



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/100334/>

Version: Published Version

---

**Monograph:**

Thompson, C, Stillwell, JCHS, Norman, PD et al. (2016) Exploring the utility of Acxiom's Research Opinion Poll data for use in social science research. Report.

<https://doi.org/10.13140/RG.2.1.3723.9922>

---

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Exploring the utility of Acxiom's Research Opinion Poll data for use in social science research

Chris Thompson, John Stillwell\*, Paul Norman, Martin Clarke  
*School of Geography, University of Leeds, Leeds, LS2 9JT, United Kingdom*

\* Corresponding author

Email: [j.c.h.stillwell@leeds.ac.uk](mailto:j.c.h.stillwell@leeds.ac.uk) Phone: 44 (0)113 343 3315 Fax: 44 (0)0113 343 3308

## Abstract

Acxiom's Research Opinion Poll (ROP), a voluntary survey designed to capture detailed information about household consumption and expenditure across Great Britain, has the potential to provide information valuable for social science research. This paper provides a review of the ROP, indicating that the survey is undertaken through a number of channels which enable Acxiom to generate over one million household responses a year. The ROP micro data collected are used in the construction of many of Acxiom's aggregate products including its geo-demographic classification system called 'PersonicX'. The ROP is found to compare favourably in areas such as sample size, geographic detail, consistency and data quality and accuracy when compared against government datasets including the 2001 Census, the Living Costs and Food Survey, the Labour Force Survey, the British Household Panel Survey, the General Lifestyle Survey and the English Housing Survey.

## Keywords

Acxiom, Research Opinion Poll, validation, microdata

## 1. Introduction

Technological advancements in data collection, data storage and the delivery of data have led to a considerable increase in the number of individual or micro-based secondary data sources, registers, databases, censuses and information systems that may be of value in social science. By definition, secondary data have already been gathered by someone other than the researcher and may not have been collected with a specific research purpose in mind (Sorenson *et al.*, 1996; Hakim, 1982). One important advantage of using secondary data sources is that they already exist and this clearly saves time spent collecting primary data. Furthermore, the costs of the project are reduced markedly, as is the waste of data, compared with collection of primary data, and the reduced likelihood of bias due to, for example, recall, non-response and the effects on the diagnostic process of attention caused by the research question (Sorenson *et al.*, 1996). In contrast, the disadvantages surrounding secondary data include limitations associated with variable selection as the survey may not cover all aspects of interest; the methods of collection are not under the control of the researcher and are often not transparent. Specific questions and categorisations of responses may not be ideal for a particular research setting. As a result, this often raises concerns over data utility and makes certain aspects impossible to validate.

The UK Census of Population, which has been delivered every ten years since 1801 (with the exception of 1941 due to World War II and 1966 when a 10% sample was taken) represents

the most comprehensive and reliable source of secondary data and has been widely used for spatial socio-demographic analysis. Historically, censuses have been the base for many of the population and socio-demographic statistics across the United Kingdom (UK), providing comparable information from the national to the local level on a range of topics, and acting as a benchmark for many other statistics. However, despite providing such a wealth of detailed information, the 2011 Census may well be the last to be administered across the UK if Cabinet Office Minister Francis Maude and the Conservative/Liberal Coalition Government decide to go through with their plans to scrap any future censuses in favour of alternative means of counting the population and collecting information about its composition. Reasons for abolishing the 2021 Census include the high costs of data collection (estimated at around £480million in 2011), a more mobile population and the increasingly complex ways in which people live make the process of taking a census more difficult, and the decadal nature of the census make the data collected less timely than would be ideal (ONS, 2011a).

Consequently, the Office for National Statistics (ONS) established the 'Beyond 2011' programme in April 2011 to take a fresh look at the alternatives to running a census in 2021. The Beyond 2011 programme (ONS, 2011a) involves a full consultation and assessment of alternative approaches in order to allow the UK Statistics Authority (UKSA) to make a recommendation to Parliament as to the best way forward in September 2014. As part of this process, there will be close collaboration with the devolved administrations in Scotland and Northern Ireland to ensure that the obligation to produce consistent UK statistics is met. The primary aim of the 'Beyond 2011' initiative will be to identify how the range of alternative data available, collected through various administrative or survey sources can be collated and used to provide detailed information about small areas and neighbourhoods that have traditionally been core outputs from the census (ONS, 2011a). The most widely used surveys which will become increasingly important in the absence of a census include the Integrated Household Survey (IHS) developed by the ONS which is comprised of the Living Costs and Food Survey (LCF), the Labour Force Survey (LFS), the General Lifestyle Survey (GLF), the English Housing Survey (EHS) and the Life Opportunities Survey (LOS). The LFS also forms part of the Annual Population Survey (APS) and the British Household Panel Survey (BHPS) now exists as part of a new longitudinal survey called Understanding Society (Buck, 2008). There is a range of administrative sources, on the other hand, from which aggregate or record level (micro) data can be produced that could be linked to produce annual or more frequent population counts. These sources include the Department of Work and Pensions' (DWP) Customer Information System (CIS), the Patient Registration Data System (PRDS) managed by the National Health Service (NHS), which holds records of all patients registered with General Practitioners (GPs), and ONS' address register which might be used either directly as a data source or as a frame for surveys.

A system that makes use of administrative sources to collate information already held about the population undoubtedly has the potential to provide a more cost-effective way to provide more frequent statistics, with reduced public burden. Nevertheless, so that this can be achieved, the 'Beyond 2011' programme will give consideration to private sector data. However, with regards to social science research, 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 sample 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, with shrinking public sector funds, over time may become increasingly useful for social science research. One company that openly advertises its data resources is Acxiom Ltd. A global leader in interactive multi-channel marketing services, 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 (Błaszczyszynski *et al.*, 2006). When the data collected through an array of sources are pooled together, the company's central database houses information on over 60 per cent of UK inhabitants including their geographic location, age, income, address, spending habits and various lifestyle choices. The main source of data which feeds this central database is the company's annual Research Opinion Poll (ROP) survey. Delivered every year across GB, the household survey, completed by an individual member of the household, provides the microdata that are the foundation for most of Acxiom's data packages and, in essence, what the company refer to as their 'holy grail'. The data represent a source of information which no other company or organisation can provide, and combined with the quick turnaround of the raw data into outputs, it means that Acxiom can provide a very sizable survey of the national population each year.

