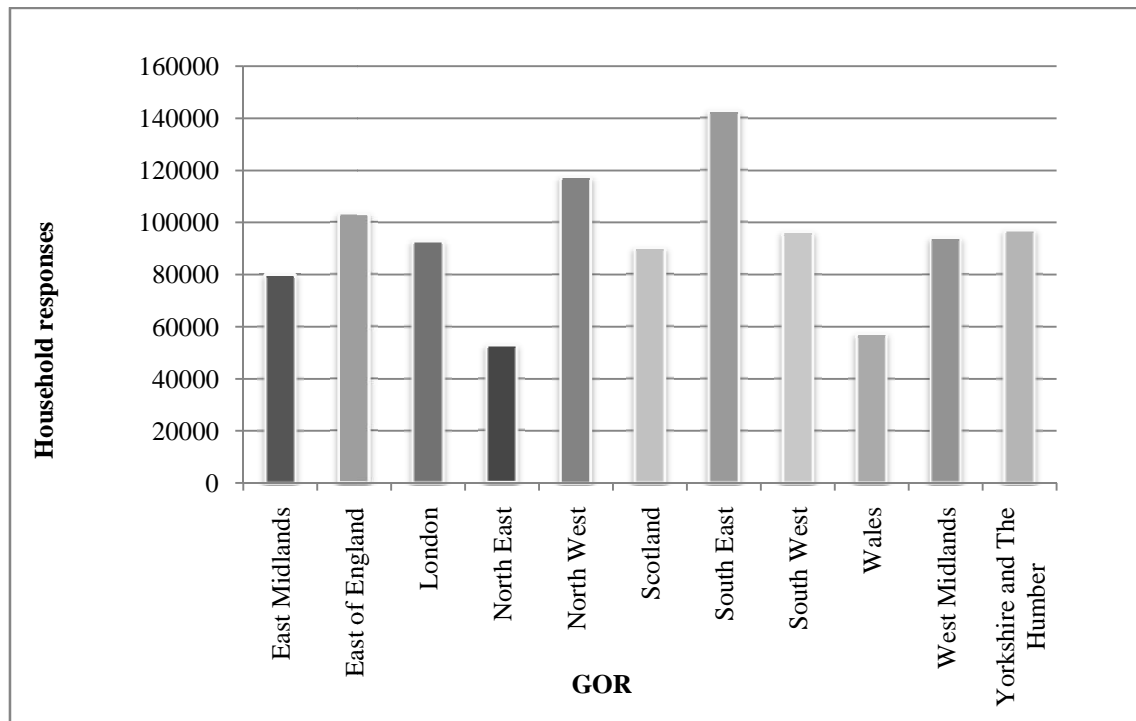
Figure 4. Raw counts of responses to ROP in 2009 by GOR
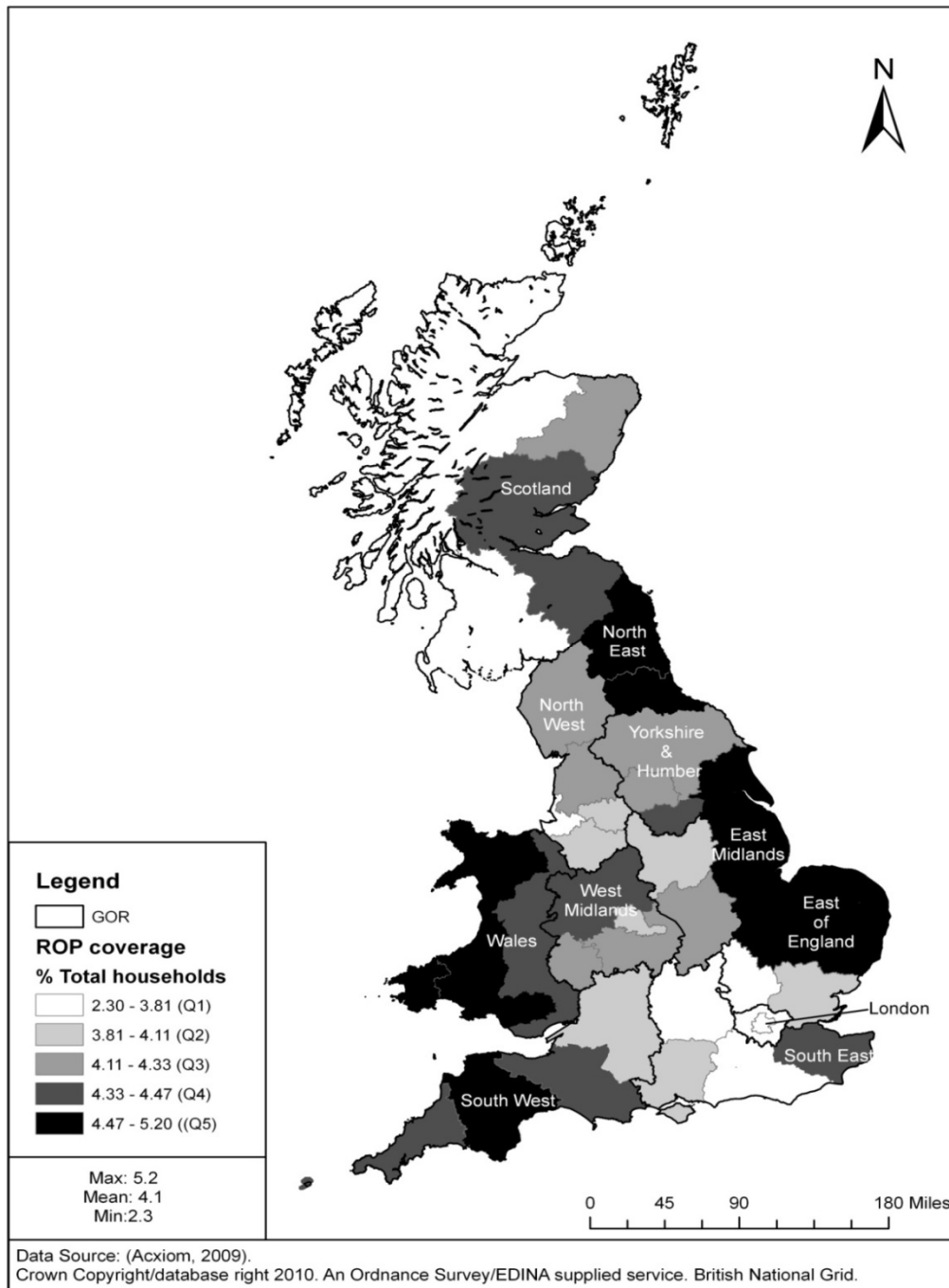
Figure 5. Household response rate to ROP in 2009 by NUTS 2 region

Figure 6 provides some further context to both Figures 4 and 5, for it displays the level of response (ROP respondents divided by total number of households) within each Office for National Statistics (ONS) Local Authority District (LAD) geodemographic area classification for 2009. It is clear that Acxiom have somewhat varying levels of success across the different area classifications. For example, the most common type of respondent is a household from the 'Coastal and Countryside' or 'Mining and Manufacturing' categories. In

comparison, 'London Centre', 'London Cosmopolitan' and 'London Suburbs' are represented the least within the ROP sample. This pattern is reflected in Figure 5 on account of the high number of responses along the East coast and relatively low numbers in and around central London. Overall, this type of bias is not unfamiliar. Using the same geodemographic technique, Frosztega (2000) found response rates in the Family Expenditure Survey (FES) to be higher for council estate residents and the affluent elderly people living in rural communities, whilst lower for wealthy executives living in inner city areas and white collar workers living in better-off multi-ethnic areas.
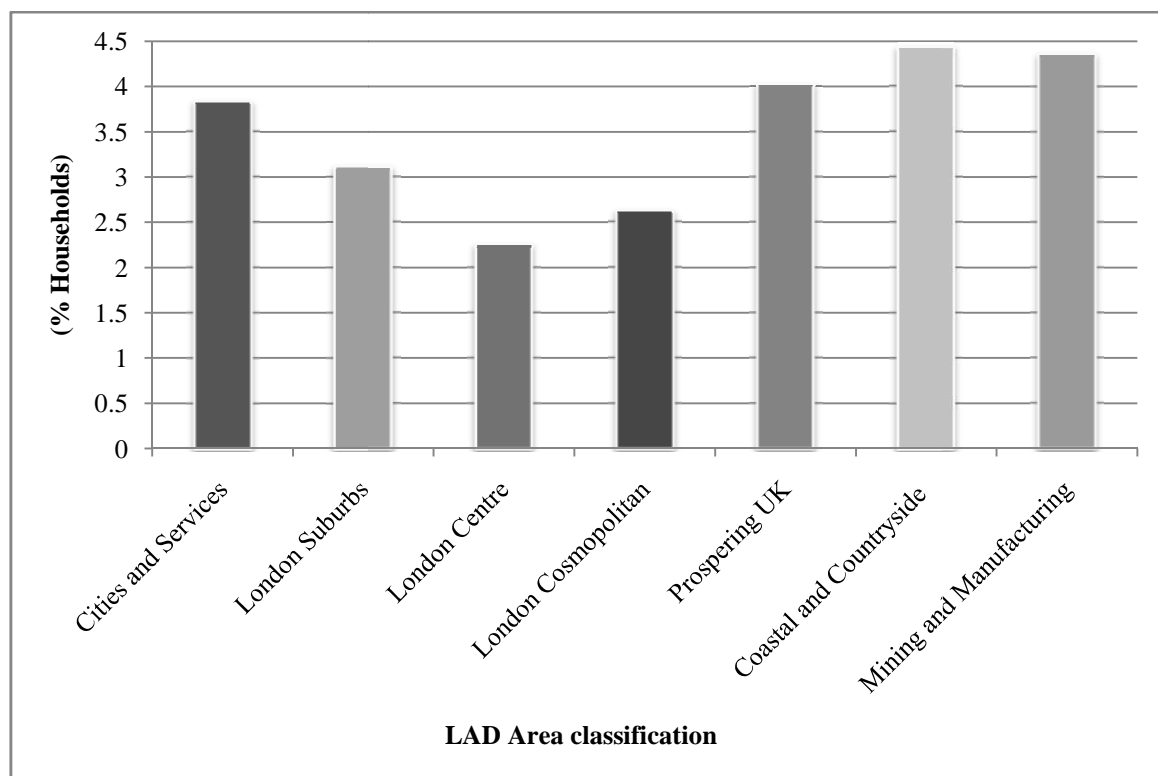
Figure 6. The percentage of responders in by LAD area classification for the 2009 ROP (ONS, 2009)

In addition to the geographic coverage, it is important within any household survey to try and achieve a representative demographic profile of respondents (Sorensen, 1996; Deaton 2000). Nevertheless, as with all optional consumer surveys, there is always going to be an inherent bias within the sample population. Figure 7(a) portrays the number of respondents by age and gender for one of the surveys distributed in 2006 (most recent complete dataset available). It is clear that females are the overriding gender and people aged between 35-55 years the most common age group. Moreover, as a means to test how representative the ROP is of GB's

population, Figure 7(b) provides a basic comparison between the age and sex of the 2006 ROP respondents against the 2006 ONS mid-year estimates (MYEs). Overall, the shape of the ROP pyramid follows a similar pattern to the proportions held in the 2006 MYEs. This is especially apparent for females, as the black outline of the MYEs sits close to the proportions of the ROP for the majority of age bands. However, it is also clear that there is a bias within the ROP towards people aged 40-70 years old, as the ROP proportions exceed those from the MYES considerably. This is predominantly prevalent for males, as the MYEs show noticeably smaller proportions for males aged 50-70 years old. In conjunction, Figure 7(b) also exemplifies the underrepresentation of people aged 18-24 years old, particularly males. Acxiom are aware of this problem and recognise that traditionally, the hardest to reach are young affluent males (Frosztega, 2000). As a result, when planning the ROP survey each year, those groups less likely to respond are over-sampled to try and increase the number of respondents. This is done through 'door-drop campaigns', a dynamic technique to top-up areas that appear to be returning low numbers of responses. Additionally, Acxiom replicates the paper-based survey through an online version of the ROP. This collects approximately 400,000 responses per year and is useful for targeting younger age groups less willing to fill in paper-based surveys. The responses are also immediately digitised which heavily reduces the processing time.

