Selectivity in Big Data Sources

An overview of methods for treating selectivity in big data sources

Methodological study

Purpose:

- **Identification of methods for treating selectivity in big data sources**
- **Initially to serve internal needs**

Methodologists:

- **Maciej Beręsewicz (Poznan University of Economics, Poznan Statistical Office)**
- **Risto Lehtonen (University of Helsinki)**

Study managed by SOGETI

Methodological study

Activity 1 - Analysis of big data sources

identification of the type of selectivity found in big data sources with focus on mobile phone data and social media

- *implication on the applicability of methods to address selectivity*

Activity 2 - Summary of selectivity treatment methods literature

Activity 3 - Analysis and evaluation of methods

Including evaluation of applicability and deeper description of the method.

What's big data *Addressing big data from a methodological point of view*

- **Defining big data:**
  - *Non-probabilistic nature*
  - Designed data and organic data
  - Big data in the context of statistical data sources

- **Mediated Data generation**
- **Variables and feature extraction**
- **Multitude of populations**

A potential statistical definition: Similarities with internet surveys Big data as an opt-in panel survey Opt-in -> self-selection from non-response research

- **Internet -> multiple frames**

Selectivity (and bias)

Sources of bias

- **Big data source specific error (technology)**
- **Unit specific selectivity (self-selection)**