However, before such a dataset can be used for 'serious research' in an academic context, an important set of issues relating to authentication and validation must be confronted. Surprisingly, despite the comprehensive use of secondary data sources in social science research, the literature concerning validation is relatively modest. Stewart (1984) and Sorenson *et al.* (1996) make an attempt to address the main issues of importance regarding the use of a secondary data through a set of questions for which answers are required. Both papers emphasise the importance of knowing the limitations of the data being used, and stress that 'any data is better than no data' is not an adequate excuse for using poor data, and is even a less adequate reason for failing to identify and assess the impact of weakness. As there is no existing documentation which discusses the appropriateness of the ROP for use in social science research, this paper represents a completely original, independent and critical appraisal of the data. To ensure an inclusive discussion, a number of the questions and criteria set by Stewart (1984) and Sorenson *et al.* (1996) for analysing secondary data sources are combined with the criteria for assessing the statistical options of data from the ONS 'Beyond 2011' programme. Through doing this, we are able to form a framework whereby we can independently validate the data recorded via the ROP on factors such as the purpose of collection, the methodology, the frequency of collection, the geography, the content and accuracy of the data, and its credibility (ONS, 2011a; Sorenson *et al.*, 1996; Stewart, 1984).

To ensure all these points are covered in the paper, we devise a set of Rs to pose the necessary questions at every stage involved in delivering the ROP survey from start to finish. The Rs are defined as:

- research area;
- responsiveness;
- records within the survey;
- representativeness; and
- robustness.

Furthermore, where possible, to provide context, comparisons will be made with the 2001 Census and some of the more established national sample surveys used in social science research (LCF, LFS, GLF, BHPS and EHS). To begin, observations will be made on the ‘research area’, covering topics such as the survey purpose, the type of information collected and the survey’s time-series capabilities. Second, the ‘responsiveness’ of the ROP will be examined, paying specific attention to the sampling framework, survey delivery, geographic coverage and the sample bias. Third, consideration will be given about the ‘records’ within the data. More specifically, we will reflect upon the data format accessibility and cost. Next, the ‘representativeness’ and of the survey data will be discussed through a univariate and bivariate analysis of common variables used in social science surveys similar to the ROP. Finally, some analysis of the ‘robustness’ of the survey data will be given by using a logistic regression model to test for a common outcome variable (employment). To close, a summary of the main findings from the paper will be given in a concluding section and recommendations will also be provided on the suitability of Acxiom’s ROP data within social science research.

## **2. Research Area**

As part of the framework created to assess the validity of the ROP data, we first consider the research area of the survey. In doing so, we will reflect (after Stewart, 1984) on the credibility of Acxiom as the data owner, the purpose the survey, what information is actually collected and the level of consistency in the survey.

### **2.1 Credibility and Survey Purpose**

In terms of data credibility, it does not matter how good the credentials of the agency responsible for collecting the data are, there must always be a degree of healthy scepticism about both the reliability and the validity of the data (Stewart, 1984). Acxiom is recognised for being a world leader in data services and has been termed by John Meyer, a former company chief executive, as “...*the biggest company you have never heard of*” (The Telegraph, 2009). This is not surprising considering Acxiom’s unique selling point is built on the collection of large volumes of sensitive consumer data across a range of topics in a number of countries across the world. For collecting survey data in the UK, the company has 20 years of experience in the design and structure of the questionnaire each year. The Data Acquisition team within Acxiom has a remit to check the design and layout of all surveys, allowing Acxiom to test the responsiveness of particular factors on an annual basis, ensuring

the various components of the survey perform in an optimum manner. The various factors 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 type of paper used. This work is crucial to the whole process as the final survey must be one which will maximise the response rate and generate the most accurate results.

Acxiom is a profit-making organisation and therefore the purpose of the survey is to collect data that other organizations will want to purchase. Consequently, it is in the company's best interest to produce data to the highest degree of accuracy possible. The main aim of the survey 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. This allows clients that utilize the data to better understand and retain their existing customers, and locate new ones. Additionally, to guarantee that the survey is profitable, Acxiom 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. Sponsored questions are not ideal for time-series analysis since once the client stops paying for their inclusion on the survey, the questions are removed. (Having a set of core questions asked annually with additional questions on specific topics is also a feature of the more conventional social surveys though; such as the Health Survey for England). Nevertheless, the majority of the questions are devised by Acxiom and asked consistently so that continuity over time for key variables can be maintained. These core questions typically feed into the construction of Acxiom's products and appear on each survey because Acxiom is committed to providing data which will support time-series analysis. Therefore, all changes to the survey that may impact on time-series analysis are stringently reviewed so that the ROP can provide a unique source of data on demographic and socio-economic changes across GB.

## **2.2 What Information is Collected?**

In addition to the survey purpose, it is essential for the secondary data analyst to establish exactly what topics the survey covers (Stewart, 1984). Table 1 indicates the number of questions and sections in each survey between 2004 and 2010. The sections are listed in the order in which they appeared on the survey for each year. 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). It is evident from Table 1 that the ROP offers a large number of questions across a

range of different areas. For example, in 2010, the survey had 141 questions spread across 29 different sections.

**Table 1. ROP questionnaire structure, 2004-2010**

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; 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.
2010	141	29	Groceries; Shopping; Coffee; Hobbies; Home; Home Improvements; Insurance; Household; Outgoings; You and Your Family; Family Health; Financial Products; Charities; Occupation; Your Local Area; Internet; Telephone; Technology & TV; Financial Planning; Environment; Research; Animal Welfare; Leisure; Tobacco; Education; Skills; Cars; Newspapers; Shopping Vouchers.

Whilst Table 1 provides an insight into the type of information collected by the ROP, it is important to understand how the types of questions offered in the ROP differ from those offered in other household surveys. The ROP essentially gathers information on household spending habits. In this respect, the Living Costs and Food (LCF) survey is the most comparable as it contains a diary on household expenditure, income, composition, size, type and location (Fortin, 1995; Blundell *et al.*, 1999; ONS, 2009a). Other available household surveys do not offer as much in the way of recording consumption and expenditure. For instance, the British Household Panel Survey (BHPS) is restricted to questions relating to expenditure on durables, housing, demographics and income (Easaw and Herav, 2009; Blundell and Etheridge, 2009). The Labour Force Survey (LFS) provides detailed information on labour market characteristics such as participation, income, training and qualifications, but nothing on consumption or expenditure (Dennett *et al.*, 2007; Blundell and

Etheridge, 2009). The primary aim of the General Lifestyle Survey (GLF) 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 GLF 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 GLF include smoking, health and consumer durables. In conjunction, the English Housing Survey (EHS) 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, 2009b). Similar to the 2001 Census Small Area Microdata (SAM), LFS and BHPS, the EHS collects little information on household consumption or expenditure.

