ADDResponse: Auxiliary Data Driven nonResponse Bias Analysis

Technical report on appending geocoded auxiliary data to Round 6 of European Social Survey (UK)

Sarah Butt[1] and Kaisa Lahtinen[2]

[1] City, University of London
[2] Formerly City, University of London

Contents

1 Introduction

There is growing interest in the potential offered by geocoded auxiliary data to enhance survey data, including the so-called multi-level multi-source (ML-MS) approach where survey data are combined with auxiliary data from a range of sources and at different levels of aggregation (Smith, 2011; Smith and Kim, 2013). Combining auxiliary and survey data can provide valuable insights both for methodological purposes (e.g. to identify sub-groups for sampling purposes or to study patterns of survey nonresponse) and for substantive analysis (e.g. to understand the effects of neighbourhood context or the built environment on attitudes and behaviour). However, utilising auxiliary data also presents a number of challenges regarding the accessibility and quality of available data. This report summaries the results of a data scoping exercise conducted in the UK to evaluate the possibilities for appending geocoded auxiliary data from a range of sources to the sample of addresses selected to participate in Round 6 of the European Social Survey (2012/13). The exercise was conducted as part of a wider research project, funded by the ESRC, exploring the potential of auxiliary data to identify and correct for possible sources of nonresponse bias in social surveys. As Smith notes, however, the potential value of combining auxiliary data with survey data is not limited to research into survey nonresponse.

This report provides technical documentation for the specific data gathering exercise conducted as part of the ADDResponse project, giving details of all of the datasets used and the variables derived from them. In addition, the report is intended as a reference guide for researchers who may be considering conducting a similar exercise as part of their own work. By documenting the different data sources used during this project and the issues encountered with them this report should provide a valuable source of information for anyone interested in making use of UK auxiliary data either for methodological or substantive analysis.

1.1 Aims of ADDResponse

The ADDResponse project appended geocoded auxiliary data from a range of different sources to UK data from Round 6 of the European Social Survey (ESS), a methodologically rigorous general population survey of public attitudes and opinion (www.europeansocialsurvey.org). The purpose of the data linkage was to explore survey nonresponse – acknowledged as a considerable and increasing problem for most general population surveys, with response rates reaching just 50%, or even less, in some cases (de Leeuw and de Heer, 2002; Massey and Tourangeau, 2013). As non-respondents may be very different from respondents, nonresponse can introduce significant bias into the conclusions drawn from survey data and there is a pressing need therefore to understand more about the extent and sources of nonresponse bias. To fully understand and address nonresponse bias requires information to be available for the full target sample, both respondents and non-respondents. In the absence of interview data being available for non-respondents, this information must be obtained from other, external, sources (Sarndal and Lundstrom, 2005; Groves, 2006). One possibility is paradata collected for all survey units as part of the survey process,

supplemented by interviewer observations (Kreuter, 2013). Another option is to exploit the large and growing amount of information that is available via pre-existing external databases (Smith and Kim, 2013).

The ESS has an established track-record of innovative research into nonresponse (Stoop et al, 2010) but has to date used pre-existing contextual data only to derive population-based post-stratification weights based on age, gender, education and region (Vehovar et al, 2013). Analysis of nonresponse bias has relied primarily on paradata e.g. call record information collected during the survey process (Billiet et al, 2007) or auxiliary information from interviewer observations (Stoop et al, 2010). The ADDResponse project provided a UK test case for using auxiliary data from external sources for nonresponse analysis.

The project had three objectives:

- To explore the opportunities that exist for matching auxiliary data from three different sources – small-area administrative data, commercial marketing data, and geocoded information from the Ordnance Survey on the physical location of sampled addresses – to data from the European Social Survey (ESS) in the UK. Each data source was evaluated in terms of:
    - What information is available which may be linked to the survey records of respondents and non-respondents;
    - Whether accurate data are available for all sampled households in the UK;

- To investigate the information that auxiliary data can provide about potential biases present in the survey data as a result of nonresponse and how these biases might vary geographically across the UK. This involved consideration of two distinct but related questions:
    - To what extent can auxiliary data be used to predict sampled households' propensity to respond?
    - How are auxiliary variables shown to predict nonresponse related to the attitudes and behaviours the survey is intended to measure?

- To assess whether weighting using auxiliary variables to adjust for the possible over or under representation of certain types of respondent in the final survey:
        - has a significant effect on survey estimates for different variables
        - leads to a reduction in nonresponse bias

Further information about the project can be found on the project website www.addresponse.org.

This report focuses on findings from the data scoping exercise conducted under the first objective.

1.2 Overview of data sources used

ADDResponse appended auxiliary data to the sample of 4,520 addresses selected from the Royal Mail Postcode Address File to participate in Round 6 of the European Social Survey in the UK. (See Chapter 2 for more details on the ESS data). The selection of auxiliary data sources was driven by two main considerations: First, in the absence of consent for data linkage from ESS target respondents (obtaining consent from non-respondents, which includes noncontacts, is not possible), all data sources needed to be publically available without the need for permissions to link. Second,

given that the ESS in the UK uses an address-based sample drawn from the postcode address files, datasets needed to include geocodes (grid references, Output Area codes etc.) to enable them to be appended to the ESS sample file.

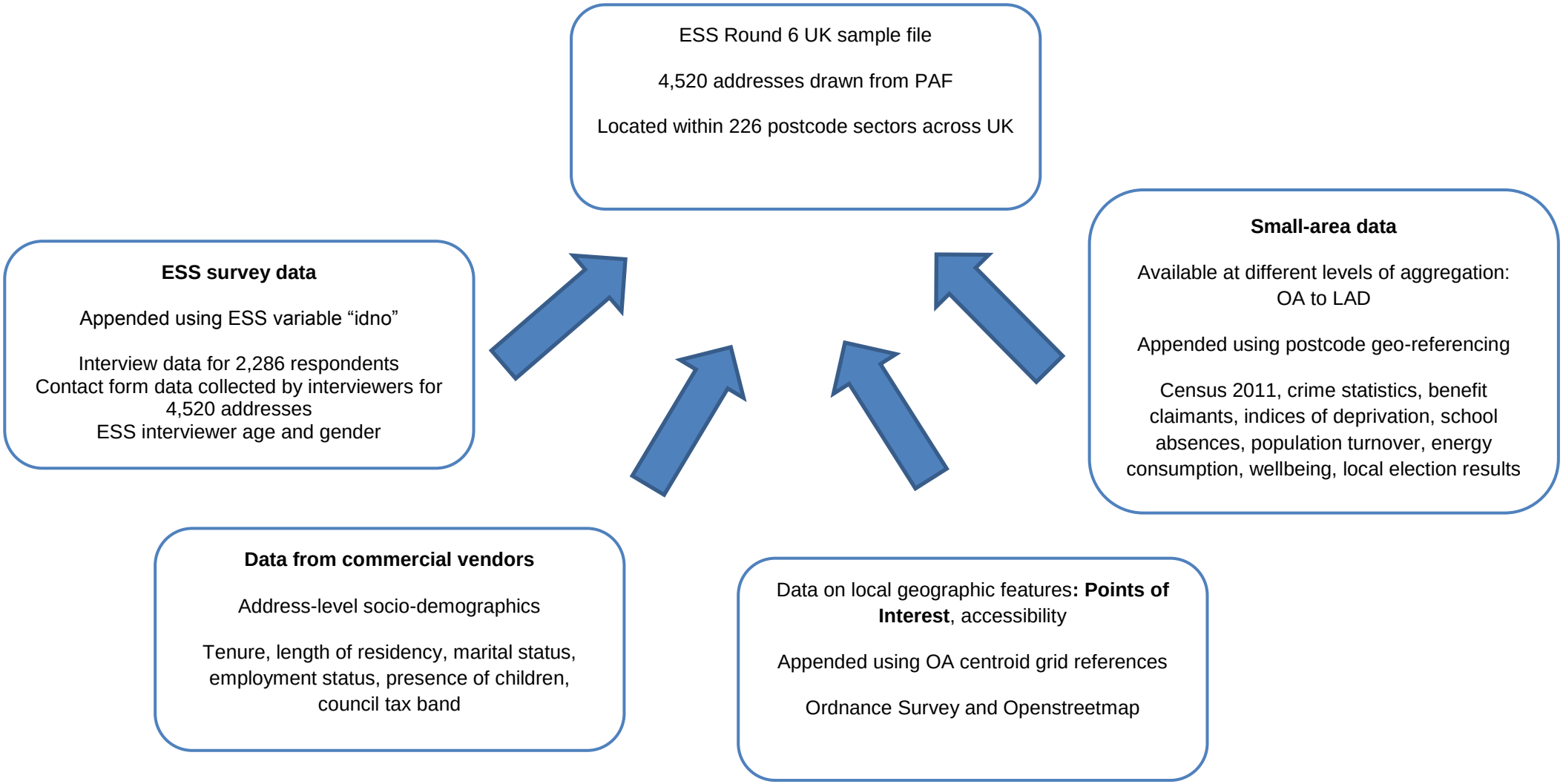
ADDResponse considers three main types of auxiliary data:

- Small-area data from administrative sources including Census 2011
- Local geographic information on the position of sampled addresses relative to local amenities
- Household level data on household and resident characteristics purchased from commercial vendors

All three of these data sources are mentioned by Tom Smith as part of his discussion of the ML-MS approach to survey data (Smith, 2011). The first type of data is already well-known and used by survey researchers for both methodological and substantive research but, given the wide and growing array of possible datasets available merits further investigation. It offers the possibility of high quality, complete data on a range of topics but only at an aggregate level. The second type of data offers rich possibilities for analysis. It is increasingly used for substantive analysis in a range of fields including geography, environmental planning and public health but has received little attention from survey methodologists. The third type of data has received attention in other countries including the US and Germany in the last few years but has not been explored systematically in the UK. It presents a valuable opportunity to access household or individual level data but there are concerns about as well as resource implications from purchasing commercial data.

The data sources and variables of each type explored as part of ADDResponse are by no means exhaustive. Decisions on which data sources and variables to include were taken by the research team in consultation with the project Advisory Group. They are intended to be indicative of the three different types of data under consideration. Given the wider project's interest in survey nonresponse we were particularly interested in factors thought to be associated with response behaviour (see Groves and Couper, 1998 for a possible model of survey nonresponse) but also selected some novel or innovative datasets even where the theoretical link was less clear in order to explore their potential.

Figure 1: Overview of ADDResponse data sources



1.3 Overview of data linkage process

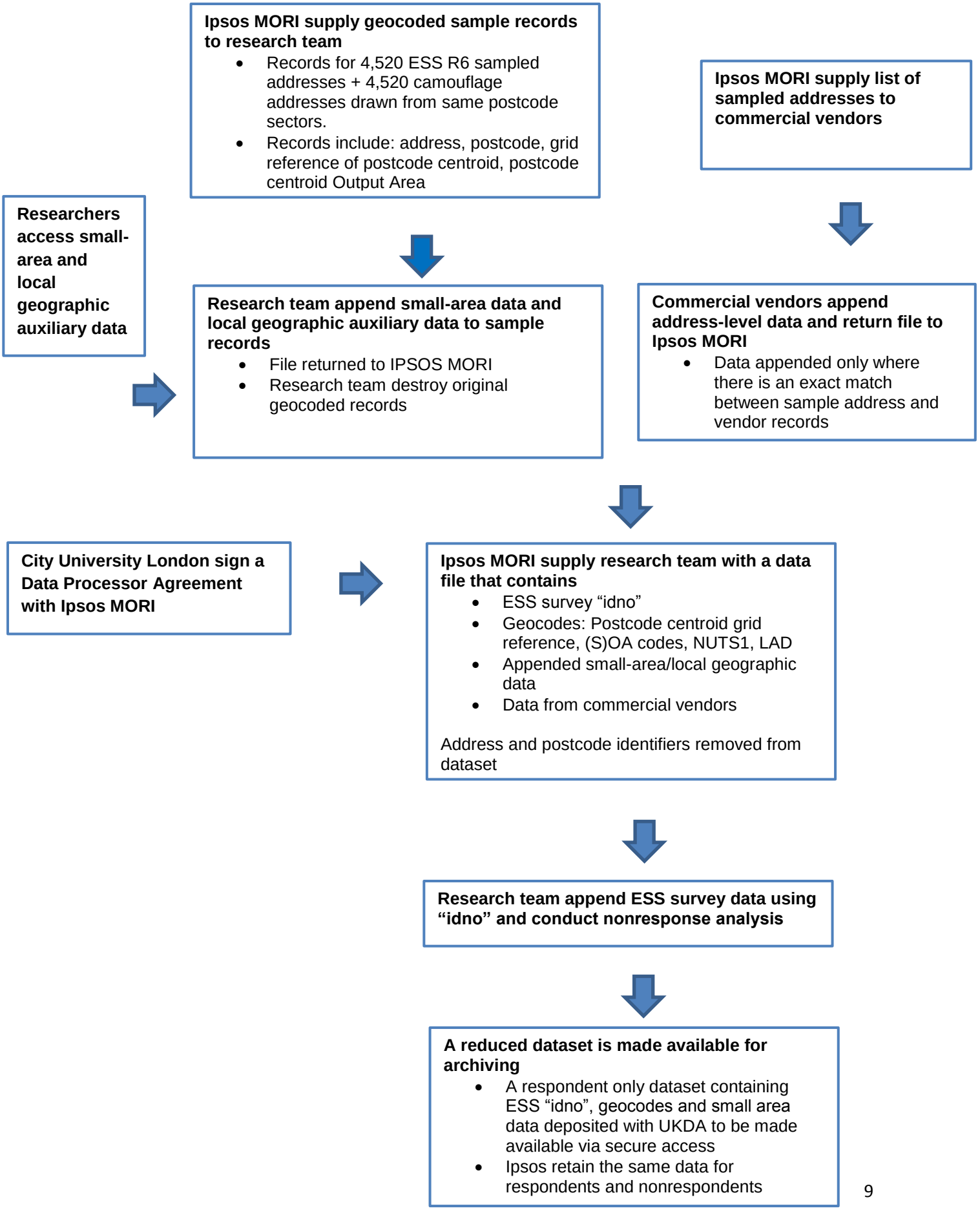
Data linkage was carried out by the research team in cooperation with Ipsos MORI, the survey agency that carried out ESS Round 6 fieldwork in the UK and remain the data controller for the sample file. Data linkage took place in a series of stages in order to minimise the risk of survey respondents and their data being identified by the research team (see Figure 2). Part way through the project a data processor agreement was put in place between Ipsos MORI and City University London which made it possible for the research team to have access to potentially identifiable data for both respondents and nonrespondents. This meant that the research team were able to explore and analyse the full range of auxiliary variables in combination with the ESS survey dataset without the need for disclosure control in the form of banding, aggregation or suppression. Nevertheless, even with the data processor agreement in place, it remains good practice to share only the minimum amount of data required at any stage of the project.

The research team were responsible for appending small-area and local geographic information. The team were first given access to a file containing the full address and postcode of all ESS sampled addresses (but no other information) in order to append geocoded auxiliary variables. This data file also contained camouflage addresses to minimise the risk of researchers identifying individual ESS participant addresses in subsequent datasets. Auxiliary data were matched using postcode and the National Statistics Postcode Lookup File (small-area data) and British National Grid coordinates (local geographic information). After the auxiliary data had been appended the file was returned to Ipsos MORI who stripped off the address information before appending ESS survey responses (and the outcome of the survey request) and returning the dataset to the research team for analysis. Because a data processor agreement was in place, the research team were allowed to retain other low level geographic identifiers (such as Output Area code) on the dataset. All data files were transferred securely using 256 encryption and password protected files. They were stored on secure networks with restricted access.

The commercial vendors were responsible for appending their data records to the sample file on behalf of the research team. Ipsos MORI provided the companies with a file listing the addresses of all sample units (but no further information) and received the data back from the companies before supplying the commercial records (minus address details) to the research team (See Chapter 5 for more details).