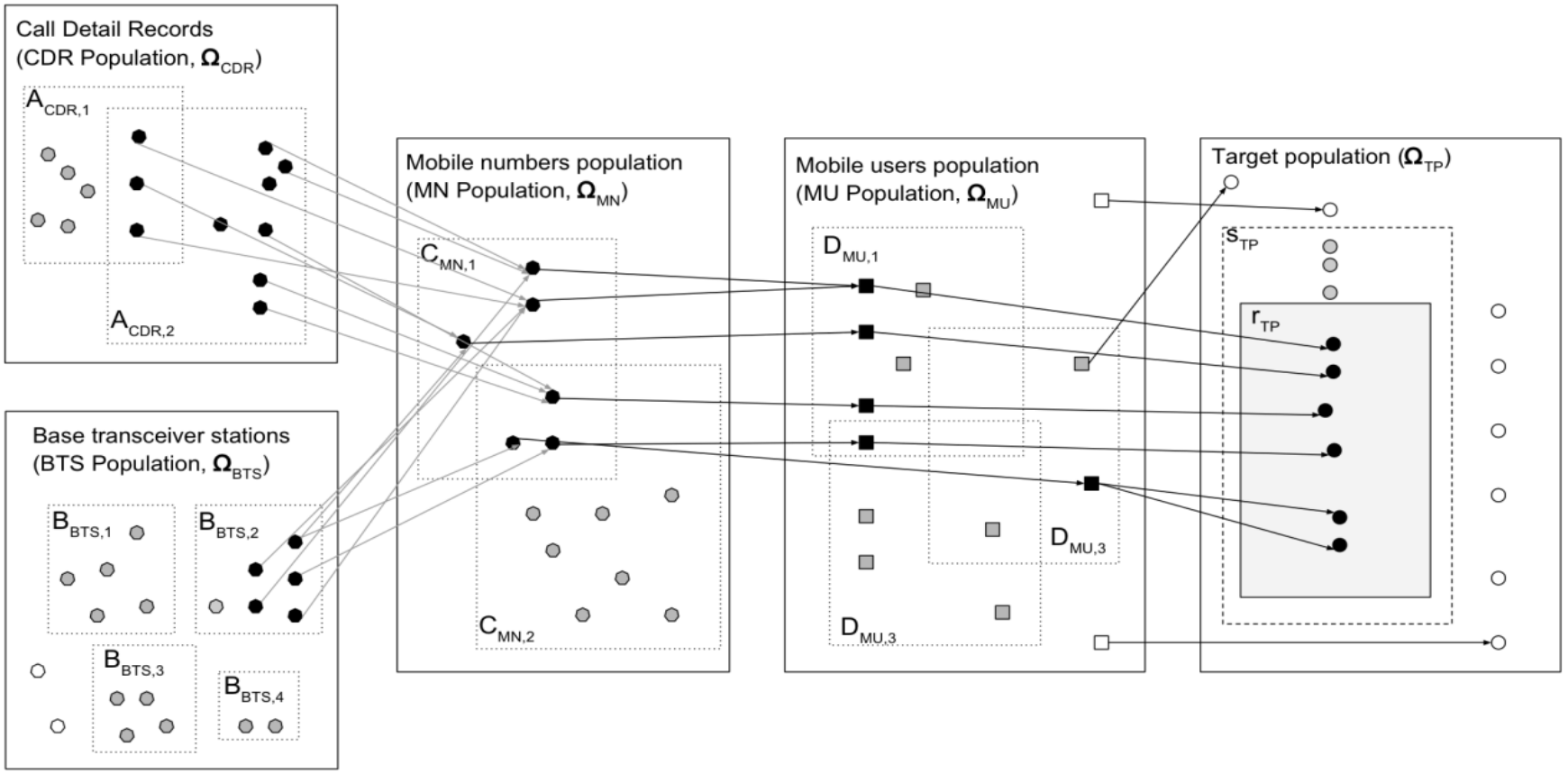
Coverage and non-response

Unit identification problem and measurement error

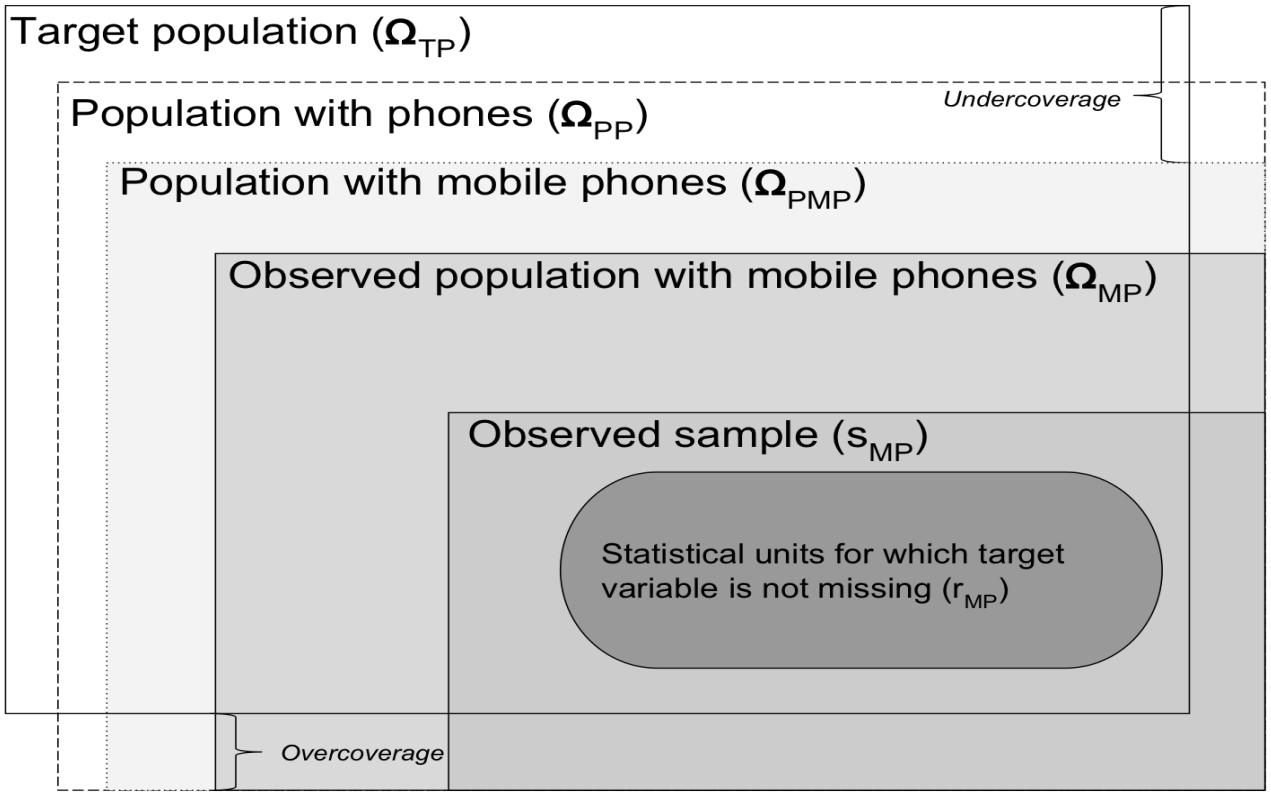
- **To be addressed before self-selection is corrected**

"Imputation" of latent target and auxiliary variables – a "measurement" error to be taken into account

Analysis of specific BD sources



Analysis of specific BD sources



Methods for correcting selectivity

Do we have independent data sources to verify the quality?

- **Compare estimates;**
- **Compare models;**
- **Link or match?**

Methods:

- **Unit level approach**
- **Domain level approach**

Unit level methods of selectivity correction (1)

Pseudo-design approach - Reweighting

- **Methods that account for existing information about auxiliary variables -> if correlated to selectivity mechanism will correct it**

  Generalized weight share method; (weighting for imperfect *frame*)

  Calibration (model-free and model calibration);

  Pseudo-empirical likelihood.

- **Methods that address directly the selectivity mechanism**

  Propensity weighting (if necessary or appropriate by subsequent adjustments)

  Two-stage weighting method (by modelling joint dstn XY in pop and data source: need for a probability based sample)

Unit level methods of selectivity correction (2)

Modelling approach

- **The basic idea is that if the models include explanatory variables correlated to the selectivity mechanism then they can correct or mitigate selectivity bias**

Role of data integration

- **Linkage**
- **Matching**

Domain level methods of selectivity correction (1)

Pseudo-design- Reweighting

- $\breve{\theta}_{cd}^{adj} = \breve{\theta}_{cd} \times a_d^{cover} \times a_d^{active} \times a_d^{share} \times a_{cd}^{cal}$
- a^{cover} **adjustment for coverage of technology**
- a^{active} **adjustment for fraction of active users**
- a^{share} **adjustment for market share data provider**
- a_{cd}^{cal} **adjustment to know population totals (e.g. background variables)**
- **Valid when MAR and corrects for coverage problems**

Domain level methods of selectivity correction (2)

Modelling approach

- **Use multiple sources**
- **Estimation of bias in big data source**

  e.g. From a sample survey (or registers)
  Direct estimation + model for sub-domains

Summary

Characteristic	Mobile	Twitter	Google Trends	Wikipedia
Unit-level corrections	possible	possible	impossible	limited
Domain-level corrections	possible	possible	very limited	limited
Existing sources	available but limited	unavailable (only for social media)	unavailable	available
Background on population	available but limited	unavailable	unavailable	limited
Paradata	available	available	unavailable	available
MNAR	unlikely	very likely	likely	likely
Measurement error in target variable	likely	very likely	very likely	very likely

Conclusions

The existence of auxiliary information is crucial

We need to understand the selectivity mechanism

There are methods applicable at individual and aggregated level

Future outcomes

A list of the relevant literature will be available online.

Working paper with relevant results of activities 1 and 2.

Contacts:
Loredana Di Consiglio: loredana .di-consiglio AT ec.europa.eu
Martin Karlberg: martin.karlberg AT ec.europa.eu
Fernando Reis: fernando.reis AT ec.europa.eu