(a) ROP by age and gender, 2006   (b). ROP and MYEs, 2006
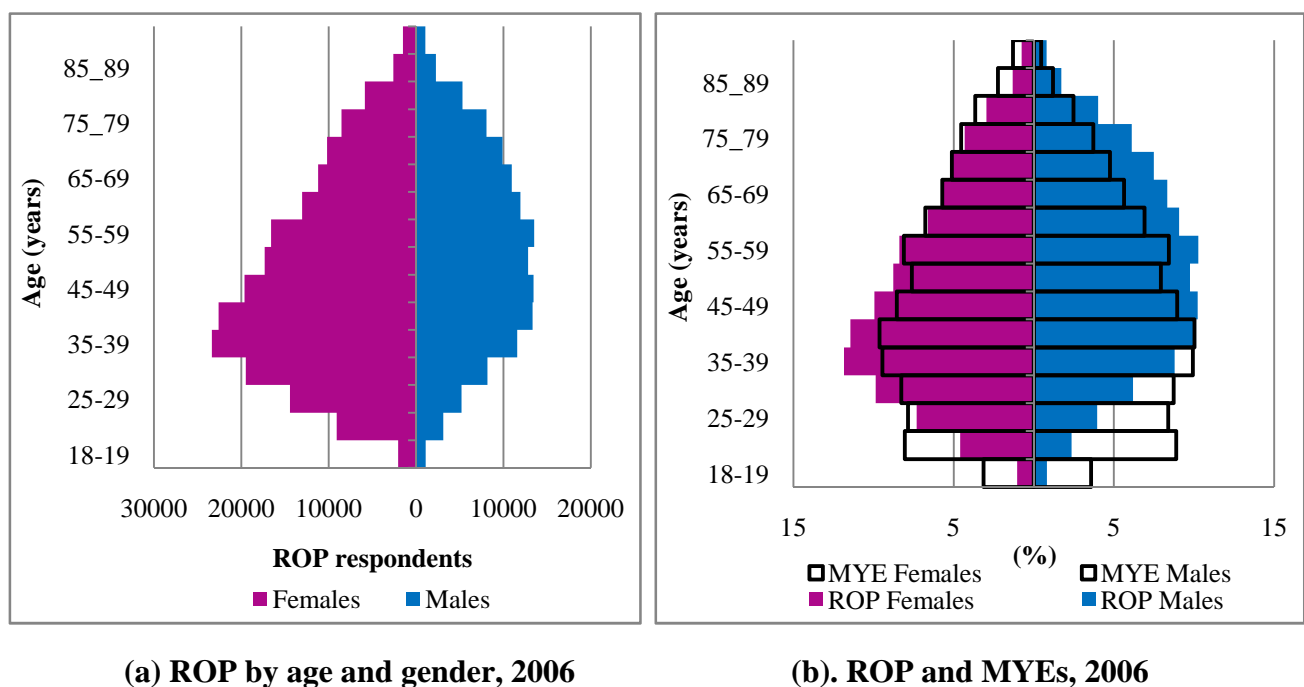
Figure 7. GB population pyramids by and gender for the 2006 ROP and 2006 MYEs

Nevertheless, despite the obvious advantages with having the ROP run online, there is the issue of households responding more than once in a year via the paper and online survey. As a way to deal with this problem, Acxiom have technology in place which allows them to create a 'single customer view' of each household that responds to the ROP. Once a household replies to the survey, it is assigned a unique identification number. Therefore, when Acxiom receive a response, they know who and where it has come from. This allows Acxiom to check for any duplicates in the same year. It is rare that this happens in the ROP sample although it is a common problem with other data sources such as product warranty cards (people buying multiple products).

3 Data

3.1 Data Processing

Once the ROP survey has been completed by a household, the form is returned via a free post envelope which comes with the survey. As stated, the return addresses are regional, which means that the responses are housed at a number of collection points in different regions across GB. After waiting for a period of approximately 8 weeks, all of the received surveys are sent off to a data processing company in Manila, Philippines. The data from the surveys have been input to a database in a number of different ways over the last 10 years. In the past, the responses were simply keyed into the computer. To reduce the likely event of errors, 'double keying' was used so that comparisons could be made between the two entries. Any differences or inconsistencies in the data would result in a survey being re-entered. However, this method was extremely inefficient, resulting in an extended wait for the final dataset. Consequently, the use of optical mark recognition (OMR) was introduced to speed up proceedings. OMR is commonly used when high-volume data entry is required (Curtis and Cobham, 2008). The documents are passed through an optical mark reader which scans the boxes on each survey and determines those in which a mark has been made. Once all responses have been scanned into the system, the data are sent back to the Acxiom data processing centre in Normanton, England. The ROP surveys completed online are also sent straight to Normanton to be combined with the paper-based responses. Once here, the data are kept under extremely high levels of security to ensure the confidentiality and privacy of the customer information. This entire process happens twice a year. The first batch of surveys

sent out in September are available as raw counts by November, then the second half distributed in the coming January are available in the same format by March.

3.2 Microdata

The results from the surveys in September and January are combined to form the entire micro dataset which is available by March. For example, the responses from September 2009 and January 2010 will form the latest micro dataset for March 2010. The microdata represent Acxiom's most detailed product. Each record within the dataset is an individual household response to the ROP. The micro datasets can be extremely large, as the most recent survey (2009) involved over 1,000,000 household responses and in excess of 1,000 possible answers to the various questions. Given the size of the annual datasets, Acxiom offers clients the choice of requesting data for specific regions to make processing more manageable. As mentioned, the microdata represent all of the raw responses given to the ROP. There is no manipulation or estimation carried out to try and fill in the gaps due to under-represented areas or demographic groups. However, because the data come with every respondent's full address, the data can be aggregated up to any geographic level. In comparison, the other datasets which Acxiom produces are restricted by both geography and the selection of variables on offer. A major advantage of using the microdata is that the records can be used to analyse relationships between any of the different variables. For instance, logistic regression models could be used to find for example, if there are any significant differences in health between demographic or socioeconomic groups (Grundy and Holt, 2009).

Figure 8 exemplifies a sample of the microdata in SPSS. The 'postcode' and 'Ethnicity' codes are self explanatory, however 'OWNRNT', 'RESTYPE' 'MARRYD' 'RDOB', 'KIDAGE' and 'GROCER' refer to household tenure, residence type, marital status, date of birth, age of children and retailer used for groceries respectively. In each case, some of the different responses are coded numerically, with each number referring to a value in an accompanying data dictionary. For example, record 1 is somebody who lives in postcode BD10 0BE, owns a detached house, born on 02/06/1944, is of white ethnic background, has no children, and shops at Asda, Marks and Spencer, Morrison's, Netto and Somerfield.

Figure 8. Example of Acxiom microdata

To provide further understanding, Figure 9 displays the information from the data dictionary held on the 'GROCER' variable from Figure 8 (Where do you shop for Groceries?). Column F and G demonstrate how the numbers equate to an answer from the survey. In addition, the client is also provided with the 'Offset', 'Field Length' and 'Value Length' information. These are crucially important for reading the data into software packages as the data are typically supplied in .DAT file format. More specifically, this means the data are held in one long horizontal line with no delimiter to separate the characters. Therefore, the 'Offset' highlights what position a particular variable starts in the data, and the 'Field Length' determines the maximum number of characters for that variable. The 'Value Length' then provides information on the number of characters in each individual answer which helps in the partitioning of multiple answers. For example, the 'GROCER' variable has a 'Value Length' of 2, which makes the number 1, 01.

	A	B	C	D	E	F	G	H
1	Name	Offset	Field Length	Category	Field Name Description	Value Description	Value	Value Length
2	GROCER	1227	36	GROCERIES	QA 01 Where do you shop for Food & Groceries	Aldi/Netto	01	2
3				GROCERIES	QA 01 Where do you shop for Food & Groceries	Asda	02	2
4				GROCERIES	QA 01 Where do you shop for Food & Groceries	Budgen	03	2
5				GROCERIES	QA 01 Where do you shop for Food & Groceries	Co-op	04	2
6				GROCERIES	QA 01 Where do you shop for Food & Groceries	Iceland	05	2
7				GROCERIES	QA 01 Where do you shop for Food & Groceries	Kwik Save	06	2
8				GROCERIES	QA 01 Where do you shop for Food & Groceries	Lidl	07	2
9				GROCERIES	QA 01 Where do you shop for Food & Groceries	Makro/Spar/Vg	08	2
10				GROCERIES	QA 01 Where do you shop for Food & Groceries	Marks & Spencer	09	2
11				GROCERIES	QA 01 Where do you shop for Food & Groceries	Morrison's	10	2
12				GROCERIES	QA 01 Where do you shop for Food & Groceries	Sainsbury's	11	2
13				GROCERIES	QA 01 Where do you shop for Food & Groceries	Sainsbury's Local	12	2
14				GROCERIES	QA 01 Where do you shop for Food & Groceries	Somerfield	13	2
15				GROCERIES	QA 01 Where do you shop for Food & Groceries	Tesco	14	2
16				GROCERIES	QA 01 Where do you shop for Food & Groceries	Tesco Metro	15	2
17				GROCERIES	QA 01 Where do you shop for Food & Groceries	Waitrose	16	2
18				GROCERIES	QA 01 Where do you shop for Food & Groceries	Corner Shop/Local	17	2
19				GROCERIES	QA 01 Where do you shop for Food & Groceries	Garage Forecourts	18	2

Figure 9. Acxiom microdata dictionary

The accuracy and degree of completeness of any data source is also an important issue (Sorenson *et al.,* 1996). One particular problem is the existence of missing data or blank fields in the data. For each single variable, *"...it should be considered whether missing information means that exposure or outcome has not taken place or whether the variable represents a missing value (*Sorensen et al., 1996, p.438). As the ROP microdata just contain the actual responses to the ROP, nothing is done to accommodate for the blank fields. Instead, these are left for the end user to decide how best to interpret the missing information. It is only during the construction of the aggregate datasets that Acxiom need to deduce the meaning of blank fields. However, due to the sheer number of respondents, blank fields are not dealt with on a case by case basis. The interpretation of the question dictates how the variable is built. For example, no tick for the income question is likely to mean that the respondent did not want to disclose their income, rather than they actually work for free. Table 2 portrays the number of responses and recorded blank fields for the postcode, age, income and accommodation variables for those households which responded in the Yorkshire and Humber region in 2009. The postcode field has a 100% response rate from all households which filled out the ROP. This is not unique to the Yorkshire and Humber sample, as Acxiom use the latest PAF to clean all of the responding household postcodes. The age and accommodation variables contain a similar amount of blank fields as only a small proportion of households decided to withhold their information. Household income is arguably a more sensitive piece of information for somebody to divulge, which explains the higher rate of blank fields for this question. Nevertheless, more than 75% of households still disclosed their annual household income.

