**Lessons from the ADDResponse project:**

**Appending geocoded auxiliary data to the European Social Survey in the UK**

Sarah Butt, City University London

**Panel on Nonresponse, data protection and research ethics**

Household NonResponse Workshop, Oslo, 31 August - 2nd September 2016

**Introduction**

The ADDResponse project (Auxiliary Data Driven nonResponse bias analysis) aimed to explore the possibility of using different sources of geocoded auxiliary data to understand and correct for nonresponse bias in the UK wave of the European Social Survey. A specific objective of the project was to explore the legal and ethical issues associated with linking auxiliary data at a low-level of aggregation with individual-level survey data. The project successfully linked around 400 variables from more than 20 different data sources to the European Social Survey (ESS) Round 6 UK sample of addresses. This paper summarises the conditions under which the data linkage took place, discusses the legal and ethical issues that needed to be addressed in relation to accessing, appending, analysing and archiving the auxiliary data and identifies factors that were instrumental in making the project a success.

Survey companies in the UK frequently append geocoded auxiliary data including commercial segmentations (e.g. ACORN, MOSAIC) or small-area census estimates to samples of addresses drawn from the Royal Mail’s Postcode Address File prior to fieldwork. These data are then used to stratify the sample and/or to produce nonresponse weights. This is unproblematic from a data protection point of view. Publically available data sources are combined for the purposes of carrying out a public opinion survey by the company responsible for conducting that survey.

The conditions under which the ADDResponse project took place were somewhat different to the usual practice described above. First, the project involved sharing data with a third party (researchers at City University London). Second, the amount of auxiliary data that was appended to the sample frame was considerably more than is usually the case and was explicitly intended to be combined with survey responses, significantly increasing the risk of deductive disclosure. Third, the research project took place two years after the initial survey data collection and went beyond the scope of what was necessary to conduct the survey. Fourth, under the terms of the project’s funding (from the UK ESRC) there was a commitment to archive datasets wherever possible. There were therefore issues around participant confidentiality, data sharing and consent that needed to be considered, both from a legal perspective as regards data protection legislation and ethically, particularly in relation to the commitment given to study participants in the letter/leaflet sent out in advance of the survey.

**The ADDResponse project**

ADDResponse (www.addresponse.org) involved appending geocoded auxiliary data to the sample of 4,520 addresses selected to take part in Round 6 of the European Social Survey (2012/13) in the UK for the purposes of exploring nonresponse bias in general social surveys. The project had three main aims: To scope available sources of auxiliary data in the UK and evaluate them in terms of data quality and accessibility (including any legal or ethical restrictions on use), to test the usefulness of auxiliary variables from the different data sources for predicting both response propensity and substantive survey variables, and to consider how these auxiliary variables might be used for weighting to correct for nonresponse bias. The project came to an end in May 2016.

Auxiliary data from three main sources were used in the project:

* 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

More information on the different data sources can be found in the ADDResponse technical report (Butt and Lahtinen, 2016). An overview of the data sources combined for the project is given in Figure 1 at the end of this paper.

Data linkage involved a collaboration between researchers at City University London and Ipsos MORI, the survey agency responsible for carrying out UK fieldwork for ESS Round 6 and, therefore, data controller of the ESS sample data. The process by which data were exchanged between the research team and Ipsos MORI is summarised in Figure 2 and the different stages of the process discussed in more detail below.

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. (ESS, 2016). A response rate of 54% was obtained.

The advance materials provided by Ipsos MORI to potential respondents provided assurances about data confidentiality. The advance letter stated that “*Your answers will be treated in strict*

*confidence in accordance with the Data Protection Act and will only be used for statistical*

*purposes*.” whilst the project leaflet used by interviewers on the doorstep stated *“your answers will be treated in accordance with the 1998 Data Protection Act. The information you give is grouped together with information provided by thousands of other people and no one will be identifiable by the results.”*  The ADDResponse project was not envisaged at the time the survey was carried out and so no reference to data linkage or sharing is made in the materials.

