Are trajectories of dataset representativeness during survey data collection generalizable?

Evidence from the 2011 Census Non Response Link Study

Jamie C. Moore

Gabriele B. Durrant

Peter W. Smith

S3RI & ADRC-E, University of Southampton

Minimising (risks of) survey non-response bias

- NR bias = that part of survey measurement error due to non-response.
- To minimise risks, previously response rates maximised.
- However, declines in responses rate and low association with measured biases.
- Now, monitoring of response within and between sample subgroups recommended.
- Monitoring often undertaken during data collection, to inform adaptations to maximise quality and / or minimise costs.

Quantifying NR bias risks: representativeness indicators

- Measure similarity to sample (representativeness) in terms of variation in sample estimated response propensities.
- Attribute information on entire sample needed.
- Two indicators:

  R indicator = 1 – 2SD

  Coefficient of Variation of response propensities *(CV) = SD / r*

  where *SD* = standard deviation of response propensities

  r = response rate

- R Indicators close to one and CVs close to zero imply high representativeness.
- Indicators (specific to attribute covariates) also decomposable to assess impacts associated with different covariates / categories.

- For a review of representativeness indicators, see Wagner J. (2012) Public Opinion Quarterly 76: 555-575.

Applying these methods

- Still few reports.
- Sample information from population register or administrative data.
    - what if such sources don’t exist?
    - can linked census data be utilised?
- Patterns of representativeness during data collection (trajectories) & phase capacity points differ between surveys.
    - not generalizable?
    - or due to different samples?

The utility of Census sample attribute information

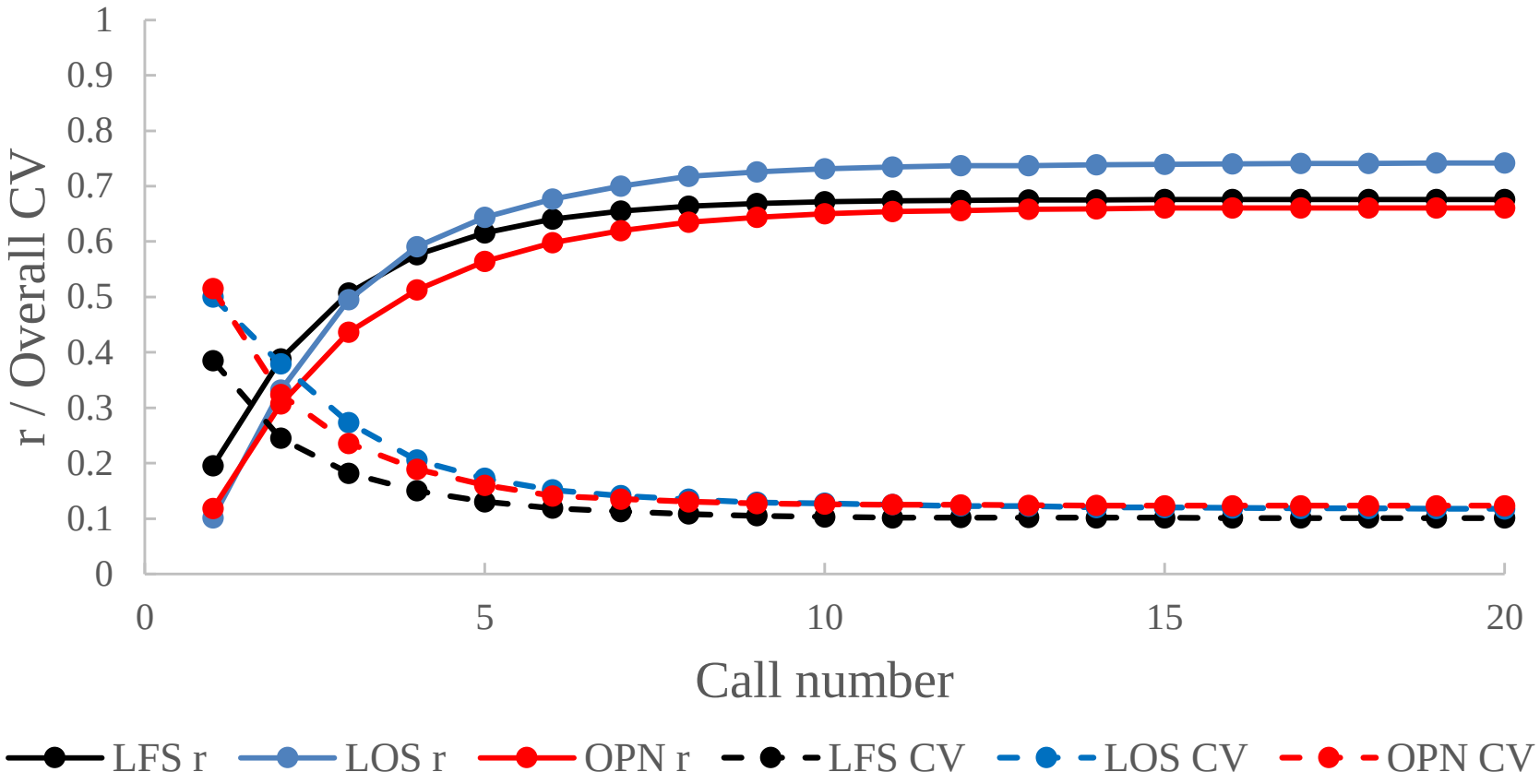
- We link call record data to census household (HH) attribute information on samples in three UK social surveys.
- Developing the ONS 2011 Census Non-Response Link Study (CNRLS).
- Surveys are the Labour Force Survey (LFS), Life Opportunities Survey (LOS) & Opinions Survey (OPN).
- Automated and (if imperfect) clerical linkage of HHs to 27th March 2011 Census records.
- Census data of great utility: >95% linkage rates & a rich suite of attribute variables available.
- Interviews <> 3 months of census day though, so question remains as to whether linkage rates drop further away.