Table 2. Blank fields for selected variables in ROP data for Yorkshire and Humber region, 2009

	Respondents	Postcode	Age	Income	Accommodation
Blanks	*n/a*	*0*	*7548*	*20237*	*8922*
Percentage	*n/a*	*0*	*8.35*	*22.40*	*9.87*
Total cases	*90378*	*90378*	*82830*	*70141*	*81456*

3.3 Aggregated Data

Once the raw microdata have been established, they are put through a rigorous process of weighting and manipulation to produce a number of different aggregated data products. The three main packages sold to clients include the Acxiom Population Estimates (APE), the Aggregated Data (AD) and PersonicX, Acxiom's segmentation profile.

3.3.1 Acxiom Population Estimates

Since the 2001 Census, some areas around GB have changed with regards to their population due to new housing developments, changing family numbers and inward and outward migration. Therefore, central to all Acxiom products is the demographic information attached to each postcode. This is known as Acxiom Population Estimates (APE) and is essentially a benchmark product. As well as being available to clients it is also used internally to ensure that other Acxiom products are representative of the GB population. To ensure that Acxiom products are applicable every year, the Product Development team re-estimates the demographic profile of each postcode in the country to form of a new APE dataset. In order to do this, the profile takes into account the latest Government MYEs and residential information from the Royal Mail PAF, Regional Trends, the 2001 Census, Population Trends, Acxiom's central database and National Statistics.

The construction of the APE requires a number of data sources and steps to be completed. The first stage involves processing externally sourced data from official sources listed above to create high-level regional estimates, which are subsequently used as controls for aggregations of low-level postcode estimates. The regional estimates are used to build a regional demographic control matrix containing five different dimensions (age band; head of household; presence of children; tenure; income). Table 3 displays eight out of the thirty control variables used in 2008 (total household, total population and six categories of household size).

Table 3. Regional level control matrix, 2008

GOR	thh08	tpop08	thhsz1	thhsz2	thhsz3	thhsz4	thhsz5	thhsz6
EA	990,004	2,280,112	310,017	354,495	151,980	120,531	40,064	12,917
EM	1,867,290	4,463,756	496,111	718,229	306,632	241,365	78,866	26,087
GL	3,278,430	7,694,483	1,098,712	1,028,681	521,218	390,218	162,546	77,055
NO	1,348,856	3,166,511	392,160	482,226	235,825	170,461	50,816	17,368
NW	2,758,538	6,354,960	903,762	921,075	453,098	325,099	112,762	42,742
SC	2,355,525	5,001,335	943,943	721,055	357,251	239,444	72,756	21,076
SE	5,116,060	11,553,688	1,765,597	1,724,818	777,704	585,201	197,264	65,476
SW	2,255,597	5,194,580	682,116	862,854	334,987	261,244	86,242	28,154
WA	1,260,583	2,993,047	353,699	459,005	216,156	157,877	55,711	18,135
WM	2,272,824	5,367,575	684,485	790,611	378,155	278,607	97,262	43,704
YH	2,245,604	5,192,102	704,267	791,058	361,217	266,205	86,190	36,667
GB	25,749,311	59,262,149	8,334,869	8,854,107	4,094,223	3,036,252	1,040,479	389,381

Next, the low-level estimates are created at postcode level using a combination of data from Royal Mail and Acxiom's InfoBase Lifestyle Universe (ILU) data, plus any historical postcode level estimates where appropriate. The ILU is Acxiom's core database which is fed by all of the company's data collection methods. This huge database stores information at an individual level on over 60% of the population. On completion of the postcode updating process, the ILU file is aggregated to postcode level. During this aggregation process, measures of the demographic variables are calculated, incorporating imputation where required through linear regression.

The low-level (postcode) and high-level (GOR) figures are finally linked through various techniques. The annual postcode roster is used to match postal, local and central government geographies to help identify the geographic areas that each specific postcode will fall into. In addition, estimates are gradually smoothed from GOR level down to postcode level using initial estimates as starting points. At each stage in the process, new targets are set for the parent area and the initial estimates are adjusted to meet the new criteria. Moreover, substantial quality control is carried out during the adjustment and calculation of population figures to ensure consistency. In a process like this, inconsistencies can occur quite readily

and it is important to check for numbers that cannot exist together within a postcode. These errors are then corrected in isolation and the estimates revised until a complete solution is available. The final information is a complete postcode dataset for GB with the total population, age distribution (age is defined in bands of approximately 10 years), total number of households, gross household income (expressed in bands of income of approximately £5,000), presence of children in the household, homeownership status, (owned, privately rented or council rented) and the size of households.

3.3.2 Aggregated Dataset

The Aggregated Data (AD) is the main product which Acxiom builds from the ROP. The dataset contains local area estimates for the total number of households owning certain products and displaying different attributes/behaviours. In short, the AD is a version of the ROP microdata which is argued to be fully representative of GB. However, in addition to the existing ROP variables, clients have the ability to build new variables based on questions provided on the ROP. In order to achieve this, the AD is dependent on two additional pieces of work being completed; the creation of the APE and the weighting of the consumer surveys, commencing after the January survey collection closes in April each year. However, whilst the APE and weighting process are being completed, a parallel process is undertaken, which is the definition of all the variables to be produced within the AD product. Initially, all the survey questions are input into a spreadsheet and then a consultation process between the client and the Acxiom Account team begins. All the variables derived from the survey questions that are required for clients are defined and outlined on this same spreadsheet. Once all variables have been signed off, the analysts add more information to this spreadsheet, including codes to define the variable and bases to smooth variables where required. Thereafter, this spreadsheet, together with the output from the APE, the survey and the associated survey weights are provided as inputs to start the processing of the AD variables.

Unfortunately, all survey responses have an inherent bias. Therefore, as mentioned above, a weighting process is used and the survey response goes through stages to remove this bias and ensure that the AD counts are representative at every level of geography. The weighting process is based on the estimates from the APE, which ultimately reconcile with published

UK statistics from the ONS and the Expenditure and Food Survey (EFS). To complete this process in a timely manner, the most efficient way of processing the survey is to split it up. The survey is split into blocks which are processed independently of each other so it is possible for them to be completed in parallel. The processing of the blocks and calculation of all the single variables are completed through the following steps.

Step 1: The initial variable definitions taken from the spreadsheet are input into the processing programs and then the weighted survey data are merged with the APE data. This creates the initial dataset which is then used for all further processing. A number of geography levels (e.g. LSOA, LAD) are utilised in this step and for each of these levels an estimate is calculated for every variable. Where there is an estimate of zero for a variable at a given geography, the estimate from the geography level above replaces the zero.

Step 2: It is important for certain variables that their overall totals equal either the total number of households or another variable. For instance, the internet provider variables should sum to the variable 'has an internet connection'. In such cases they are adjusted at a high geography level, therefore, creating smoothing targets at this high level.

Step 3: Those variables that need to add up to other variables, or to total households, are then adjusted at each geography level, making sure that at each stage they equal the higher level targets produced in step 2. The process also ensures each geographic level estimate is an integer. For example, the total number of households in postcode TW11 9AD is 25; therefore this becomes the target for the survey data. Table 4 demonstrates how Acxiom smooth the survey counts at postcode level so that the four categories for the variable 'number of cars' sum to the target of 25 households, in order to retain the shape of the distribution.

Table 4. Smoothing process for example postcode

Postcode	0 cars	1 car	2 cars	3+ cars	Households
Survey count for TW11 9AD	*5*	*6*	*3*	*1*	*15*
Survey proportion	*5/15*	*6/15*	*3/15*	*1/15*	
Target Counts for TW11 9AD	*25*(5/15) = 8*	*25*(6/15) = 10*	*25*(3/15) = 5*	*25*(1/15) = 2*	*25*

Step 4: Any variable where the summed total is greater than the total number of households is adjusted down to align with the total households figure. Checks are made to ensure all

variables at all levels of geography are consistent. The final processed data and counts are then output, summary information is added to the spreadsheet and data for the different blocks are collated together prior to dispatch. Finally, the aggregated dataset is sent to the client for evaluation to ensure that it meets expectations. The lowest level of geography for which the product is available is the LSOA level. Figure 10 provides an example of how the AD is provided to clients. The LADs are listed down the left hand side, with the counts for each variable in each of the subsequent columns going from left to right For example, based on the data in Figure 11, LAD 00AA has a population of 68,294,735 households, 79 households that shop at Aldi, 649 households that shop at Asda, 94 households that shop at Budgens and 95 households that shop at CO-OP.

	A	B	C	D	E	F	G
1	lga	B1	B4	sh002	sh003	sh004	sh005
2	00AA	6829	4735	79	649	94	95
3	00AB	174249	72720	10736	43469	484	4400
4	00AC	326494	135744	10919	42449	5461	4716
5	00AD	236156	99265	11317	45219	516	12727
6	00AE	268879	103719	9021	50951	1566	2692
7	00AF	312519	136155	7155	23150	498	14498

B1 = Population
B2 = Households
Sh002 = shop at Aldi
Sh003 = Shop at Asda
Sh004 Shop at Budgens
Sh005 = Shop at Co-op

