Informing Non-Response Bias Model Creation in Social Surveys with Visualisation

Kaisa Lahtinen[*] City University London

Aidan Slingsby[†] City University London

Jason Dykes[†] City University London

Sarah Butt[*] City University London

Rory Fitzgerald[*] City University London

Abstract

Through an ongoing process of co-design and co-discovery we are developing and using visualization to explore large amounts of auxiliary data from unfamiliar sources to understand non-response bias in social surveys. We present auxiliary data in their geographical contexts and show how this can complement traditional data analysis and provide a more comprehensive understanding of the data. This is helping select variables for non-response modelling. These processes are not just limited to non-response analysis, but have potential to be used in wider quantitative analysis in social science.

Keywords: correlation, variable selection, social science, survey response, geovisualization

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces; J.4 [Social and Behavioural Sciences];

1 Overview, Context & Data

We report on an on-going process of co-design and co-discovery in which visualization complements and enhances traditional statistical analysis and results in new graphical approaches that provide insight. These preliminary designs are the result of our experiments with visualization to address non-response in social surveys, one of the major challenges facing social science research [3].

Our focus is the European Social Survey (ESS) – a major cross-national survey of public attitudes – and in particular the potential for using auxiliary data to understand survey non-response bias. The key issue here is that if certain groups are inadequately represented due to their lack of engagement the survey results will be biased against their views. We hope to use auxiliary data to help explain and model non-response to ultimately address such bias.

The survey dataset consists of a geographically structured sample of 4,520 addresses selected to take part in Round 6 of the ESS in the UK 2012-13. These occur within 226 primary small sampling units (PSUs) from which 20 households have been sampled. These are shown in the inset map in Fig. 1 as 226 sets of 20 points. These data have been linked through their geographical locations to auxiliary data from administrative sources, commercial consumer profiling, official sources and open-source sources. Selected based on existing social theory, the resulting 401 candidate variables are considered as proxies for a range of social phenomena that might be associated with survey non-response. However, their value for non-response analysis has to be evaluated empirically.

2 The Challenge

Using these auxiliary datasets to understand survey non-response bias requires social scientists to interpret and evaluate these variables of different types (numeric, categorical, ordinal). Some of these data are at different geographical scales, have been drawn from a range of different and sometimes unfamiliar sources and have different geographical characteristics. Traditionally, data analysis uses well-known descriptive and inferential statistical methods. However, in this project the vast, various and largely unknown nature of the data makes this a significant challenge.

Using data visualization, We aim to develop and evaluate interactive visualization to support and enhance variable selection and broader analysis of survey non-response. In so doing, we aim to improve both analysis and understanding in social science and knowledge of applied visualization.

3 Visualization in Use

Discussions during which survey methodology and visualization approaches have been demonstrated have been used to understand the three main stages of the current workflow for analysis and the ways in which visualization may help:

i. understanding data distributions;
ii. assisting in variable selection;
iii. interpreting the results of statistical analysis and modeling

The 226 sets of 20 tightly clustered household locations (see inset map in Fig. 1) required novel cartographic design (Fig. 1) to present the data in a non-occluding manner that preserves the two-level hierarchy, yet retains much of its geographical distribution.

3.1 Understanding Data Distribution

As the auxiliary data are rich and unfamiliar to survey researchers, visualization that supports rapid comparison has proved useful for general validity checks, understanding geographical variation and establishing relationships between variables. Our specifically-designed maps help use see the spatial and statistical distributions for all 401 variables. In Fig. 1 (left) each square is a Primary Sampling Unit (PSU) in a non-overlapping geographical layout. Each PSU split into vertical bars slices that show the proportions of those addresses surveyed that resulted in responses (red), non-response (blues) and for which no eligible residents were identified (yellow). The PSUs in the maps in Fig. 1 (right) show the statistical and approximate geographic distribution of auxiliary data for each of the 20 sampled households. Four of the 401 variables are shown.

3.2 Assisting in Variable Selection

Survey analysts need to identify which of the variables are most likely to be useful in analysis of non-response. This involves identifying the most effective variables to measure each theoretical construct. Multiple candidates may be available from competing data sources or encoded in different ways. One example is the *Point of Interest* data available through the Ordnance Survey's commercial product and the freely available OpenStreetMap. Both sources cover comparable points, which may also be encoded in various ways – e.g.distance to closest point; number of points within distance; number of points within irregular area; density of points per area; proportion of points of interest of a particular type.

Whilst theory and statistical significance testing, for example of bivariate correlations, provide possible routes to select between these variables, visualization is adding value by helping analysts

[*] Centre for Comparative Social Surveys, City University London; e-mail: {kaisa.lahtinen | sarah.butt.1 | r.fitzgerald}@city.ac.uk.

[†] giCentre, Department of Computer Science, City University London; e-mail: {a.slingsby | j.dykes}@city.ac.uk