Quantifying representativeness trajectories

- Interviews face to face, up to 20 attempts.
- Successful HH member interview = response (else non-contact or refusal).
- For each survey, compute *CV*s and partial variants at each call (1 to 20).
- Estimate response propensities using / compute partial indicators for 10 HH attribute covariates:

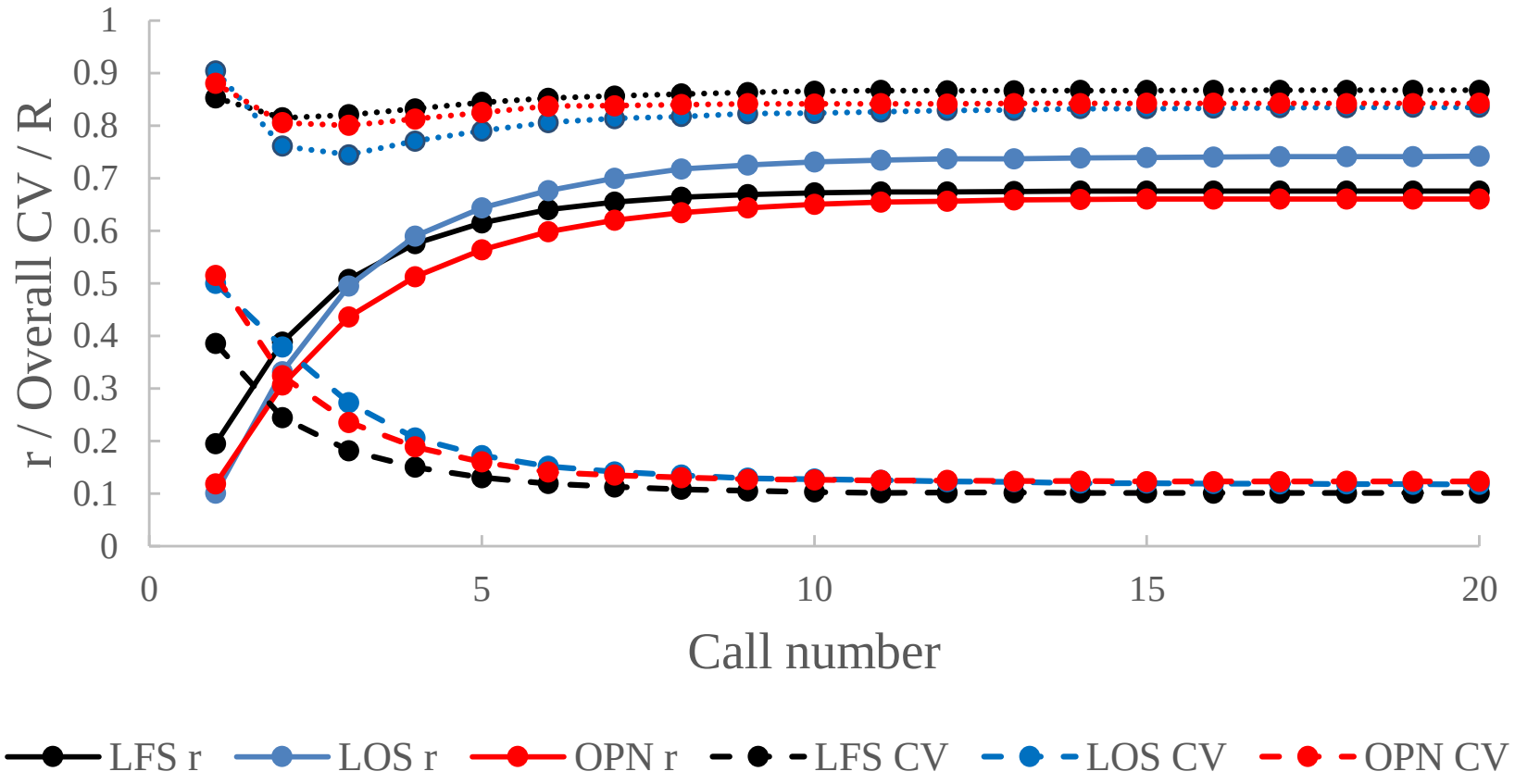
  HH Economic Status, HH Structure, Accommodation type, Tenure type, Cars available, Ill Health individual in HH, Impaired individual in HH, Retiree in HH, English fluency in HH, Located in London / SE.