One limitation of the staged approach to data linkage is that the research team needed to decide on the auxiliary variables to be appended to the sample file, and how they should be coded, before gaining access to the survey data or being able to carry out exploratory analysis. This meant that many variables were added to the dataset “just in case” and the final selection could perhaps have been more selective if

Figure 2: ADDResponse data linkage process



1.4 Accessing ADDResponse data

One of the objectives of the ADDResponse project was to evaluate the availability of different auxiliary data sources, both for use by the original research team and possible reuse by other researchers. This report demonstrates that the research team were able to obtain and combine auxiliary data from a wide variety of different sources. However, there are several constraints on making the data more widely available.

One constraint is that some of the auxiliary data was made available to the research team under license and cannot be archived for wider use. This is the case for the data obtained from Ordnance Survey as well as data purchased from commercial vendors.

A second constraint is the disclosure risk which arises from the combination of individual-level survey data and low-level auxiliary data making it possible to identify specific addresses within the dataset. ESS participants were assured in the survey materials that it would not be possible to identify them from the data. In addition, there are particular concerns about sharing data on survey nonrespondents, who did not even consent to participate in the original survey. The usual solution, to produce an anonymised version of the original research dataset for archiving, was not considered feasible in this instance. The added value of the combined ADDResposne dataset comes in the combination of multiple different data sources plus the presence of low level geocodes such as grid reference to enable spatial analysis. This added value would be undermined significantly if the data were to be edited to minimise disclosure risk sufficiently to consider the dataset anonymised.

A dataset containing small-area data, geocodes and ESS "idno" has been made available via the UKDA under secure access conditions.[3] ESS "idno" can be used to link these data with ESS Round 6 data for the UK (survey responses, contact form data and information on the survey interviewers). Given the additional sensitivities about making data on nonrespondents available, the dataset available via the UKDA contains records for respondents only. Researchers interested in accessing ADDResponse data for nonrespondents may contact Ipsos MORI and submit a request to enter into a data processor agreement to access these data.[4] Ipsos MORI will consider all requests on a case by case basis based on the strength of the research case made. Further details of how to submit a request are available via the ADDResponse (www.addresponse.org) website.

All of the auxiliary data used in this project are publically available. Researchers interested in accessing and using similar data for their own research projects can do so by going directly to the original source (links are provided in the report wherever possible).

[3] https://www.ukdataservice.ac.uk/use-data/secure-lab/about

[4] The UK Data Protection Act (1998) allows for the processing of personal data provided that this is done lawfully and fairly and that at least one of the conditions for processing set out under Schedule 2 of the act is met. Processing of personal data was allowed under the ADDResponse project (and potentially other similar research projects) as – given its focus on survey nonresponse and improving survey practice - it was deemed to be in the "legitimate interests" of the data controller.

1.5 Outline of this report

The aim of the ADDResponse data scoping exercise was to evaluate the availability and suitability of geocoded auxiliary data for appending to address-based survey samples in the UK. These data could then be used for exploration of survey nonresponse or other methodological research or for substantive analysis of survey data.

Three related questions were addressed:

- What data are available to append to address-based survey samples using geocodes? Issues considered around availability include: any limits on access e.g. license arrangements, cost (if any), and scope for data reuse
- How easy or difficult is it to append data to the survey sample? What level of resources is required?
- How suitable are the data? Issues considered around data quality include: timeliness, completeness, geographic coverage, accuracy and appropriateness (e.g. unit of analysis).

This report discusses the outcome of that data scoping exercise, providing full documentation of all of the datasets used and the variables derived as well as issues encountered during the process. Chapters 2 to 5 of this report provides a detailed commentary on the different data sources investigated as part of the ADDResponse project: ESS data, small-area data, local geographic data and data from commercial vendors. Each chapter provides details of the specific data sources used, how these were appended to the main dataset, how variables were derived for analysis and any challenges encountered regarding data access, linkage or quality.

Chapter 6 concludes by summarising some of the main issues encountered during the project and the implications for future research combining auxiliary data and survey data.

Details of all of the variables included in the ADDResponse dataset are provided in the excel file "ADDResponse variable listing.xls" made available alongside this report.

1.6 Acknowledgements

Many people were involved in the production of the ADDResponse dataset. These include project investigators Rory Fitzgerald, Aidan Slingsby and Jason Dykes (City, University of London) and Chris Skinner (LSE); academic collaborators Rainer Schnell and Kathrin Thomas (City, University of London); Gideon Skinner and Sarah Tipping from Ipsos MORI, and the project Advisory Group (Chris Brunsdon, NUI Maynooth; Andy Falllows, ONS; Alun Humphreys, NatCen Social Research and National Coordinator of ESS UK; Peter Lynn, ISER at University of Essex; Patten Smith, Ipsos MORI; Tom Smith, NORC at the University of Chicago; Ineke Stoop, Netherlands Institute for Social Research (SCP) ; Richard Webber, Origins UK). Thanks are also due to the people responsible for the original auxiliary datasets who answered questions about the data.

ADDResponse - Auxiliary Data Driven Nonresponse Bias Analysis Project was funded by the Economic and Social Research Council grant ES/L013118/1.

2 European Social Survey Data

Established in 2001, the ESS is a biennial cross-national survey of public attitudes and opinions. Data are collected from a representative sample of adults aged 15 and over in between 20 and 30 countries each round. Consisting of a core questionnaire that remains the same in every round alongside round-specific rotating modules, the face-to-face survey covers many topics including: satisfaction with democracy, political trust, citizen engagement, attitudes towards immigration, subjective wellbeing, health inequalities, attitudes to work and family, attitudes on climate change and energy use and views on the welfare state.

This project used UK data from Round 6 of the European Social Survey (2012/13). Fieldwork for ESS Round 6 was conducted in the UK by Ipsos MORI between September 2012 and February 2013.

Ipsos MORI retain the sample file for ESS Round 6 which contains address and postcode information plus additional geographic variables (e.g. Output Area) for all productive and unproductive cases. The full sample file containing 4,520 records, including address and postcode, was made available to the research team so that they could append auxiliary data. Once auxiliary data had been appended address and postcode information were stripped off of the file but other geographic identifiers were retained for analysis.

ADDResponse appended a range of ESS interview and fieldwork data to the original sample records. These ESS data are publically available via the ESS website (www.europeansocialsurvey.org). Data were appended to the sample records using the variable "idno", a unique 6-digit reference number assigned to each case and available in each ESS dataset.

2.1 ESS sample file

The sample for the UK ESS is drawn from the Royal Mail Postcode Address File and uses a three-stage clustered sample design. At the first stage a sample of 226 postcode sectors distributed across the UK were selected as Primary Sampling Units (PSUs). A sample of 20 addresses was then randomly selected from within each PSU giving a total sample of 4,520 addresses. At the final stage, the interviewer randomly selects a respondent on the doorstep, selecting from among adults aged 15 and over resident at the address.

The ESS sample file initially provided to the research team to append auxiliary data included: address, postcode, Primary Sampling Unit identifier (Postcode Sector), British National Grid coordinates, and Output Area of the postcode centroid. The dataset supplied for analysis post-data linkage was stripped of address and postcode variables but contained: Primary Sampling Unit identifier (Postcode Sector), British National Grid coordinates, and Output Area of the postcode centroid. It also contained post-code level ACORN classifications for every address. ACORN is a segmentation tool produced by CACI which classifies the UK population into demographic types.[5] ACORN segments households, postcodes and neighbourhoods into six categories, 18 groups and 62

[5] http://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf.

types. Finally, it contained the ESS unique reference number "idno" to enable other ESS datasets to be appended to the sample file.

2.2 ESS main data file

A total of 2,286 productive interviews were achieved in ESS Round 6, a response rate of 53%.[6] Interview data are available for these 2,286 cases. As part of ADDResponse, these data were used to examine associations between auxiliary variables and substantive survey outcomes and thereby identify potential sources of bias in the event of differential nonresponse. Other researchers may wish to use the combination of contextual auxiliary variables and data from ESS respondents to examine substantive research questions.

The dataset used for ADDResponse was the UK country file, a subset of the ESS integrated file covering all participating countries:
ESS Round 6: European Social Survey Round 6 Data (2012). Data file edition 2.1. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

Full survey documentation, the ESS Round 6 questionnaire and data protocol are available via the ESS website:
ESS Round 6: European Social Survey (2016): ESS-6 2012 Documentation Report. Edition 2.2. Bergen, European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC

2.3 ESS contact form data

ESS interviewers are asked to complete a Contact Form for each sampled address recording details of the date, time and mode of all survey contact attempts and the outcome of those contact attempts. They also record interviewer observations regarding the type and condition of the property and the surrounding neighbourhood (European Social Survey, 2012). Call patterns and fieldwork efforts are known to influence response rates (Kreuter and Kohler, 2009) and significant associations have been found between variables recorded via interviewer observation (e.g. type of dwelling, barriers to entry) and response propensity (Stoop et al, 2010; Blom, 2012; Fuchs et al, 2013). ADDResponse provided an opportunity to compare the relative value of survey paradata and auxiliary data from external sources in predicting survey nonresponse.

Contact form data, including a record of the interviewers who worked on each case, are available via the ESS website along with full documentation, including copies of the contact form. The dataset used for ADDResponse was:
ESS Round 6: European Social Survey Round 6 Data (2012). Data from contact forms, edition 2.0. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

[6] Base for response rate calculation excludes 265 ineligible addresses (e.g. commercial premises, derelict properties or properties under construction)

Contact forms were completed for all 4,520 addresses in the UK. Interviewer observations are available for nearly all (97%) of all UK sampled addresses

The contact form data were used to construct outcome variables used for nonresponse analysis i.e. to indicate whether an address resulted in a productive interview, a refusal, a non-contact, or was unproductive for another reason. A sampled address may be visited multiple times, resulting in multiple, potentially inconsistent, outcomes. There are two ways a final outcome code to the address, either based on the outcome of the last contact attempt or by constructing a priority system of visit outcomes and selecting the outcome with the highest priority (see Blom, 2013).

The ESS uses a combination of these two approaches to assign final outcome codes as reported in the National Technical Summary (NTS), assigning codes mainly on the basis of the last contact attempt but prioritising refusals. ADDResponse used priority coding to ensure that the classification of respondents and nonrespondents best reflects the de facto response behaviour observed over the course of fieldwork. It ensures, for example, that only addresses where no contact was made at any point during fieldwork (rather than just at the last visit) are counted as noncontacts. A productive interview at any contact attempt was given priority over ineligibility which was given priority over a refusal which in turn was given priority over non-contact. Details of the priority coding used and the resulting variables are given in the appendix to this report. [7]

2.4 ESS interviewer data

ESS interviewers are asked to complete a short questionnaire after every interview recording information on the conditions under which the interview took place, e.g. was there anyone else in the room, did the respondent have difficulty answering the questions. The interviewer questionnaire data file also contains a unique identifier for the interviewer who conducted each interview (INTNUM) and variables provided by the fieldwork agency giving the interviewer's age and gender. Interviewer age and gender can be matched with ESS contact form data using the variable "INTNUM" thereby providing information on the age and gender of the interviewer who made each contact attempt. Such information may be useful given the large literature on how interviewers can influence response rates (Durrant et al, 2010) as well as survey data quality (Pickery and Loosveldt, 2004).

The dataset used for ADDResponse was:
ESS Round 6: European Social Survey Round 6 Data (2012). Data from Interviewer's questionnaire, edition 2.0. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

[7] Priority coding and the NTS outcome codes result in similar distributions although there are fewer noncontacts and more refusals under priority coding.

3 Small-area data

A wide variety of small-area data from administrative sources is made publically available under the Open Government License for public information. Data can be downloaded from a number of sources. Official statistics can be downloaded from www.gov.uk/government/statistics. There are also a number of dedicated data repositories including NOMIS - which provides access to labour market statistics (www.nomisweb.co.uk), digital.nhs.uk which provides access to health data, and https://data.police.uk which provides open access to local data on police reported crime. The devolved administrations for Wales (statswales.gov.uk), Scotland (www.gov.scot/topics/statistics) and Northern Ireland (www.nisra.gov.uk) also produce their own statistics.

The small-area data used in ADDResponse are only a small subset of what is available. In selecting which data sources to explore, ADDResponse focused on variables which it was felt might plausibly be associated with survey nonresponse based on existing literature. Small-area data were intended to provide measures of neighbourhood characteristics (e.g. crime rates, deprivation) which might be associated with nonresponse and also to serve as proxies for household or individual characteristics which might influence nonresponse (e.g. age, household size). Priority was given to data which covered the whole of the UK (or at least Great Britain) and which were available for 2011 -2012 i.e. the year immediately preceding ESS fieldwork. Data was appended at the lowest possible level of aggregation.

3.1 Data linkage

The small-area data were appended to the ESS sample using geographic referencing. The National Statistics Postcode Lookup (NSPL) produced by ONS and freely available to download from their Open Geography Portal was used to obtain a full range of geographic codes for each sampled address.[8] These include: NUTS1 region, Local Authority District (LAD), Middle Super Output Area (MSOA), Lower Super Output Area (LSOA), Output Area (OA), 2003 CAS ward, and electoral ward.[9] These geographic codes were then used to append the different small-area statistics (see below for full details of data linkage by data source).

NSPL enables postcode records to be assigned to the correct administrative unit (e.g. Output Area, Local Authority, Health Authority). Postcodes are allocated to output areas by plotting each postcode's centroid directly into the Output Area boundaries. All postcodes in an output area are then allocated to the same higher geography using a best-fit method. NSPL is ONS' primary product for producing statistics and ensures that all higher level statistics produced are built from Output Areas.

New versions of ONS postcode products are released quarterly. ADDResponse used the November 2014 NSPL which includes versions using both 2001 and 2011 census codes (ONS, 2014a).

[8] http://geoportal.statistics.gov.uk/

[9] For further details of UK administrative, census and other geographic units see: https://www.ons.gov.uk/methodology/geography/ukgeographies

3.1.1 Issues encountered with data linkage

Having the NSPL available to access the UK's comprehensive system of geocodes makes combining small-area data with other geo-referenced data (such as survey sample files) relatively straightforward. However, a number of issues were encountered. One issue was that census area codes changed between 2001 and 2011. Statistics providers did not always specify which set of codes were used in their data releases. Fortunately, both sets of codes – for 2001 and 2011 - were available via the 2014 NSPL.[10] There was also an issue with appending data at local authority level as some local authority codes had change between when the data was produced (2011-12) and when the NSPL was issued (2014), either as a result of changes in the legal structure of local authorities or boundary changes. Where codes had changed over time, historic codes were identified and appended to the ADDResponse dataset using the Code History Database (CHD) (ONS, 2016a). This is a database containing all previous area codes (for a variety of geographical units) as well as information on the dates that changes occurred and the relationships between different geographical hierarchies. It proved very useful in cross-checking the data linkage and resolving issues with problem cases.

3.1.2 Quality checks

A number of quality checks were carried out after appending the small-area data to the sample file. The number and distribution of missing cases was checked to verify whether the linkage had worked as anticipated and that any missing data was genuinely missing. Summary statistics for each variable were examined to verify that the distributions e.g. maximum and minimum values were as expected. Randomly selected observations from the combined dataset were also checked against the original source data.

These checks helped to identify a number of mismatched cases which were the result of changes/mismatches in area codes between the source data and the NSPL. Once identified, these mismatches were resolved wherever possible by tracing the correct area codes (see above).

3.2 Using small-area data: Issues to be aware of

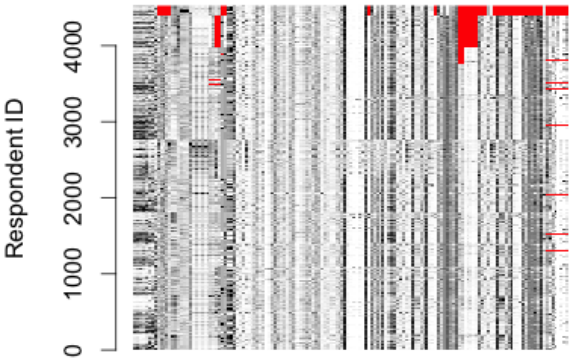
Small-area data from government sources are a valuable source of auxiliary data. Increasing amounts of data are being published open-source via government websites. These data are often classified as Official Statistics so have been quality assured and there is usually no problem with missing data (see Figure 3 below). Many statistical series publish data on a yearly, or even quarterly or monthly basis, making them a timely source of data.[11] Trend data from previous years is usually retained making this a particularly useful source of data for appending to survey data post-collection for secondary analysis. However, small-area data do present some challenges.

