Mobile Phone Data for Official Statistics: Elements for a Production Framework

BigSurv 2018

David Salgado (coord.) (INE, Spain)

Marc Debusschere (Statistics Belgium, Belgium)

Ossi Nurmi, Pasi Piela (Tilastokeskus, Finland)

Elise Coudin, Benjamin Sakarovitch (INSEE, France)

Sandra Hadam, Markus Zwick (DESTATIS, Germany)

Roberta Radini (ISTAT, Italy)

Martijn Tennekes (CBS, Netherlands)

Ciprian Alexandru, Bogdan Oancea (INSSE, Romania)

Elisa Esteban, Soledad Saldaña, Luis Sanguiao (INE, Spain)

Susan Williams (ONS, UK)

Barcelona, 24-26 October, 2018

Overview

1. Deliverable **5.1**:

   Current **status of access** to mobile phone data in the ESS (July, 2016)

2. Deliverable **5.2**:

   Guidelines for the access to mobile phone data within the ESS (May, 2017)

3. Deliverable **5.3**:

   Proposed elements for a **methodological framework** for the production of official statistics with mobile phone data (Feb, 2018)

4. Deliverable **5.4**:

   Some **IT elements** for the use of mobile phone data in the production of official statistics (March, 2018)

5. Deliverable **5.5**:

   Some **quality aspects and future prospects** for the production of official statistics with mobile phone data (May, 2018)

+ Meeting **Minutes** + **Github** Page

Big Data definition revisited

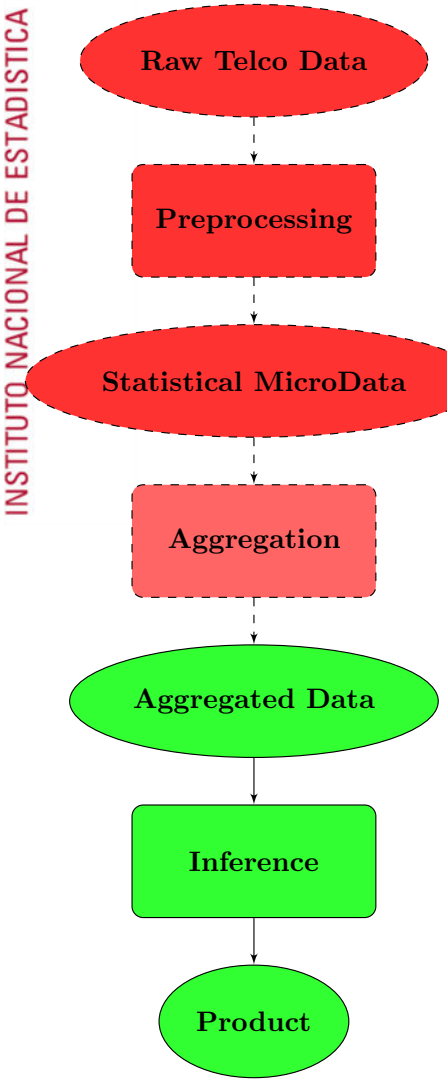
Big Data ***for Official Statistics***

- refer to **third people** and not to data holders;
- are **central in their economic activity**;
- **lack statistical metadata** (since they are generated for very different purposes).

UNECE (wider) Definition for **Admin Data**:

“Data collected by sources **external to statistical offices**.”

