European Social Survey

Using geocoded auxiliary data to predict nonresponse in address-based samples: Are household- level commercial data any better than aggregate-level census data?

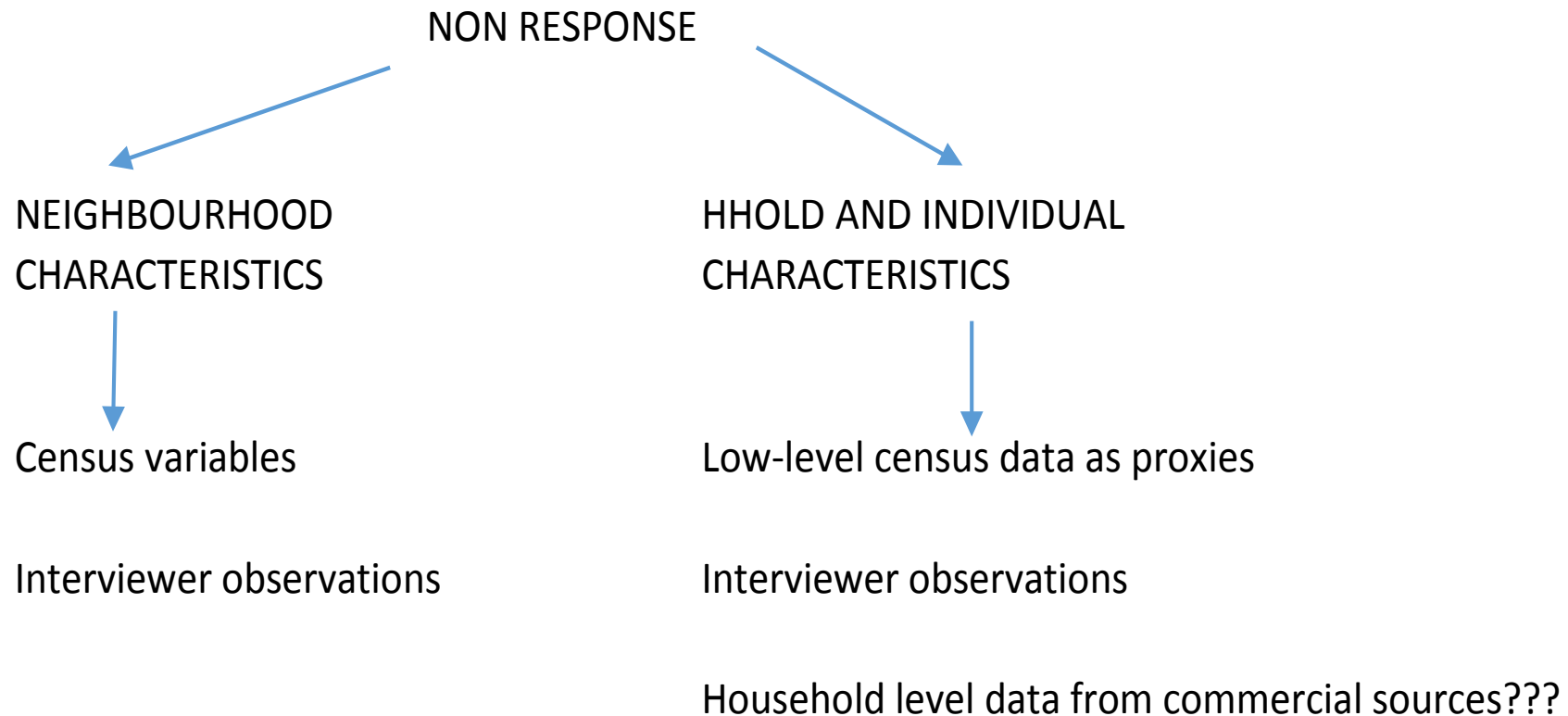**Enhancing survey data with geocoded auxiliary data**

ESRA, Reykjavik
15 July 2015

Kaisa Lahtinen

Sarah Butt

Rory Fitzgerald

City University London

Research Question

Can auxiliary data provide useful information with which to understand and correct for nonresponse bias in general social surveys?