**Accessing auxiliary data**

Accessing auxiliary data did not present a significant challenge in the context of this project. From the outset the project restricted attention to geocoded auxiliary data that was publically accessible without the need for consent. This choice was driven first by the fact that the ESS in the UK uses an address-based sample frame which contains no information on specific individuals. Second, given the focus on nonrespondents, it was not practical to obtain consent for data linkage.

There is a large and growing amount of geocoded auxiliary data in the UK which is available in the public domain and which can be accessed by researchers without the need for subject’s consent. The primary source of data is the wide array of small-area statistics produced by government agencies and available for reuse under the terms of the Open Government License.[[1]](#footnote-1) There is also a growing amount of crowdsourced georeferenced data such as OpenStreetMap which is available via the Open Database License[[2]](#footnote-2) well as other data (such as that provided by Ordnance Survey) which can be licensed for research or commercial purposes. Finally, data is available to purchase from commercial vendors, though it should be born in mind that there is a financial cost involved in obtaining these data and license conditions may stipulate that data can only be used and retained for one to two years.

**Appending and analysing auxiliary data**

As part of the ADDResponse project the research team were able to successfully append nearly 400 variables from over 20 different data sources to UK survey data from ESS Round 6. This required the research team to have access to address records for the ESS sample and resulted in the production of a potentially disclosive final dataset. Two different approaches to data sharing between Ipsos MORI and the research team were considered: the first was to explore means of disclosure control to avoid having to share personal data with the research team. The second was to set up a data processor agreement between Ipsos MORI and City University London to permit the sharing of personal data. Ultimately, two factors were instrumental in facilitating data linkage and access to the linked data for analysis. The first was the perceived value of the research project - especially as regards its scoping of different auxiliary data sources - for survey practice, thereby justifying the setting up of a data processing agreement on the grounds of legitimate interests. The second was the close working relationship established between the different parties involved (City University London and Ipsos MORI) both before and during the project.

The starting assumption was that it should be possible to carry out data linkage without an undue risk of disclosure, thereby avoiding legal or ethical concerns around data sharing. Data linkage took place in a series of steps with the amount and type of data available to non-Ipsos MORI researchers restricted at each stage. In the first stage of the process, for example, the research team were provided with a list of addresses in order to append the auxiliary data but no survey data or indicators of which were responding/nonresponding addresses. The appended auxiliary data were then returned to Ipsos MORI to strip off identifiers and add on the survey data whilst the research team destroyed the list of addresses supplied. The unique identifier on the dataset was not the same as the “idno” variable on ESS datasets and, as an additional precaution, the dataset for linkage contained an additional 4,520 camouflage addresses from the same Primary Sampling Units as the real addresses to minimise the risk of cases being re-identified at a subsequent stage.

The data linkage process worked smoothly under these conditions. One limitation of the staged approach was 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 led to some auxiliary variables being coded in multiple ways or included in the dataset “just in case” increasing the number of variables required. This was a particular issue with this project where the express purpose was to explore the potential of multiple data sources. However, it need not be a problem for other projects if the list of auxiliary variables required is pre-defined.

The original intention had been for Ipsos MORI to carry out necessary disclosure control before returning the combined auxiliary and survey dataset to the research team. However, it became apparent that, with the large amount of low level auxiliary data appended plus the research team’s desire to make use of geographic identifiers to carry out spatial analysis, it would not be feasible to minimise the risk of disclosure control sufficiently without significantly compromising the analytic value of the combined auxiliary and survey data. The solution was to set up a data processor agreement (DPA) between Ipsos MORI (the controller) and City University London (the processor) thereby enabling the research team to have access to potentially identifiable i.e. personal data without the need for disclosure control (though exact addresses were removed).