As a way to assess the suitability of the information collected through the ROP for more wider social science research (not just household consumption), we provide a list of key variables that would traditionally be used to describe the main attributes of a given population in Table 2. To ensure a comprehensive list was compiled, a combination of the primary variables selected for the ONS OAC classification by Vickers and Rees (2007) and the assessment of population and migration statistics by Raymer *et al.* (2012) are used. Additionally, on account of the wealth of expenditure information recorded in the ROP, a number of key expenditure variables are examined across the selected surveys as well. It is clear from Table 2 that the ROP performs weakest with the demographic variables. Whilst the main variables are collected (age, gender, etc.), there is a lack of information being gathered on the respondent's country of birth, religion and sexual identity. Their omission is probably due to the sensitive nature of having these questions on a voluntary survey; even the 2001 Census did not contain any questions on the last of these variables. In comparison, the surveys which make up the IHS contain a wealth of demographic variables, in particular the LFS which includes all of the major variables selected for comparison.

**Table 2. Common variables associated with social science research**

		Census 2001	IHS	IHS	IHS	IHS	US
	ROP	SAM)	GLF	LCF	LFS/APS	EHS	BHPS
<b>Demographic</b>							
Age/DOB	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nationality	<i>Partial</i>		Yes	Yes	Yes	Yes	
Country of birth			Yes	Yes	Yes	Yes	Yes
Year arrived in UK			Yes	Yes	Yes	Yes	
Ethnicity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Religion		Yes	Yes	Yes	Yes	Yes	Yes
Sexual identity			Yes	Yes	Yes	Yes	
First language	<i>Partial</i>				Yes		Yes
<b>Housing and household composition</b>							
Marital status	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Length of time at address	Yes		Yes	Yes	Yes	Yes	Yes
Previous address	Yes						Yes
Number of cars/vans	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Total number in household	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dependent children	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tenure	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of rooms		Yes	Yes	Yes	Yes	Yes	Yes
Number of bedrooms	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Type of accommodation	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Type of family unit	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Central heating		Yes	Yes				Yes
Internet connection	Yes			Yes			Yes
<b>Socioeconomic</b>							
LLTI and general health	<i>Partial</i>	Yes	Yes	Yes	Yes	Yes	Yes
Smoking	Yes		Yes	Yes	Yes	Yes	Yes
Qualifications	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Socio-economic class (NS-Sec)		Yes	Yes	Yes	Yes	Yes	
Drinking	Yes		Yes	Yes			Yes
Expenditure	Yes			Yes			Yes
Debt	Yes		<i>Partial</i>	<i>Partial</i>	<i>Partial</i>		<i>Partial</i>
Hobbies	Yes		Yes	Yes	<i>Partial</i>		<i>Partial</i>
Financial products	Yes		<i>Partial</i>	Yes			
Shopping channels	Yes			<i>Partial</i>			
Wellbeing and opinions	<i>Partial</i>						<i>Partial</i>
Charity contributions	Yes			<i>Partial</i>			
Holiday destination	Yes			Yes			Yes
Area classification	Yes			Yes			
Geography (LAD and below)	Yes	Yes			Yes		Yes
<b>Employment</b>							
Hours worked			Yes	Yes	Yes		Yes
Currently studying	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Economic activity	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupational group	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Income	Yes		Yes	Yes	Yes	Yes	Yes
Pension scheme membership	Yes		Yes	Yes	Yes	Yes	Yes
Benefit entitlement	Yes		Yes	Yes	Yes	Yes	Yes
Location of employment	Yes				Yes		

As the ROP primarily collects information at household level, it performs strongly on the selected variables for housing and household characteristics. More specifically, there are only two variables from other surveys which do not exist on the survey (number of rooms and central heating). The absence of a central heating variable is not uncommon as it is also not available across many of the other surveys. In addition, the ROP also includes information which the other surveys do not. For example, it records the respondent's previous address and whether the household has an Internet connection. Both these variables provide an element of added value to the ROP. Having information on the previous address of the respondent will allow for detailed insights into internal migration at a time when migration is a topical issue (Travis, 2011, Thomas *et al.*, 2012). Furthermore, with average weekly value of Internet retail sales in August 2011 rising to £536.5 million and making up approximately 9.6 per cent of total retail sales (ONS, 2011b), it will be important in social science research to know which households have an Internet connection. Moving onto the socioeconomic variables, the ROP once again performs strongly. The ROP provides all the main variables such as qualifications, smoking and health, whilst also offering information on expenditure, shopping channels used, holiday destinations, hobbies and debt. It is only really the LCF which can match the ROP in its range of socioeconomic indicators as the other surveys are limited in this area. The final section in Table 2 covers the employment variables which provide information on the country's labour force. Overall, all the surveys including the SAM provide most of the variables likely to be used for comparison. Unsurprisingly, the LFS has the most complete coverage. The SAM covers the least amount of variables as the 2001 Census did not record household income on the survey. Alternatively, the ROP has a good range of employment variables and even records the location of the respondent's place of work. In the context of journey to work analysis, this variable which would be extremely useful as it can be difficult to obtain this information from other sources.

### **2.3 Consistency**

In addition to the range and suitability of the questions asked on a survey, we must also consider the consistency and any substantial changes which may have occurred over time (Stewart, 1984). This is because, for time-series analysis, the consistency of the questions asked on any survey is crucial. The ROP has evolved substantially since the early 1990s as its commercial utility has increased. Furthermore, to ensure the ROP collects relevant and as up-to-date information as possible, new sections and questions are regularly introduced. 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). Nonetheless, Acxiom recognises that, commercially, it makes sense to have a consistent dataset and has 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, this can be alleviated by aggregating to higher levels of geography.

The ROP is not the only survey to have some consistency issues; the other surveys mentioned have also 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 which have changed over time. For instance, the Integrated Household Survey (IHS) developed by the ONS in 2008 integrates the LCF, LFS, GLF and the EHS. The aim of the IHS is to bring together a number of key questions asked across a range of social surveys conducted by ONS. This is achieved through a set of ‘core’ questions asked in the individual surveys which are then deposited on the IHS, while a number of ‘bolt on’ questions which are not included in the IHS are reported in the individual surveys (Raymer *et al.*, 2011). However, following the first reported results for 2009/10, some ‘core’ questions were dropped and it has been reported that the GLF will be phased out from autumn 2012. As Walthery (2011, p.3) comments, “...it is expected that the composition of the IHS will be flexible with some surveys leaving the IHS and others entering each year”. In conjunction, the LCF component of the IHS has evolved considerably since 2001. Originally, the LCF 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 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 EHS is comprised of the former surveys, the Survey for English Housing and the English House Condition Survey (EHCS). The BHPS also went through a major change in 2009 as it has now been replaced by a new longitudinal survey called Understanding Society (Buck, 2010). The GHS has also been changed to offer longitudinal data since 2006. Being longitudinal the data are based on the same sample every year, which means that one can construct measures of change, for example in household structure, residential mobility, income, employment history and health measures (ESDS, 2009c).