Overall representativeness trajectories



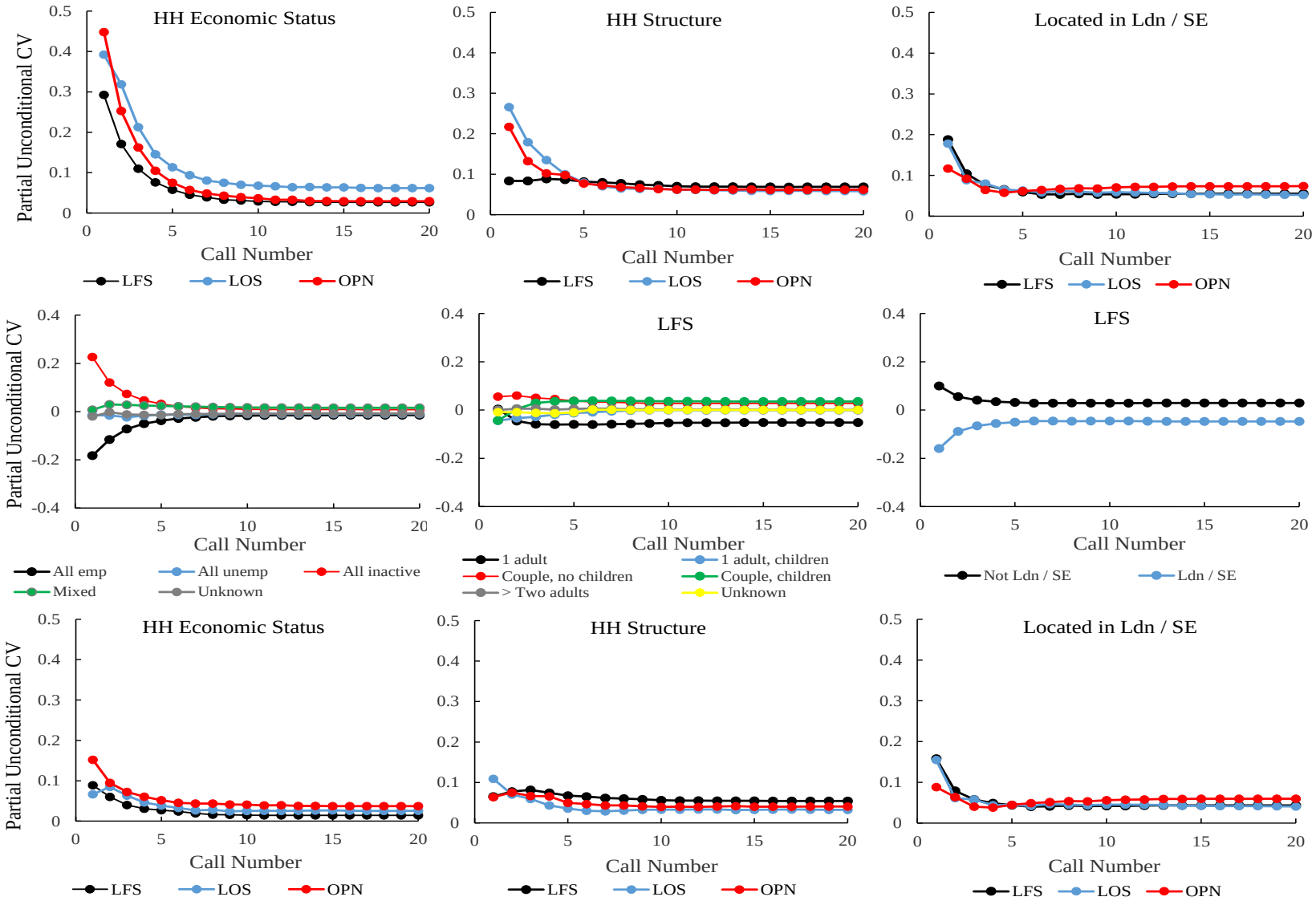
- LFS *N* = 18997, final *r* = 65.7%; LOS *N* = 6469, final *r* = 70.1%; OPN *N* = 6249, final *r* = 64%.

Overall R indicators



- Low *r* = limited potential for response propensity divergence and *R* indicator positive bias.

Partial by covariate (category) *CV*s



Phase capacity points

- When indicator within 0.02 of minima (adaptive strategy) or previous value (responsive); points earlier when threshold increased.
- Overall:

Survey	Adaptive point (% calls saved)	Responsive point (%) calls saved
LFS	6 (7.6%)	5 (12.2%)
LOS	8 (15.2%)	7 (18.2%)
OPN	6 (13.4%)	6 (13.4%)

- By covariate (adaptive strategy):

Survey	HH Economic Status	HH structure	Located in Ldn / SE
LFS	6	1	4
LOS	7	6	4
OPN	7	5	4

Summary

- The census is a high quality source of attribute information for assessing dataset representativeness when surveys are within 3 months of its date.
- Its utility further from this date though, needs to be studied.
- Representativeness increases at a decreasing rate over call records similarly in the three surveys.
- Sources of non-representativeness are under-representation of economically active HHs, HHs located in London / SE, and single adult HHs.
- Despite the above, data collection phase capacity points differ between surveys.
- Hence, in this case we recommend caution when generalising strategies.

Further / other work

- Similar analyses at the level of the individual respondent (rather than HH).
  - Interviewer effects and proxy responses.
- Actual biases in survey answers during data collection.
  - MSEs as well.
  - Negotiating access to linked survey data for this.
- Record linkage.
  - Methods here can be used to evaluate linked datasets compared to sources.
  - Apparent there are two components to dataset representativeness: who consents to linkage, & who, given linkage methods, is linkable.

Acknowledgements

This research was funded by the ESRC National Centre for Research Methods, Workpackage1 (grant reference number ES/L008351/1) and the ESRC Administrative Research Centre for England (ADRCE) (grant reference number ES/L007517/1). This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.