Figure 10. Aggregate dataset for GB LADs, 2009

3.3.3 Micro vs Macro

As mentioned in section 3.3.2, the raw responses from the ROP are used in the methodology to build Acxiom's AD. Therefore, it is important to try and understand how much estimation and manipulation of the original ROP data actually takes place. Figure 11 displays basic bivariate linear regression relationships for a selection of four variables taken from the ROP. The scatter plots exemplify the relationship between the values held in the ROP microdata against the corresponding value produced for the AD. The regression models have been run for the responses received in all 694 MSOAs within the Yorkshire and Humber region for 2009 (n=694). MSOAs were chosen as the desired geography because they represent one of the smallest spatial scales offered by Acxiom for the AD. Acxiom argue that because the AD dataset is modelled from the bottom-up, it retains low-level (MSOA) variations across all of its variables. Consequently, despite small sample sizes, there should still be a strong relationship between corresponding variables in the ROP microdata and the AD.
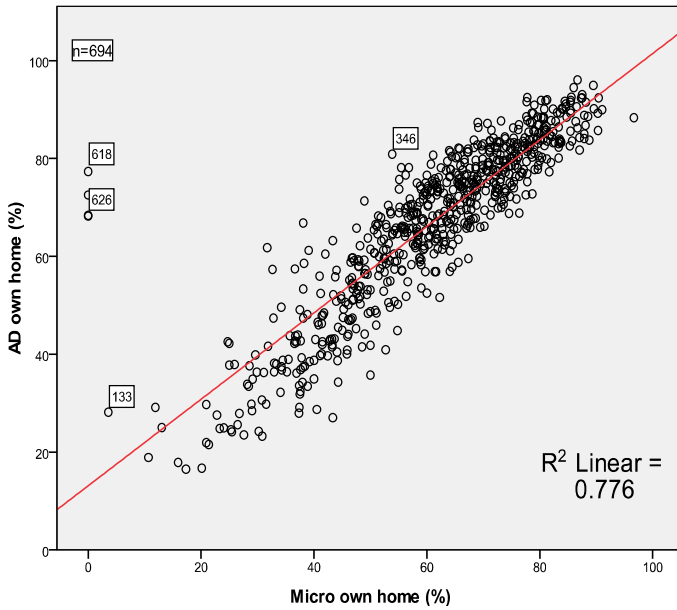
(a) Owner occupied house
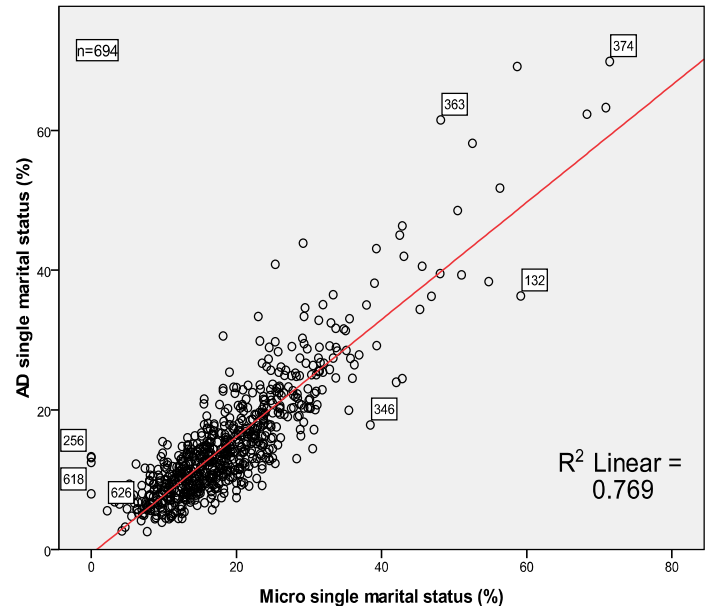
(b) Single marital status
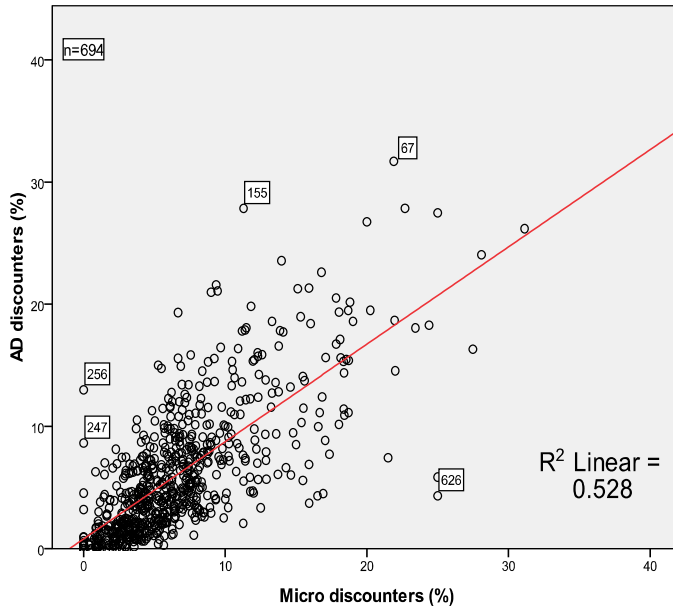
(c) Household main shops is at discount retailer
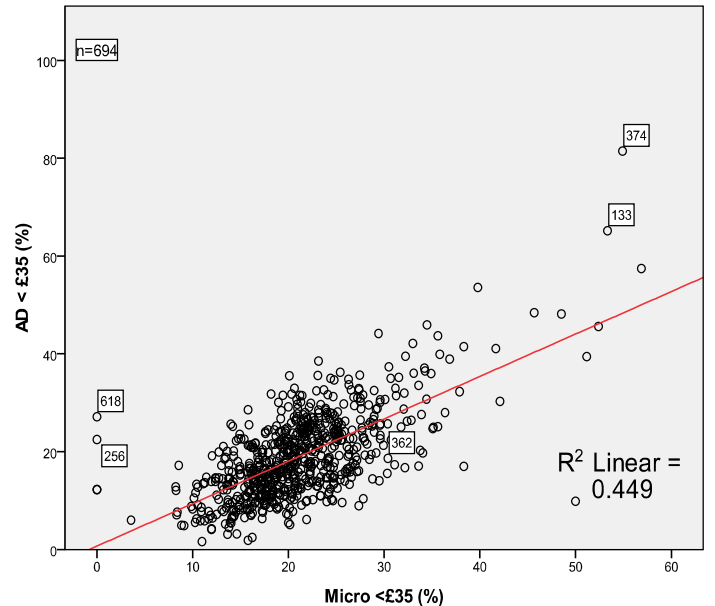
(d) Household weekly grocery spend is less than £35

Figure 11. Scatter plot analysis of Acxiom's ROP and AD using data from the Yorkshire and Humber region, 2009

Figure 11(a) shows that the number of owner occupied houses within the ROP microdata and AD are very similar at MSOA level within Yorkshire and Humber. The linear regression model shows an R^2 value of 0.776 which indicates that the ROP microdata can account for 77.6% of the variation in the AD. A similar situation can be seen in Figure 11(b) for the number of single people living in Yorkshire and Humber (R^2=0.769).   The strong

relationships highlight that the ROP microdata can be trusted for these two variables, as they clearly heavily influence the values produced in the AD. This is encouraging, for it would be expected that at such low-levels of geography, exaggerations caused by small numbers in the sample would have to be brought back in line with values seen at higher spatial scales.

Figure 11(c) highlights the relationship between values held in the ROP microdata and AD for the number of households which shop at discount stores. The R^2 value of 0.528 for this variable is less significant than the previous two. Although more than 50% of the variation in the AD can be accounted for by the ROP microdata, the remainder is not explained. This is arguably a result of Acxiom using additional data sources within the production of the AD. Acxiom do this to remove bias in the sample and ensure that the AD are in line with comparative packages such as the 2001 Census, EFS and MYEs. Similarly, the same can be said for Figure 11(d), which portrays the relationship between the two datasets for households which spend less than £35 a week on groceries. In Figure 11(d), the low R^2 value of 0.449 shows that less than 50% of the variation in the AD can be accredited to the ROP microdata. This is arguably a result of a much smaller sample for this question, for households can be more reluctant to disclose information about their spending and income than about other activities or behaviour.

Despite the close relationships between the microdata and AD, there are a number of outlying MSOAs which have conflicting sets of data. Table 5 lists a number of the outlying cases which are highlighted by their case number in Figure 11.

Table 5. A selection of the main outliers in Figure 11

Case	MSOA	WARD	LAD	Description
132	E02001640	Netherthorpe	Sheffield	Student area
133	E02001641	Netherthorpe	Sheffield	Student area
256	E02002265	Calder Valley	Calderdale	Very rural
346	E02002355	Weetwood	Leeds	Student area
362	E02002371	Chapel Allerton	Leeds	Young professionals
363	E02002372	Headingley	Leeds	Student area
374	E02002383	Headingley	Leeds	Student area
618	E02005742	Ingleton & Clapam	Craven	Very rural
626	E02005750	Great Ayton	Hambleton	Very Rural

It is evident that a number of the outlying MSOAs appear multiple times across the graphs in Figure 11. Furthermore, the majority of them are either located in areas heavily populated by students or much more rural and wealthier parts of the region. This is not surprising, as it was discussed in section 2.3 that the ROP is underrepresented by people under the age of 24 years and those households with an income in the higher earning brackets. At a low geographic unit such as MSOA, this will undoubtedly cause small number problems within the data. More specifically, the small sample will cause exaggerated values in the microdata (Goodchild and Gopal, 1989), which is why the AD is constructed using weights calculated from other datasets and not just the ROP alone.

3.4 PersonicX

Acxiom also uses the data captured in the ROP to produce its unique geodemographic area classification named PersonicX. Geodemographic classifications encapsulate variations in neighbourhood characteristics such as demographics, socio-economic status, wellbeing and housing tenure. Brown (1991) states that they are *"...a shorthand label for both the development and application of area typologies that have proved to be powerful discriminators of consumer behaviour and aids to market analysis"* (Brown 1991, p.221). The ROP survey data are combined with data from Acxiom's ILU and the 2001 Census (OA level) to produce a behaviourally optimised segmentation for the whole of GB. The integration of the ROP data gives Acxiom a distinct competitive advantage over traditional geodemographic classifications such as MOSAIC (Experian), ACORN (CACI), CAMEO (EuroDirect) and OAC (census), which only really provide an insight into the socio-economic and demographic makeup of neighbourhoods (Goss, 1995). These types of area classification assume that consumer behaviour can be predicted according to 'who you are and where you live'. In comparison, the ROP data allow PersonicX to accurately predict consumer behavior such as holiday destination, online activity and spending patterns within individual clusters. Additionally, as well as being able to differentiate consumers by their behaviour, the geographic level provided is finer than any of the other segmentation products available both publically and commercially. The wealth of data and information which Acxiom assembles has allowed them to produce area classifications at household and individual level. So far, Acxiom have produced a total of five different area classification profiles, each at varying geographic levels (Table 6).

Table 6. PersonicX packages

Personicx Solution	Used for	Clusters	Level
PersonicX v2	*One to one marketing and customer insight*	*52*	*Household*
PersonicX GEO	*Spatial and catchment area analysis*	*60*	*Postcode*
PersonicX Financial	*Financial specific solution for banks, credit cards*	*50*	*Individual*
PersonicX Onliners v2	*Identifies online behaviours and internet use.*	*24*	*Individual*
PersonicX International	*Customer intelligence on a global basis*	*n/a*	*Varies*

4 Comparative Surveys

So far, this paper has concentrated solely on describing the production of the various Acxiom products rather than comparing them with other datasets. The latter is very important to establish data credibility and therefore, the remaining section will involve a comparison of data from the Acxiom ROP with data from some of the major national surveys: the Labour Force Survey (LFS); the Expenditure and Food Survey (EFS); the Small Area Microdata (SAM) based on the 2001 Census; the General Household Survey (GHS); the British Household Panel Survey (BHPS); and the Survey of English Households (SEH). It is understood that a number of these surveys are now recognised under new names; however, for the purpose of this study they shall be referred to by the above. There are also other household surveys for which data are available, but those selected are most similar to the Acxiom ROP with regard to the types of questions asked of the respondents. The Acxiom AD have also been included to allow for comparisons to be made against the distributions found in the Acxiom microdata.