### **3. Responsiveness**

Various factors impact on the responsiveness of the ROP data. These includes factors such as the response rate (Sorenson *et al.*, 1996), the sampling framework used to collect the data and the sampling unit (Stewart, 1984). Knowledge of the sampling framework will help provide an indication of the extent to which the population sampled is likely to correspond to the ‘true’ population. The level of geographic detail and coverage will also be assessed, as this forms a major component of the ONS ‘Beyond 2011’ assessment (ONS, 2011a). Any potential biases in the sample will also be analysed so that potential researchers using the data can make allowances for their likely effects (Stewart, 1984).

#### **3.1 Survey Delivery**

The ROP is delivered in the form of a survey to households across GB, because when dealing with a large sample, the questionnaire is an indispensable tool when primary data are required about people, their behaviour, attitudes and opinions (Hay, 2005). Although the primary

sample unit is the household, the ROP also collects information about families and individuals. The survey is rolled out twice a year, initially in September and then in the following January. September and January are chosen specifically because extensive research by ROP found that the greatest response rate occurs in these months. During this time of year, respondents are more likely to complete the survey forms 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 months.

The survey is delivered through two channels. The main channel is direct mail which provides a controlled and reliable method to survey a large number of households (Bradburn, 2004). To ensure all parts of the country are surveyed, Acxiom use a variety of sources, with the national Postcode Address File (PAF) providing the main source for the sampling framework. The PAF overcomes problems of under-representation of specific subgroups because it samples addresses, not people, and, unlike the Electoral Register, does not depend upon self-registration (Raper *et al.*, 1986); it is also frequently updated and has a coverage of 28 million households in the UK (Royal Mail, 2012). It is restricted, however, to those addresses which receive fewer than 25 pieces of mail a day, which means it excludes some private residences that act as small businesses (Stewart, 1984). The second channel is the Internet, as the ROP is replicated online to reach respondents less willing to fill in paper-based surveys. The responses are also immediately digitised which heavily reduces the processing time. Despite the advantages of using two channels, there is an issue (although rare) of households responding more than once in a year via the paper and online survey. Therefore, Acxiom has technology in place which allows the company 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 and can check if it has duplicates in the same year.

During the collection process, Acxiom uses a number of techniques to entice more responses and improve the quality of the data. For instance, 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. In addition, the survey predominantly contains closed questions because these are easy for respondents to answer, to code and to standardise and the data lends itself easily to statistical analysis (Fink, 1995). Open-ended questions are generally 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). 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 means that personal information such as age, ethnicity, income and health are left to the end of the questionnaire. Respondents are also reassured that they do not have to provide answers to these more sensitive sections of the questionnaire. Incentives are also used as participants are offered the chance to receive both financial rewards and prizes upon completion of the questionnaire. However, this may also increase the number of false records as respondents rush through the survey just to have a chance of receiving a prize. As a result, Acxiom use the positioning and wording of certain questions to 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 identify whether or not any questions relating to the household’s online shopping habits from home have been ticked.

Once the ROP survey has been completed by a household, the form is returned via a free post envelope which comes with the survey. The return address is regional, which makes the survey appear more personalised to local areas and the responses can be housed at a number of collection points in different regions across GB. After waiting for a period of approximately eight weeks, all of the received surveys are sent off to a data processing company in Manila, Philippines. In the past, the responses were simply keyed into the computer and 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 and scan the survey questionnaires on the computer. OMR is commonly used when high-volume data entry is required (Curtis and Cobham, 2008). 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. 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 Accuracy and Degree of Completeness**

The sample size and the frequency of any survey are also crucial indicators of its reliability and utility. Thus, Figure 1 demonstrates the average number of household responses received for each of the household surveys mentioned. In the context of household surveys, the Small Area Microdata (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 (CCSR, 2010) for all countries of the UK, with 2.96 million cases, 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 1. The SAM aside, with an annual sample of around 1,100,000 households, the ROP is the largest annual survey in GB and the largest population study outside of the Census of Population. 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). The GHS, BHPS EHS and LCF all have similar samples sizes between 5,500 and 25,000 households (Dennett *et al.*, 2007). Like the LFS, the LCF is also run on a quarterly basis, providing an advantage over the ROP with regards to the potential offered for time-series analysis of seasonal variations.

### **Figure 1 about here**

In addition to the sample size, it is also important to know the demographic profile of any secondary data (Sorensen *et al.*, 1996; Deaton, 2000), since all surveys contain an inherent bias within the sample population. For example, with regard to the ROP, the Household Reference Person (HRP) that fills out the majority of the questions must be a minimum of 18 years old. However, there are questions which record the information of other members of the family, including children. In comparison, the HRP for the LCF only has to be 16 years old, but again, 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) as well as institutional populations. In the same way as the ROP, the BHPS is based only on adults.

As a way to identify any demographic bias in the ROP, Figure 2(a) portrays the age structure of all the respondents recorded in the January 2009 ROP survey, the percentage of respondents by age from the 2001 Census and the age ranges for 2009 ONS Mid-Year Estimates (MYEs). Overall the three datasets show a consistent trend of high proportions in young children, low proportions in young adults, high percentages in older adults and low percentages in the elderly. It is worth noting, however, that the percentages for each age group from the ROP fit closer to the 2009 mid-year estimates proportions than the 2001 Census data. This is encouraging, as it shows that structural changes (ageing population) occurring in the population are picked up in the ROP data (ONS, 2010). However, when compared to the 2009 MYEs, it is clear from Figure 2(a) that the ROP has an under-representation of people in the age groups below 40 years old and there is over-representation for the age groups between 50 and 75 years old. Additionally, Figure 2(b) exemplifies the level of bias within the sample by dividing the number of people in each age and gender category by the total sample. In Figure 2(b), it is evident that there is an over-representation of females in the sample, especially for ages between ages of 40 and 70 years. In comparison, men provide a smaller part of the sample, with the most difficult group of all to capture being young males aged 18 to 24 years. This is not unique to the ROP, as Frosztega (2000) recognises this group as traditionally the hardest to reach in sample surveys. Moreover, because the ROP is essentially a household survey, Figure 2(c) displays the HRP population by age and gender for the 2009 ROP and the 2001 Census. The results in Figure 2(c) are encouraging, as the population pyramids for the two datasets are relatively consistent. For instance the gender differences in Figure 2(b) are not as defined and the proportions for younger respondents are far more representative of the actual population. Nevertheless, there

is a noticeable non-response bias for the very elderly. Again, this is an issue documented as a widespread problem (Redpath, 1986; Holt and Elliott, 1991), but which rarely has a significant effect on analysis. On account of the varying levels of bias, those groups less likely to respond are over-sampled to try and increase the number of respondents through ‘door-drop campaigns’ and through the online ROP, which is useful for targeting younger age groups.

