Case for Support

This project will explore the extent to which the predictive power of various forms of "Big Data" can be harnessed to overcome the impact of survey nonresponse - arguably one of the most serious challenges facing survey research today. Nonresponse is 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, Stoop, Billiet, Koch and Fitzgerald, 2010; Massey and Tourangeau, 2013). Given that non-respondents may differ systematically from respondents, this risks introducing significant bias into the numerous inferences regularly drawn from survey data by academics, policy makers, journalists and others (Groves, 2006; Peytchev, 2013). There is a pressing need to understand more about the extent and source of nonresponse bias and how we can take steps to respond, for example through weighting or responsive fieldwork design (Groves and Peytcheva, 2008; Peytchev, 2013). To fully understand and address nonresponse bias requires information to be available for the full target sample. In the absence of interview data being available for non-respondents, the key is to identify sources of auxiliary data which can provide information on both respondents and non-respondents (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 (Stoop et al, 2010; Kreuter, 2013). However interviewer generated paradata is necessarily limited, costly to collect and subject to interviewer error (Matsuo et al, 2010; Olson, 2013). Therefore, another option is to exploit the large and growing amount of information that is available via pre-existing external databases (Smith and Kim, 2013). Looking for new ways to use existing data is cost effective and, in the absence of interviewers, the approach is particularly suitable for use on mixed–mode and/or web surveys which are likely to become increasingly prevalent in future.

Research on this topic is timely. Opportunities to supplement survey data with external data – for both methodological and substantive research - are increasing. The amount of data available is expanding rapidly with the emergence of “Big Data”. In the UK the growth in transaction data from the private sector is complemented by the government’s agenda to make increasing amounts of administrative data publically available (HMSO, 2012). The launch of the ESRC funded Administrative Data Research Network in 2013 marks a major new investment in this direction. Meanwhile, software developments, including the growth of GIS, facilitate more accurate and efficient data linkage via postcodes or other geo-identifiers and provide new means to explore spatial variation in survey data.

However, using auxiliary data from pre-existing sources also presents a number of challenges. There are concerns over the coverage, accuracy and timeliness of external databases, the extent to which data which is often highly aggregated can characterise sampled households, and the increased likelihood of deductive disclosure as a result of combining different data sources (Smith and Kim, 2013). As with any study of nonresponse bias, appropriate auxiliary variables need to be identified i.e. variables which not only predict survey response but are also correlated with survey variables of interest (Little and Vartivarian, 2005; Groves, 2006). Given that different survey estimates may be prone to different biases, especially on a general social survey that has multiple measurement aims, thought must be given to how data from multiple sources and available at multiple levels of aggregation can be combined most effectively (Kreuter and Olson, 2011). As the opportunities for linking auxiliary data increase, a systematic investigation into the usefulness of such data for tackling nonresponse bias is required (Massey and Tourangeau, 2013).

This study will make a major contribution to that process by carrying out a systematic investigation into the feasibility of appending auxiliary data from multiple sources and at multiple levels of aggregation to UK data from the European Social Survey (ESS). These data will then be evaluated in terms of their usefulness for understanding and correcting for nonresponse bias using weighting. Like other general social surveys, the ESS, a survey of public attitudes and behaviour across more than 30 countries, has considerable nonresponse. The UK response rate is typically around 55%. 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, Slavic and Kralj, 2013). Analysis of nonresponse bias has relied primarily on paradata (Billiet, Phillipens, Fitzgerald and Stoop, 2007) or auxiliary information from interviewer observations (Stoop et al, 2010). The UK can now provide a test case for matching other forms of data to the sample and, if successful, this might provide a stimulus for similar developments in other ESS countries.

This project has the potential to contribute significantly to our understanding not only of survey nonresponse bias but also the statistical tools available to remedy it, thus generating more robust data to better understand public attitudes and behaviour. Combining the expertise of a multi-disciplinary team of survey methodologists, statisticians and geographic information (GI) specialists, the project has three strands.

Strand 1: Scoping sources of auxiliary data available to append to the UK ESS sample

The first strand (months 1 to 6) will be to conduct a scoping study of the sources of auxiliary data available which can be appended to the ESS sample. The intention is to look beyond aggregate-level census variables and standard geo-demographic classifications commonly used in analyses of nonresponse bias, often with limited results (e.g. Lynn et al, 2012; Biemer and Peytchev, 2013), and consider data from a wider range of sources and at different levels of aggregation. Picking up on some of the potential challenges in relation to auxiliary data identified above, the project will need to address the following questions:

- What information is available from different auxiliary data sources which can be matched to the ESS sample records of respondents and non-respondents?
- Is accurate information available for all sampled addresses in the UK?
- How successfully can the risk of identification or disclosure be minimised whilst retaining the usefulness of the data for secondary analysis?