[10] Only 2011 codes are available in subsequent NSPL files.

[11] One exception to this of course is the decennial census. The majority of the demographic variables used in ADDResponse came from the 2011 Census. These data were timely for this specific project, fieldwork for which took place in 2012-13. However, census data will not always be as useful given that areas can change over a 10 year period (as some of the variables in the ADDResponse dataset show (see Section 3.3.3)).

One significant issue is that many statistics are produced separately for England, Wales, Scotland and Northern Ireland rather than for the UK as a whole. This means that data will not necessarily be available for the whole of the UK. Figure 3 below highlights in red cases for which small-area observations were missing. Cases are ordered by country (Northern Ireland at the top, England at the bottom. Nearly all variables were available for England and Wales but were sometimes missing in the case of Scotland and, especially, Northern Ireland. Another issue with country-specific data, as discussed above, increase the complexity of data linkage. Perhaps most importantly, where data are available for all four countries, the way the data have been collected or statistics produced may vary. Separate Indices of Deprivation are produced for England, Wales, Scotland and Northern Ireland for example using different data and covering different time periods. Crime figures are reported under different category headings in different countries. Care must be taken when combining data for the four countries.

Figure 3: Pattern of missing data on small area variables



Note: Red indicates missing values

Another issue to bear in mind is that these data are aggregate data, some of it available only at relatively high levels of aggregation such as local authority district (LAD). Depending on the research problem at issue this need not be a problem; aggregate variables may be most appropriate for studying neighbourhood contextual effects for example. However, researchers should remain aware of the "modifiable areal unit problem"(MAUP) and fact that the level of aggregation at which variables are measured will affect both the point estimates and correlations observed between variables (Gehlke and Biehl, 1934; Openshaw, 1983). Particular care must be taken when using aggregate-level variables as proxies for individual or household characteristics to avoid the so-called ecological fallacy (Robinson, 1950). Relationships which are observed (or which fail to be observed) at an aggregate level may not be a true reflection of relationships at the individual level. For example just because survey nonresponse may be higher in areas with high unemployment, we cannot conclude that unemployed people are less likely to respond; it may be that it is employed people that are less likely to respond in areas of high unemployment due to contextual effects.

3.3 Government data used in ADDResponse

This section summarises the different sources of government-produced small-area data used for ADDResponse. All of the data sources listed in this section are crown copyright and made available under the terms of the Open Government License.[12]

3.3.1 2011 Census Urban Rural classification

Source	Adapted from: Department for Environment Food and Rural Affairs (2013) Scottish Government (2012a)
Data retrieved from	National Statistics Postcode Lookup (November 2014) https://data.gov.uk/dataset/national-statistics-postcode-lookup-uk-nov-2014
Time period	2011
Geographic coverage	Great Britain
Level of aggregation	Output Area

Separate urban-rural classifications are produced for the constituent countries of the UK. Those for England/Wales and Scotland (though not Northern Ireland) were included as part of the 2014 NSPL. The England/Wales classification has ten categories and the Scotland classification eight categories. In England/Wales OAs are classified as urban if they are allocated to a built up area with a population of 10,000 or more. In Scotland OAs are classified as urban if they are in a settlement of more than 3,000 people. Urban and rural OAs are then further sub-divided based on the type of settlement and the local context i.e. drive time to the nearest large settlement (Scotland) or the density of the surrounding population (England/Wales).

For ADDResponse, in addition to the full classification, we also derived a collapsed three-fold classification as follows:

England/Wales	Scotland	ADDResponse
Urban: Major conurbation Urban: Minor conurbation Urban: City and Town	Urban: Large urban areas Urban: Other urban areas Urban: Accessible small towns Urban: Remote small towns Urban: Very remote small towns	Urban
Rural: Town and fringe	Accessible rural	Mixed
Rural: Village	Rural: remote rural Rural: very remote rural	Rural

[12] http://webarchive.nationalarchives.gov.uk/20160105160709/http://nationalarchives.gov.uk/doc/open-government-licence/version/3/

Rural: Hamlets and isolated dwellings		

3.3.2 Output Area Classification 2011

Source	Classification produced via a collaboration between ONS and University College London. More information: http://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications
Data retrieved from	National Statistics Postcode Lookup (November 2014) https://data.gov.uk/dataset/national-statistics-postcode-lookup-uk-nov-2014
Time period	2011
Geographic coverage	UK
Level of aggregation	Output Area

The 2011 Census OAC is used to group together geographic areas according to key characteristics common to the population in those areas. These groupings or clusters are derived using 2011 population census data. OAs are classified into 8 supergroups, 26 groups and 76 subgroups. Further information on how the classification was produced, and the census variables used to produce it, is available via the ONS website.[13]

Similar classifications have also been derived for other geographic units such as local authorities.

3.3.3 Census

Source	Adapted from: Office for National Statistics, National Records of Scotland, Northern Ireland Statistics and Research Agency (2016)
Data retrieved from	http://www.ons.gov.uk/census/2011census/2011ukcensuses/ukcensusesdata (release 13 June 2014)
Time period	2011
Geographic coverage	UK
Level of aggregation	Output Area, Lower Super Output Area, Middle Super Output Area

Source	Adapted from: Office for National Statistics (2011)
Data retrieved from	http://dx.doi.org/10.5257/census/aggregate-2001-2 (England and Wales) http://www.nisra.gov.uk/census/2001%20Census%20Results/Key%20Statistics/KeyStatisticsforSuperOutputAreas.html (Northern Ireland) Scottish Data supplied to research team on CD

[13]http://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications

Time period	2001
Geographic coverage	UK
Level of aggregation	Lower Super Output Area, Middle Super Output Area

Census 2011 is the main source of data on the socio-demographic profile of local areas and potential respondents within the ADDResponse dataset. OA level data for the whole of the UK was obtained through ONS' data release: Key and Quick Statistics for Output Areas in the UK (13 June 2014). Variables giving the proportion of the population per OA with a certain characteristic were derived from the counts available in the release. Most variables were appended at OA level using 2011 OA codes; we were interested in data at the lowest level of aggregation possible as the variables were intended as proxies for individual or household characteristics.

Some variables (e.g. unemployment rate, ethnic fractionalisation) were also intended as measures of local context. These variables were included at multiple levels of aggregation (OA, LSOA and MSOA) to enable us to explore the most appropriate level of aggregation at which to analyse them. Data were appended at OA level and then aggregated to higher level geographies using the data.table package in R.[14]

MSOAs are not found in Northern Ireland and so data at this level of aggregation is missing for Northern Ireland. All other census data are complete.

Change 2001-2011

There was interest in analysing how recent change in an area's demographic profile (e.g. if it has become more ethnically diverse, been gentrified etc.) might influence behaviour. For selected variables we therefore calculated the change in an LSOA/MSOA profile 2001-2011. Variables from the 2001 census were appended in the same way as the 2011 variables and change variables created by subtracting the 2001 value from the 2011 value. Only the change variables (not 2001 variables) were retained on the final dataset.

In calculating change variables we faced the challenge that some census boundaries changed between 2001 and 2011 to take account of population changes meaning that there was not necessarily a one to one match between 2001 SOAs and 2011 SOAs. This was resolved in the following way: In Northern Ireland there was no change in SOA boundaries 2001-2011 so it was straightforward to calculate the difference between 2001 and 2011 figures (NISRA, 2013). In Scotland all 2001 census areas were mapped onto equivalent 2011 areas using a lookup file (NRS, 2014). England and Wales were treated in line with guidance from ONS on comparing census areas over time (ONS, 2013a). SOAs were flagged as having stayed the same (U), been split (S), merged (M) or changed completely (X) since 2001. For those areas staying the same the comparison was straight forward. For areas which had merged (31 LSOAs and 23 MSOAs), 2001 areas were added together to create areas corresponding to those in 2011. For areas which had been split (78 LSOAs and 36 MSOAs), 2011 areas were added together to create areas which would correspond to those in 2001. Areas which had changed completely (7 LSOAs and 11 MSOAs) were set to missing in line with ONS guidelines.

[14] http://cran.r-project.org/web/packages/data.table/data.table.pdf

Details of all of the census variables included in the ADDResponse dataset are provided in the excel file "ADDResponse variable listing.xls" made available alongside this report.

Hard to count (HTC) classification

A Hard to Count (HTC) classification was used to classify areas in advance of the 2011 census, according to their expected relative difficulty of enumeration. The classification was used to direct the number of enumerators in the field. Separate HTC measures were constructed for England and Wales, Scotland and Northern Ireland. The England and Wales HTC classified LSOAs from 1 to 5 based on their predicted non-response propensity, 1 being the easiest to enumerate (ONS, 2010). The Scotland HTC classified datazones from 1 to 5 (1 being the easiest to enumerate) depending on the predicted difficulty of obtaining a response in the census. It takes into account a number of factors known to affect response rates, including the proportion of students and privately rented dwellings, and the datazone's ranking in the Scottish Index of Multiple Deprivation (NRS, 2013). The HTC for Northern Ireland comprises 8 HTC strata based on deprivation and urban/rural status (NISRA, 2015a).

LSOA level data was obtained on request from Office for National Statistics, NISRA and National Records of Scotland and merged onto the ADDResponse dataset using 2011 LSOA, Datazone or Super Output Area codes.

3.3.4 Population turnover

Source	Adapted from: Office for National Statistics (2016b)
Data retrieved from	https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/localareamigrationindicatorsunitedkingdom
Table/file reference	Sheet: 2012
Time period	2011-2012
Geographic coverage	UK
Level of aggregation	LAD

Data on population flows into and out of local authorities was appended to the ADDResponse dataset to provide an indicator of how stable the resident population within the area is likely to be. Data were appended using LAD codes and the NSPL.[15]

Two variables are available: turnover rate per 1,000 population (international migration) and turnover rate per 1,000 population (internal migration). Figures show the volume of international or

[15] Codes for Glasgow City, Northumberland, Welwyn Hatfield and Stevenage differed between the NSPL and the source statistics and had to be bridged using the Code History Database.

internal population movement within the UK and are calculated as the sum of in and out-migration (in the year to June) per 1,000 resident population.

International inflows are measured in terms of the number of persons arriving or returning from abroad to take up residence in a country for a period of at least 12 months whilst outflows are defined in terms of persons leaving their country of usual residence to take up residence in another country for a period of at least 12 months. Estimates are compiled based on the International Passenger Survey, Labour Force Survey and Home Office figures on asylum seekers and their dependents. Internal population flows measure movement between constituent countries within the UK or between local authorities within countries. Since internal moves are not recorded formally, information obtained from the NHS Central Register (NHSCR) and GP Patient Registers is used as a proxy.

3.3.5 Crime statistics

Source	Adapted from: Home Office (2016)
Data retrieved from	https://data.police.uk/
Time period	September 2011-August 2012
Geographic coverage	England and Wales
Level of aggregation	LSOA

Source	Adapted from: Home Office (2012)
Data retrieved from	http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do (Neighbourhood Statistics - > Crime and Safety -> Notifiable offences recorded by the police)
Table/file reference	Notifiable Offences Recorded by the Police 2011/12 C150311_2454_2011SOA_LA
Time period	April 2011 - March 2012
Geographic coverage	England and Wales
Level of aggregation	LAD

Source	Adapted from: Scottish Government (2012b)
Data retrieved from	http://www.gov.scot/Publications/2012/06/1698
Table/file reference	Recorded crime in Scotland, 2011-12 (Table 7)
Time period	April 2011 - March 2012

Geographic coverage	Scotland
Level of aggregation	LAD

Source	Adapted from: NISRA (2015b)
Data retrieved from	http://www.ninis2.nisra.gov.uk/public/Theme.aspx?themeNumber=131&themeName=Crime%20and%20Justice
Table/file reference	Recorded Crime (administrative geographies) Sheet: LGD
Time period	April 2011 - March 2012
Geographic coverage	Northern Ireland
Level of aggregation	LGD

Two sets of crime figures were appended to the ADDResponse dataset, LSOA-level estimates available for England and Wales and LAD level estimates covering the whole of the UK. In both cases, figures are based on police figures for reported crimes. Variables report the number of crimes per 10,000 population.

LSOA data

Data.police.uk, is an open access data base of recorded crimes for every police force area in England, Wales and Northern Ireland, available monthly going back to December 2010. The website's custom download facility was used to create a dataset of all crimes reported in the period September 2011-August 2012. Each crime record in England and Wales includes a record of the LSOA in which it was reported as occurring, allowing it to be appended to the ADDResponse dataset.[16] Crime records were aggregated to provide counts per LSOA using the data.table package in R and then appended to the ADDResponse dataset using LSOA codes. Finally, crime rates per 10,000 population were calculated by dividing crime counts by 2012 mid-year population estimates provided by ONS (ONS, 2013b).

Separate figures are reported for five different categories of crime: "violence and sexual offences", "burglary", "criminal damage and arson", "public order" offences, and "antisocial behaviour order" (ASBOs).

The data provider notes some limitations with the data available via data.police.uk especially as regards the accuracy of the location data provided. It is estimated that geo-referencing accuracy varies between 60 and 97% across forces.

[16] Crime records in Northern Ireland are not allocated to an LSOA and so could not be used here. Each crime is assigned a set of coordinates to provide the 'exact' location where it occurred. We considered using these coordinates to produce crime maps but decided against it for two reasons. First, LSOA data are easier to produce and analyse and consistent with how other variables in the ADDResponse dataset have been defined. Second, there is some doubt over the accuracy of location coordinates. Coordinates do not record the exact location of the crime but rather the location of the nearest 'mapped point'. A lack of mapped points in some, especially rural, areas can distort the data.

LAD data

LAD crime figures are available for the whole of the UK. However, figures are produced separately for England and Wales, Scotland, and Northern Ireland. There are some differences in data collection methodology and definitions across the four countries. For example, Scotland figures include all police recorded crimes whereas those for England and Wales cover notifiable offences which will exclude some minor offences, even if these were reported to the police. Importantly, the categories for which crimes are recorded also vary between countries and it is not always possible to create comparable categories across all four countries. The way in which categories have been mapped between England/Wales, Scotland and Northern Ireland is shown below.

England and Wales	**Scotland**	**Northern Ireland**
Violence with injury + Violence without injury	Non-sexual crimes of violence	Violence with injury + Violence without injury
Sexual Offences	Sexual crimes	Sexual Offences
Domestic Burglary	-	Domestic Burglary
Criminal damage and arson	-	Criminal damage

Data were appended to the ADDResponse data file using LAD codes.[17] Data are missing for Taunton Deane and South Somerset local authorities (74 cases).

The Scottish Government publishes figures on crimes per 10,000 population. Similar rates were calculated for England, Wales and Northern Ireland by dividing the raw offences counts by population figures (ONS, 2013c).

3.3.6 Indices of deprivation

Source	Adapted from: Department for Communities and Local Government (2011)
Data retrieved from	https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010
Time period	2010
Geographic coverage	England
Level of aggregation	LSOA

Source	Adapted from: Stats Wales (2011)
Data retrieved from	https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2011/WIMD2011

[17] Codes for Northumberland and Stevenage Unitary Authorities differed between the NSPL and the source statistics and had to be bridged using the Code History Database.

Time period	2011
Geographic coverage	Wales
Level of aggregation	LSOA

Source	Adapted from: Scottish Government (2012c)
Data retrieved from	http://www.gov.scot/Topics/Statistics/SIMD/DataAnalysis/Background-Data-2012
Time period	2012
Geographic coverage	Scotland
Level of aggregation	LSOA

Source	Adapted from: NISRA (2010)
Data retrieved from	http://www.nisra.gov.uk/deprivation/nimdm_2010.htm
Time period	2010
Geographic coverage	Northern Ireland
Level of aggregation	SOA

Deprivation covers a broad range of issues and refers to unmet needs caused by a lack of resources of all kinds, not just financial. The Indices of Deprivation attempt to measure a broad concept of multiple deprivation, made up of several distinct dimensions, or domains, of deprivation including income, health and education. In addition, an overall summary index of multiple deprivation (IMD) is produced.