The DPA was justified on the grounds that the project and its use of personal data was in the “legitimate interests” of the data controller i.e. Ipsos MORI. Clause 33 of the UK Data Protection Act 1998[[3]](#footnote-3), the research exemption, allows for personal data to be processed for research purposes other than those for which it was originally obtained, provided that the personal data are not processed to support measures or decisions relating to particular individuals and are not processed in such a way that substantial damage or distress may be caused to any of the data subjects. However, it is still necessary for data to be processed lawfully and fairly and meet at least one of the “conditions of processing” set out in Schedule 2 of the act. Usually in survey research, consent forms the basis of lawful and fair processing. However, in the absence of consent for data linkage - and any consent from nonrespondents to use their data - the alternative condition of legitimate interests was used. Condition 6.1 states that processing of personal data may take place if:

*The processing is necessary for the purposes of legitimate interests pursued by the data controller or by the third party or parties to whom the data are disclosed, except where the processing is unwarranted in any particular case by reason of prejudice to the rights and freedoms or legitimate interests of the data subject.*

The ADDResponse project’s focus on possible solutions to the problem of survey nonresponse, arguably one of the biggest problems facing survey research today, along with its wider aim of scoping auxiliary data sources which could also be appended to survey data for other purposes e.g. sampling was considered by the data controller to meet this condition.

The DPA contained strict provisions on data security (data to be transferred via password protected encrypted files, to be stored on secure restricted-access servers etc.), the use to which the data could be put, and restrictions on linking additional data sources or using the data for anything other than the specified research purpose. The original DPA was set up for one year and later extended for a further 12 months.

**Archiving auxiliary data**

Good research practice, as well as the terms of the funding received for the ADDResponse project, require that research data should be archived and made available for further use wherever possible. Thinking about how best to meet this requirement posed some challenges for the ADDResponse project.

First, some of the auxiliary data was accessed under licence, the terms of which prohibit reuse of the data outside of that license. This was not a major limitation however. Although it meant that the data from commercial vendors and Ordnance Survey could not be archived, all other data sources were available without restriction and could be archived for reuse.

The second, more serious constraint, was the potentially disclosive nature of the final ADDResponse dataset. The obligation to make research datasets available needs to be offset against the obligation to maintain participant confidentiality and treat data in accordance with the Data Protection Act. Over and above any legal obligation under the Data Protection Act, as data controller Ipsos MORI also have a clear ethical responsibility to participants given the assurances provided in the advance survey materials. This responsibility is particularly salient in the case of nonrespondents who did not consent even to be part of the original ESS survey. It restricts the sharing of personal data with any third party, including data archives themselves.

The usual solution, to produce an anonymised version of the original research dataset for archiving, was not considered feasible in this instance. All of the data sources used in ADDResponse are already publically available; 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.

The solution agreed upon was to make a respondent only dataset containing small-area data, geocodes and ESS “idno” (to allow linkage to ESS survey and Contact Form data) available via the UKDA under secure access conditions.[[4]](#footnote-4) This service is provided for data which are considered too detailed, disclosive or confidential to be made available even under a standard special license agreement. The inclusion of multiple low level auxiliary variables plus very low level geographic identifiers mean that the ADDResponse data fall into this category. Data cannot be downloaded but controlled access to data is provided via remote connection or safe room to UK-based accredited researchers on completion of appropriate training and signature of a user agreement. Importantly, the Data Controller, Ipsos MORI, retains control over who is allowed to access the data and for what purpose. Even though the original purpose of the data linkage was nonresponse analysis, the linked auxiliary data provides plenty of scope for substantive analysis of ESS survey response. Ipsos MORI will retain another dataset containing the same set of auxiliary variables but for respondents and nonrespondents and consider applications to access the data under a data processor agreement similar to that which was put in place with City University London. As was the case for ADDresponse, any application would need to demonstrate that access to and processing of the personal data serves the legitimate interests of Ipsos MORI i.e. had the potential to contribute significantly to survey practice.