**Figure 2 about here**

In addition to age-gender bias, geography also represents an important facet of any secondary dataset (ONS, 2011a). Table 3 provides information on the geographic coverage of selected surveys for comparison, along with the level of geographic detail assigned and available for each of the household responses. The LFS, LCF, SAM and BHPS cover the whole of the UK whereas the ROP and BHPS exclude Northern Ireland, and the EHS is run for England only. When comparing the lowest level of geography assigned to each of the household respondents, the ROP comes out as superior by a long way. The ROP household data are captured at address level. 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 3. Geographic coverage and most detailed level of geography available for household surveys**

Household Survey	ROP	LFS	SAM	LCF	GHS	EHS	BHPS
Geographic Coverage	GB	UK	UK	UK	GB	England	UK
Lowest level of Geography	Address	GOR	LAD	GOR	GOR	LAD	LAD

**3.3 Missing Data**

Another important issue associated with the completeness of any data source is the existence of missing or blank fields in the data (non-response). 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 stated, sample surveys provide a biased representation of the total population unless. In some of the government surveys (e.g. BHPS, LCF, GLF, SEH, LFS) missing data are dealt with by using assigned weights to correct for the non-equal probability of selection of respondents, and differential response rates within the group of selected individuals/households. In terms of the ROP, no weighting is conducted at the individual level. This is because of the large sample it generates, which means that even small associations will give statistically significant results during analysis (Sorensen *et al.*, 1996). Instead, the blank fields are left for the end user to decide how best to interpret the missing information. Nevertheless, Acxiom’s aggregate products 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 PersoniX, Acxiom’s geodemographic segmentation profile. It is not possible to discuss the weighting process for this procedure remains confidential. This problem is not unique, as Sorenson *et al.* (1996) recognise it as one of the major issues when using any secondary data source. However, we can confirm that the weights are calculated using published UK statistics from the ONS such as the 2001 Census, MYEs and the LCF. Table 4 displays the level of non-response bias for selected variables. 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 per cent of households still disclosed their annual household income.

**Table 4. Blank fields for selected variables in ROP data for GB, 2009**

	<b>Respondents</b>	<b>Postcode</b>	<b>Age</b>	<b>Income</b>	<b>Accommodation</b>
<b>Blanks</b>	<i>n/a</i>	0	7,548	20,237	8,922
<b>Percentage</b>	<i>n/a</i>	0	8.35	22.40	9.87
<b>Total cases</b>	90,378	90,378	82,830	70,141	81,456

#### 4. Records

Due to the fact that survey data are secondary sources, there are a number of considerations and questions which must be addressed with regard to accessibility, confidentiality, the format of the data and the possibility of record linkage with other datasets (ONS, 2011a; Sorensen *et al.*, 1996).

##### 4.1 Accessibility and Confidentiality

Sorensen *et al.* (1996) recognise the importance of financial costs when using secondary data. Unfortunately, because Acxiom is a private organisation, its products are only available at a cost. However, as the company provides bespoke data packages, the company are flexible in terms of the cost. Furthermore, with regard to academic use there could be an opportunity to use the data for research purposes at little or no cost through an agreed partnership similar to the one with the School of Geography at the University of Leeds. This is the obvious drawback when comparing the ROP to the data collected through the various ONS surveys, as the data are available at no cost to the majority of academic institutions.

The confidentiality of a dataset is crucial when considering the appropriateness of a dataset for research, especially where information about the general public is concerned. People will be less likely to relinquish personal information if there is a worry that the data might be acquired by a third party. This is why public acceptability and risk form a fundamental component of the Beyond 2011 assessment. The Census Act and the ONS’ commitment in general make clear that details about any one individual are never divulged (ONS, 2011a). Therefore, to guarantee respondent information collected through the ROP survey is kept confidential, the data are kept under the highest levels of security at the data processing

centre in Normanton. Furthermore, extreme care is taken when the data are delivered to clients. The data can be accessed a number of ways, however due to the large size of the data files, File Transfer Protocol (FTP) is often the favoured method and is the standard network protocol used to transfer files from one host to another host over the Internet. To connect to the FTP site, a user name and password is given to the client beforehand. Once the data have been successfully downloaded a second password is required to access the folder containing the data. This type of online system whereby a username and password is required is common practice. For instance, census data and the components of IHS can be accessed through the Economic and Social Data Service (ESDS) hosted by the University of Essex. Parts of the LFS can also be downloaded through nomisweb and EuroStat.

#### **4.2 Data Format and Record Linkage**

The format of any dataset is an important factor to consider as the construction of any survey database is a socially negotiated exercise (Stewart, 1984). Survey records can be formatted or structured in such a way that their use is made difficult for research (Sorensen *et al.*, 1996). For example, the data files may not be compatible with certain software packages or the data might contain an inappropriate format of variables (age bands). The ROP microdata can be delivered in a range of formats to suit a variety of software packages (.dat, .txt, .sav, .csv and .xls). Figure 3 provides a sample of the microdata in SPSS. The 'postcode' and 'Ethnicity' codes are self-explanatory; however 'OWNRNT', 'RESTYPE' 'MARRYD' 'RDOB' and 'KIDAGE' refer to household tenure, residence type, marital status, date of birth, and age of children 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 and has no children.

#### **Figure 3 about here**

As stated, one of the reasons for exploring the use of data from administrative sources is because of the potential benefits that can be gained from record linkage (Sorenson *et al.*, 1996). The ONS Methodology Directorate (MD) has a team dedicated to working on record linkage methodology which has become involved in many record linkage projects. For example, there is a project to link the APS database to Individual Learner Record data (Heasman, 2008). However, because record linkage involves combining data on a respondent captured in multiple surveys through a common identifier (e.g. date of birth, address or National Insurance Number) it can be problematic. For instance, the recorded data must be standardised across datasets if it is to be matched, otherwise considerable cleaning of the data is required. Regarding the ROP, record linkage forms a key part of Acxiom's business model, for the data collected in the ROP is linked to data collected from an array of other sources. This is then pooled into one central database where each respondent is given a unique identifier. Record linkage is possible with the ROP because very detailed information is collected such as name, date of birth and a complete address. As using a person's name may breach confidentiality issues, using a combination of the address and date of birth may be more appropriate.

