Proceedings of the 22nd GSS Methodology Symposium 2017
Methodology: Insight; Innovation; Implementation; Impact

ADRC-E: Making a Methodological Impact

Peter W. F. Smith[1]

Abstract

As part of its remit, the Administrative Data Research Centre for England is undertaking a programme of methodological research. In this paper I will describe a few examples from this programme including work on: guidance about the information that needs to be made available about the data linkage process by data providers, data linkers, analysts and researchers; a new approach to linkage based upon weights derived using a scaling algorithm; and the representativeness of surveys using a unique data set linking call record paradata from three UK social surveys to census auxiliary attribute information on sample households.

Key Words: ADRC-E; ADRN; GUILD; Record linage; Survey representativeness.

1. Outline

I will begin by giving an overview of the Administrative Data Research Network (ADRN) and in particular the Administrative Data Research Centre for England (ADRC-E), indicating what it is doing to facilitate administrative data research. I will then briefly describe three projects which are impacting or have the potential to impact considerably on administrative data research. All three have recently been published, so further details on them are readily available. The first project presents some guidance on the information that should be provided at various stages of the data linkage pathway from data collection to reporting. The second is a new more efficient method for record linkage with the desirable properties of probabilistic methods without some of their drawbacks. The third is a piece of research using survey data linked to census data to assess the representativeness of the surveys and how this changes as the data are collected.

2. Administrative Data Research Network (ADRN)

The ADRN is funded by the Economic and Social Research Council (ESRC) and was instigated in 2013. The setting-up of the ADRN was a recommendation of the Administrative Data Taskforce (2012). They recommended a centre in each of the four countries of the UK, because of their different administrative data environments. The Network also contains the Administrative Data Service (ADS), which has a co-ordinating role, being a central point of entry for researchers, leading on public engagement, outreach, policies and procedures, and hosting the ADRN website. Each Centre is also a partnership between academic institutions and national statistical organisations.

3. About the Administrative Data Research Centre for England (ADRC-E)

The ADRC-E is led by the University of Southampton (UoS) and run in collaboration with University College London (UCL), the London School of Hygiene and Tropical Medicine, the Institute for Fiscal Studies and the Office for National Statistics (ONS), with the three London institutions collectively known by the consortium as the Bloomsbury Group. Its Director is Peter Smith (UoS), with a Deputy Director based in each of the three groups: UoS, Bloomsbury and ONS. They are Dave Martin (UoS), Ruth Gilbert (UCL) and Lucy Vickers (ONS); day-to-day operations are overseen by an Assistant Director (Operations), Emma White (UoS).

[1]Peter W. F. Smith, Administrative Data Research Centre for England and the University of Southampton, P.W.Smith@soton.ac.uk

4. The other parts of the Network

The ADS is led by the University of Essex and run in collaboration with the Universities of Manchester and Oxford, with acting Director Tanvi Desai. ADRC Northern Ireland is led by Queen's University Belfast and run in collaboration with University of Ulster and the Northern Ireland Statistical Research Agency. Its Director is Dermot O'Reilly. ADRC Scotland is led by the University of Edinburgh and run in collaboration with the Universities of Aberdeen, Dundee, Glasgow, Herriot-Watt, St. Andrews and Stirling, and Scottish Government. Its Director is Chris Dibben. ADRC Wales is led by Swansea University and run in collaboration with Cardiff University and Welsh Government. Its Director is David Ford.

5. Administrative Data Research Centres

As part of their remit, Administrative Data Research Centres are required to:

- Provide state-of-the-art facilities for research access to de-identified administrative data by accredited researchers.

ADRC-E has facilities that can be accessed from secure labs in both the University of Southampton and the Farr Institute in Bloomsbury, and partially funds the ONS Virtual Microdata Laboratory (VML) with access points in Titchfield, Newport and Pimlico.

- Provide data management and statistical analysis support functions for external researchers accessing the data.

ADRC-E has administrative staff, data scientists, statisticians, etc. to support researchers using ADRC-E facilities across the three groups: UoS, Bloomsbury and ONS.

- Conduct original research using linked administrative data and related analytical and methodological approaches.

ADRC-E currently has over 50 research projects underway. We have given over 20 presentations at international research conferences and hosted the 2016 ADRN Conference. Our 2017 return to the ESRC included 92 publications in journals or conference proceedings.

- Engage in training, capacity building and public engagement.