4.2 Questions and Consistency

For time-series purposes, the consistency of the questions asked on any survey is crucial. All of the continuous surveys mentioned offer time-series capabilities in so much as they are collected either quarterly or annually. However, many of the surveys have undergone a number of administrative and methodological changes which can affect time-series analysis. For instance, many of the selected surveys amalgamate to form modules within larger,

centralised surveys. The Integrated Household Survey (IHS) developed by the ONS integrates the EFS, LFS, GHS and the SEH (NDS, 2010). The EFS in particular has evolved considerably since 2001. Originally, the EFS was created in 2001 through combining the National Food Survey (NFS) and the Family Expenditure Survey (FES). It was then renamed the Living Costs and Food (LCF) survey in 2008. As a result of these changes, time-series data on certain variables can be problematic. Moreover, as much of the information in the EFS is collected through a written diary, privacy reasons prevent access to the entire dataset which means certain variables are not available from one year to the next (ONS, 2009a; ESDS, 2009a). In addition, the LFS now forms part of the Annual Population Survey (APS) and the SEH one half of the English Housing Survey (EHS). The BHPS also went through a major change in 2009 as it has now been replaced by a new longitudinal survey called Understanding Society (NDS, 2010; Buck, 2010).

The ROP survey is also not without its consistency problems as it too has seen some changes over the years. The survey has been running for over 20 years and has evolved substantially to prevent it becoming out-of-date. Nonetheless, Acxiom recognise that commercially it makes sense to have a consistent dataset and have therefore made every effort since 2004 to keep the questions and methodology consistent. Conversely, Acxiom cannot control for the sponsored questions paid for by external organisations. Once a company decides it no longer wants a question on the ROP, Acxiom will usually withdraw the question. Additionally, at small area scales such as postcode and Lower Super Output Area (LSOA), small number problems have the potential to make the data quite spiked on some of the variables from one year to the next. However, because the ROP data offer the benefit to aggregate up to higher levels of geography, these local changes are easily verified at regional or national levels.

The range of questions and information gathered from the ROP has already been discussed in section 3. However, it is important to understand how the types of questions offered on the ROP differ from those offered on other household surveys. The ROP is essentially a consumption based survey as it gathers information on household spending habits. In this respect, the EFS is the most comparable as it contains questions on household expenditure, income, composition, size, type and location (Fortin, 1995; Blundell *et al*., 1999; ONS, 2009a). Due to the wide range of variables on offer, the EFS data are used heavily in determining the basket of goods that constitute the Retail Price Index (RPI). The other household surveys do not offer as much in the way of recording consumption and

expenditure, although they still collect valuable household level information. For instance, the BHPS is restricted to questions relating to expenditure on durables, housing, demographics and income (Easaw and Herav, 2009; Blundell and Etheridge, 2009). Additionally, the LFS provides detailed information on labour market characteristics such as participation, income, training and qualifications, but again nothing on consumption or expenditure (Dennett *et al*., 2007; Blundell and Etheridge, 2009). The primary aim of the GHS has been to document the major changes in households, families and population which have occurred over the last 30 years. The main themes within the survey are household and family information, housing tenure and accommodation, consumer durables including vehicle ownership, employment, education, health and use of health services, smoking and drinking, income and demographic information (National Statistics, 2003). The primary areas of consumption recorded in the GHS surround smoking, health and consumer durables. In conjunction, the SEH collects data on the type of accommodation, household and personal characteristics, tenure, second homes, moves, repossessions, satisfaction with the accommodation and area, waiting lists for council or housing association housing, owner occupation, social sector tenants, and private renters (ONS, 2009b; ESDS, 2009c). Similarly to the Census, LFS and BHPS, the SEH collect little information on household consumption or expenditure.

4.1 Sample Size and Frequency

The sample size and the frequency of any survey are crucial indicators of its reliability and utility. Figure 12 demonstrates the average number of household responses received for each of the household surveys mentioned. In the context of household surveys, the SAM in fact has the greatest sample size with just short of 3,000,000 household responses. Nevertheless, because the SAM represents a 5% sample of individuals drawn from the 2001 Census (CSSR, 2010), it is only a one-off static measure in time. Therefore, because the SAM cannot be used for time-series analysis, it is excluded from Figure 12. The SAM aside, it is without question that the ROP generates a much greater response than any of the other national surveys. The ROP is delivered twice a year which totals to an annual sample of around 1,100,000 households. Additionally, as parts of the survey also capture information on both the household reference person and their partner, this increases the sample size for certain variables to over 2,000,000 individuals. The LFS has the next largest sample, with each quarterly wave based on 60,000 household responses covering 126,000 individuals. This

gives the LFS an annual household sample of about 240,000 households (Rees *et al.*, 2002; Blundell and Etheridge, 2009). In addition, the GHS, BHPS SHE and EFS all have similar samples sizes between 5,500 and 25,000 households (Dennett *et al.*, 2007). Moreover, like the LFS, the EFS is also run on a quarterly basis, providing a clear advantage over the ROP with regards to the greater potential it offers for more detailed time-series analysis of seasonal variations.

	ROP	LFS	EFS	SEH	GHS	BHPS
■ Survey	1100000	240000	26000	25000	10000	5500
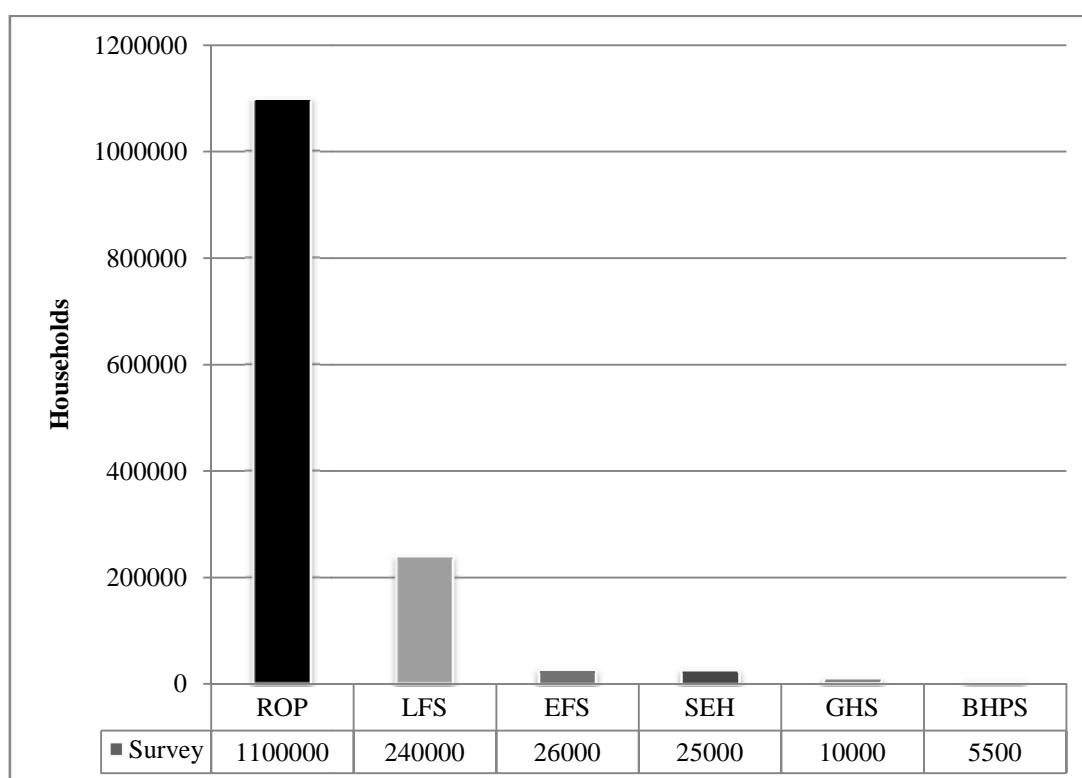
Figure 12. Annual sample size for continuous household surveys in GB

The survey sample is not only the total number of responses; demography and geography also have a part to play (Sorensen, 1996; Gibson *et al.*,1999; Deaton, 2000). From a demographic perspective, the surveys selected also offer quite different levels of detail. The ROP can only be answered by respondents who are a minimum of 18 years old. To answer the EFS survey, respondents only have to be 16 years old, but some parts of the survey also provide information on children between 7 and 15 years old (ESDS, 2009a). The LFS also includes 16+ year olds, but it has a cap of 65 years which means that socioeconomic data on the very elderly are not collected. The SAM provides the most comprehensive demographic coverage, as it includes information on the entire family (all ages). In the same way as the ROP, the BHPS is based only on adults. However, the BHPS differs in that it is a longitudinal survey

which follows the same representative sample every year. The GHS has also recently been changed to offer longitudinal data since 2006. The advantage of having longitudinal data is that one can *"...construct measures of change, for example in household structure, residential mobility, income, employment history and health measures"* (ESDS, 2009b).

In addition to demographic coverage, Table 7 provides information on the geographic coverage of the various datasets, along with the level of geographic detail assigned to each of the household responses. The LFS, EFS, SAM and BHPS cover the whole of the UK whereas the ROP and BHPS exclude Northern Ireland, and the SEH is run for England only. Additionally, when comparing the lowest level of geography assigned to each of the household respondents, the ROP comes out on top by a long way. The ROP household data are released at address level (no aggregation). As this is the lowest form of geographic detail, the ROP microdata are free from the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). Furthermore, the data can be aggregated up to any other set of geographic units (administrative or census). In comparison, the SAM and BHPS are both available at LAD level while all of the other continuous household surveys only provide household data at Government Office Region (GOR) level.

Table 7. Geographic coverage and level of geography for household surveys

Household Survey	ROP	LFS	SAM	EFS	GHS	SEH	BHPS
Geographic Coverage	*GB*	*UK*	*UK*	*UK*	*GB*	*England*	*UK*
Lowest level of Geography	*Address*	*GOR*	*LAD*	*GOR*	*GOR*	*County*	*LAD*

4.3 The Data

This final sub-section concentrates on providing a comparison between some of the core variables in the Acxiom datasets and those in other public household surveys. Figure 13 provides a group of bar graphs displaying the proportions of households in each survey according to household tenure, accommodation type, age of head of household, ethnicity, household income and car ownership within Yorkshire and Humber for 2006. The year 2006 was chosen as this was the most recent dataset available for the GHS. The 2001 Census proportions have also been included as a benchmark. Although the 2001 Census data are temporally inconsistent with the 2006 survey data, GORs such as Yorkshire and Humber are

likely to have remained reasonably consistent on a number of these core characteristics. Confidence intervals of 95% based on survey sample size have also been included (error bars) to help provide a measurement of reliability for the survey proportions. Furthermore, as a result of the differences discussed in section 4.2, some of the graphs in Figure 13 do not have a complete set of values for each of the surveys.

Figure 13(a) and (b) display the survey proportions for household tenure and accommodation type respectively. It is clear that for both variables similar proportions are found within the categories for each of the different datasets. Nevertheless, it could be argued that the EFS and Acxiom microdata have an overrepresentation of owner occupied houses and an underrepresentation of privately rented accommodation. In comparison, the Acxiom AD compares very well with the 2001 Census, a possible consequence of using data from the 2001 Census in the creation of the AD. This would not be a problem with the more recent datasets, having phased out its use in their methodologies for a number of years. Therefore, it is more likely that the EFS, BHPS and Acxiom microdata are just uncovering recorded changes in housing since 2001, such as the growth in home ownership (National Statistics, 2003). A similar situation can be seen in Figure 13(b), where the percentage of households living in flats across all the surveys appears to be lower than in the 2001 Census. This is arguably because it is much harder to gain access to and encourage people living in flats to answer the surveys. Conversely, the Acxiom AD has a closer fit with the 2001 Census data, re-emphasising that adjustments in the aggregation process make the data a much more representative sample at the regional level.