## 5. Representativeness and Robustness

This final section concentrates on providing a comparison between some of the core variables in the ROP against those in selected sample surveys, as it is relevant when analysing a secondary dataset to know the distribution of the data for key variables (Sorenson *et al.*, 1996). To ensure a comprehensive analysis of the data, comparisons will be given of both univariate and multivariate distributions. Additionally, logistic regression models with employment as the common outcome variable will be given to test for robustness in the data. It should be noted that whilst any time period could have been chosen for analysis, data from 2005 was selected so that realistic comparisons from the annual surveys could be made back to the SAM. Using the most recent data would have been problematic given the various demographic and socioeconomic changes over the last 10 years.

### 5.1 Univariate Analysis

Figure 4 contains a group of bar graphs displaying the proportions of households in each survey according to household tenure, marital status, ethnicity, gross annual income and government office regions within GB. Confidence intervals of 95 per cent based on survey sample size have also been included (error bars) to help provide a measurement of reliability for the survey proportions. First of all, it is clear from Figure 4(a) that similar proportions are found within the different tenure categories for each of the datasets. The only noticeable difference is what appears to be a slight overrepresentation of owner occupied households in the BHPS. However, this may in fact just be reflecting the growth in home ownership since 2001. Figure 4(b) displays the percentage of HRP's by marital status. Once again, all sources capture the same patterns in terms of the overall internal distribution. The Acxiom micro data, EFS, GHS, SHE and BHPS all have very similar figures. The LFS and the SAM exemplify slightly higher proportions for single HRP's. It is worth noting however that the confidence intervals associated with both Acxiom datasets (across all graphs in Figure 4) are much smaller than the other surveys on account of the large sample size. Only the SAM has smaller error bars.

It is evident from Figure 4(c) that the HRP ethnicity proportions have a more diverse pattern than any of the other core variables. The vertical axis on this graph has been altered to range from 70 to 100 per cent to account for the overwhelming percentage of white people in GB, which makes the differences appear slightly exaggerated. Moreover, the confidence intervals are coloured differently to help distinguish between overlapping error bars. Unfortunately, the BHPS and Acxiom AD (can be produced as a custom variable on request) do not provide the ethnicity of the HRP so cannot be compared. In comparison, the Acxiom microdata and GHS have much lower levels of Asian and Black respondents. This is surprising considering these two ethnic groups have witnessed the most growth since 2001, albeit small (ONS, 2006). One can assume that the ROP has a 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; Sheldon *et al.*, 2007). Nevertheless, when considering the size of the ROP sample, the absolute counts of non-white respondents are still much higher than other surveys.

Figure 4(d) 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. The Acxiom datasets show good comparability with the EFS, GHD and SEH. 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). Figure 4(f) shows the proportions within GORs and demonstrates a high level of consistency across all surveys apart from the BHPS (SEH not displayed as only for England). Once again, this is encouraging for using the Acxiom data as the geography of the British population also appears to be captured reliably within the data.

**Figure 4 about here**

## **5.2 Multivariate Analysis and Logistic Regression**

A means of assessing the utility of the ROP for social science research is to see whether in an example piece of research, a result emerges such that you would conclude a very similar answer. One of the advantages of microdata is the versatility to recode variables and here we harmonise as closely as possible a small number of socio-demographic variables across the ROP and other social surveys.

First, as an example, we cross-tabulate tenure with marital status (Table 5) to see how people in different personal circumstances are distributed across different ownership situations. Note that tenure is not identified in a consistent way across the surveys (particularly in the LFS) so some differences will emerge. There is however, a close correspondence for house owners by marital status such that a descriptive analysis of the pattern would provide the same conclusion with couples being the largest percentage followed by ‘others’ and then those who are single. Whilst there is more variation of percentages in the rental tenures there is also some consistency with other data sets. For private renters, the largest percentages are for those who are single; for council renters, similar percentages are apparent when comparing persons who are single and those in the other category with couples showing the lowest percentages. Note that the confidence intervals are narrower in the surveys which have the largest samples, the ROP and the SAM. Overall, whichever survey used, similar conclusions would be made about marital status and tenure though the reported percentages may vary and there are some definitional differences which affect results.

**Table 5. Cross-tabulation of tenure and marital status for GB, 2005**

Tenure	Survey	Marital Status		
		Single	Couple	Other
		<b>Owner</b>	ROP	51.06 (50.85,51.27)
	SAM	50.07 (49.96,50.17)	82.82 (82.73,82.90)	58.24 (58.13,58.34)
	EFS	47.27 (46.51,48.03)	81.45 (80.86,82.05)	60.07 (59.32,60.81)
	SEH	45.77 (44.92,46.63)	82.13 (81.47,82.78)	58.71 (57.86,59.55)
	LFS	51.69 (51.18,52.20)	85.17 (84.80,85.53)	60.77 (60.27,61.27)
<b>Private Rent</b>	ROP	24.92 (24.47,25.36)	8.48 (8.19,8.77)	15.95 (15.58,16.33)
	SAM	23.90 (23.68,24.12)	5.99 (5.87,6.12)	12.30 (12.13,12.47)
	EFS	22.54 (20.82,24.26)	7.09 (6.03,8.14)	8.62 (7.46,9.77)
	SEH	23.91 (21.96,25.86)	7.40 (6.20,8.59)	8.47 (7.20,9.75)
	LFS	46.92 (46.11,47.74)	13.94 (13.38,14.51)	37.59 (36.80,38.38)
<b>Council Rent</b>	ROP	24.01 (23.60,24.42)	9.89 (9.60,10.18)	21.49 (21.10,21.88)
	SAM	26.02 (25.84,26.19)	11.18 (11.05,11.30)	29.45 (29.27,29.64)
	EFS	30.17 (28.83,31.52)	11.44 (10.51,12.38)	31.30 (29.94,32.65)
	SEH	30.30 (28.77,31.83)	10.46 (9.44,11.48)	32.80 (31.24,34.37)
	LFS	1.37 (0.46,2.29)	0.88 (0.14,1.61)	1.63 (0.63,2.63)

Note: Data are percentages across marital status. 95% confidence intervals are in brackets