Our 2017 return to the ESRC also included 68 engagement activities. In both 2014-15 and 2015-16, ADRC-E delivered at least 20 days of training and is doing the same in 2016-17. We have successfully organised and delivered training courses to around 600 researchers on various aspects of administrative data. ADRC-E has delivered eight ADRN Accreditation Training (now SURE Training) courses for ADRC-E staff and ADRN researchers across the government and academic sectors.

- Work in collaboration with other elements of the ADRN.

ADRC-E is heavily involved in all aspects of the ADRN and I currently direct the Network and chair the Directors Group meetings. ADRC-E members actively participated in the former Management Group and ADRN Working Groups, and continue to participate actively in the Directors Group, the Operations Group and their Subgroups and Task Teams. We have contributed to the development of ADRN strategy, principles and policies, the Approvals Panel and the Administrative Data Service (ADS) self-review processes, gateway reviews and the mid-term review.

6. GUILD: Guidance for information about linking data sets

Joint work across the ADRC-E consortium (Bloomsbury: Ruth Gilbert, Gareth Hagger-Johnson, Katie Harron, Harvey Goldstein; UoS: Li-Chun Zhang, Peter Smith; and ONS: Rose Lafferty) and beyond (Chris Dibben, ADRC Scotland) has produced some guidance for information about linking data sets. Also, there were contributions from a team of UK experts and attendees at an international workshop on data linkage.

6.1 Guild: Why?

The data linkage pathway is fragmented and involves different teams from data collection through linkage to analysis and reporting. There can be little exchange of information between the teams, some of which may be a result of the need to separate processes, e.g., the linkage of identifiers and the data analysis of the functionally anonymised linked attribute data. However, analysts and report writers need to have information on earlier processing in order to assess its impact on their results. Therefore, linkers need to provide information on source of linkage errors and who are affected, since linkage errors maybe larger for subgroups of interest. If data providers have awareness of the methods used downstream and the impact of data quality on linkage errors and study results, then they may be able to take action to enhance the data quality in useful ways. The overall aim of providing the GUILD guidance (Gilbert et al., 2017) was to enhance the quantity and analysis of linked data. Hopefully, it will also promote further discussion of the issues.

6.2 Guild: What?

The GUILD guidance divides the data linkage pathway into four areas: data provision, data linkage, data analysis and reporting study findings, and recommends what information should be provided at each stage. For example, data providers should provide information on the population and geographical coverage of the data and how any opt-outs have been dealt with. They should also share any information which might help the linkage process such has how the data were collected and cleaned and whether any disclosure control has been applied. Data linkers should provide a description of the linkage process, record-level indicators of the linkage process, aggregate-level linkage results, generic reports of linkage accuracy, description of any disclosure control performed and an overview of data linkages they have undertaken. Data analysist should account for data linkage error, and study reports where possible should contain information from all three stages. See Gilbert et al. (2017) for full details.

7. Record linkage

Members of ADRC-E's Bloomsbury Group have developed a new more efficient method for record linkage with the desirable properties of probabilistic methods without some of their drawbacks (Goldstein, Harron and Cortina-Borja, 2017).

7.1 The problem

Most probabilistic approaches to record linkage are based on the theory formulated by Fellegi and Sunter (1969). The probability of the observed agreement/non-agreement of p identifiers for the lth pair of records is

$$P(y_{l1}, \dots, y_{lp}) = \left(\prod_{j=1}^{p} m_j^{y_{lj}} (1 - m_j)^{1-y_{lj}} \right) \pi + \left(\prod_{j=1}^{p} u_j^{y_{lj}} (1 - u_j)^{1-y_{lj}} \right) (1 - \pi), \qquad (1)$$

where $y_{lj} = 1$ (0) if the jth identifier agrees (disagrees). It is the sum of a term corresponding to the situation where the record pair is a true match and a term where the record pair is not a true match, and depends on the probabilities that an identifier agrees in these two situations (latent classes): m_j and u_j, respectively. Here π is the probably that a record pair is a true match. The product of $P(y_{l1}, \dots, y_{lp})$ over record pairs is assumed to be a likelihood and

maximised to obtain estimates of m_j and u_j, typically using the Expectation Maximisation (EM) algorithm. Once the m_j and u_j are estimated, some function of them (e.g., the log-ratio of the two products in Equation (1)) is used to give a score to each agreement/non agreement profile. These scores can be used to classify record pairs as matches or non matches or even 'undecideds'.