Separate Indices are produced for England, Wales, Scotland and Northern Ireland. Data were appended to the ADDResponse dataset using LSOA codes and the NSPL. The English data were appended using LSOA codes from 2001. Data for other countries were appended using LSOA codes from 2011.

Comparing deprivation measures across the four countries is not straightforward. Domain indices for each country are constructed using slightly different measures and refer to different time points. Individual domains are assigned different weights in the construction of the overall IMD measure. Given differences in the way the deprivation indices are constructed it is not possible to compare raw deprivation scores across domains or countries. Because ranks are assigned on a country by country basis nor is it possible to conduct UK-wide analysis using the ranks; the number of LSOAs in each country is different and therefore being ranked 100th in Scotland is not necessarily equivalent to being ranked 100th in England.

However, it is possible to construct broadly comparable measures of deprivation for each country across the following seven domains as well as an overall IMD. In order to create measures of deprivation that can be compared across the four countries of the UK we created deprivation percentiles by first ordering LSOAs from most to least deprived based on ranks and then assigning each LSOA to the appropriate percentile within countries. The 1st percentile includes the most deprived LSOAs and the 100th percentile the least deprived LSOAs in each country.

Deprivation Index	**England**	**Wales**	**Scotland**	**Northern Ireland**
IMD	Summary IMD	Summary IMD	Summary IMD	Summary IMD
Income	Income domain	Income domain	Income domain	Income domain
Employment	Employment domain	Employment domain	Employment domain	Employment domain
Health	Health and disability domain	Health domain	Health	Health and disability domain
Education	Education, skills and training domain	Education domain	Education domain	Education, skills and training domain
Housing	Living environment domain	Housing domain	Housing domain	Living environment domain
Crime	Crime domain	Community safety domain	Crime domain	Crime and disorder domain.
Access	Barriers to housing and services domain	Geographical access to services domain	Access	Proximity to Services domain

3.3.7 School absences

Source	Adapted from: Department for Education (2012)
Data retrieved from	http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do (Neighbourhood Statistics -> Education, Skills and Training -> Pupil Absence in Schools by Gender in England (Referenced by Location of Pupil Residence))
Table/file reference\	Pupil Absence in Schools by Gender in England (Referenced by Location of Pupil Residence), 2011/2012 E970312_2716_GeoPolicy_LSOA E970312_2716_GeoPolicy_LA
Time period	2011/12
Geographic coverage	England
Level of aggregation	LSOA , LAD

Source	Adapted from: Stats Wales (2012)

Data retrieved from	https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Absenteeism/absenteeismbypupilsofcompulsoryschoolageinmaintainedsecondaryandspecialschools-by-localeducationauthority-year
Table/file reference	Absenteeism by pupils of compulsory school age in maintained secondary and special schools by local education authority and year
Time period	2011/12
Geographic coverage	Wales
Level of aggregation	LAD

Source	Adapted from: Scottish Government (2013)
Data retrieved from	http://www.gov.scot/Topics/Statistics/Browse/School-Education/AttendanceAbsenceDatasets/attab2013
Table/file reference	Attendance and Absence 2012/13
Time period	2012/13
Geographic coverage	Scotland
Level of aggregation	LAD

Source	Adapted from: NISRA (2015c)
Data retrieved from	http://www.ninis2.nisra.gov.uk/public/Theme.aspx?themeNumber=130&themeName=Children%20Education%20and%20Skills
Table/file reference	Attendance rates for post primary pupils by pupil residence (administrative geographies) Sheet: LGD
Time period	2012/13
Geographic coverage	Northern Ireland
Level of aggregation	LGD

Data on school absences were included as a possible indicator of parental engagement, which in turn may be correlated with their willingness to participate in surveys. Data cover secondary school absenteeism which is generally higher than, but highly correlated with, primary school absenteeism at local level. The Department for Education focuses on two key measures of absence: proportion of total half day sessions missed due to absence and proportion persistent absentees (pupils count as a persistent absentee if they miss around 15% of sessions). These two measures are highly correlated at OA level in England (DfE, 2011). The focus here is on proportion of total absences (authorised + unauthorised) as this figure (unlike proportion of persistent absentees) is available across the UK. Sub-

national figures are calculated based on the pupil's area of residence rather than where they go to school. Data are appended using LAD codes and the NSPL.[18] Data for England only are available at LSOA as well as LAD level and were similarly appended using LSOA codes and the NSPL.

Figures are produced separately for the four countries of the UK. Absence figures are compiled slightly differently across the four countries of the UK and refer to slightly different time periods. In Wales for example figures are for all secondary school age pupils (11-15) and include pupils of this age attending special schools. In Scotland figures are for all maintained secondary schools, including pupils in S6 (aged 16-17). Scottish data are biennial only and were not available for 2011/12. In Northern Ireland figures are for post primary school pupils, years 8-12 which corresponds to compulsory education in NI. Data were not available for 2011/12 so data from 2012/13 were used. All statistics are based on school returns and the quality of reporting may vary by school/local authority within countries

3.3.8 Out of work benefit claimants

Source	Adapted from: Department for Work and Pensions (2016)
Data retrieved from	https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp
Table/file reference	DWP Benefits -> WPLS-> benefit claimants - working age client group DWP Benefits -> WPLS-> benefit claimants - working age clients for small areas Figures are calculated as rates/proportions and cover total claimants
Time period	Nov 2011, Feb 2012, May 2012, Aug 2012
Geographic coverage	Great Britain
Level of aggregation	2003 CAS ward, LAD

Source	Adapted from: Nomis (2016)
Data retrieved from	https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp
Table/file reference	Jobseeker's Allowance -> Jobseeker's Allowance with rates and proportions
Time period	Nov 2011, Feb 2012, May 2012, Aug 2012
Geographic coverage	Great Britain
Level of aggregation	2003 CAS ward, LAD

Source	Adapted from: Department for Communities (NI) (2012)
Data retrieved from	https://www.communities-ni.gov.uk/topics/benefits-statistics

[18] Codes for Northumberland and Stevenage Unitary Authorities and Welwyn Hatfield differed between the NSPL and the source statistics and had to be bridged using the Code History Database.

Table/file reference	Client Group Analysis Table WA11 (Benefit claimant numbers) Table WA10 (mid-year population figures to calculate rates)
Time period	Feb 2012, May 2012, Aug 2012
Geographic coverage	Northern Ireland
Level of aggregation	LGD

Benefits data were intended to provide a proxy measure for workless households and so measures focus on out of work benefits paid to people of working age. Data are available separately for Great Britain and Northern Ireland. Data are published quarterly; we summed data from Nov 2011, Feb 2012, May 2012 and Aug 2012 to produce 12-month figures for 2011/12.[19] Data are available at LAD level for Great Britain and Northern Ireland and by 2003 CAS ward for Great Britain only. Data were matched using LAD/2003 CAS ward codes and the NSPL (Nov 2014).[20]

Data for Great Britain were downloaded from the nomis website (www.nomisweb.co.uk). Data on the proportion of the resident population aged 16-64 claiming any out of work benefits are supplied by the Department for Work and Pensions (DWP) and collected via the Work and Pensions Longitudinal Study covering 100% of benefit claimants. Data on the proportion of the resident population aged 16-64 claiming Jobseeker's Allowance are derived from figures supplied by Jobcentre Plus local offices. These are not official statistics for unemployment but are the only data available for small areas.

Data for Northern Ireland were obtained from the Northern Ireland Benefits Statistics Summary produced by the Department for Social Development. Figures break down the number of claimants across a range of benefits (including Jobseeker's Allowance) and are provided alongside mid-year population estimates for the working age population allowing rates to be calculated. Calculations for the proportion of the working age population (16-64 for men and 16-59 for women) claiming out of work benefits includes claimants of Jobseeker's Allowance, Employment Support Allowance, Lone parent benefit, other income related benefits.[21]

3.3.9 Energy consumption

Source	Department of Energy and Climate Change (2015a)
Data retrieved from	https://www.gov.uk/government/statistics/lower-and-middle-super-output-areas-electricity-consumption
Table/file reference	LSOA/MSOA 2011 historic electricity consumption (Sheet: MSOA domestic electricity estimates 2011)
Time period	2011

[19] NI data were a 3 quarter average Feb 2012, May 2012, Aug 2012.
[20] Codes for Northumberland and Stevenage Unitary Authorities and Welwyn Hatfield differed between the NSPL and the source statistics and had to be bridged using the Code History Database.
[21] In the breakdown figures, individuals are allocated to only one statistical group, based on the main reason they are claiming benefits, so there is no double counting of those on multiple benefits.

Geographic coverage	Great Britain
Level of aggregation	MSOA

Source	Department of Energy and Climate Change (2015b)
Data retrieved from	https://www.gov.uk/government/statistical-data-sets/experimental-sub-national-domestic-electricity-consumption-statistics-for-northern-ireland-2009
Table/file reference	Northern Ireland sub-national domestic electricity consumption 2008-2013 (Sheet: 2011)
Time period	2011
Geographic coverage	Northern Ireland
Level of aggregation	Local government district

Data on energy consumption may be of interest as a proxy for household size and/or socio-economic resources. The government (formerly Department of Energy and Climate Change, now Department for Business, Enterprise and Industrial Strategy) publishes sub-national figures on domestic gas and electricity consumption for Great Britain. Similar data on electricity consumption are available for Northern Ireland. Electricity and gas data are based on real consumption recorded from meters which is then aggregated upwards to local areas. The measure we include in our dataset is average domestic electricity consumption (Kwh per meter). Data are at MSOA level for Great Britain and Local Government District in Northern Ireland.[22] Some care is needed when interpreting these data given that electricity consumption is likely to depend on the availability of gas central heating, something which varies across areas. A more accurate measure of energy consumption would perhaps combine gas and electricity consumption and/or at least control for the availability of mains gas within an area.[23]

Data for England, Wales and Scotland were matched using 2001 MSOA codes obtained from the NSPL. Data from Northern Ireland were matched using Local Government District (LGD) codes, also available via the NSPL.

Data for some MSOAs in England were omitted from the published statistics because of disclosure risk. These omissions led to missing data for 13 cases in the full ADDResponse dataset (including nonrespondents).

[22] Data for Great Britain are also now available at LSOA level via www.gov.uk (though not in the ADDResponse dataset).
[23] The census provides estimates of the proportion of households with access to mains gas but this information is not available in the ADDResponse dataset.

3.3.10 Personal wellbeing

Source	Adapted from: Office for National Statistics (2014b)
Data retrieved from	http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/wellbeing/measuring-national-well-being/personal-well-being-in-the-uk--2013-14/sb-personal-well-being-in-the-uk--2013-14.html
Table/file reference	Reference Table 1: Personal Well-being estimates geographical breakdown, 2013/14 (Excel sheet 646Kb)
Time period	2013-14
Geographic coverage	UK
Level of aggregation	LAD

Policy makers are paying increased attention to subjective wellbeing and its potential impact on a wide range of personal and social outcomes. Feelings of wellbeing and a positive emotional state have for example been linked to helping behaviour, which may in turn translate into survey participation (Groves et al, 1992). ONS publish four headline measures of personal wellbeing based on data collected via the Annual Population Survey (APS), the largest constituent survey of the Integrated Household Survey. The sample size of the 12 month APS dataset is approximately 165,000 adults aged 16 and over and living in residential accommodation in the UK. Survey respondents are asked to rate how happy they felt yesterday, how anxious they felt yesterday, whether they feel their life is worthwhile and whether they are satisfied with life, all on a 0-10 scale. Responses are then averaged to provide estimates of national wellbeing.

In 2013/14 ONS published sub-national estimates of wellbeing - down to LAD level - for the first time. Despite the fact that these data cover a period after the behaviour (participation in ESS Round 6 2012/13) that we wish to study, and are at a high level of aggregation it was considered worthwhile to include them on an exploratory basis. We include variables for the average (mean) rating (0-10) for each of the four wellbeing measures.

Data were appended to the ADDResponse dataset using LAD codes and the NSPL.[24]

3.4 Small-area data from other sources

As well as government statistics, several other sources of small-area data were also explored as part of ADDResponse.

[24] Codes for Northumberland and Stevenage Unitary Authorities differed between the NSPL and the source statistics and had to be bridged using the Code History Database.

3.4.1 Local elections data

Source	Local Elections Centre, University of Plymouth
URL:	http://www1.plymouth.ac.uk/research/ceres/tec/Pages/default.aspx
Time period	2010-2013
Geographic coverage	Great Britain
Level of aggregation	Electoral ward (2014)

The Local Elections Centre at the University of Plymouth provided the research team with ward level data on turnout and party vote shares in local elections in response to a bespoke request. Data refer to the election that took place closest to 2012 and cover the period 2010-13.

Producing a dataset for the whole of Great Britain was a time consuming task. Local elections take place at different times both across and within local council areas and it was necessary to piece together a dataset based on the results of the nearest election to 2012. Separate data files were supplied for Wales, Scotland, London, other metropolitan areas within England and English council districts. A look up file was provided enabling us to map the electoral ward and local council codes provided in the electoral files to ONS census wards. Census wards were then used to append the electoral data to the ESS sample file using the NSPL. Identifying the correct census ward was not always straightforward as matching was done using text strings and the strings available in the different sources were not always a perfect match (e.g. use of "&" and "and" in ward names).

There is some missing data; 7% or 310 cases from the full dataset (including nonrespondents) have missing data on vote share and 5% or 225 cases have missing data on council control. This is for a variety of reasons. Changes in electoral boundaries mean it was not always possible to map electoral wards to the census wards given in the NSPL. This was the case for addresses in our sample located in Hartlepool and Swindon Unitary Authorities. Some data are missing where certain wards within an authority follow a different electoral cycle to the majority (and so data were not supplied) or were uncontested at the relevant election.

3.4.2 Personality traits

Source	University of Cambridge. Department of Psychology, British Broadcasting Corporation (2015)
Data retrieved from:	http://dx.doi.org/10.5255/UKDA-SN-7656-1.
Time period covered	2009-11
Geographic coverage	Great Britain
Level of aggregation	LAD

An innovative academic study which mapped personality types across Britain according to the Big Five Inventory (BFI) has published data on average personality traits for each LAD in Great Britain (Rentfrow et al, 2015). The Big Five personality traits are: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness (Goldberg, 1992; John and Svrivastrava, 1999).

Data were collected via a large scale opt- in web survey – the BBC Big Personality Test 2009-11 – with just over 386,000 respondents aged 18+. Respondents completed the Big Five Inventory (BFI) of 44 short statements coded on a 5-point Likert scale (strongly agree to strongly disagree). The BFI was used to calculate respondents' scores on each of the 5 personality dimensions and local authority averages were calculated from the means of the un-weighted scale scores of respondents who reported living in a local authority at the time at which they participated in the survey. Variables in the ADDResponse dataset are expressed in terms of the local authority personality trait T-scores on each dimension. Data were appended using LAD code and the NSPL. [25]

The data collection methodology for the personality study has clear limitations - an opt- in web survey may not necessarily be representative of the whole population - and data are at a high level of aggregation. Nevertheless, it was considered a novel source of contextual data worth exploring. The study authors posit for example that personality traits at local authority level are correlated with both health and political outcomes (conservative vote share) (Rentfrow et al, 2015).

Data from the BBC Big Personality Test 2009-11 have been archived with the UKDA [SN: 7656] and are available for reuse.

4 Local geographic information

In addition to using small-area data to explore the socio-demographic context in which ESS sample units are located, Geographic Information Systems (GIS) can be used to locate sample units in relation to the built environment (e.g. transport networks or local environments), natural geography (e.g. rivers, mountains) or other environmental features such as noise or air pollution levels. There is a growing body of research linking survey and geospatial data to explore how environmental factors influence health and wellbeing (see, for example, van Praag and Barsma, 2005; Welsch, 2006) or how the built environment influences attitudes and behaviour. The prevalence of fast-food outlets has been linked to higher rates of child obesity for example (Fraser and Edwards, 2010), whilst the presence of an "evening economy" may be associated with a stronger perception of antisocial behaviour (Taylor et al, 2014). The 2007 INSPIRE Directive which provides a framework for sharing data to support environmental policies across the EU has facilitated increased use of spatial data, particularly cross-nationally.[26]

The main way in which ADDResponse used geospatial data was to explore opportunities to locate ESS sample units in relation to "Points of interest" or local amenities. This information may be relevant to studying respondent behaviour in two ways: First, the prevalence of certain types of amenities may serve as an indicator of the socio-demographic profile of the area with, for example, fast-food outlets more prevalent in more deprived and ethnically diverse areas (Bodicoat et al, 2014). Second, proximity to certain types of amenity may have a direct bearing on individual

[25] Codes for Glasgow City, Northumberland, Welwyn Hatfield and Stevenage differed between the NSPL and the source statistics and had to be bridged using the Code History Database.
[26] http://inspire.ec.europa.eu/

attitudes and behaviour, with, for example, access to green space tending to be associated with higher wellbeing and increased social contact (Maas et al, 2009).