The process in a nutshell



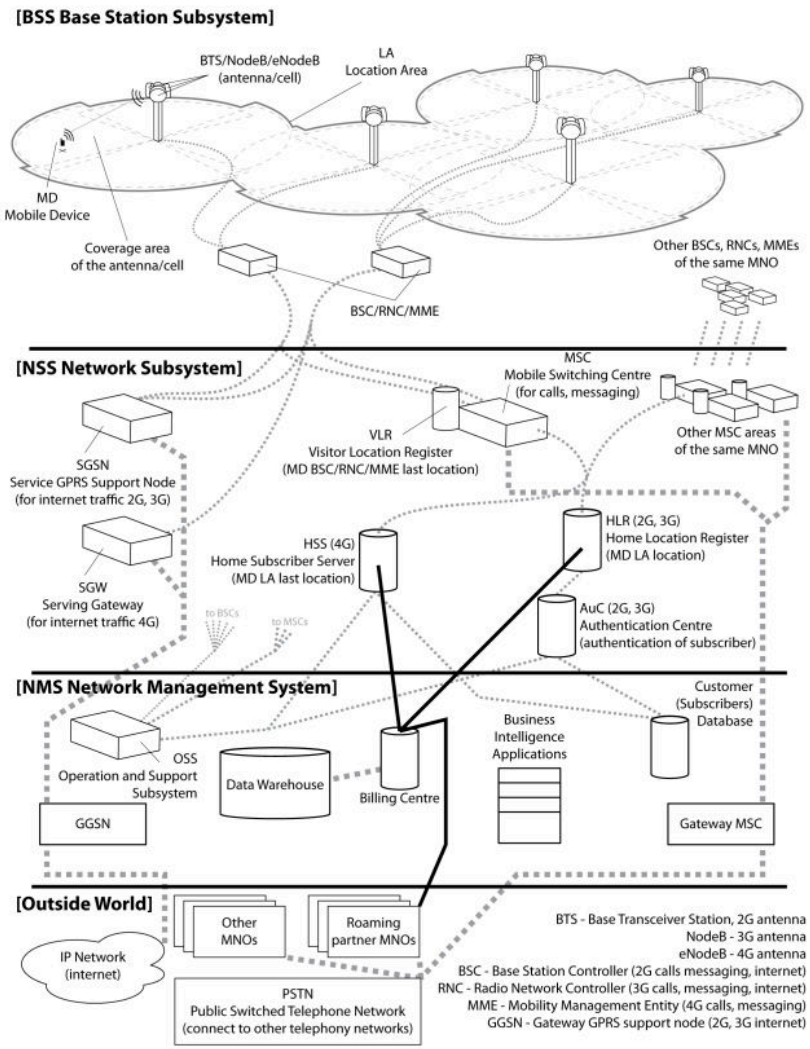
MobPhone Data (WP5)

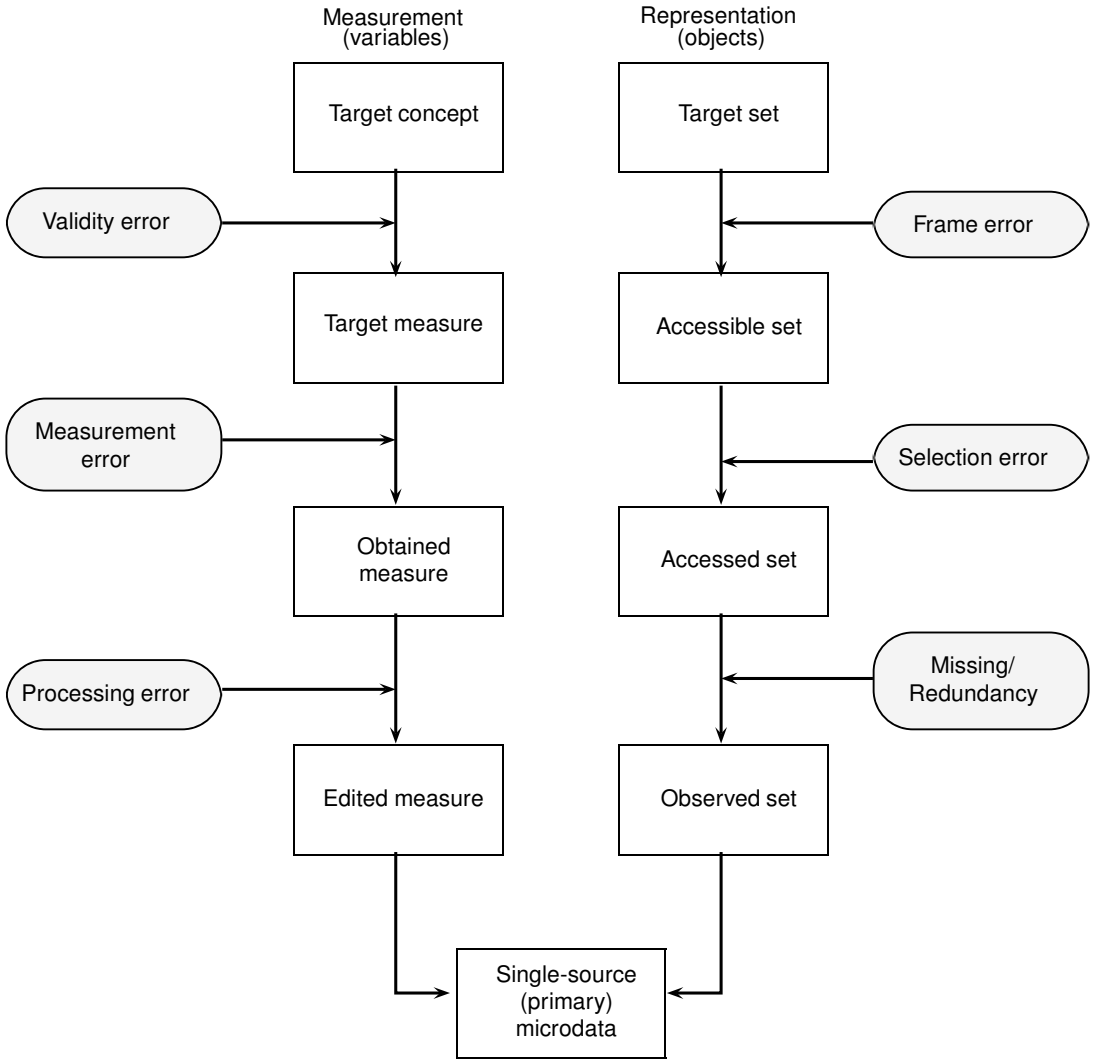
- Phase 1: Raw Telco Data Generation
- Phase 2: Statistical MicroData Generation
- Phase 3: Aggregated Data Generation
- More phases: Inference