This method has some limitations. It assumes that for any pair of records, agreement/non-agreement on each of the p identifiers is independent. This might not be true, since errors in identifiers for a single record resulting in non-agreement for a true match might be correlated. It also assumes that record pairs are independent, which is unlikely to be true, since each record is involved in many pairs. Furthermore, iterative algorithms, such as the EM algorithm, are required to estimate the m_j and u_j. These are computationally and storage-intensive, so struggle with large datasets, and convergence is not guaranteed.

7.2 The solution

The solution presented by Goldstein, Harron and Cortina-Borja (2017) is to use a less computationally demanding method to calculate the scores (weights), based on agreement categories for each identifier with K agreement categories in total. It is a scaling approach that uses ideas from correspondence analysis to assign weights to each agreement category, based on a model originally discovered by Fisher (1938) and subsequently rediscovered by many others including Healy and Goldstein (1976). It is computationally efficient, based on counting the numbers of pairs of records in each identifier agreement category and each pair of identifier agreement categories. It is also storage-efficient, requiring only a $K \times K$ matrix. The algorithm is not iterative, so convergence is not a problem. Independence between identifiers is mirrored in the basic implementation, since the extent of agreement on each identifier is given a separate weight. However, it is not required, since agreement categories can be based simultaneously on two identifiers.

8. Survey Representativeness

Members of ADRC-E's Southampton Group have used survey data linked to census data to assess the representativeness of the surveys and how it changes as the data are collected (Moore, Durrant and Smith, 2017).

When collecting a survey we want to minimise nonresponse bias. One way is to minimise nonresponse, but this is problematic with diminishing returns from repeated attempts to secure a response and there is evidence that this has little effect on nonresponse bias. Therefore, focus has switched to monitoring subgroups of surveys to ensure they are properly represented using indices, such as the coefficient of variation of the response propensities and dissimilarity index. These are calculated as the data are collected to inform adaptations in data collection strategies to maximise quality and/or reduce costs.

Moore, Durrant and Smith (2017) use linked call record data from the three UK social surveys to census household attribute information to assess representativeness over calls. The three surveys are the Labour Force Survey, the Life Opportunities Survey and the Opinions Survey. Results suggest that representativeness does not increase after 6 to 8 calls. Stopping here would reduce the number of calls made by 7 to 15 percent. Based on the findings of this work, the ONS has reduced the maximum calls made to a survey subject from 20 to 13 calls and is conducting tests to confirm if further reductions to between 6 and 8 calls do not result in impacts on survey data quality. This will result in substantial financial savings.

9. Contacting us

To find out more about these projects and all the projects being supported by the ADRN visit adrn.ac.uk. The ADS can be contacted by email (help@adrn.ac.uk) and telephone (01206 873435). For more detail about ADRC-E visit www.adrn.ac.uk/centres/england or email adrce@soton.ac.uk. You can also follow us on Twitter @ADRC_E.

Acknowledgements

The research highlighted in this paper was funded by the Economic and Social Research Council, grant reference number ES/L007517/1, establishing the Administrative Data Research Centre for England (ADRC-E). The ADRC-E is led by the University of Southampton and run in collaboration with University College London, the London School of Hygiene and Tropical Medicine, the Institute for Fiscal Studies and the Office for National Statistics (ONS). The findings, interpretations and conclusions expressed in this paper are entirely those of the author.

References

Administrative Data Taskforce (2012). The UK Administrative Data Research Network: Improving Access for Research and Policy. Available online: [30/08/2017] https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-administrativedatataskforcereportdecember201_tcm97-43887.pdf.

Fellegi, I. and Sunter, A. (1969). A theory for record linkage. Journal of the American Statistical Association, 64, 1183–1210.

Fisher, R.A. (1938). Statistical Methods for Research Workers, 7th edition. Edinburgh, Oliver & Boyd.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L-C., Smith, P., Dibben, C. and Goldstein, H. (2017). GUILD: Guidance for information about linking data sets. Journal of Public Health, 1–8. DOI: 10.1093/pubmed/fdx037.

Goldstein, H., Harron, K. and Cortina-Borja, M. (2017). A scaling approach to record linkage. Statistics in Medicine, 36, 2514–2521.

Healy, M.J.R. and Goldstein, H. (1976). An approach to the scaling of categorised attributes. Biometrika, 63, 219–229.

Moore, J.C., Durrant, G.B. and Smith, P.W.F. (2017). Dataset representativeness during data collection in three UK social surveys: generalizability and the effects of auxiliary covariate choice. Journal of the Royal Statistical Society, Series A. Available online: [30/08/2017] http://onlinelibrary.wiley.com/doi/10.1111/rssa.12256/pdf.