4.1 Geospatial datasets used

The following geospatial datasets were explored as part of ADDResponse:

4.1.1 Ordnance Survey Points of Interest

Source	Ordnance Survey (OS)
URL	https://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html
Time period	September 2014. Dataset updated quarterly.
Geographic coverage	Great Britain
Level of aggregation	Location relative to postcode centroid

Ordnance Survey Points of Interest is a location-based gazetteer of all public and private owned businesses, leisure facilities and educational establishments in Great Britain, compiled and managed by PointX on behalf of OS. Over 4 million points of interest (POIs) of over 600 different types are recorded in the database along with their address and other location information. POIs are classified into a three-tier hierarchy comprising nine groups, 52 categories and over 600 classes. The nine groups are: Accommodation, eating and drinking; commercial servicea; attractions; sport and entertainment; education and health; public infrastructure; manufacturing and production; retail; and transport.

The database comprises data from around 150 suppliers though the major ones are Ordnance Survey (42% of POIs), 118 Information (34%) and Department for Transport (9%). The database is updated quarterly.

ESS sample addresses can be located in relation to the various POIs. Different measures can be derived including counts of the number of a particular POI within a given area, distance to the nearest POI of a particular type or the ratio of one type of POI to another within a given area (Macyntire et al, 2008). The research team made a selection of which POIs to focus on and which metrics to use to analyse them.[27] Decisions were informed by a combination of existing research using POIs to analyse behaviour and theories of survey nonresponse. There was particular interest in using POIs to capture aspects of the local environment not captured by other administrative sources. For example, much small-area data focuses on deprivation whereas POIs could potentially be used to identify more affluent areas via amenities such as golf courses that have been found to be associated with higher house prices (Giles-Donovan and Corti, 2002) or a higher concentration of specialist food shops, found in the USA to be associated with more affluent, predominantly white areas (Moore and Diez Roux, 2006).

[27] The decision making process entailed four members of the research team each making an independent selection of POIs before coming together to discuss those choices and make a final selection.

Five different POI metrics were used to derive variables depending on the POI under consideration. All metrics take the postcode centroid as the reference point for calculation[28]:

A: Prevalence within MSOA: Prevalence measures were used for POIs i) whose presence may be indicative of a certain type of area e.g. gambling outlets, pawnbrokers, discount stores, fast food outlets, frozen food stores being more common in deprived areas (Macyntire et al, 2008) ii) where a higher concentration might influence behaviour or attitudes e.g. a higher concentration of pubs, bars and nightclubs leading to more perceived antisocial behaviour (Taylor et al, 2014). A measure based on administrative boundaries i.e. MSOA was chosen for comparison with other auxiliary variables (e.g. from the census) and so that alternative measures of POI per population could be derived as required.

B: Prevalence within an 800m radius: An alternative measure of prevalence based on physical boundaries rather than administrative boundaries. 800 meters - or 10 minutes' walk - is the scale commonly used by town planners and local authorities for resource allocation.

C: Prevalence within a 10km radius: An alternative measure of prevalence based on physical boundaries rather than administrative boundaries, this time assuming a radius within 15 minutes driving time.

D: Ratio: (e.g. proportion of eating and drinking venues within an MSOA which are fast-food outlets, proportion of food shops within an MSOA that are frozen food stores): This measure provides another way of characterising areas over and above straightforward prevalence.

E: Distance to nearest: This measure was used for POIs such as golf courses, green space, water, and transport links which are considered to be desirable to have close by and whose proximity may be judged beneficial, for example, contributing to wellbeing or rising house prices (Luttik, 2000; Gibbons and Machin, 2005; Maas et al, 2009).

F: Present within a 1km radius: This measure was used for POIs such as waste facilities or industrial plants which are likely to create negative externalities (noise, pollution etc.) and so have a negative effect on health and wellbeing or house prices through proximity (de Voor and de Groot, 2011).

Details of all of the OS POI variables included in the ADDResponse dataset are provided in the excel file "ADDResponse variable listing.xls" made available alongside this report.

Data were made available to the research team by Ordnance Survey under a Research Data Agreement for the nominal cost of £1 (enabling a contract to be put in place). The data can be used for the purposes of the ADDResponse project but cannot be distributed beyond that.

[28] The postcode centroid is the mean grid reference of all addresses in that postcode, snapped to the nearest property - the geographic centre of the postcode

4.1.2 OpenStreetMap Points of Interest

Source	OpenStreetMap (crowdsourced)
URL	http://www.openstreetmap.org
Time period	Dataset continuously updated. Dataset downloaded March 2015
Geographic coverage	UK (Worldwide coverage available)
Level of aggregation	Location relative to postcode centroid

OpenStreetMap (OSM) is a global open-source map generated from crowd-sourced information. It contains information about transport networks (roads, railways, paths, waterways, cycle routes) and geographical features found along those routes. The majority of information is contributed by members and updates are made to maps within minutes of having been logged. In addition to volunteered information, OSM has agreements with several internet companies including Yahoo and Microsoft Bing to use information from their satellite imagery whilst government data on road networks has been imported into OSM in several countries.

OSM has worldwide coverage which makes it a particularly useful source of data for cross-national surveys such as the ESS. Levels of coverage do vary across countries however: coverage is relatively good in North America and Europe (especially the UK and Germany) but tends to be less comprehensive in other parts of the world (Neis and Zielstra, 2014). As of March 2015 there was a community of over 2,000,000 registered users. [29]

POIs for analysis were selected according to a similar process to that used with the OS POI database. OSM contains different POI listings to the OS database; many are less detailed but others (such as places of worship) are classified in more detail.[30] Details of all of the OSM POI variables included in the ADDResponse dataset are provided in the excel file "ADDResponse variable listing.xls" made available alongside this report.

The dataset is made freely available under the Open Database License.[31]

4.1.3 AddressBase

Source	Ordnance Survey
URL	https://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-products.html
Time period	Dataset updated every six weeks. Dataset supplied November 2014
Geographic coverage	Great Britain
Level of aggregation	Record matched to exact address

[29] http://wiki.openstreetmap.org/wiki/Press_Kit
[30] Examining how sampled addresses are located relative to different places of worship may, for example, suggest something about the ethnic and/or religious makeup of the local area.
[31] http://opendatacommons.org/licenses/odbl/

Ordnance Survey AddressBase products make it possible to append additional information about a property to a single address and to map the information. There are three products - AddressBase, AddressBase Plus and AddressBase Premium - containing increasing amounts of information. AddressBase Plus contains over 28 million addresses from the Royal Mail Postcode Address File plus additional records from local authorities, the Valuation Office Agency and OS. Among other information it classifies addresses as commercial or residential and identifies multi-occupancy dwellings. It also contains OS MasterMap® topographic identifiers (TOID®) enabling other OS data to be appended to address records.

One of the main features of AddressBase - classifying properties into residential or commercial - will not be required by surveys using the PAF as this information is already taken into account when the sample is drawn. However, being able to identify whether a property is multi-occupancy - through the number of "child" addresses associated with a "parent" address could prove useful. A variable to this effect was included on the ADDResponse dataset.

AddressBase Plus records were also used to create a measure of "generalised accessibility" as an alternative to population density. Generalised accessibility estimates the distance from the sampled address' postcode centroid to *n* locations of individual addresses at their actual address location. The further the distance to the nearest n addresses the more remote/less accessible an address can be considered. Measures of generalised accessibility at different scales were created - distance in meters to the nearest 200, 500, 1000, 2000, 5000 and 10000 addresses.

AddressBase Plus covers mainland Great Britain. A separate product, AddressBase Islands, covers Northern Ireland, the Channel Islands and the Isle of Man. However, the geographic coverage afforded by AdressBase Plus was considered sufficient for the exploratory analysis carried out as part of the ADDResponse project.

AddressBase Plus is made available free of charge to public sector institutions under the Public Sector Mapping Agreement (the PSMA). Different arrangements are in place for commercial organisations, although it is possible to download data for a free trial to determine whether the data meet business needs. The PSMA provided access to the research team but means that the data cannot be reproduced or archived by the team for reuse.

4.1.4 OS MasterMap®Integrated Transport Network™ Layer

Source	Ordnance Survey (OS)
URL	https://www.ordnancesurvey.co.uk/business-and-government/products/itn-layer.html
Time period	Dataset updated every six weeks. Dataset supplied November 2014
Geographic coverage	Great Britain
Level of aggregation	Record matched to exact address

The OS MasterMap® Integrated Transport Network™ Layer is a dataset mapping out more than 550,000 Km^2 of Great Britain's road networks. We were originally interested in using this information to calculate travel times between sampled addresses and other POIs. However, whilst technically

possible, this proved beyond the programming resources available for this particular project. The dataset was used as a source of information regarding the type of road - Motorway, A road, B road, minor road, local street, alleyway, private roads – publicly accessible, private road – restricted access, pedestrianised streets on which each sampled address - was located.

OS MasterMap® is made available free of charge to public sector institutions under the Public Sector Mapping Agreement (the PSMA). Different arrangements are in place for commercial organisations, although it is possible to download data for a free trial to determine whether the data meet business needs. The PSMA provided access to the research team but means that the data cannot be reproduced or archived by the team for reuse.

4.1.5 GeoLytix

Source	GeoLytix
URL	http://geolytix.co.uk/blog/open-supermarket-locations/
Time period	Dataset downloaded 19th March 2015
Geographic coverage	UK
Level of aggregation	Location relative to postcode centroid
Copyright	Open Supermarkets © GeoLytix copyright and database right 2015 Contains Ordnance Survey data © Crown copyright and database right 2015 Contains Royal Mail data © Royal Mail copyright and database right 2015 Contains National Statistics data © Crown copyright and database right 2015

Commercial firms may also hold geospatial data which they are happy to make available. GeoLytix, for example, is a consultancy firm which specialises in helping firms with customer targeting and location planning and which makes georeferenced data available for these purposes. This includes an "Open Supermarkets" dataset (now renamed Retail Points) which provides the name and exact location of all supermarket retailers in the UK and is regularly updated to reflect store opening/closures/renaming. Knowing where a sample point is in relation to different types of stores (e.g. budget supermarkets vs. Waitrose) may provide an indication of the type of area and, for example, whether it has been subject to recent gentrification (the "Waitrose effect".)[32]

The Open Supermarkets database is available open access via the GeoLytix website, covered by the Open Government License for Public Sector Information.[33]

[32] http://www.bbc.co.uk/news/magazine-24629300
[33] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/

4.2 Data linkage

POI data were appended to the ADDResponse sample file using British National Grid coordinates. Coordinates available within the ADDResponse dataset refer to the sampled address' postcode centroid. Selected POIs were extracted from the full OS POI database (supplied on CD as an ASCI pipe separated text file) using MySQL and then converted into a shapefile using ArcGIS software. A shapefile is an Esri vector data storage format for storing the location, shape, and attributes of geographic features, including spatial coordinates such as grid references. Spatial joins were used to derive variables for analysis - e.g. distance to nearest or number of POIs within a given radius - based on the POIs' relative location to one another. [34] A similar procedure was used to append OSM POIs. OSM POIs were extracted from the online database using the OSM-API and then converted into a shapefile using ArcGIS. OSM POIs were originally georeferenced using a different system of coordinates than British National Grid (WGS84 Bounds used by GPS). However, it was possible to convert OSM coordinates into British National Grid coordinates using the GDAL (Geospatial Data Abstraction Library) translator library.[35] GeoLytix data were downloaded from the company's website, converted into a shapefile using ArcGis and appended to the ADDResponse dataset using British National Grid coordinates.

Some of the POI measures used for ADDResponse are expressed in terms of POIs within an MSOA. Administrative boundaries were overlaid on the POI data, downloaded from the UK Data Service Census Support.[36]

AddressBase Plus records were appended to the ADDResponse dataset using postcode, followed by text string matching on the full address within postcode. This achieved a match rate of 93%. The remaining cases are unlikely to be missing at random but are instead likely to be atypical addresses such as flats or other multi-occupancy dwellings. A variable indicating whether an exact match was found between the sample file address and AddressBase records may therefore prove useful in subsequent analysis.

The OS MasterMap® ITN Layer was converted into a shapefile using interpOSe.[37] Information on road type was appended to the ADDResponse dataset via AddressBase Plus. Each feature within OS MasterMap® is assigned a unique TOID® (Topographical Identifier) or series of TOIDs, typically associated with a polygon that represents the area on the ground the feature covers, using British National Grid coordinates. AddressBase records contain the TOID® for the road on which each address is located enabling road type information to be appended to addresses.

Once records had been appended summary statistics and maps of the POI and other geospatial data were generated to check that the data had been appended correctly. Extreme values were checked to ensure that they were correct. This preliminary analysis did not reveal any errors in the dataset. However, it did suggest that some of the measures we had originally planned to use were meaningless given the distribution of POIs and that alternative measures needed to be derived. For

[34] http://desktop.arcgis.com/en/arcmap/10.3/tools/analysis-toolbox/spatial-join.htm
[35] http://www.gdal.org
[36] https://census.edina.ac.uk/easy_download.html
[37] https://digimap.edina.ac.uk/webhelp/digimapgis/arcgis/importing_mastermap/using_interpose.htm

example, it became clear that a measure of POIs per Output Area would be limited by the very high proportion of OAs containing no POIs (of the selected type) so instead we focused on measures giving the number of POIs within an 800m radius. In addition, prevalence measures based on counts were very highly correlated both with one another and with population density as multiple POIs tend to be concentrated in urban centres. This suggested that measures of relative concentration such as ratios of one POI type to another may perhaps be more useful in analysis. (See Section 4.3 for further details).

4.3 Using geospatial data: Issues to be aware of

The use of geospatial data raises a number of issues relating to the quality and completeness of the datasets and to their usefulness in analysis.

4.3.1 Data quality

Ordnance Survey POI data are collated from over 150 different primary sources and so some caution is required when using the data. Based on a sample of 61,849 POIs drawn from the March 2010 dataset, it is estimated that 95% of POIs are located within 17.5 meters of their given location (Ordnance Survey, 2014). However, there may be gaps in the dataset and/or - depending on how the original provider has classified the POI - the same actual POIs may appear in different groups. As part of the production process, multiple instances of what is believed to be the same feature are removed but it is possible that individual instances of what are, in fact, the same type of feature can be reported in different classes. No information on how POIs are classified is provided beyond the name of the category and so there is room for ambiguity over the meaning/content of these categories. A study comparing the OS POI dataset against local authority records of amenities found evidence of duplicates within the POI dataset as well as possible under coverage, especially in rural and deprived areas (Burgoinne and Harrison, 2013). Nevertheless, whilst it may be feasible to obtain more accurate local authority data for small scale studies of particular municipalities or local authorities, given their accessibility OS data are likely to represent the best source of data on local amenities for a country wide study.

OS POI data arguably still provide higher quality data than crowd-sourced data from OSM (Hacklay, 2010) although the quality of crowd-sourced data are likely to improve over time as the number of contributors grows and updates and corrections to the data are made.[38] There is considerable interest in the growth of crow-sourced GIS data, what Goodchild (2008) terms “volunteered geographic information” or VGI. The reliance on volunteers without formal training in mapping and not subject to centralised controls raises questions about the accuracy and completeness of such data (see Elwood et al, 2012 for an overview of issues associated with this type of data). Data quality can be very good for some features; road networks have received particular attention, for example, with Haklay (2010), estimating an average of 80% overlap between OSM and official sources as regards motorway placement in England and up to 69% coverage in some areas. However, quality and coverage are variable, the majority of contributions tend to be made by a small core of “power” users, and there tend to be coverage blackspots (Neis and Zipf, 2012). Coverage tends to be highest in more affluent areas with a better educated population where

[38] Though increased popularity also comes with risks of data vandalism (see Neis and Zielstra, 2014).

volunteers are more active and diligent, for example, and where communities are more close-knit and social trust is higher (Haklay, 2010; Kesler and de Groot, 2013). For this reason, levels of data coverage for VGI may in themselves be useful metadata, indicating something about the socio-demographic make-up of the local area.