**Conclusions**

The ADDResponse project demonstrates that there is scope within the UK for researchers to access and append a variety of geocoded auxiliary data to survey data even in the absence of consent for data linkage provided that access to personal data can be justified on the grounds of legitimate interest. Nevertheless, access to such data must necessarily remain restricted and under the control of the survey agency that owns the data, limiting the scope for subsequent data archiving. Ethical obligations to study participants are just as important as legal obligations under the Data Protection Act and must be taken fully into account.

**Questions for discussion**

* One concern throughout the project was that linking auxiliary and survey data was not anticipated at the start of fieldwork and study participants were not notified of the intention. If similar data linkage exercises were planned for future rounds of the ESS, (how) should the advance materials make reference to this?
* How much of a concern is the restriction on archiving auxiliary data used for nonresponse analysis alongside the survey data?
* Is the value of auxiliary data for understanding and tackling survey nonresponse sufficient to justify similar data processor agreements under the “legitimate interests” condition?

**References**

Butt, S. and K. Lahtinen (2016, forthcoming) *ADDResponse: Auxiliary data driven nonresponse bias analysis. Technial report on appending geocoded auxiliary data to Round 6 of European Social Survey (ESS)* available via [www.addresponse.org](http://www.addresponse.org).

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 available via [www.europeansocialsurvey.org](http://www.europeansocialsurvey.org)

**Figure 1: Overview of ADDResponse data sources**

ESS Round 6 UK sample file

4,520 addresses drawn from PAF

Located within 226 postcode sectors across UK

**Small-area data**

Available at different levels of aggregation: OA to LAD

Appended using postcode geo-referencing

Census 2011, crime statistics, benefit claimants, indices of deprivation, school absences, population turnover, energy consumption, wellbeing, local election results

**ESS survey data**

Appended using ESS variable “idno”

Interview data for 2,286 respondents

Contact form data collected by interviewers for 4,520 addresses

ESS interviewer age and gender

**Data from commercial vendors**

Address-level socio-demographics

Tenure, length of residency, marital status, employment status, presence of children, council tax band

Data on local geographic features**: Points of Interest**, accessibility

Appended using OA centroid grid references

Ordnance Survey and Openstreetmap

**Figure 2: ADDResponse data linkage process**

**Ipsos MORI supply geocoded sample records to research team**

* Records for 4,520 ESS R6 sampled addresses + 4,520 camouflage addresses drawn from same postcode sectors.
* Records include: address, postcode, grid reference of postcode centroid, postcode centroid Output Area

**Ipsos MORI supply list of sampled addresses to commercial vendors**

**Researchers access small-area and local geographic auxiliary data**

**Commercial vendors append address-level data and return file to Ipsos MORI**

* Data appended only where there is an exact match between sample address and vendor records

**Research team append small-area data and local geographic auxiliary data to sample records**

* File returned to IPSOS MORI
* Research team destroy original geocoded records

**Ipsos MORI supply research team with a data file that contains**

* ESS survey “idno”
* Geocodes: Postcode centroid grid reference, (S)OA codes, NUTS1, LAD
* Appended small-area/local geographic data
* Data from commercial vendors

Address and postcode identifiers removed from dataset

**City University London sign a Data Processor Agreement with Ipsos MORI**

**Research team append ESS survey data using “idno” and conduct nonresponse analysis**

**A reduced dataset is made available for archiving**

* A respondent only dataset containing ESS “idno”, geocodes and small area data deposited with UKDA to be made available via secure access
* Ipsos retain the same data for respondents and nonrespondents

1. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ [↑](#footnote-ref-1)
2. <http://opendatacommons.org/licenses/odbl/1.0/> [↑](#footnote-ref-2)
3. https://www.gov.uk/data-protection/the-data-protection-act [↑](#footnote-ref-3)
4. https://www.ukdataservice.ac.uk/use-data/secure-lab/about [↑](#footnote-ref-4)