Figure 13(c) displays the age distribution of the household reference person (HRP) (headship) for each of the surveys. On the whole, the surveys are quite close with these proportions, although there are some notable differences. For example, the 2001 Census, EFS and GHS depict Yorkshire and Humber as having much smaller numbers of head of households aged 18-24 years old compared with the Acxiom datasets, the BHPS and SEH. This may be highlighting the increase in single person households and the large increase in accommodation provided by the region's universities since 2001 (ONS, 2010a). The confidence intervals suggest that based on the ROP sample size, the Acxiom data and SEH are more likely to be correct, as the other surveys have much wider error bars (greater uncertainty). As expected, on account of the ageing UK population (ONS, 2010b), the 65+ age band holds the greatest proportion of households across all the surveys.
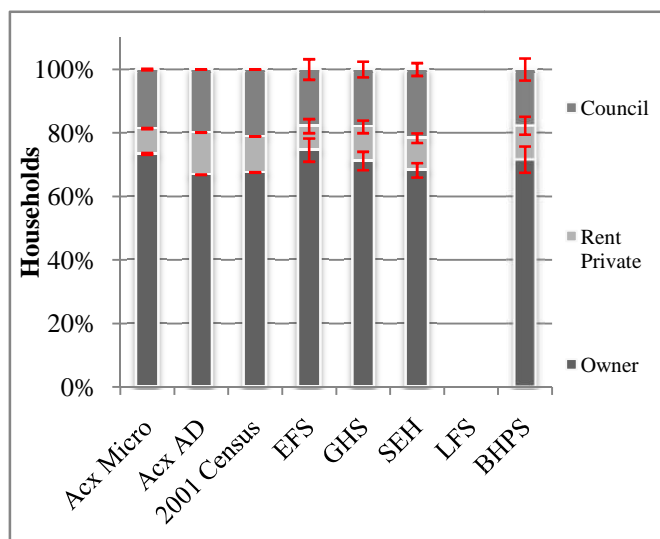
(a) Tenure
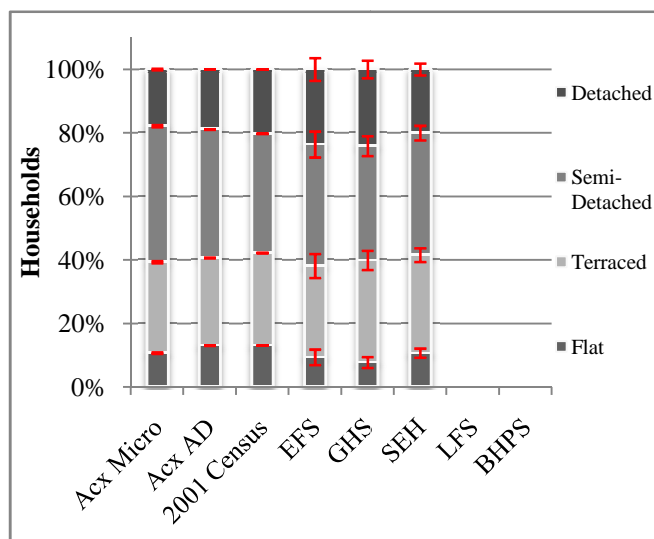
(b) Accommodation
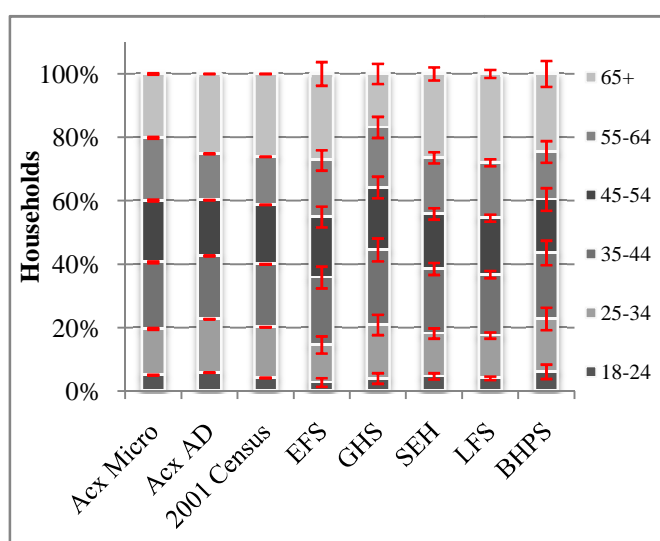
(c) HRP age
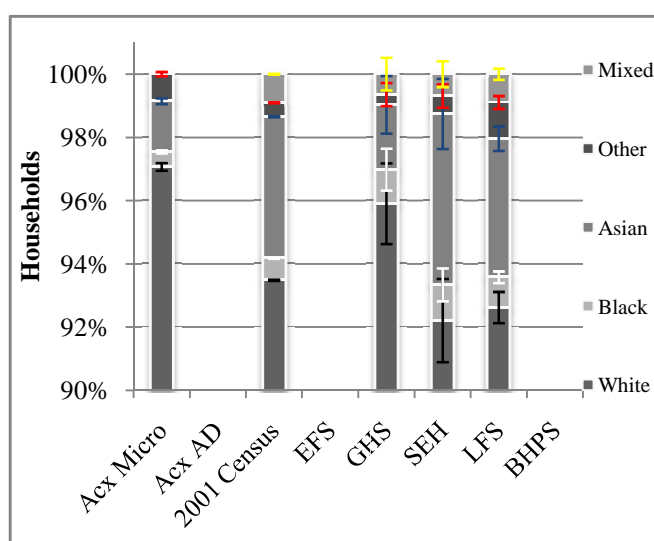
(d) HRP ethnicity
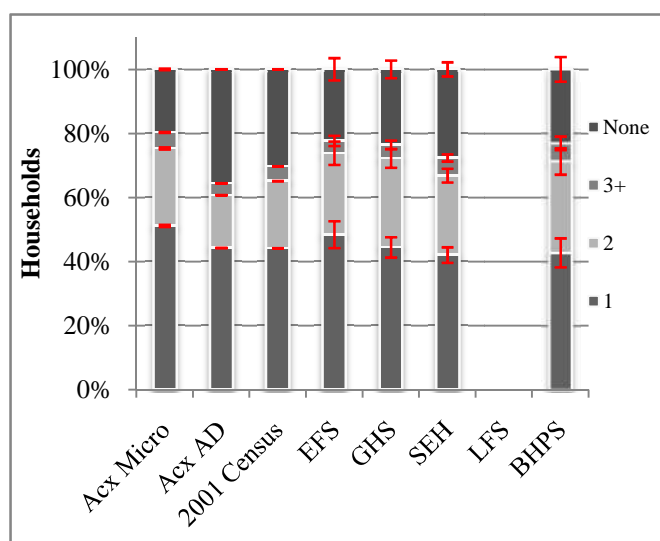
(e) Car ownership
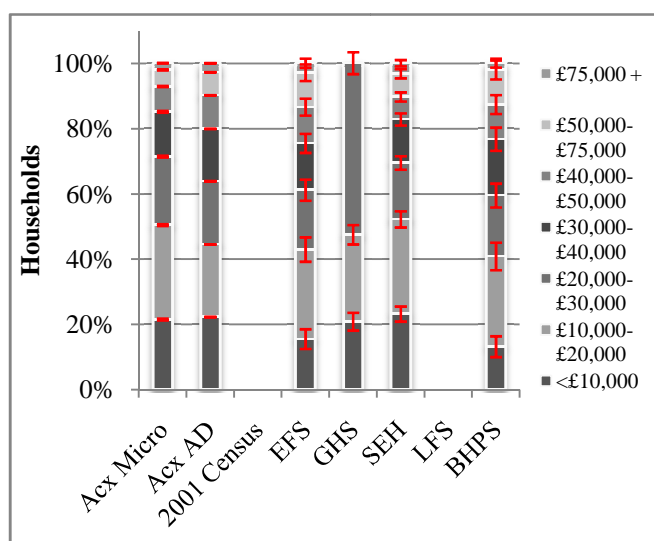
(f) Gross annual income

Figure 13. Comparison between Acxiom data from ROP, continuous national surveys and 2001 Census in 2006 for Yorkshire and Humber, 2006

It is evident from Figure 13(d) that the HRP ethnicity proportions have a more diverse pattern than any of the other core variables. The y axis on this graph has been altered to range from 90-100% to account for the overwhelming percentage of white people in Yorkshire and Humber, which makes the differences appear slightly exaggerated. Moreover, the confidence intervals are coloured differently to help distinguish between overlapping error bars. Unfortunately, the EFS, BHPS and Acxiom AD do not provide the ethnicity of the HRP so cannot be compared. However, as the ethnicity variable is included in the ROP, ethnicity proportions can be created on request from the client. It is clear that the SEH and LFS provide a similar representation to the 2001 Census. In comparison, the Acxiom microdata and GHS have much lower levels of Asian and Black respondents living in Yorkshire and Humber. This is surprising considering these two ethnic groups have witnessed the most growth since 2001, albeit small (ONS, 2006). Therefore, despite the smaller error bars in the ROP, because the LFS, SEH and GHS exemplify much closer proportions, one can assume that the ROP has a considerable bias towards white households in the sample. Ethnic minorities are much harder to engage in voluntary surveys on account of the language barrier and the fact they can be far more marginalised from mainstream society (Gibson *et al.,* 1999).

Figure 13(e) portrays the proportions for the number of cars owned by each household. As with Figures 13(a) and 13(b) the bar graphs indicate a high level of consistency between the different surveys. However, the data from the 2006 surveys would imply that there has been an increase in car ownership since the 2001 Census in Yorkshire and Humber. Figure 13(f) represents the proportion of households in each of the various annual household income bands. The 2001 Census did not ask an income question so no comparison can be made. Moreover, the GHS only offers a limited set of income bands for the respondent to choose from, hence the wider top band. This aside, the Acxiom datasets show good comparability with the EFS and SHE. Furthermore, it is evident when comparing the Acxiom microdata to the AD that there has been some adjustment to increase the number of households in the top income band. The top earning income group is the hardest to reach with these types of optional surveys (Gibson *et al.,* 1999); therefore it is not surprising that the proportions are lower than the BHPS and EFS. The income data are regarded as one of Acxiom's strongest variables. This is evidenced by the fact that Figure 14 displays geographic distributions consistent over a four year period (2006-2009). The South and South East have a much higher percentage of households earning £75,000+, along with parts of North Yorkshire and