4.3.2 Deriving variables for analysis

One of the main challenges associated with using geospatial data - particularly POI data - may be to derive suitable variables for analysis. A selection must first be made from among the 600 plus POIs available and then decisions made about the best metric for those POIs. A range of different decisions must be made including whether to focus on measures of prevalence or distance, whether distances are Euclidian i.e. "as the crow flies" or based on travel times by road, the scale of the area to be considered and whether to be guided by administrative boundaries or distances when defining neighbourhoods (Macyntire et al, 2008). The choice of metric will depend on the POIs of interest and the specific research question to be answered.

It should be borne in mind that the majority of POIs are concentrated in densely populated, urban areas leading to skewed distributions across the country and high correlations between straightforward counts or distance measures for different POIs and between these measures and population density (see Figure 4 below for cartograms comparing the distributions of selected POIs across ESS PSUs). This again puts the onus on the researcher to think carefully about their research question and the appropriate selection of variables. Many, though by no means all, studies exploring the effects of the built environment using POI data focus on a particular town or city (see, for example, Macyntire et al, 2008; Fraser and Edwards, 2010) and indeed this sort of local level analysis, comparing relative distributions within a particular urban area, may be more straightforward and informative than trying to conduct comparative analysis across the whole country (it also makes data verification more straightforward - see Section 4.3.1 above).

4.3.3 Timing

The geospatial datasets mentioned here are frequently updated to reflect changes in amenities on the ground. Historical data are not available so any data obtained provides a snapshot of the locality at a particular point in time. This means that the data are likely to be most useful if appended to survey records either immediately before or immediately after data collection rather than at a subsequent point in time for secondary analysis.

Figure 4: Distribution of OS POIs across ESS Primary Sampling Units

POIs: From Top left clockwise: Number of pawnbrowkers within 800m, number of discount stores within 800m, number of gambling outlets within 800m, population density within LSOA, number of fastfood outlets within 800m, number of pubs, bars, clubs within 800m.

5 Data from commercial vendors

One potential source of data from the commercial sector is the vast array of information commercial companies hold about individuals and households through transactional data i.e. billing information, store loyalty cards etc. However, these data are likely to cover only the subset of the population that is registered with that particular company. Another option would be to use directory listings such as 192.com or uklocalarea.com. However, at present it is not possible to carry out batch processing and obtain data for large numbers of records simultaneously; these data sources may be better suited to running enquiries for specific sample records. A third option - and the one pursued by ADDResponse - is to purchase data from what Dugmore (2010) calls "value added resellers" such as Experian who collate and resell information provided by other organisations for consumer marketing purposes and/or for credit referencing. These companies' databases offer both broad population coverage and access to a wide array of variables.

Purchasing such data appears attractive because, unlike the more usual sources of small-area auxiliary data, it provides access to individual or household level information. Furthermore, datasets are frequently updated and so offer access to timely data. However, such data are also likely to present some challenges regarding completeness, accuracy and transparency. Several studies have considered the possibility of using such data in survey research, either for exploring survey nonresponse or for targeted sampling, in a US context (Pasek, et al, 2014; Sinibaldi et al, 2014; West et al, 2015) and in Germany using Microm data (Schrapler et al, 2010). ADDResponse represents the first systematic exploration of appending such data to survey records in a UK context.

5.1 Datasets used

ADDResponse purchased data from two "value added resellers": Experian and Callcredit. There are other companies within the UK and worldwide who perform similar functions and hold similar databases including CACI and Equifax. Experian is perhaps the most well-known of such companies in the UK. We wanted data from a second company to serve as a comparison and investigate the extent to which data sources varied across companies. Callcredit was selected as that comparator following helpful preliminary exchanges with their data team.

The research team obtained data dictionaries from each company and selected a subset of variables from among the large amount of data held. Variables available cover a range of topics - socio-demographics, consumer behaviour, financial status - and are a combination of segmentation or classification variables (e.g. Mosaic), variables providing information on specific demographic characteristics e.g. age, gender, employment status, and propensities defining likely consumption and spending habits. When selecting variables, priority was given to variables giving information on specific characteristics rather than segmentation or propensity variables, the true interpretation or derivation of which was not always clear. We also prioritised variables available from both companies (for comparison purposes) and variables where the rate of missing data was relatively low. There was interest in exploring some of the consumer behaviour variables (e.g. newspaper readership, preferred supermarket, eco purchasing) as these may tell us something interesting about

the household. However, many of these variables were missing for 50% or more of addresses and so were not viable.

Variables selected include: Number of adults in household, age, gender and employment profile of residents, whether children present in the household, marital status, household income, property value, property council tax band, tenure, length of residency, and whether anyone in the household is experiencing financial stress. We also included the major geodemographic segmentations produced by each company; Mosaic (Experian) and CAMEO (Callcredit). Further details of all the commercial variables included in the ADDResponse dataset are provided in the excel file "ADDResponse variable listing.xls" made available alongside this report.

5.1.1 Experian ConsumerView

Source	Experian ConsumerView (2015)
URL	http://www.experian.co.uk/marketing-services/products/consumerview.html
Time period	Dataset supplied 17/02/2015. Experian records updated monthly
Geographic coverage	UK
Level of aggregation	Address

Experian's ConsumerView database covers over 49 million adults (18+) in the UK and contains over 500 variables (Experian, 2015). This includes Mosaic, Experian's comprehensive consumer classification which divides the population into 15 groups and 66 types (Experian, 2014).[39] Mosaic classifications are available at both postcode and household level. ConsumerView data are drawn from a combination of the Edited Electoral Roll, including updates from the monthly rolling register, Experian's proprietary data assets, partnerships with other data owners and other compliant data sources. This includes drawing on census summary statistics and other open data sources such as the Land Registry. The most common demographic and household variables are updated monthly with other variables updated six monthly. All records included in the database are known to have been active at the registered address in the past 12 months.

Many variables are derived using data from a number of sources and statistical modelling rather than reporting actual data values. Limited information about the ConsumerView database and the origin of the variables it contains is made publically available. However, some further information about how specific variables are produced was supplied on request. One of Experian's proprietary data assets, the Canvasse Lifestyle Survey Programme, collects data from individuals and households through a series of mail, telephone and online surveys covering a range of consumer behaviours. These surveys, which include self-reported demographic variables, can be used to construct statistical models to predict demographic characteristics where direct information is not available. Age for example is predicted by a combination of forename, household composition, length of

[39] Two versions of Mosaic are available - MOSAIC UK and MOSAIC Public Sector. The labels and ordering of the categories varies across the two classifications with the former aimed at commercial companies and the latter at public sector service provision. However, at least at the group level, ESS sampled addresses were clustered in the same way under both classifications. This suggests that either classification could be used interchangeably.

residency, and postcode Mosaic. Tenure and the presence of children in the household are both predicted through household pixel which contains information on age, household composition, length of residency, company directorships, type of shareholdings and type of property.

Data were purchased under license. The license holder is usually allowed access to the data for 12 months although we were able to negotiate access for 24 months for the purposes of this research project.

5.1.2 Callcredit Define database

Source	Callcredit Information Group (2015)
URL	http://www.callcredit.co.uk/products-and-services/consumer-marketing-data
Time period	Dataset supplied 17/02/2015. Call Credit records updated every two weeks.
Geographic coverage	UK
Level of aggregation	Address

Callcredit InformationGroup make available a range of demographic, financial and consumer behaviour variables through their Define database of over 46 million marketable contacts aged 18+. This includes CAMEO, a family of household and postcode level segmentations covering different aspects of financial, consumer and lifestyle behaviour.[40]

Data were purchased under license. Licenses usually allow access to the data for 12 months. However, we were able to negotiate permanent access to the data supplied for the purposes of this research project.

5.2 Data linkage

Commercial data records were appended to the ADDResponse dataset using address. The commercial vendors were responsible for appending the data on behalf of the research team. Ipsos MORI sent Experian and Callcredit files containing the 4,520 sampled addresses along with a unique reference number for each case (not ESS "idno") which would enable returned records to be combined with the remaining ADDResponse data. No other data besides address and a randomised reference number was included in the files sent to the companies. All data were transferred securely as csv files and accompanied by a variable listing and codebook. The turnaround time was very quick (a couple of weeks once the specification and access arrangements were agreed) and contacts at the commercial companies were very responsive and happy to help with enquiries.

Commercial data records are at the individual level and contain a mixture of postcode, household and individual level information whereas the ESS sample file is address-based. The companies first

[40] http://www.callcredit.co.uk/products-and-services/consumer-marketing-data/segmentation-analysis. A mistake in the variable specification sent to CallCredit meant that we did not obtain access to the overall segmentation variable CAMEO UK as originally intended but did have access to CAMEOFinancialGroups and CAMEOIncomeGroups and CAMEO International, all of which are postcode level classifications.

identified all individual level records held for adults aged 18 and over where there was an exact match between the address on the ESS sample file and the address in the commercial data record. Experian were able to match at least one record for 4,093 addresses (91%). Callcredit were able to match at least one record for for 3,778 addresses (84%). In total, Experian matched 8,164 individual records and Callcredit 7,670. However, because the ESS sample is address-based and does not contain records of the individuals resident at each sampled address there is no way to assess the completeness of the person level match.

Before returning the dataset to Ipsos MORI, the commercial companies then transformed their individual-level records into an address-level data file. Experian did this by converting individual-level variables into household arrays i.e. they provided variables indicating the number of people per address recorded as having a certain characteristics e.g. being unemployed, male or aged 35-54. Callcredit took a different approach and instead provided a series of variables for each address giving the age of person 1, age of person 2….age of person n etc. The research team subsequently converted these variables to household arrays, in line with the Experian variables.[41]

5.3 Issues with commercial data

Being able to access individual or household level data across the UK is potentially valuable. However, there are a number of caveats associated with using data provided by "value added resellers".

5.3.1 Completeness

The commercial databases have wide coverage, with records for over 46 million adults across the UK. Furthermore, as discussed in Section 5.2 matching records to the ESS sample on the basis of address resulted in a relatively high match rate of between 80 and 90% addreeses. Nevertheless, there is some missing data. The extent of missing data at address level varies significantly between the two companies' databases and, especially in the case of Callcredit, depending on the variable in question (Table 1). There is also likely to be missing data within households - if not all resident individuals appear in the commercial databases. Without a sample of named individuals there is no way for ADDResponse to determine this. However, as discussed in Section 5.3.2 below, there is a lack of agreement between the datasets as regards the number of adults resident at each address suggesting that missing data at the individual level is likely to be an issue.

Callcredit data on commercial habits (e.g. newspaper readership) has a particularly high rate of missingness (around 50%) and so was not considered for this project whilst indicators of whether an individual has a landline and/or mobile telephone had 60-70% missing data.

[41] Experian were not willing to supply data in the format provided by Callcredit. Because the match records Ipsos MORI provided were for addresses, Experian were only prepared to supply address-level data in return.

Table 1: Percentage of addresses missing data in commercial databases

	Experian %	Callcredit %
Number of adults	9.4	16.4
Children in household	9.4	26.2
Age/gender of residents	9.4	16.4
Marital status	10.0	-
Employment status	9.4	-
Tenure	9.4	21.9
Council tax band	11.4	20.2
Segmentation variables	9.4	16.4

Base (all sampled cases) = 4,520. % missing for all commercial variables is given in the accompanying file "ADDResponse variable listing.xls"

Analysis suggests that data are not missing from the commercial databases at random. Perhaps reassuringly, data are particularly likely to be missing for ESS sampled addresses later found to be ineligible (e.g. commercial premises, derelict properties or properties under construction) as well as addresses which resulted in no contact during fieldwork (Table 2). The same groups who prove hard to reach for surveys also tend to be missing from other data sources. Missingness may therefore prove to be useful auxiliary information in itself (Smith, 2011) and may, for example, be useful to screen for ineligible addresses prior to fieldwork or for identifying potentially hard to reach targets. It should be noted, however, that it was not necessarily the same cases that were missing from both commercial databases. Only 50% of the addresses missing any Experian record were also missing a Callcredit record whilst only 28% of addresses missing any Callcredit record were also missing an Experian record.

Table 2: Percentage of addresses missing data in commercial databases, by survey outcome

	Survey response outcome				
	Ineligible	Non contact	Other nonresponse	Refusal	Completed interview
	%	%	%	%	%
Experian					
Missing	43.8	17.6	8.7	7.1	5.9
Not missing	56.2	82.4	91.3	92.9	94.1
Callcredit					
Missing	48.7	22.9	20.3	14.5	12.6
Not missing	51.3	77.1	79.7	85.5	87.4
N	265	306	231	1432	2286

Base: All sampled cases. % missing based on "Number of adults in household" variable which indicates the presence of any records for that address present in the database

An examination of levels of missing data among ESS survey respondents by their characteristics (Table 3) reveals few systematic differences. Differences on the basis of country of birth, education or income are non-significant, perhaps because survey respondents are already a self-selected group with the most marginal in society not covered by the survey. There is some evidence that commercial databases are more likely to miss out on people not living with a partner and people living in single person households (not surprisingly as the more people there are living at the address, the greater the probability of finding at least one matching record).

Table 3: Percentage of addresses missing data in commercial databases, by respondent characteristics

	Experian % missing	Callcredit % missing	N
All ESS respondents	9.4	16.4	2286
Household size			
1	11.8	20.2	642
2	4.5	10.2	831
3	2.3	10.3	348
4	2.0	9.4	299
5+	5.5	6.1	165
Marital status			
Living with a partner	3.4	10.6	1219
Not living with a partner	8.8	15.0	1066
Born in country			
Yes	5.9	12.4	2020
No	6.0	14.7	266
Income			
Lowest quintile	8.8	14.7	510
2nd quintile	5.3	10.3	339
3rd quintile	5.2	11.5	288
4th quintile	4.5	11.4	332
Highest quintile	3.4	15.2	328
Education			
Less than secondary education	7.7	11.7	545
Lower secondary	4.0	10.9	604
Upper secondary/vocational	6.3	10.9	607
Tertiary	6.3	18.6	442

Base: ESS respondents. % missing based on "Number of adults in household" variable which indicates the presence of any records for that address present in the database.

5.3.2 Accuracy

There are several sources of potential inaccuracies in the commercial data including inaccuracies in the original source data, errors in attributing values to the correct individual or address and errors which occur from deriving model estimates rather than using actual values (Pasek et al, 2014).

Experian estimate the accuracy of some of their key modelled variables as follows:

- Age (10 year bands) 70% to 90% accurate

- Tenure — 86% accurate (owner occupied properties); 43% accurate (private renters); 65% accurate (social renters)
- Children present in household — 81% accurate (no children); 51.5% accurate (children)
- Employment status — 55% accurate
- Financial stress — 63% accurate

One way to assess accuracy is to make comparisons across the two commercial data sources. Identifying which of the two sources provides the "best" estimate is difficult in the absence of a gold standard data source to compare against. However, if discrepancies exist this suggests reason to be cautious about one or both datasets. Comparing the small number of variables which are common across the two commercial data sources used for ADDResponse reveals significant differences between the two. Differences are particularly noticeable on the key variable of number of adults in the household i.e. number of matched records per address (Table 4). This suggests further reason to be cautious about the completeness as well as the accuracy of the two datasets as they may be missing records for some individuals resident at sampled addresses. Although it is not possible to know which of the two sources, if either, is correct, in the case of tenure for example, comparing the distribution of the commercial databases against the 2011 census suggests that, at least in the aggregate, Experian data provide a more realistic picture with the Callcredit data underestimating private renters.