The project will explore auxiliary data from three main sources:

Small-area contextual data: It is widely recognised that neighbourhood contextual data can provide relevant information about response propensity, either by providing proxy information on the socio-demographic characteristics of the household (Campanelli et al, 1997) or because neighbourhood characteristics such as population density, crime rates and poverty can themselves influence people’s willingness to participate in surveys by affecting feelings of collective efficacy and social trust (Groves and Couper, 1998; Johnson et al, 2006). This project is an opportunity to explore the wide range of small-area administrative data being made available at increasingly low levels of aggregation by government departments and other public bodies in the UK. The optimal level of aggregation for contextual information used in nonresponse analysis will be investigated, taking into account the balance between disclosure risk and data utility.

Commercial marketing data: There is growing interest in the possibility of exploiting commercial marketing data in survey research (Blohm and Koch, 2013; Pasek, 2013). These data can provide information about households and their residents including demographic characteristics, property type and length of residence. Preliminary studies, mostly in the US, suggest that insights from these data may be limited, not least because of concerns over data completeness and accuracy (Pasek, 2013). Nevertheless, given the potentially rich variety of data available at household level these data are worthy of investigation in a UK context. This project will verify the quality of the underlying data empirically by studying the incidence and patterns of missing data and, where similar information is available from other sources (interviewer observation, ESS survey responses), making comparisons with commercial data.

Geocoded information from the Ordnance Survey (OS): on the physical location of sample units: Using geo-location indicators (e.g. grid references) to link to GIS-databases such as those administered by the OS allow sampled addresses to be located in relation to rich descriptions of geography including transport networks and local amenities. Such information – which can provide indicators of social as well as physical exclusion - is increasingly recognised as predicting social attitudes and behaviour (ODPM, 2003; Comber et al, 2011). This project will consider the possibility that such contextual information may also shed light on response patterns, for example, if proximity to local amenities serves as a proxy for social cohesion, which in turn is related to survey participation, or if indicators of physical remoteness are related to interviewer calling patterns and noncontact rates.

The question of how to maximise the amount of auxiliary information available for analysis whilst minimising the risk to respondent anonymity through disclosure will be addressed during the data matching process. We will consider different methods of statistical disclosure control, such as recoding, perturbation and record swapping (Skinner, 2009; Hunepool et al, 2012). The effectiveness with which these different approaches preserve respondent anonymity without loss of information will be evaluated by comparing the performance of different versions of the same auxiliary variables in nonresponse analyses.

Strand 2: Using auxiliary variables to identify nonresponse bias in the UK ESS

The second strand of the research (months 7 to 14) will be to take those auxiliary variables successfully matched to ESS survey data at stage one and use them to investigate possible sources of nonresponse bias in the UK ESS. In order to do this, two related but distinct research questions will be considered:

- To what extent can auxiliary data be used to predict sampled households’ propensity to respond?
- How are any auxiliary variables identified as predictors of survey response associated with different survey outcomes?

In previous studies, the focus has generally been on answering the first of these questions (Campanelli et al, 1997; Johnson et al, 2006; Lynn et al, 2012). Much less attention has been paid to the second question. However, to be useful for nonresponse adjustment auxiliary variables must be correlated both with response propensity and survey outcomes (Little and Vartivarian, 2005; Groves, 2006; Olson, 2013).

Logistic regression models will be used to predict sample units' propensity to respond. The relative predictive power of models including auxiliary data will be compared against models based on survey paradata and interviewer observations as previously used in ESS nonresponse analyses (Stoop et al, 2010). One question to address during the analysis will be how the factors associated with nonresponse analysis vary depending on the type of nonresponse i.e. non-contact vs. non-cooperation. A second will be how the usefulness of auxiliary data varies depending on the level of aggregation. A third question will be whether and how the predictors of nonresponse vary geographically. The possibility that contextual factors influence not only the overall response rate in an area but also the profile of those who respond, will be investigated using geographically weighted regression. This technique generates a different model for each location for which data is available allowing geographic comparisons (Brunsdon et al, 1996). We will develop visualization techniques, such as those used to show uncertainty error in the OAC geo-demographic classifier (Slingsby et al., 2011) to explore, present and communicate these geographically varying results and relate them to the most significant local explanatory variables.