Cheshire in the North West. This also reflects the consistency of the income data as it is clear that the spatial patterns have remained pretty stable since 2004.
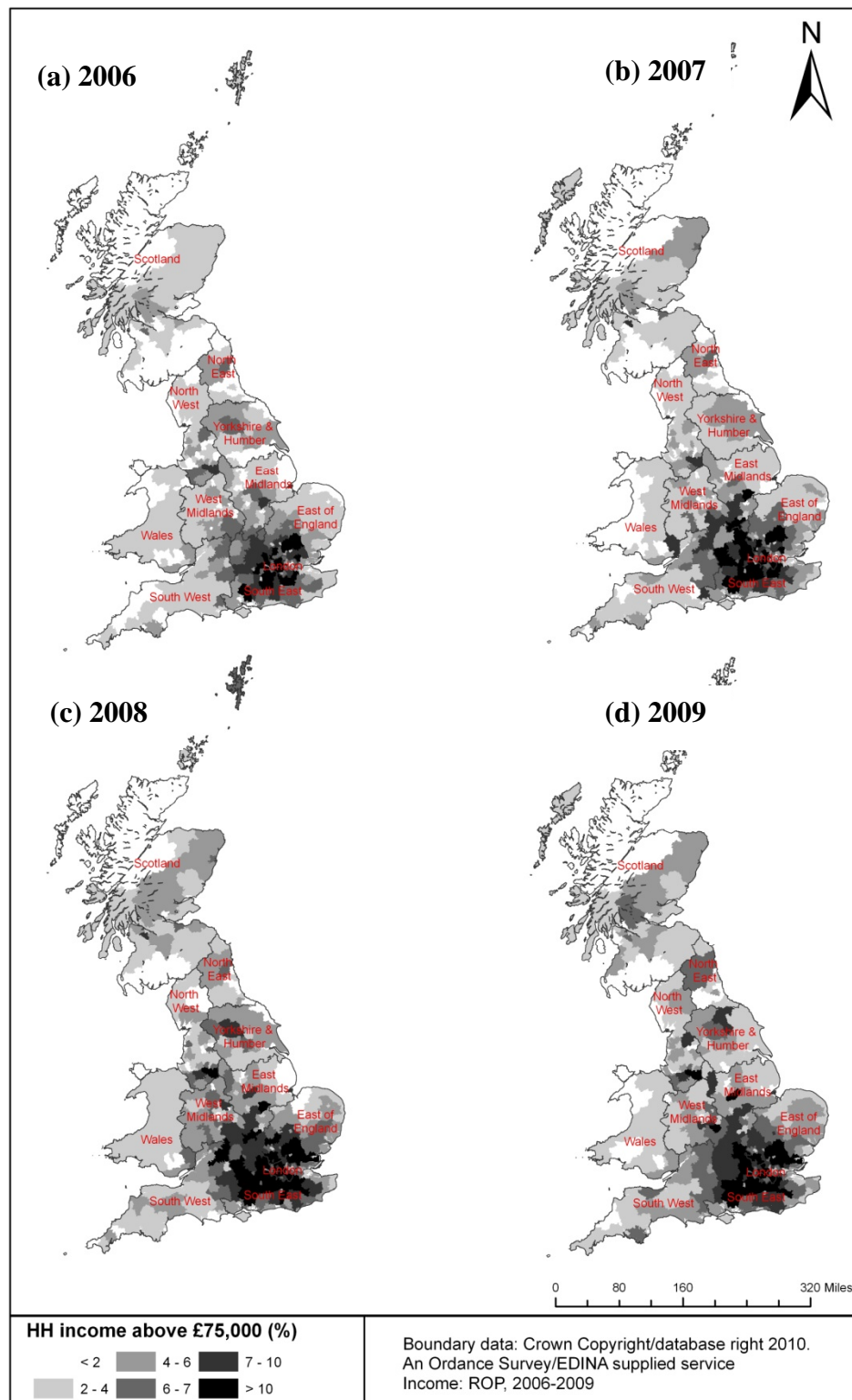
Figure 14. Percentage of households in each LAD with a combined annual income of £75,000

Figure 13 is useful for gaining an understanding about the variations in core demographic and housing variables; however it does not highlight variations in household consumption and expenditure. Consequently, Figure 15 provides the data for those households which have an internet connection by income group from the EFS (a) and ROP (b) respectively. The data are also displayed from 2005 to 2008 to highlight consistency issues within the datasets. It would have been possible to use 2009 and 2010 data from the ROP; however the latest EFS data available for this variable are from 2008. As with Figure 13, the data are based on households within the Yorkshire and Humber region. Furthermore, for both sets of data, confidence intervals have been calculated for the total percentage of households which have an internet connection. The year-on-year EFS data suggest that the total percentage of households with an internet connection has been steadily increasing from 2005 to 2008. Nevertheless, on viewing the breakdown by the various income groups, it would appear that some of the income groups have actually seen a decline in the percentage of households with an internet connection from 2007 to 2008. For instance, the both the £30-40,000 and £50,000+ income groups show declining rates since 2007. This is concerning, as one would have expected this high earning income group to have one of the highest levels of growth. The actual levels within the data are slightly more as expected. For instance, the general pattern suggests that as household income increases so does the likelihood of having an internet connection.

In comparison, the ROP data have produced much smaller error bars which promotes greater confidence in the data. In addition, all of the different income bands follow a similar level of growth to the total number of households in Yorkshire and Humber. This is encouraging as it highlights the year-on-year consistency within the ROP. Stable levels of growth at such high confidence intervals allow for more reliable conclusions. As with the EFS, there is a positive relationship between household income and internet connection. This is not surprising, because if a household has an internet connection, it would be reasonable to assume the household has greater disposable income to afford a computer. One major difference between the two surveys is the level of internet penetration between comparable income groups. On average, the ROP data has greater percentage of households with an internet connection. For example, from 2005 to 2008, the total percentages for the EFS ranged from 49% to 62%, whereas the ROP range from 57% to 78%. Based on the consistency seen in the ROP data and the high level of uncertainty in the EFS data, it seems more likely that the ROP values are a more accurate reflection of reality.
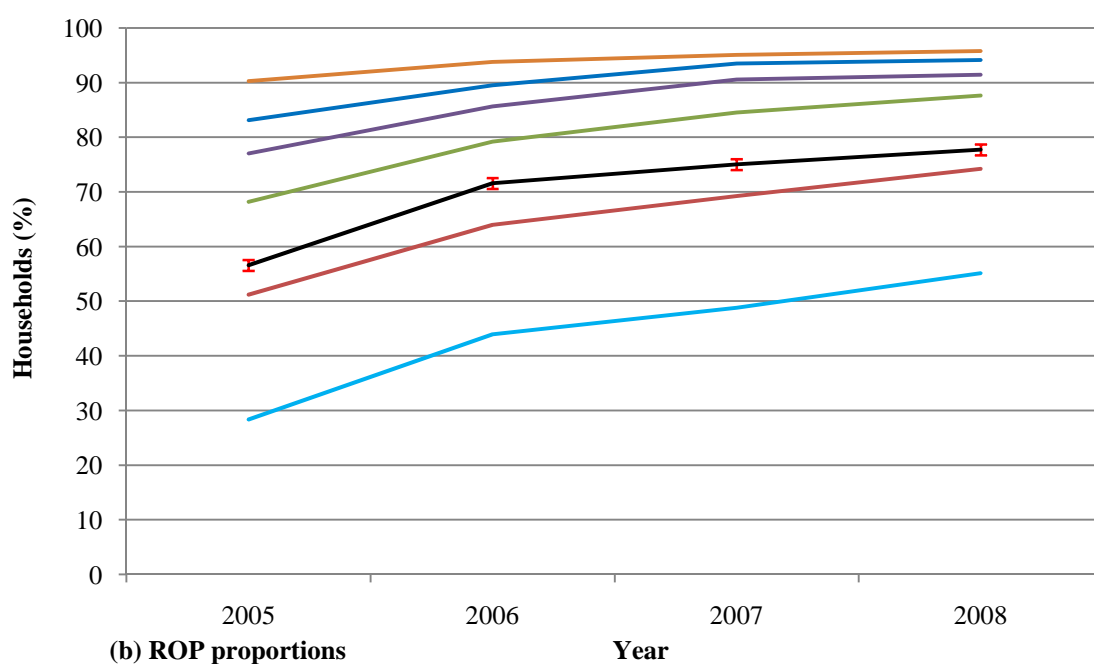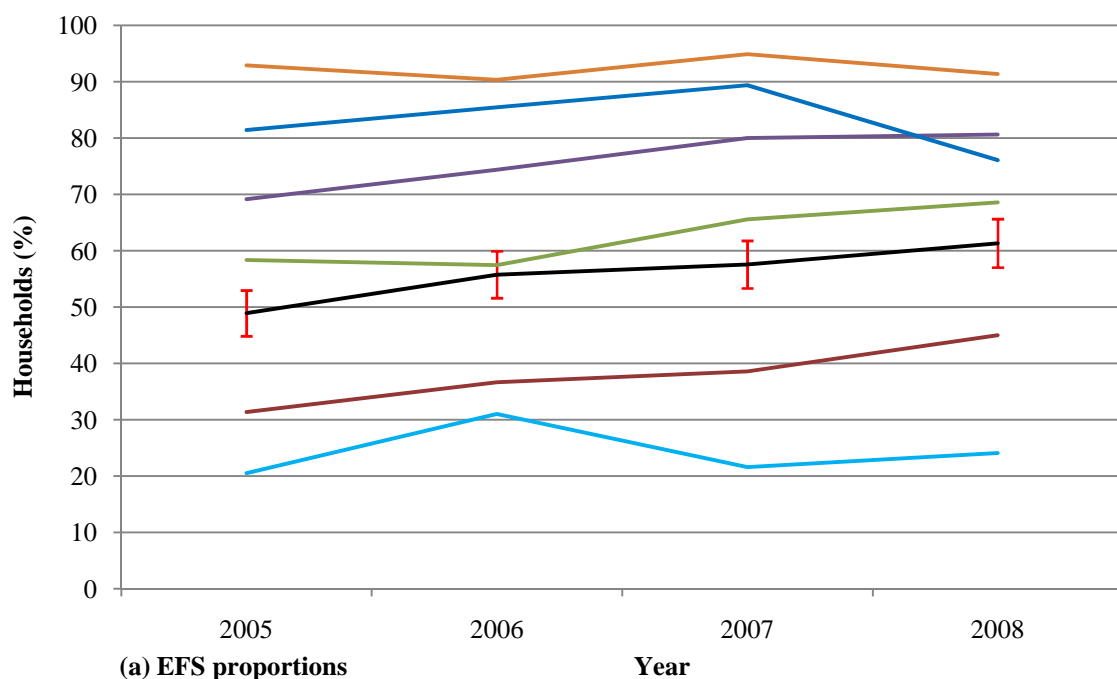
Figure 15. Households with an internet connection by income group, EFS and ROP, 2005-2008

5 Conclusions

The aim of this paper has been to provide a comprehensive review of Acxiom's ROP. Attempts were made to cover a number of the points raised by Stewart (1984) for analysing secondary data sources. Initially, attention was given to the production and delivery of the ROP across GB. This annual exercise represents an enormous task for Acxiom. The company has developed a highly stringent design process to ensure that it can maximise household response rates and increase geographic and demographic penetration. As a result, Acxiom is able to deliver the largest annual optional household survey (in numbers) outside of the census, which in comparison is only run once every 10 years.