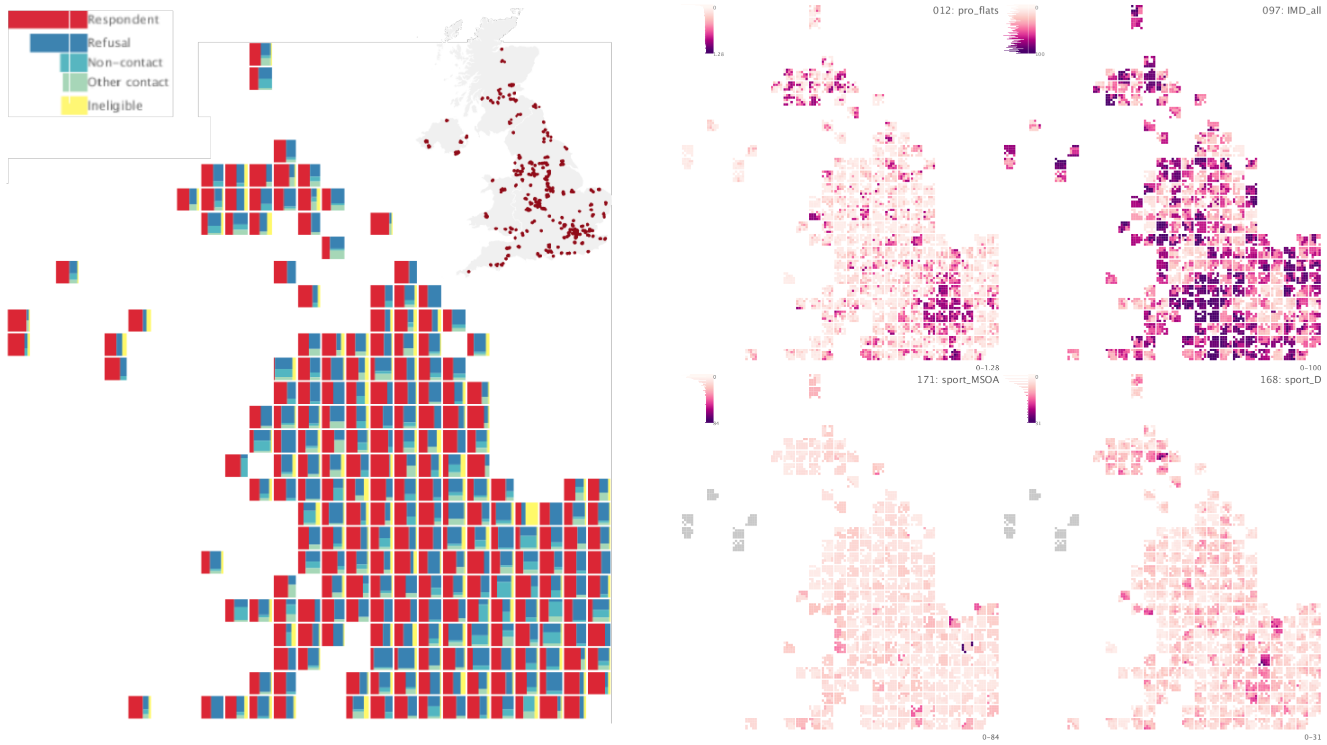

Figure 1: Fixed-size cartograms that show characteristics of the 226 PSU sets of 20 tightly clustered households (see inset map) in a non-occluding manner. *Left*: Proportions of the PSU household sample that responded, did not respond and were ineligible. *Right:* Auxiliary data by household (small squares) within PSUs (large squares). From top left to bottom right: 'proportion of apartments', 'Index of Multiple Deprivation (IMD)', 'sporting outlets within MSOA', and 'sporting outlets within 800m'. Legends contain histograms of statistical distributions.

understand the geographical inter-relationships between variables and the geographical relationship between auxiliary variables and survey response. Maps were used for multiple purposes during the variable selection. For example we could evaluate whether variables were simply re-illustrating the urban-rural divide or whether they were capturing more subtle neighbourhood differences. Being able to see nuanced differences between variables at low geographical levels was helpful in deciding between different encodings of the same variables. We know from previous research that geographically clustered variables can help understand how survey response patterns vary across PSUs [1]. Clustered variables can be selected using measures of spatial autocorrelation. However, being able to eye-ball variables that are strongly or weakly clustered using maps and comparing the geography of distributions may help us understand them and select variables with complimentary distributions.

3.3 Interpreting the Results of Statistical Analysis

Survey (non)response is a complex issue. The drivers of non-response bias are likely to vary geographically. It is possible to take account of this variation in statistical modelling through geographically weighted models [2]. Successfully interpreting the results of the statistical output – and pulling out the key findings for survey practitioners to inform future data collection – is more difficult. We are in the process of developing interactive graphics through which geographically varying model parameters and residuals can be compared in support of this activity.

4 Conclusion

Our initial work strongly indicates that visualization has potential. Co-design and co-discovery are helping the social scientists and visualization scientists in our research team to understand their domains and develop mutually beneficial approaches. Visualization – such as that shown in Fig. 1 – is helping us make sense of the the survey and auxiliary data in concert – stimulating discussions about modelling and outputs and helping with variable selection in the context of the established data regarding response types. It is having impact on the processes used in survey non-response analysis and modelling in the case of the ESS and new graphics are being generated to enhance, speed up and improve this process – enabling researchers to interact with data in a different way as they build and interpret their models. Task and sub-task analysis are ongoing in light of discovery and knowledge sharing as the response modelling process continues. These visualisation techniques and workflows are not limited to non-response analysis, but are applicable for boarder quantitative analysis in social science.

Acknowledgments

We wish to thank Prof. Chris Skinner and Prof. Chris Brunsdon for their advice. This work is part of the 'Auxiliary Data Driven Nonresponse Bias Analysis (ADDReponse)' project funded by the UK Economic and Social Research Council (grant ES/L013118/1).

References

[1] P. Biemer and A. Peytchev. Using Geocoded Census Data for Nonresponse Bias Correction: An Assessment. *Journal of Survey Statistics and Methodology*, 1(1):24–44, May 2013.

[2] A. S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester, England ; Hoboken, NJ, USA, 1st edition, Oct. 2002.

[3] R. M. Groves. Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5):646–675, Jan. 2006.