Table 4: Level of agreement between two commercial databases

Variable	Agreement rate between Experian and Callcredit %	Kappa statistic of inter-source reliability (95% CI)	N
Number of adults in household (1,2,3,4,5+)	54.9	0.35 (0.32,0.37)	3564
Children present in household (Yes/No)	76.5	0.37 (0.33,0.40)	3208
Tenure (Owner, private rent, social rent)	74.5	0.41 (0.38,0.43)	3360
At least one person in household aged 18-25 (Yes/No)	82.5*	0.33 (0.29, 0.36)	3564
At least one person in household aged (Yes/No)	80.5**	0.57 (0.54, 0.60)	3564

Base: Cases present in both commercial databases.
* Majority of addresses (78%) were recorded as having no one of that age present
** Majority of addresses (64%) were recorded as having no one of that age present

Another way to evaluate the accuracy of the two datasets is to compare values within the commercial databases against self-reported survey responses (Pasek et al, 2014; West et al, 2015). Although survey estimates are themselves not without error, we would normally expect survey responses to demographic questions to be relatively accurate and any mismatch to cast doubt on the external data. Making comparisons between survey and commercial data is complicated in this instance; the timing mismatch between survey data collected in 2012/13 and commercial data obtained in 2015 (see Section 5.3 below) means that any discrepancy may be due to actual changes

in household circumstances or composition rather than errors in the commercial data. Furthermore, most ESS survey variables (with the exception of household size and age and gender of residents) relate to the respondent only - plus partner in the case of employment/education - whereas the commercial databases cover the whole household. This may lead to further discrepancies between the data sources without necessarily indicating errors in the commercial data. Nevertheless, comparing ESS survey data with demographic data obtained from the commercial databases may still be indicative of quality.

Table 5 suggests a relatively high level of correspondence between values in the commercial databases and responses given in the ESS, at least for broad yes/no indicators. We compare responses given by ESS respondents regarding marital status, whether they (or their partner) is retired, whether there are children present in the household and whether anyone in the household is aged 50+ to information on these same characteristics provided by the commercial companies. Agreement rates for whether children are present in the household (West and Kreuter, 2013) are similar to those found for interviewer observations of the same characteristic for example.

Whether data are considered sufficiently accurate to be useful depends on the purpose for which they are required. For example, if there is interest in sampling certain subgroups within the population, commercial data might still be useful for an initial screening of households or individuals likely to fall within those subgroups even if there is a degree of inaccuracy. Drawing a larger gross sample to allow for false positives in the commercial data and a high ineligibility rate in the field may still prove more cost effective than identifying subgroups through other means such as focussed enumeration (Barron et al, 2015).

Table 5: Level of agreement between commercial database and ESS survey responses

	Experian			Callcredit		
Variable	**Agreement rate between Experian and ESS %**	**Kappa statistic of inter-source reliability (95% CI)**	**N**	**Agreement rate between Experian and ESS %**	**Kappa statistic of inter-source reliability (95% CI)**	**N**
Children present in household* (Yes/No)	74.4	0.40 (0.36,0.44)	2150	71.6	0.29 (0.24,0.33)	1798
At least one person in household aged 50+ (Yes/No)	91.1	0.82 (0.79,0.85)	2150	81.1	0.60 (0.56,0.64)	1997
Marital status (Married/Not)	77.8	0.56 (0.52,0.60)	2143	-	-	-
At least one person in household retired (Yes/No)	89.2	0.76 (0/73, 0.79)	2150	-	-	-

Base: Cases present in both commercial database and ESS respondent dataset

* Commercial databases cover children 0-17 whereas ESS does not specify an age limit so it is not a like for like comparison

5.3.3 Timing

Commercial databases are frequently updated and it is not possible to negotiate access to historical records. This means that any data purchased provide a snapshot of individuals resident at addresses at the time of purchase. Studies wishing to make use of data from commercial databases should therefore make sure to purchase these data at the time the sample is drawn rather than subsequently.

Concurrent data purchase was not possible for ADDResponse, a project which was initiated post fieldwork; we purchased data in February 2015 to append to survey data collected in 2012/13. Analysis of Experian's "length of residency" variable suggests that at least 31% of all addresses in the ESS sample had undergone a change in household composition since ESS fieldwork took place; in 16.5% of cases the entire household had been resident for less than two years and in 14.5% of cases at least one person in the household had been resident at the address for less than two years. Furthermore, even in households where everyone has been resident since 2012 circumstances may have changed. The data still have the potential to be useful as, even if the specific individuals are not the same, similar types of people tend to reside at an address over time (at least in the short run). Nevertheless, the timing mismatch should be borne in mind when considering findings from the commercial data, particularly accuracy evaluations based on comparisons between variables in the commercial and survey datasets. Results *may* have been more consistent if the two datasets had been concurrent.

5.3.4 Transparency

Smith (2011, 393) notes that "many databases are 'black boxes' that do not disclose how they are constructed and what rules are followed." As discussed above, it is possible to obtain some information about how segmentations are created or variables modelled on request. However, this information remains fairly superficial whilst serving to emphasise how complex the derivation of even the simplest variables such as age can be. This may not be too big a problem if variables are required solely for prediction - as may be the case for example when constructing response propensity weights for surveys - but is more problematic for analysts interested in explanation.

5.3.5 Cost

Unlike the other data sources used as part of ADDResponse which are available free of charge, there was a financial cost associated with purchasing the commercial data. Cost was dependent on the number of records matched and the variables selected; segmentation variables such as Mosaic were most expensive to purchase. Costs varied significantly between the two companies but in both cases were small relative to the costs of conducting face-to-face survey fieldwork. There are limitations to the commercial data as discussed above and the potential for such data to add value to survey data requires further investigation of specific applications. Nevertheless, the costs associated with purchasing such data are unlikely to prove prohibitive if the data are found to be useful.

6 Appending auxiliary data: Lessons learned

The final chapter of this report reflects on some of key lessons that emerged from the ADDResponse project regarding the process of appending geocoded auxiliary data from external sources to address-level survey records. In particular we revisit three questions:

- What data are available to append to address-based survey samples using geocodes?
- How easy or difficult is it to append data to the survey sample?
- How suitable are the data for analysis?

6.1 Accessing auxiliary data

- There is a large and growing amount of auxiliary data available to survey researchers, which can be linked to survey data at address level or above without the need to obtain specific consent. This includes a wide array of data published by government departments, public bodies and local authorities. Other research projects such as the BBC's Big Personality Test or work carried out by the Local Elections Centre at the University of Plymouth may also be willing to share auxiliary data. Overall, the ADDResponse project was able to append around 400 auxiliary variables from 20 different data sources to the UK ESS sample.
- Much of the data available, including data from government sources, is freely available open access without restrictions. Other data may only be available under license for a limited time period (the terms may vary depending on whether data are requested for academic or commercial purposes).
- Data can also be purchased from commercial vendors such as Experian. Purchasing these data does incur a cost but any costs incurred are likely to be small relative to the overall costs of conducting fieldwork for a large scale survey.
- Auxiliary data should be accessed and appended to the sample at the time the sample is drawn, prior to fieldwork. Many sources of auxiliary data, especially commercial and environmental data, are frequently updated and it is not always possible to obtain historical data after the fact. Appending data when the sample is drawn maximises the changes that the time period covered by the auxiliary data will correspond to the time period covered by the survey data.
- Whilst accessing data to append to survey records may be relatively straightforward as part of the process of conducting a survey; accessing or sharing combined auxiliary and survey datasets for secondary analysis is likely to be more problematic. This is because of the risk of deductive disclosure posed by the combination of individual survey data and low level geographic identifiers. Secondary access to such data may be possible via a data processor agreement with the data controller, provided that the research can be considered in the "legitimate interests" of the data controller (i.e. the survey agency) and so can be justified under Section 2 of the UK Data Protection Act (1998).
- One challenge, not specifically addressed by the ADDResponse project, but relevant to surveys such as the ESS is the availability (or lack thereof) of comparable auxiliary data cross-nationally. Many data sources will be country-specific and the availability – and definition and measurement – of different auxiliary variables will vary. Eurostat and the INSPIRE

geoportal provide a wide variety of European data but often only country or regional (NUTS1/NUTS2) level. Volunteered Geographic Information (VGI) such as provided by OpenStreetMap often has cross-national reach though coverage is likely to vary across countries.

6.2 Data linkage

- The UK's well-established system of administrative geocodes and the growth of GIS means that appending auxiliary data to geographically referenced survey data is relatively straightforward.
- However, even using geocodes, it is important not to underestimate the amount of time needed for data linkage, especially when combining data from multiple sources each of which may use a different set of geocodes. Geocodes are also subject to change over time – and vary across the four countries of the UK – which can further complicate the process, especially as information on the geocodes used is not always readily apparent in data documentation.
- All of the data used on ADDResponse were available as downloadable databases and could be appended to the survey data in one batch. However, other auxiliary data sources e.g. google maps or address search databases such as 192.com may not allow batch processing and only allow a limited number of manual searches per day. Obtaining data from these data sources will be very labour intensive, therefore, and may not be feasible for large scale surveys.
- Metadata from the linkage process e.g. information on boundary changes and/or missing data points can in itself be a useful source of auxiliary information. For example, addresses missing from the commercial databases may indicate "hard to reach" or otherwise disenfranchised individuals. In our study, addresses missing from the commercial databases were more likely to be non-respondents to the ESS (particularly ineligible or non-contact addresses).
- The combination of individual level survey data and auxiliary data at low levels of aggregation does raise data protection considerations by creating an increased risk of deductive disclosure. Any linked data must, therefore, be stored securely and access may need to be restricted.

6.3 Suitability of data

- Just because auxiliary data are available does not necessarily make them useful; careful thought should be given to whether and how the data can be related to the underlying theoretical concepts and problems of interest. Ultimately, any conclusions regarding the suitability of auxiliary data will depend on the specific research question(s) to be addressed. Acceptable levels of missingness and/or possible misclassification, the appropriate level of aggregation, and the theoretical and empirical relevance of specific variables will all vary depending on the purpose for which the data are required.

- ADDResponse was interested in auxiliary data primarily as a source of information on nonrespondents in order to explore possible drivers of survey nonresponse. Auxiliary data were found to be of limited use for this purpose (see www.addresponse.org) for this purpose. However, there may be other ways in such data could be exploited either for methodological purposes e.g. to sample certain sub-populations or for substantive analysis e.g. to explore the effect of local context on attitudes and behaviour.
- Small-area data from government sources offer a potentially valuable source of auxiliary data; there is usually little problem with missing data, many data sources are official statistics and so have been quality assured, and data are available for multiple time points. The caveat is that these data are available at aggregate level only. Thought should be given to the most appropriate level of aggregation for the research question of interest and particular care taken when attempting to draw inferences about individuals from these aggregate data.
- Particular care should be taken when conducting UK-wide analysis as data sources, variable definitions and the timing of data collection can vary between the four constituent countries of the UK, thereby limiting comparability.
- Crowd-sourced data e.g. OpenStreetMap is likely to become an increasingly important source of auxiliary data and both coverage and accuracy will potentially increase over time as databases are continually added to and technological advances improve the accuracy of geolocation. Nevertheless, users should remain aware of potential gaps and biases arising from the self-selected nature of these data.
- Commercial databases provide an opportunity to obtain auxiliary data at address and/or individual level. Such data should be approached with caution. There is a relatively high level of missing data on many variables, estimates can vary between data sources and the provenance of modelled estimates is not always obvious. Nevertheless, if the right variables are selected, the data may prove useful for identifying certain types of household. Analysis of the data appended to ADDResponse suggests, for example, that commercial databases may prove similarly accurate to interviewer observations when it comes to predicting the presence of children in a household.

Overall, ADDResponse has demonstrated that it is possible to combine UK survey data and low level auxiliary data successfully for analysis. This includes a number of data source which have not previously received much attention from survey methodologists in the UK such as data from commercial vendors and Points of Interest data. The data matching process can be time consuming and is not without its challenges. Specific data sources should be further investigated and evaluated in the context of a specific purpose e.g. nonresponse analysis, sampling, substantive analysis to determine their usefulness. Nevertheless, these data certainly provide opportunities to survey researchers.

7 References[42]

Barron, M., Davern, M., Montgomery, R., Tao, X., Wolter, K. M., Zeng, W., ... & Black, C. (2015). Using Auxiliary Sample Frame Information for Optimum Sampling of Rare Populations. *Journal of Official Statistics*, *31*(4), 545-557.

Billiet J, Philippens M, Fitzgerald R & Stoop, I (2007) "Estimation of Response bias in the European Social Survey: using information from reluctant respondents in Round One", *Journal of Official Statistics*, 23: 135-162.

Blom, A.G., 2012. Explaining cross-country differences in survey contact rates: application of decomposition methods. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175, 217–242.

Blom, A. G. (2013). Setting priorities: Spurious differences in response rates. *International Journal of Public Opinion Research*, edt023.

Bodicoat, D. H., Carter, P., Comber, A., Edwardson, C., Gray, L. J., Hill, S., ... & Khunti, K. (2015). Is the number of fast-food outlets in the neighbourhood related to screen-detected type 2 diabetes mellitus and associated risk factors?. *Public health nutrition*, *18*(09), 1698-1705

Burgoine, T., & Harrison, F. (2013). Comparing the accuracy of two secondary food environment data sources in the UK across socio-economic and urban/rural divides. *International journal of health geographics*, *12*(1), 1.

De Leeuw, E. and de Heer, W. (2002) "Trends in household survey nonresponse: a longitudinal and international comparison, in Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little. R.J.A. (eds.) *Survey* Nonresponse, New York: John Wiley and Sons

Department for Education (2011) *A profile of pupil absence in England DFE-RR 171*.
Available via: https://www.gov.uk/government/uploads/system/uploads/...data/.../DFE-RR171.pdf

Department for Education (2012) Pupil Absence in Schools by Gender in England (Referenced by Location of Pupil Residence), 2011/2012
Available via: http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do
(Neighbourhood Statistics -> Education, Skills and Training -> Pupil Absence in Schools by Gender in England (Referenced by Location of Pupil Residence))
This information is licensed under the terms of the Open Government Licence.

Department of Energy and Climate Change (2015a) LSOA/MSOA 2011 historic electricity consumption
Available at: https://www.gov.uk/government/statistics/lower-and-middle-super-output-areas-electricity-consumption
This information is licensed under the terms of the Open Government Licence.

Department of Energy and Climate Change (2015b) Northern Ireland sub-national domestic electricity consumption 2008-2013

[42] All weblinks were active as of 31st August 2016

Available at:
https://www.gov.uk/government/statistical-data-sets/experimental-sub-national-domestic-electricity-consumption-statistics-for-northern-ireland-2009
This information is licensed under the terms of the Open Government Licence.

Department for Communities (NI) (2012) Benefits Statistics Summary Publication (National Statistics)
Available from:
https://www.communities-ni.gov.uk/topics/benefits-statistics
This information is licensed under the terms of the Open Government Licence.

Department for Communities and Local Government (2011) English Indices of Deprivation 2010
Available via:
https://www.gov.uk/government/statistics/english-indices-of-deprivation-2010
This information is licensed under the terms of the Open Government Licence.

Department for Environment, Food and Rural Affairs (2013) *The 2011 Rural Urban Classification for Small Area Geographies: A User Guide and Frequently Asked Questions (v1.0).*
Available via:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/239478/RUC11user_guide_28_Aug.pdf

Department for Work and Pensions (2016) Work and Pensions Longitudinal Study (WPLS) Benefit Claimants – working age client group
Available via: https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp
This information is licensed under the terms of the Open Government Licence.

De Vor, F., & De Groot, H. L. (2011). The impact of industrial sites on residential property values: A hedonic pricing analysis from the Netherlands. *Regional Studies*, *45*(5), 609-623.

Dugmore, Keith. 2010. "Information Collected by Commercial Companies: What Might Be of Value to Official Statistics? The Case of the UK Office for National Statistics." German Data Forum (RatSWD).

Durrant, G.B., Groves, R.M., Staetsky, L., Steele, F., 2010. Effects of interviewer attitudes and behaviors on refusal in household surveys. Public Opinion Quarterly nfp098.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the association of American geographers*, *102*(3), 571-590.