The response models produced will be used to generate predicted response rates for each possible ESS sample point (i.e. postcode sector) in the UK. The effectiveness of these predictions will be evaluated against the actual response rates achieved in the different postcode sectors sampled for ESS Round 7 (fieldwork late 2014). Having an effective model to predict response rates down to sample point level will provide a more nuanced picture of response patterns than a simple extrapolation of national or regional response rates on past surveys. This information may be used to direct future fieldwork more effectively.

The final stage of the nonresponse analysis will be to take those auxiliary variables shown to be correlated with nonresponse – at either a national or local level - and investigate whether and how these variables are associated with survey outcomes. Analysis will focus on the correlation between auxiliary variables and items in the ESS core questionnaire, such as wellbeing, political evaluations, trust, civic participation, immigration and fear of crime.

Strand 3: Weighting for nonresponse bias using auxiliary variables

The third strand (months 11 to 16) will be to investigate the use of different weighting methods to correct for nonresponse bias, drawing on the auxiliary variables studied in the previous stages of the research. The following broad questions will be addressed using modern weighting methods:

- What is the effect on ESS survey estimates of including new auxiliary variables in nonresponse weights? How does the effect of these weights vary depending on the nature of the survey variables and estimates?
- Will the inclusion of new auxiliary variables in nonresponse weights reduce nonresponse bias?

Weighting methods can be classified into sample-based methods or population-based methods, depending on whether information on the auxiliary variables is available for the sample (respondents and non-respondents) or the population (Brick and Montaquila, 2009). This project will focus on sample-based methods consistent with the form of auxiliary data being investigated. However, the combination of sample-based methods with existing population-based post-stratification weighting approaches used with the ESS (Vehovar et al, 2013) will be considered in order to investigate the additional effect of sample-based weighting. Among sample-based methods, propensity weighting will be contrasted with generalized regression (GREG) weighting, a form of calibration weighting (Brick and Montaquila, 2009). The response propensity weights follow naturally from the earlier research exploring the dependence of response on the auxiliary variables. The GREG weights follow naturally from the research exploring the dependence of the survey variables on the auxiliary variables. It is anticipated that differences between the two methods will be more substantial for survey variables which are well predicted by the auxiliary variables (Micklewright et al., 2012).

A first research question will be to ask whether the use of new forms of auxiliary information leads to any change in the weighted estimates, measuring change both relative to the value of the estimate and relative to its standard error. We will explore how these changes depend on the nature of the survey variables and estimates and how the standard errors depend on changes in the weights. To judge whether the changes lead to a reduction in nonresponse bias we will undertake a limited simulation study along the lines of Biemer and Peytchev (2013), where the ESS respondents are treated as the population and an induced 'nonresponse' variable is created, e.g. by treating respondents who either refused initially or who required a certain number of call attempts as non-respondents. The weights will then be applied to the 'induced respondent' data and the weighted estimates compared with the full sample data.

One challenge will be missing data on the auxiliary variables, which may be quite substantial for some of the commercial data. If such data are to be used to construct nonresponse weights a decision must be made whether to use just the observed values or to impute missing values of the auxiliary variables. This project will compare the two approaches.

A further methodological question will be to consider how spatial analysis methods - for example the possible use of geographically weighted regression within a generalized regression estimator - might contribute to the construction of weighting adjustments.

Currently, the extent to which the sources of auxiliary data to be studied in this project can provide the means of identifying and correcting for survey nonresponse bias is unknown. It is anticipated that these multiple data -and the application of new techniques to study them - will reveal new insights which can be used to enhance survey data quality. However, even if the project ultimately finds that the scope for using auxiliary data in this way is limited, this knowledge will in itself represent an important advance in our understanding.

Data sources

Auxiliary data will be matched to survey paradata and interview data from Round 6 of the European Social Survey collected in the UK in 2012/13. Data will be matched for all sampled addresses (minus ineligibles) giving a total sample of 4,281 addresses across the UK. Matching will be conducted by Ipsos Mori in collaboration with the research team. Steps will be taken to ensure data confidentiality is maintained (see Data Management Plan).

Geo-coded administrative data at different levels of aggregation – where possible down to output area level (covering an average of 125 households) - will be matched onto sampled addresses using postcode and the ONS postcode directory which matches postcodes to the administrative areas into which they fall. The nonresponse analysis will explore the optimal level of aggregation for predicting response propensity. The final list of variables will be agreed during the scoping stage of the project. Drawing on existing theories of nonresponse, it will include information on the socio-demographic profile of the area as well as the social and physical environment e.g. crime rates, population density. Open access data from a range of sources will be considered including: the 2011 census (ONS), benefit claimant data (HMRC), Indices of Multiple Deprivation (DCLG), data on energy usage (DECC) and recorded crime figures (Police.uk) amongst others. ESS sample sizes are not sufficient to allow separate analyses of all four countries within the UK. Some analysis might have to be limited to England and Wales due to a lack of comparable administrative data across the UK.