**Table 6. Modelled odds ratios of being employed in GB, 2005**

Independent variables		Survey				
Variable	Category	ROP	SAM	EFS	SEH	LFS
Age	16-24 (Reference)	1.00	1.00	1.00	1.00	1.00
	25-29	2.30	1.81	2.32	2.18	2.08
	30-39	2.41	1.91	2.49	2.52	1.65
	40-49	2.43	1.91	2.57	2.52	1.50
	50-59	1.29	<i>1.02</i>	<i>1.11</i>	1.31	0.64
	60-74	0.09	0.04	0.07	0.06	0.04
	75+	0.02	0.00	0.00	0.00	0.00
Marital Status	Single (Reference)	1.00	1.00	1.00	1.00	1.00
	Couple	2.01	2.57	2.34	2.59	2.08
	Other	0.93	0.82	<i>1.01</i>	<i>1.10</i>	<i>0.95</i>
Tenure	Council Rent (Reference)	1.00	1.00	1.00	1.00	1.00
	Private Rent	2.00	2.09	4.34	3.50	0.34
	Owner	4.35	7.48	10.29	8.74	2.16
Ethnicity	White (Reference)	1.00	1.00	1.00	1.00	1.00
	Black	0.96	0.87	<i>0.89</i>	<i>0.95</i>	<i>0.89</i>
	Asian	0.75	0.39	0.56	0.52	0.49
	Other	1.09	0.59	<i>1.04</i>	0.60	0.67

Note: Odds ratios in italics are not significantly different from the reference category

### 5.3 Logistic Regression

In a similar manner to Stillwell *et al.* (2010), we can build on the above using logistic regression to investigate the likelihood of employment for individuals controlling for age-group, marital status, tenure and ethnicity. With a binary outcome of employed/not employed, equivalent models were run using samples derived from the different surveys.

The outputs of the binary logistic regression models include the odds ratio of being employed for each category of a variable compared with a reference/base level. In Table 6, the reference level is the first category for each variable. An odds ratio greater than one means that persons of the particular category are more likely to be employed than the base level and *vice versa* for odds ratios of less than one.

In terms of age, compared with persons aged 16-24, the logistic regression model using the ROP shows increasing odds ratios up to age 40-49 and then the odds ratios decrease for age 50-59 but this group are still more likely to be employed than the base level. The oldest two age-groups are less likely than persons aged 16-24 to be employed. All the differences from the reference category are statistically significant. A very similar pattern is evident for the odds ratios derived from the other surveys although in the SAM and EFS the differences between those aged 50-59 and the base category are not significant and in the LFS persons of this age-group are less likely to be employed than those aged 16-24.

Across all the surveys, persons who are in couples are significantly more likely to be employed than persons who are single. In the ROP and SAM, persons in the 'other' category of marital status are significantly less likely to be employed than those who are single. For the EFS, SHE and LFS, the odds ratios for the other group are not significantly different to the single category.

The pattern for tenure is consistent across the surveys with those in private rental property and who are home owners progressively more likely to be employed than persons living in council rented property. The exception is the LFS which is the survey in which the tenure variable is recorded differently.

Compared with the White ethnic group, the Black and the Asian groups are less likely to be employed. The pattern is slightly different for the Other ethnic group with the ROP and EFS showing that this group are slightly more likely to be employed than the White group and the SAM, SHE and LFS having this group slightly less likely. It should be noted that the differences by ethnic group are not necessarily significant since the surveys do not necessarily have large numbers of non-White ethnic groups in their samples.

We can observe from these logistic regression models that, in the main, very similar conclusions can be drawn regarding the relationship between age-group, marital status, tenure

and ethnic group and the likelihood of being employed. Minor differences in results will relate to sample size and to survey purpose and coverage.

## **6. Conclusions**

The aim of this paper has been to provide a comprehensive review of Acxiom's ROP to assess whether it is fit for purpose in academic research. In order to achieve this, a framework was devised by combining the criteria set by Stewart (1984) and Sorenson *et al.* (1996) for analysing secondary data sources with the standards for assessing the statistical options of data from the ONS 'Beyond 2011' programme. In doing so, we believe this paper forms an independent validation of the data recorded via the ROP on factors such as the research area, responsiveness, records within the survey, representativeness and robustness.

Initially, attention was given to the research area of the ROP. Given the nature of the lifestyle questions on consumption and expenditure, it was found that the ROP survey is closely related to the LCF survey. However, as it also collects data on a number of 'core' variables, the 2001 Census and other administrative sources compared favourably with regards to subject matter. Furthermore, whilst some issues were raised over the inclusion of 'sponsored questions', the commitment by Acxiom to ensure compatibility over time means the majority of the ROP data can be used for time-series analysis. Additionally, to ensure the ROP collects relevant and as up-to-date information as possible, new sections and questions are regularly introduced. In comparison, some of the other surveys, noticeably the LCF survey have undergone considerable structural changes, thus, highlighting that the issue of consistency is not solely an issue for commercial data sources only.

A discussion of the production and delivery of the ROP across GB has demonstrated how Acxiom has developed a highly stringent design process to ensure that it can maximise household response rates and increase geographic and demographic penetration. For example, as the survey data are recorded at postcode level, the ROP has a greater level of detail than any of the other sample surveys mentioned. Furthermore, by combining paper-based surveys with an online version, the company 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. Nevertheless, despite Acxiom's best efforts, analysis found that when compared to the 2009 MYEs, the ROP has a slight under-representation of people in the age groups below 40 years old and an over-representation for the age groups between 50 and 75 years old, especially females. Still, this was argued to be a widespread issue regarding sample surveys, and one which rarely has a significant effect on analysis.

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 represent the raw responses from the ROP and provide an excellent source of data and primary focus for this paper. However, in the final section on representativeness and robustness, both the Acxiom microdata and Acxiom AD compared positively with official datasets. In particular, the Acxiom AD was found to sit well with the 2001 SAM data on household tenure, marital status of the HRP and the location of respondents by GOR - highlighting the accuracy in the

estimation Acxiom undertakes with the ROP data. Conversely, there were concerns with the reliability of the ethnicity variable, given that the ROP struggles to gain responses from ethnic minorities (again not uncommon). Where the ROP and other surveys differed to the 2001 SAM, 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. Furthermore, logistic regression models exemplified that overall, consistent conclusions can be drawn regarding the relationship between selected independent 'core variables' and the likelihood of being employed.

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, and with the support of the Beyond 2011 programme, 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.

### **Acknowledgements**

The first author is a PhD student funded by the Economic and Social Research Council in partnership with Acxiom Ltd. The authors are grateful to Acxiom for the provision of micro and aggregated data derived from their annual ROP.

### **References**

- Blaszczynski, 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](http://www.cs.put.poznan.pl/wkotlowski/research/DAWAK06.pdf)).
- 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](http://www.econ.umn.edu/~fperri/papers/uk2.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), 570-597.
- Bradburn, N., 2004. *Asking Questions: The Definitive Guide to Questionnaire Design for Market Research, Political Polls and Social and Health Questionnaires*. John Wiley, San Francisco.
- Buck, N., 2008. Understanding Society: The UK Household Longitudinal Study, In Praise of Panel Surveys. ESRC, Swindon. pp.24-25.