Once the survey has been distributed across GB, Acxiom are able to achieve a response in excess of one million households. This figure dwarfs any of the other samples sizes gained in comparable surveys such as the EFS, LFS, GHS, SEH and BHPS. Consequently, because the sample size is so large, Acxiom has confidence in the accuracy and reliability of its sample products. Furthermore, Acxiom are able to secure a greater response rate from the younger (18-24 years) demographics by replicating the ROP online. This gives Acxiom a distinct competitive advantage over rival organisations as this demographic are notoriously difficult to gain responses from. In addition to the delivery of the ROP, considerations were given to the processing of the raw data and methodologies behind the building of Acxiom's data packages. In order to increase turnaround time and reduce the likelihood of human error, Acxiom aims to make every stage of the processing as efficient as possible. Therefore, the latest OMR technologies are employed to convert the paper-based surveys into actionable information, allowing Acxiom to deliver the data to clients within a couple of months of being collected.

On the back of the ROP, Acxiom is able to produce a number of data packages both at household (micro) and geographic (aggregate) levels. The microdata represents the raw responses from the ROP and provides an excellent source of information for interaction analysis, logistic regression and cross-sectional analysis. Additionally, the APE and AD are both argued to be fully representative of the total households in GB. A review of the methodologies behind the APE and AD were then provided to try and identify how the missing data are modelled. Simple linear regression charts were run to identify the level of influence the ROP has on the production of the AD. It was found that the AD does depend on

the microdata heavily, however on certain variables were sample sizes are smaller, Acxiom also utilise a wealth of official external data sources in the AD estimation process. However, further detail and clarity is required in this section as certain parts of the methodology behind the AD and APE remain unclear. This is primarily because Acxiom is reluctant to release full details of its methodologies for confidentiality reasons. This problem is not unique, as Sorenson *et al* (1996) recognise it as one of the major issues when using any secondary data source. In conjunction, Acxiom's geodemographic segmentation package, PersonnicX, was found to be streets ahead of its competitors. Acxiom integrate a number of the variables from the ROP which allow them to produce a behavioural consumer-based classification. Moreover, the level of geography available is down to the individual which is far more refined than the classifications that MOSAIC (Experian), ACORN (CACI), CAMEO (EuroDirect) and OAC (2001 Census) provide.

A comparative analysis was undertaken of the Acxiom datasets against the EFS, LFS, GHS, BHPS, SEH and the 2001 Census. Compared to the other continuous public surveys (excluding the SAM from the 2001 Census) the ROP sample size is much greater. However, the EFS and LFS were found to offer greater time-series capabilities for seasonal trends; the ROP is conducted twice a year, whereas the LFS and EFS are run quarterly. Additionally, from a temporal consistency point of view, concerns were raised over changes to the production of the EFS and sponsored questions being removed year-to-year from the ROP. Nevertheless, the large range of questions asked on the ROP allows for a greater variety in information gathered. For instance, the LFS is restricted to labour market statistics and the BHPS, SEH and GHS have limited questions on consumption and expenditure. The EFS does offer a wealth of information on consumer spending habits, although the small sample size and poor level of geographic detail restrict its analytical capabilities.

In addition, comparison of core socioeconomic variables between the selected sources indicated that both the Acxiom microdata and the Acxiom AD compare favourably with the other datasets. In particular, the Acxiom AD was found to sit well with the 2001 Census data on household tenure, accommodation and age of HRP. This is encouraging as it highlights the accuracy in the estimation Acxiom undertakes with the ROP data. Conversely, there are concerns with the reliability of the ethnicity variable, given that the ROP struggles to gain responses from ethnic minorities. Where the ROP and other surveys differed to the 2001 Census, confidence intervals were utilised to identify the reliability in the trends. Due to the gulf in sample sizes, the ROP was shown to have the lowest levels of potential error.

Comparisons were also made between the EFS and ROP on the consumption variable "has internet connection" by income group. The comparative analysis highlighted the problems with small sample sizes as the EFS data produced rather inconsistent patterns. Conversely, the ROP data demonstrated a degree of reliability reflected in the small error bars and stable year-on-year growth.

In conclusion, even with the shortcomings mentioned, there is no doubting that the ROP provides an excellent source of up-to-date information on consumer behaviour and expenditure patterns, with massive potential for use in academic research. Moreover, by helping to reshape our understanding of a wide range of human behaviours, the data has the potential to help formulate long-term policy decisions across a wide range of areas across the social sciences. On this basis, it is without question that commercial data sources such as Acxiom's ROP will become ever more apparent in social science research. In the past, official sources of secondary data such as government surveys have been considered to have greater dependability and credibility. However, even official government data has its issues, often presented in a way to support hidden agendas (Lancaster, 2005). Acxiom recognises the growing potential of the data, so in spite of increasing postage and raw materials costs, the company is committed to maintaining its extensive survey programme over the coming years. Furthermore, there are no other organisations which are currently able to provide the same level of consistency, volume, geographic detail and reliability in the data it collects.

Bibliography

Acxiom (2009) About Acxiom.
http://www.acxiom.co.uk/about_us/Pages/AboutAcxiom.aspx [Accessed: 21/10/2009].

Blaszczy´nski, J. (2006) Mining direct marketing data by ensembles of weak learners and rough set methods, pp. 218–227 (available online: www.cs.put.poznan.pl/wkotlowski/research/DAWAK06.pdf).

Blundell, R., Pashardes, P. and Weber, G. (1993) What do we learn about consumer demand patterns from microdata, *The American Economic Review*, 83, (3), pp.570-597.

Blundell, R. and Etheridge, B. (2009) Consumption, income and earnings inequality in the UK, Institute for Fiscal Studies and Department of Economics, pp.1-51. (Available online: www.econ.umn.edu/~fperri/papers/uk2.pdf)

Bradburn, N. (2004) Asking Questions: *The Definitive Guide to Questionnaire Design for Market Research, Political Polls and Social and Health Questionnaires*, San Francisco: John Wiley.

Brown, P. (1991) Exploring geodemographics. In Masser, I. and Blakemore, M. (eds), *Geographic Information Management: Methodology and Applications*, Longman, London, pp. 221-258.

CSSR (2010) Small Area Microdata.
http://www.ccsr.ac.uk/sars/2001/sam/ [Accessed: 21/04/2010].

Curtis, G. and Cobham, D. (2008) *Business Information systems; Analysis, Design and Practice*, Essex: Pearson Education Limited.

Deaton, A. (2000) The analysis of household surveys a microeconomic approach to development policy, London John Hopkins: University Press.

Dennett, A. Duke-Williams, O. and Stillwell, J. (2007) Interaction data sets in the UK: An audit, Working Paper 07/05, School of Geography, University of Leeds, Leeds. (Available online: http://www.geog.leeds.ac.uk/research/wpapers.html)

Dunn, K. (2005) Interviewing. In: Hay, L. (eds) *Qualitative Research Methods in Human Geography*, Melbourne: Oxford University Press.

Easaw, J. and Herav, S. (2009) Are household subjective forecasts of personal finances accurate and useful? A directional analysis of the British Household Survey Panel, *Journal of Forecasting*, 28, pp.667-680.

ESDS (2009a) Expenditure and Food Survey.
http://www.esds.ac.uk/government/efs

ESDS (2009b) General Household Survey.
http://www.esds.ac.uk/government/ghs/faq/ [Accessed: 02/01/2010].

ESDS (2009c) General Household Survey.
http://www.esds.ac.uk/government/seh/ [Accessed: 02/01/2010].

Fink, A. (1995) *How to Ask Survey Questions*, Thousand Oaks: SAGE Publications.

Fortin, N. (1995) Heterogeneity biases, distribution effects and aggregate consumption: An empirical analysis using stratified microdata, *Journal of Applied Econometrics*, 10, pp.287-311.

Frosztega, M. (2000) Income distribution data for Great Britain: robustness assessment report, Department of Social Security. (Available online: http://statistics.dwp.gov.uk/asd/hbai/frsrar2.pdf).

Gibson, J., Koepsall, T., Diehr, P. and Hale, C. (1999) Increasing response rates for mailed surveys of Medicaid clients and other low-income populations, *American Journal of Epidemiology*, 149 (11), pp.1057-1062.

Goodchild, M. and Gopal, S. (1989) *Accuracy of Spatial Databases,* USA: CRC Press.

Goss, J. (1995) We know who you are and we know where you live: the instrumental rationality of geodemographic systems, *Economic Geography*, 71 (2), pp.171-198.

Grundy, E. And Holt, G. (2001) The socioeconomic status of older adults: How should we measure it in studies of health inequalities?, *Epidemiol Community Health,* 55, pp.895–904.

Hay, I. (2005) *Qualitative Research Methods in Human Geography*, Melbourne: Oxford University Press.

Lancaster, G. (2005) *Research Methods in Management: a Concise Introduction to Research in Management and Business Consultancy*, Oxford: Elsevier.

Langley, P. (2008) Sub-prime mortgage lending: a cultural economy, *Economy and Society*, 37 (4), pp. 469-494.

Morse, J.M. (1994) *Critical Issues in Qualitative Research Methods*. Thousand Oaks: Sage Publications.

National Statistics (2001) National Food Survey http://www.statistics.gov.uk/ssd/surveys/national_food_survey.asp [Accessed: 02/01/2010]

National Statistics (2003) General Household Survey http://www.statistics.gov.uk/ssd/surveys/general_household_survey.asp [Accessed: 21/05/2010].

National Statistics (2009) Integrated Household Survey http://www.statistics.gov.uk/cci/nugget.asp?id=936 [Accessed: 21/10/2009].

Nesvetailova, A. and Palan, R. (2008) A very North Atlantic credit crunch: Geopolitical implications of the global liquidity crisis, Journa*l of International Affairs,* 62 (1), pp.165-185.

The National Data Strategy (2010) UK strategy for data resources for social and economic research 2009-2012, pp.1-3

Ongena, Y. And Dijkstra, W. (2009) Preventing mismatch answers in standardised survey interviews, *Quality and Quantity*, 20, pp.30-44.

Office for National Statistics (2006) Population Estimates by Ethnic Group http://www.statistics.gov.uk/statbase/product.asp?vlnk=14238 [Accessed: 21/05/2010].

Office for National Statistics (2009a) Expenditure and Food Survey http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361&More=Y [Accessed: 02/01/2010]

Office for National Statistics (2009b) Survey for English Households http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361&More=Y [Accessed: 02/01/2010]

Office for National Statistics (2010a) Single person households. http://www.statistics.gov.uk/cci/nugget.asp?id=350 [Accessed: 02/01/2010]

Office for National Statistics (2010b) Ageing across the UK. http://www.statistics.gov.uk/CCI/article.asp?ID=2418&Pos=5&ColRank=2&Rank=224 [Accessed: 02/01/2010]

Openshaw, S. (1984) *The Modifiable Areal Unit Problem*, Concepts and Techniques in Modern Geography 38, Norwich: Geobooks.

Rees, P., Martin, D. and Williamson, P. (2002) *The Census Data System*, Chichester: Wiley & Sons.

Stewart, D. (1984) *Secondary Research: Information Sources and Method*s, Beverly Hills: Sage.

Sorensen, H., Sabore, S. and Olsen, J. (1996) A framework for evaluation of secondary data sources for epidemiological research, *International Journal of Epidemiology*, pp.435-442.

Telegraph (2009) Acxiom: the company that knows if you own a cat or if you're right-handed.http://www.telegraph.co.uk/finance/newsbysector/retailandconsumer/5231752/Acxiom-the-company-that-knows-if-you-own-a-cat-or-if-youre-right-handed.html [Accessed Online: 22/11/2009].