Common geo-demographic classifications will be appended to the dataset including CACI's ACORN, Experian's MOSAIC classification and the OAC classification based on census data. Data will be purchased from two firms which provide data for commercial marketing purposes - Experian and CallCredit - and matched to the sample file at household level. Data will include tenure type, marital status, employment status, number and age of children, ethnic background, property value and length of residence. Identifying information such as name, telephone number or email address will not be included nor will sensitive information, for example on a households' financial situation.

Ordnance Survey data will be appended for each sample unit using the grid reference of the property closest to its unit postcode centroid. Postcode centroids are available through OS CodePoint, which is accessible through our institutional subscription to EDINA's Digimap service. The "OS MasterMap Integrated Transport Network" (a topologically consistent network of roads and pathways from which distances to amenities can be derived) is also available through Digimap We will derive variables indicating i) Euclidian distance ii) travel time to the nearest town and amenities such as parks, schools and community centres. Rather than giving exact times/distances, variables will indicate whether the address is within a given distance (e.g. 5km) or travel time (e.g. 10 minutes). The nonresponse analysis will explore the optimal size/distance metric for predicting response propensity.

Project Advisory Group and Consultation

A project advisory group (AG) will be established to provide a range of expertise on the use of auxiliary data and survey nonresponse. To maximise its relevance and reach, the group will include experts from the USA, Europe and the UK and engage non-academic stakeholders as well as academic experts. Chaired by ESS Director and project PI Rory

Fitzgerald, the AG will be made up of members of the project team plus: Dr Ineke Stoop, SCP Netherlands, an expert on survey nonresponse and chair of the ESS Balanced Response Rate Group; Dr Tom W. Smith, NORC, PI of the US General Social Survey who has published widely on the use of multi-source multi-level auxiliary data in survey research; Dr Peter Lynn, an experienced survey methodologist based at the Institute for Social and Economic Research, Essex and working on Understanding Society; the National Coordinator of the UK ESS; Patten Smith, Director of Research Methods at Ipsos MORI and chair of the Survey Research Association; a representative from ONS, and Richard Webber from Kings College London who developed the MOSAIC geo-demographic classification. The group will meet three times during the project (in Months 1, 8 and 14) to: advise on auxiliary data to be appended to the sample file, provide expert advice on the analysis of nonresponse bias, and comment on project outputs and dissemination opportunities.

In addition, towards the start of the project, members of the team will meet with key personnel from the administrative and business and local government data research centres being established as part of the ESRC's Big Data Network from 2013 to tap their expert knowledge of data sources and linkage issues.

Outputs and dissemination (Months 10 to 18)

This project has the potential to contribute significantly to our understanding of survey nonresponse bias and the statistical tools available to remedy this, to improve survey data collection and generate more robust data. Lessons learnt will enhance general social surveys in the UK and internationally. This will benefit the wide range of stakeholders involved in the funding, collection, and analysis of survey data and those who rely on the insights it provides. This includes academics across the social sciences, commercial survey agencies, policy makers, charities and, ultimately, the general public. The research team will build on existing links to members of the international survey community and work closely with the Advisory Group to ensure that the outputs meet the needs of both academic and non-academic stakeholders. The project will produce the following outputs:

- A project website giving a summary of project aims and key findings (including predicted response rates for UK sample points) plus links to all data on the ESS website;
- At least three articles by the team intended for publication in peer-reviewed survey methods and GI/visualization journals including: i) a descriptive article discussing the quality of the different auxiliary data sources and data linkage ii) an article reporting the results of the nonresponse analysis and different weighting strategies to correct for bias iii) an article on the spatial variation in nonresponse bias based on GWR analysis;
- Presentations given at academic conferences including the 2015 meetings of i) the European Survey Research Association and the American Association of Public Opinion Research, two of the leading associations for survey researchers and attended by academic and non-academic survey practitioners ii) GISRUK, the national focus for GIS research and IEEE VIS, the world's premier forum for advances in scientific and information visualisation;
- A workshop with ESS National Coordinators to discuss the feasibility of using similar approaches to correct for nonresponse bias in other European countries;

- A one day workshop for academic and non-academic stakeholders to discuss the implications of project findings for survey practice.

It is also intended to make the nonresponse weights and as much of the matched data as possible available to the research community via the ESS Data Archive under conditions of special license. A technical report giving details of auxiliary data sources and matching procedures will also be made available.