ESS Round 6: European Social Survey Round 6 Data (2012). Data file edition 2.1. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 6: European Social Survey Round 6 Data (2012). Data from contact forms, edition 2.0. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 6: European Social Survey Round 6 Data (2012). Data from Interviewer's questionnaire, edition 2.0. (UK only) NSD - Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 6: European Social Survey (2016): ESS-6 2012 Documentation Report. Edition 2.2. Bergen, European Social Survey Data Archive, NSD - Norwegian Centre for Research Data for ESS ERIC

European Social Survey (2012). *ESS Round 6 Project Instructions (CAPI*). London: ESS ERIC Headquarters, Centre for Comparative Social Surveys, City University London

Experian (2014) Mosaic UK Brochure
Available via: www.experian.co.uk/assets/marketing-services/brochures/mosaic_uk_brochure.pdf

Experian (2015) Experian Marketing Services: Consumerview
Available via: www.experian.co.uk/assets/marketing-services/documents/FS_ConsumerView.pdf

Fraser, L. K., & Edwards, K. L. (2010). The association between the geography of fast food outlets and childhood obesity rates in Leeds, UK. *Health & place*, *16*(6), 1124-1128.

Fuchs, M., Bossert, D., Stukowski, S., 2013. Response rate and nonresponse bias-impact of the number of contact attempts on data quality in the European Social Survey. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 117, 26–45.

Gehlke, Charles E., and Katherine Biehl. "Certain effects of grouping upon the size of the correlation coefficient in census tract material." *Journal of the American Statistical Association* 29.185A (1934): 169-170.

Geolytix (2015) Open Supermarkets
© GeoLytix copyright and database right 2015
Contains Ordnance Survey data © Crown copyright and database right 2015
Contains Royal Mail data © Royal Mail copyright and database right 2015
Contains National Statistics data © Crown copyright and database right 2015

Gibbons, S., & Machin, S. (2005). Valuing rail access using transport innovations. *Journal of urban Economics*, *57*(1), 148-169.

Giles-Corti, B., & Donovan, R. J. (2002). Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment. *Preventive medicine*, *35*(6), 601-611.

Groves, R.M. (2006) "Nonresponse rates and nonresponse bias in household surveys", *Public Opinion Quarterly*, 70: 646-75

Groves, R.M. and Couper, M.P. (1998) *Nonresponse in Household Interview Surveys*, New York: John Wiley and Sons

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475-495.

Haklay, M. (2010) "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B Planning and Design*, no. 37: 682–703.

Home Office (2012) Notifiable Offences Recorded by the Police, 2011/12
Available via:

http://www.neighbourhood.statistics.gov.uk/dissemination/Download1.do
(Neighbourhood Statistics - > Crime and Safety -> Notifiable offences recorded by the police)
This information is licensed under the terms of the Open Government Licence.

Home Office (2016) Data on recorded offences provided to https://data.police.uk
This information is licensed under the terms of the Open Government Licence.

Keßler, C., & de Groot, R. T. A. (2013). Trust as a proxy measure for the quality of volunteered geographic information in the case of OpenStreetMap. In *Geographic information science at the heart of Europe* (pp. 21-37). Springer International Publishing.

Kreuter, F. (ed) (2013) *Improving surveys with paradata: Analytic uses of process information* New York: John Wiley and Sons

Kreuter, F., & Kohler, U. (2009). Analyzing contact sequences in call record data. Potential and limitations of sequence indicators for nonresponse adjustments in the European Social Survey. *Journal of Official Statistics*, *25*(2), 203.

Luttik, J. (2000). The value of trees, water and open space as reflected by house prices in the Netherlands. *Landscape and urban planning*, *48*(3), 161-167.

Maas, J., Verheij, R. A., de Vries, S., Spreeuwenberg, P., Schellevis, F. G., & Groenewegen, P. P. (2009). Morbidity is related to a green living environment. *Journal of epidemiology and community health*, *63*(12), 967-973.

Macintyre, S., Macdonald, L., & Ellaway, A. (2008). Do poorer people have poorer access to local resources and facilities? The distribution of local resources by area deprivation in Glasgow, Scotland. *Social science & medicine*, *67*(6), 900-914.

Massey, D.S. and Tourangeau, R. (2013) "The nonresponse challenge to surveys and statistics", *Annals of the American Academy of Political and Social Science*, 645:6-27

Moore, L. V., & Diez Roux, A. V. (2006). Associations of neighborhood characteristics with the location and type of food stores. *American journal of public health*, *96*(2), 325-331.

National Records of Scotland (2013) *Census 2011: Release 1B - how the 2011 Census population estimates were obtained*. Available via: www.scotlandscensus.gov.uk/documents/censusresults/release1b/rel1bmethodology.pdf

National Records of Scotland (2014) *Geography – Background Information – 2011 Census: Linkage Matrix.* Available via: http://www.nrscotland.gov.uk/files/geography/2011-census/backgroundinformationmatrix.pdf

Neis, P., & Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet*, *6*(1), 76-106.

Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project—The case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, *1*(2), 146-165.

Nomis (2016) Jobseeker's allowance with rates and proportions. Available at: https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp
This information is licensed under the terms of the Open Government Licence.

Northern Ireland Statistics and Research Agency (2010) Northern Ireland Multiple Deprivation Measure 2010. Available via: www.nisra.gov.uk/deprivation/archive/Updateof2005Measures/NIMDM_2010_Report.pdf
This information is licensed under the terms of the Open Government Licence.

Northern Ireland Statistics and Research Agency (2013) *A guidance note on comparisons of Census outputs from the 2001 and 2011 Censuses for geographic areas within Northern Ireland.* Available via: http://www.nisra.gov.uk/archive/census/2011/comparisons-of-census-outputs-from-the-2001-and-2011-censuses-for-geographic-areas.pdf

Northern Ireland Statistics and Research Agency (2015a) *Northern Ireland Census 2011 General Report* Available via: www.nisra.gov.uk/archive/census/2011/evaluation/general-report.pdf

Northern Ireland Statistics and Research Agency (2015b) Recorded Crime 2000-2014 (Administrative geographies) Available at: www.ninis2.nisra.gov.uk/Download/Crime%20and%20Justice/Recorded%20Crime%20(administrative%20geographies).ods
This information is licensed under the terms of the Open Government Licence.

Northern Ireland Statistics and Research Agency (2015c) Attendance rates for post-primary pupils by pupil residence (Administrative geographies) 2012-14 Available at: www.ninis2.nisra.gov.uk/Download/Children%20Education%20and%20Skills/Attendance%20Rates%20for%20Post-Primary%20Pupils%20by%20Pupil%20Residence%20(administrative%20geographies).ods
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics (2010) *Predicting patterns of household non-response in the 2011 Census. Census Advisory Group paper AG (09)17.* Available at: https://www.ons.gov.uk/file?uri=/census/2011census/howourcensusworks/howwetookthe2011census/howweprocessedtheinformation/coverageassessmentandadjustmentprocesses/htc-for-web-v26-220210_tcm77-189762.pdf.

Office for National Statistics (2011): 2001 Census aggregate data (Edition: May 2011). UK Data Service. DOI: http://dx.doi.org/10.5257/census/aggregate-2001-2
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics (2013a) Census Geography lookups Available via: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/index.html
Lookup products are supplied under the Open Government Licence. Contains National Statistics data © Crown copyright and database right 2013

Office for National Statistics (2013b) Super output area mid-year population estimates, mid 2012 Available via:
http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/sape/soa-mid-year-pop-est-engl-wales-exp/mid-2012/index.html
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics (2013c) Population Estimates for UK, England and Wales, Scotland and Northern Ireland, Mid-2011 and Mid-2012
Available at:
http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/mid-2011-and-mid-2012/index.html
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics (2014a) National Statistics Postcode Lookup (UK) Nov 2014
Available via: https://data.gov.uk/dataset/national-statistics-postcode-lookup-uk-nov-2014
Lookup products are supplied under the Open Government Licence. Contains National Statistics data © Crown copyright and database right [2014]. Contains Ordnance Survey data © Crown copyright and database right [2014], Contains Royal Mail data © Royal Mail copyright and database right [2014]

Office for National Statistics (2014b) Personal wellbeing in the UK 2013/14
Available at:
http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/rel/wellbeing/measuring-national-well-being/personal-well-being-in-the-uk--2013-14/sb-personal-well-being-in-the-uk--2013-14.html
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics (2016a) Code History Database February 2016
Available via: https://data.gov.uk/dataset/code-history-database-uk-feb-2016
Licensed under Open Government Licence. Contains National Statistics data © Crown copyright and database right [2016].

Office for National Statistics (2016b) Local Area Migration Indicators Suite
Available via:
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/migrationwithintheuk/datasets/localareamigrationindicatorsunitedkingdom
This information is licensed under the terms of the Open Government Licence.

Office for National Statistics; National Records of Scotland; Northern Ireland Statistics and Research Agency (2016): 2011 Census aggregate data. UK Data Service (Edition: June 2016). DOI: http://dx.doi.org/10.5257/census/aggregate-2011-1.
This information is licensed under the terms of the Open Government Licence.

Ordnance Survey (2014) Ordnance Survey Points of interest database: User guide and technical specification v3.3. available at: https://www.ordnancesurvey.co.uk/docs/user-guides/points-of-interest-user-guide.pdf
© Crown copyright and database right [2014]. All rights reserved. Licence number 100034829.
This product includes data licensed from PointX ©Database Right/Copyright 2014

Openshaw, Stan (1983). "Multivariate analysis of census data: the classification of areas." In D Rhind (ed) A census user's handbook London Methuen, pp 243-263.

OpenStreetMap (2015) Points Of Interest Database
© OpenStreetMap contributors. OpenStreetMap database is made available under the Open Database License: http://opendatacommons.org/licenses/odbl/1.0/. Any rights in individual contents of the database are licensed under the Database Contents License: http://opendatacommons.org/licenses/dbcl/1.0/

Pasek, J., Jang, S. M., Cobb, C. L., Dennis, J. M., & Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *Public Opinion Quarterly*, *78*(4), 889–916.

Pickery, J., & Loosveldt, G. (2004). A simultaneous analysis of interviewer effects on various data quality indicators with identification of exceptional interviewers. *Journal of Official Statistics*, *20*(1), 77.

Rentfrow, P. J., Jokela, M., & Lamb, M. E. (2015). Regional personality differences in Great Britain. *PloS One*, *10*(3), e0122245.

Robinson, W.S. (1950). "Ecological Correlations and the Behavior of Individuals". American Sociological Review. American Sociological Review, Vol. 15, No. 3. ***15*** *(3): 351–357. doi:10.2307/2087176. JSTOR 2087176*

Sarndal, C.E. and Lundstrom, S. (2005) *Estimation in surveys with nonresponse*, Chichester: Wiley

Schräpler, J. P., Schupp, J., & Wagner, G. G. (2010). Individual and Neighborhood Determinants of Survey Nonresponse–An Analysis Based on a New Subsample of the German Socio-Economic Panel (SOEP), Microgeographic Characteristics and Survey-Based Interviewer Characteristics.

The Scottish Government (2012a) Scottish Government Urban Rural Classification 2011-12
Available via:
http://www.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification/Urban-Rural-Classification-2011-12

The Scottish Government (2012b) Recorded crime in Scotland 2011-12
Available via: http://www.gov.scot/Publications/2012/06/1698
This information is licensed under the terms of the Open Government Licence.

The Scottish Government (2012c) Scottish Index of Multiple Deprivation 2012
Available via: http://www.gov.scot/Topics/Statistics/SIMD/DataAnalysis/Background-Data-2012
This information is licensed under the terms of the Open Government Licence.

The Scottish Government (2013) Attendance and Absence 2012/13
Available via: http://www.gov.scot/Topics/Statistics/Browse/School-education/AttendanceAbsenceDatasets/attab2013
This information is licensed under the terms of the Open Government Licence.

Sinibaldi, J., Trappmann, M., & Kreuter, F. (2014). Which is the better investment for nonresponse adjustment: purchasing commercial auxiliary data or collecting interviewer observations? *Public Opinion Quarterly*, nfu003.

Smith, T. W. (2011). The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys. *International Journal of Public Opinion Research*, *23*(3), 389-402.

Smith, T.W. and Kim, J. (2013) "An assessment of the Multi-level Integrated Database approach", *Annals of the American Academy of Political and Social Science*, 645:185-221

Stats Wales (2011). Welsh Index of Multiple Deprivation 2011 by rank and local super output area
Available via: https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/Archive/WIMD-2011/WIMD2011
This information is licensed under the terms of the Open Government Licence.

Stats Wales (2012). Absenteeism by pupils of compulsory school age in maintained secondary and special schools by local education authority and year
Available via: https://statswales.gov.wales/Catalogue/Education-and-Skills/Schools-and-Teachers/Absenteeism/absenteeismbypupilsofcompulsoryschoolageinmaintainedsecondaryandspecialschools-by-localeducationauthority-year
This information is licensed under the terms of the Open Government Licence.

Stoop, I., Billiet, J., Koch, A. and Fitzgerald, R. (2010) *Improving Survey Response: Lessons learned from the European Social Survey*, Chichester: Wiley

Taylor, J., Twigg, L., & Mohan, J. (2014). Understanding neighbourhood perceptions of alcohol-related anti-social behaviour. *Urban Studies*, 0042098014543031.

University of Cambridge. Department of Psychology, British Broadcasting Corporation. (2015). *BBC Big Personality Test, 2009-2011: Dataset for Mapping Personality across Great Britain*. [data collection]. UK Data Service. SN: 7656, http://dx.doi.org/10.5255/UKDA-SN-7656-1.

van Praag, B.M.S., Baarsma, B.E., 2005. Using happiness surveys to value intangibles: the case of airport noise. Economic Journal
115, 224–246.

Vehovar, V., Slavec, A. and Kralj, M. (2013) "Country-specific quality control checks for ESS weighting procedures" DACE ProjectWP12 mid-term report

Welsch, H., (2006). Environment and happiness: valuation of air pollution using life satisfaction data. Ecological Economics 58,
801–813.

West, B.T. and Kreuter, F., 2013. Factors affecting the accuracy of interviewer observations evidence from the National Survey of Family Growth. *Public opinion quarterly*, p.nft016.

West, B. T., Wagner, J., Hubbard, F., & Gu, H. (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, smv004.

8 Appendix: Priority outcome codes used for ADDResponse

Table A1: Priority ranking of ESS response outcomes

Call outcome	Classification	Outcome code
Interview - complete	Respondent	11
Interview – incomplete	Refusal	12
Address not residential (institutional)	Ineligible	21
Address not residential (business)	Ineligible	22
Address not yet built	Ineligible	23
Address derelict/demolished	Ineligible	24
Address unoccupied	Ineligible	25
Address not traceable	Ineligible	26
Ineligible – other	Ineligible	27
Refusal by respondent	Refusal	31
Refusal by proxy	Refusal	32
Refusal before household selection	Refusal	33
Appointment with respondent not realised	Refusal	34
Appointment with someone else in household not realised	Other contact	41
Mentally/physically unable	Other contact	42
Language barrier	Other contact	43
Unavailable during fieldwork period	Other contact	44
Other contact – undefined	Other contact	45
No contact	Noncontact	51

Table A2: ESS Round 6 UK fieldwork outcomes

	ESS National Technical Summary		ADDResponse priority outcome codes	
	N	**%**	**N**	**%**
Total sample	4520	100	4520	100
Ineligibles	212	4.7	265	5.9
Total eligible sample[1]	4308	95.3	4255	94.1
Total eligible sample	4308	100	4255	100
Respondents	2286	53.1	2286	53.7
Non-respondents	2022	46.9	1969	46.3
Contact	3798	88.2	3949	92.8
Non-contact	483	11.2	306	7.2
(Unknown eligible)	(27)			
Refusal	1268	29.4	1432	33.7
Total contacted sample	3798	100	3949	100
Refuse	1268	33.4	1432	36.3
Co-operate	2286	60.2	2286	57.9
Other (e.g. ill, away, language barrier)	244	6.4	231	5.8

[1] NTS includes 27 "address not traceable" cases in eligible sample (following AAPOR RR 1). Priority coding excludes these unknown eligibility cases from sample (following AAPOR RR 5).