Contribution to the discussion of "A comparison of sample survey measures of earnings of English graduates with administrative data"

Paul A. Smith, S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ. p.a.smith@soton.ac.uk

I welcome the general approach to the use of linked administrative data under appropriate access conditions. In this response I will focus on the quality of the compared data sources, rather than the substantive results. This paper presents an interesting example where the data quality from both the linked administrative and corresponding survey sources are subject to errors. In this case the data are linked deterministically through National Insurance numbers (NINos), and while there may be some minor issues related to temporary NINos, the linkage is likely to be effective. However, since the analysis depends on the linkage, some follow up of cases which did not link might have been useful, to understand whether they were errors, or just cases that should not link.

It is well recognised that administrative data are not collected directly for statistical purposes, and therefore that some definitions are not ideal for the analyses we would like to do. However, there has been less investigation of measurement errors and 'respondent' interactions with administrative systems, so the errors in those systems are less well known statistically. However, survey data are also subject to biases and other errors, and in this particular example the response rates in the Labour Force Survey have been declining, which makes it potentially more susceptible to selection effects. However, the LFS questionnaire is specifically designed and targeted at high quality measurement of variables of interest.

This paper presents very interesting observations on the differences in the two approaches. It is not clear which is best, as there are errors on both sides. In a perfect word I would wish for a "total error" evaluation (like Total Survey Error (Groves & Lyberg 2010) but adapted where necessary for administrative data). Perhaps the eventual arbiter would be to match the LFS data with tax information (which has been done experimentally with PAYE, see ONS 2018). The LFS does not collect NINo, however, so in this case the match would not be deterministic and it would be important to account for linkage error in analyses (Smith & Chambers 2018).

References

Groves, R.M. & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly* **74** 849-879.

ONS (2018) *Exploring the use of external data to assess for observed bias in Labour Force Survey estimates: interim findings*. Available from https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/exploringtheuseofexternaldatatoassessforobservedbiasinlabourforcesurveyestimatesinterimfindings (accessed 13 November 2018).

Smith, P.A. & Chambers, R.L. (2018) Discussion of "Statistical challenges of administrative and transaction data" by David J. Hand. *Journal of the Royal Statistical Society, Series A*. 181 585. doi: 10.1111/rssa.12315.