Are household-level commercial data any better at predicting nonresponse than aggregate-level census data?

Motivation

Commercial data is available for respondents and nonrepsondents at household/individual level …

NON RESPONSE

NEIGHBOURHOOD CHARACTERISTICS

HHOLD AND INDIVIDUAL CHARACTERISTICS

Census variables

Low-level census data as proxies

Interviewer observations

Interviewer observations

Household level data from commercial sources???

Motivation

… but there are concerns about data quality …

- Lower accuracy compared to self-reported data

- Completeness of commercial data (Pasek et al 2014)

- Lower accuracy compared to interviewer observations (Sinibaldi 2014)

- Useful for predicting eligibility, but limited usefulness when predicting other response patterns (West et al 2015)

Motivation

… but there are concerns about data quality …

- Lower accuracy compared to self-reported data

- Completeness of commercial data (Pasek et al 2014)

- Lower accuracy compared to interviewer observations (Sinibaldi 2014)

- Useful for predicting eligibility, but limited usefulness when predicting other response patterns (West et al 2015)

… and resource implications

- Cost of the data

- Access and archiving of the data

Our study

First in-depth examination of commercial data in UK

Part of a wider project exploring auxiliary data and nr bias

Auxiliary Data Driven nonResponse bias analysis (ADDResponse)

Examine nonresponse to ESS R6 in UK (2012-13)

- General social survey.

- Address based sample with respondent selection on doorstep. RR 53%

- Potential extensions to other ESS countries in future

Our study

Model response propensity as function of:

- ESS interviewer observations

- 2011 Census data (at different levels of aggregation)

- Household level variables from commercial companies

  - Experian

  - Household classifications (MOSAIC)

Explore:

- Separate models for contact and co-operation

- Whether commercial data predictive of nonresponse on its own

- Whether commercial data improve model fit/remain significant alongside census/interviewer observations

European Social Survey

Quality issues with commercial data

European Social Survey

There are some quality issues

- Missing data
  - Highest amongst ineligibles followed by non-contacts
  - Potential predictor of nonresponse?

- Divergence between two commercial data sources

- Timeliness
  - Commercial data from 2015. ESS conducted 2012-13
  - 36% movers in the past 3 years
  - Potential predictor of nonresponse?

Results

Commercial data is predictive of nonresponse

Contact

- Significant: Missingness, household size, owner occupation, marital status, income

- Not significant: children, full-time employment, financial stress, movers

- Pseudo adj. R2 0.107

Results

Commercial data is predictive of nonresponse

Contact

- Significant: Missingness, household size, owner occupation, marital status, income

- Not significant: children, full-time employment, financial stress, movers

- Pseudo adj. R2 0.107

Co-operation

- Significant: Missingness, household size, full-time employment

- Not significant: children, owner occupation, marital status, financial stress, income, movers

- MOSAIC predicting well

- Pseudo adj. R2 0.018

Nested models: Contact

Model 1 and 2: Interviewer observations and census data

- Interviewer observations (for type of property) strongly predictive
- Adding census variables only improve the fit a slightly

Model 3: Interview observations, census, and commercial data

- Best fitting model
- Census variables and interviewer observations remain significant.
- Some of the commercial variables remain significant

Nested models: Co-operation

Model 1 and 2: Interviewer observations and census data

- Interviewer observation and census variables both significant

- Poor fit

Model 3: Interview observations, census, and commercial data (+MOSAIC)

- Interviewer observations and census remain significant

- MOSAIC doing majority of the work

- Movers also important

- Poor fit

Conclusions

Commercial data has potential to add value

- Predictive of nonresponse
  - More for contact than co-operation

- Provides additional information beyond census data

- Helpful to understand nonresponse mechanisms
  - Not sure about bias yet

- Do findings extend to other data sources?

European Social Survey

CONTACT

✉ kaisa.lahtinen@city.ac.uk

€ www.europeansocialsurvey.org

www.addresponse.org