- CCSR, 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*. Pearson Education Limited, Essex.
- Deaton, A., 2000. *The Analysis of Household Surveys A Microeconomic Approach to Development Policy*. John Hopkins University Press, London.
- 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*. Oxford University Press, Melbourne.
- 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, 667-680.
- ESDS, 2009a. Expenditure and Food Survey. <http://www.esds.ac.uk/government/efs>
- ESDS, 2009b. Survey English Housing. <http://www.esds.ac.uk/government/seh/> [Accessed: 02/01/2010].
- ESDS, 2009c. General Household Survey. <http://www.esds.ac.uk/government/ghs/faq/> [Accessed: 02/01/2010].
- Fink, A., 1995. *How to Ask Survey Questions*. SAGE Publications, Thousand Oaks.
- Fortin, N., 1995. Heterogeneity biases, distribution effects and aggregate consumption: An empirical analysis using stratified microdata. *Journal of Applied Econometrics* 10, 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), 1057-1062.
- Hakim, C., 1982. *Secondary Analysis in Social Research: A Guide to Data Sources and Methods with Examples*. Allen and Unwin, London.
- Hay, I., (Ed.). 2005. *Qualitative Research Methods in Human Geography*. Oxford University Press, Melbourne.
- Heasman, D., 2008. *Record Linkage Development in the Office for National Statistics*, ONS, London.
- Holt, D. and Elliot, D., 1991. Methods of weighting for unit nonresponse. *The Statistician* 40, 333-342.
- Lancaster, G. 2005. *Research Methods in Management: a Concise Introduction to Research in Management and Business Consultancy*, Elsevier, Oxford.
- Langley, P., 2008. Sub-prime mortgage lending: a cultural economy, *Economy and Society*, 37 (4), 469-494.

- Morse, J.M., 1994. *Critical Issues in Qualitative Research Methods*. Sage Publications. Thousand Oaks.
- National Statistics, 2003. *General Household Survey*  
[http://www.statistics.gov.uk/ssd/surveys/general\\_household\\_survey.asp](http://www.statistics.gov.uk/ssd/surveys/general_household_survey.asp) [Accessed: 21/05/2010].
- Nesvetailova, A. and Palan, R., 2008. A very North Atlantic credit crunch: Geopolitical implications of the global liquidity crisis, *Journal of International Affairs*, 62 (1), 165-185.
- Norman P, Marshall A, Thompson C, Williamson L & Rees P. (2011) Estimating detailed distributions from grouped sociodemographic data: 'get me started in' curve fitting using nonlinear regression. *Journal of Population Research* 29(2): 173-198 DOI: 10.1007/s12546-012-9082-9
- Ongena, Y. and Dijkstra, W., 2009. Preventing mismatch answers in standardised survey interviews. *Quality and Quantity* 20, 30-44.
- ONS, 2006. *Population Estimates by Ethnic Group*  
<http://www.statistics.gov.uk/statbase/product.asp?vlnk=14238> [Accessed: 21/05/2010].
- ONS, 2009a. *Expenditure and Food Survey*  
<http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361&More=Y> [Accessed: 02/01/2010]
- ONS, 2010. *Ageing across the UK*.  
<http://www.statistics.gov.uk/CCI/article.asp?ID=2418&Pos=5&ColRank=2&Rank=224>  
 [Accessed: 02/01/2010]
- ONS, 2011a. *Beyond 2011* <http://www.ons.gov.uk/ons/about-ons/what-we-do/programmes--projects/beyond-2011/index.html>
- ONS, 2011b. *Retail Sales* <http://www.ons.gov.uk/ons/publications>
- ONS. 2009b. *Survey for English Households*  
<http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=361&More=Y> [Accessed: 02/01/2010]
- Openshaw, S., 1984. *The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography* 38. Geo Books, Norwich.
- Raper, J., Rhind, D. and Shepherd (1992) *Postcodes: The New Geography*, Longman, Harlow.
- Raymer, J. Rees, P. and Blake, A., 2012. *Conceptual Framework for UK Population and Migration Statistics. Report for ONS*. <http://www.ons.gov.uk/ons/guide-method/method-quality/imps/latest-news/conceptual-framework/index.html>
- Redpath, R., 1986. *Family Expenditure Survey: a second study of differential response, comparing Census characteristics of FES respondents and nonrespondents*. *Statistical News* 72, 13-16.
- Rees, P., Martin, D. and Williamson, P., 2002. *The Census Data System*. Wiley & Sons, Chichester.

- Royal Mail, 2012. The Postcode Address File [http://www.royalmail.com/marketing-services/address-management-unit/address-data-products/postcode-address-file-paf?campaignid=paf\\_redirect](http://www.royalmail.com/marketing-services/address-management-unit/address-data-products/postcode-address-file-paf?campaignid=paf_redirect) [Accessed 02/03/2012].
- Sheldon, H., Graham, C., Potheary, N. and Rasul, F., 2007. Increasing response rates amongst black and minority ethnic and seldom heard groups. Picker Institute, pp.1-75.
- Sorensen, H., Sabore, S. and Olsen, J., 1996. A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology*, 435-442.
- Stewart, D., 1984. *Secondary Research: Information Sources and Methods*. Sage, Beverly Hills.
- Stillwell, J., Norman, P., Thomas, C. and Surridge, P., 2010. Spatial and social disparities. In: Stillwell, J., Norman, P., Thomas, C. and Surridge, P. (Eds.) *Understanding Population Trends and Processes Volume 2: Spatial and Social Disparities*. Springer, Dordrecht, pp.1-15.
- The 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/Axiom-the-company-that-knows-if-you-own-a-cat-or-if-youre-right-handed.html> [Accessed Online: 22/11/2009].
- Thomas, C., Gould, M. and Stillwell, J., 2012. Exploring the potential of microdata from a large commercial survey for the analysis of demographic and lifestyle characteristics of internal migration in Great Britain. *Working Paper 12/3*, School of Geography, University of Leeds, Leeds.
- Travis, A., 2001. UK net migration hits record high. *The Guardian*, 24 November.
- Vickers, D. and Rees, P., 2007. Creating the National Statistics 2001 Output Area Classification, *Journal of the Royal Statistical Society, Series A* 170 (2), 379-403.
- Walthery, P., 2011. *Introductory Guide to the Integrated Household Survey*, ESDS Government.