Admin Data (Reid et al., 2017)

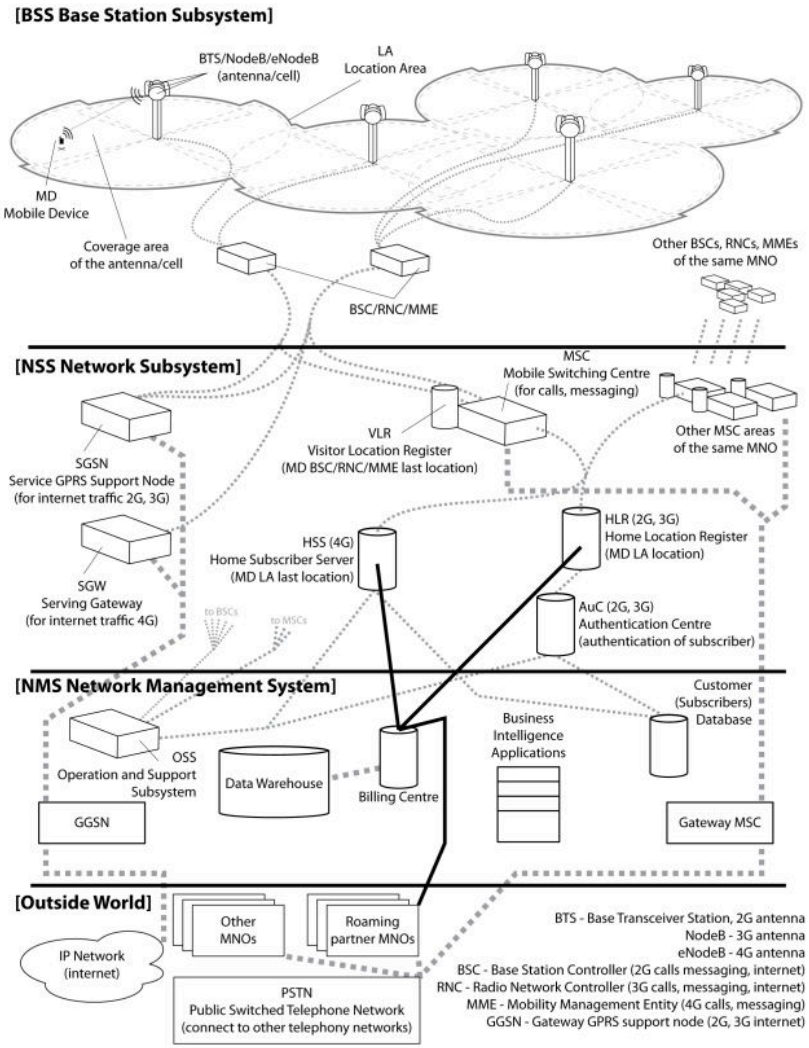
- Phase 1: Admin Data Generation
- Phase 2: Statistical MicroData Generation
- Phase 3: Inference to Target Population

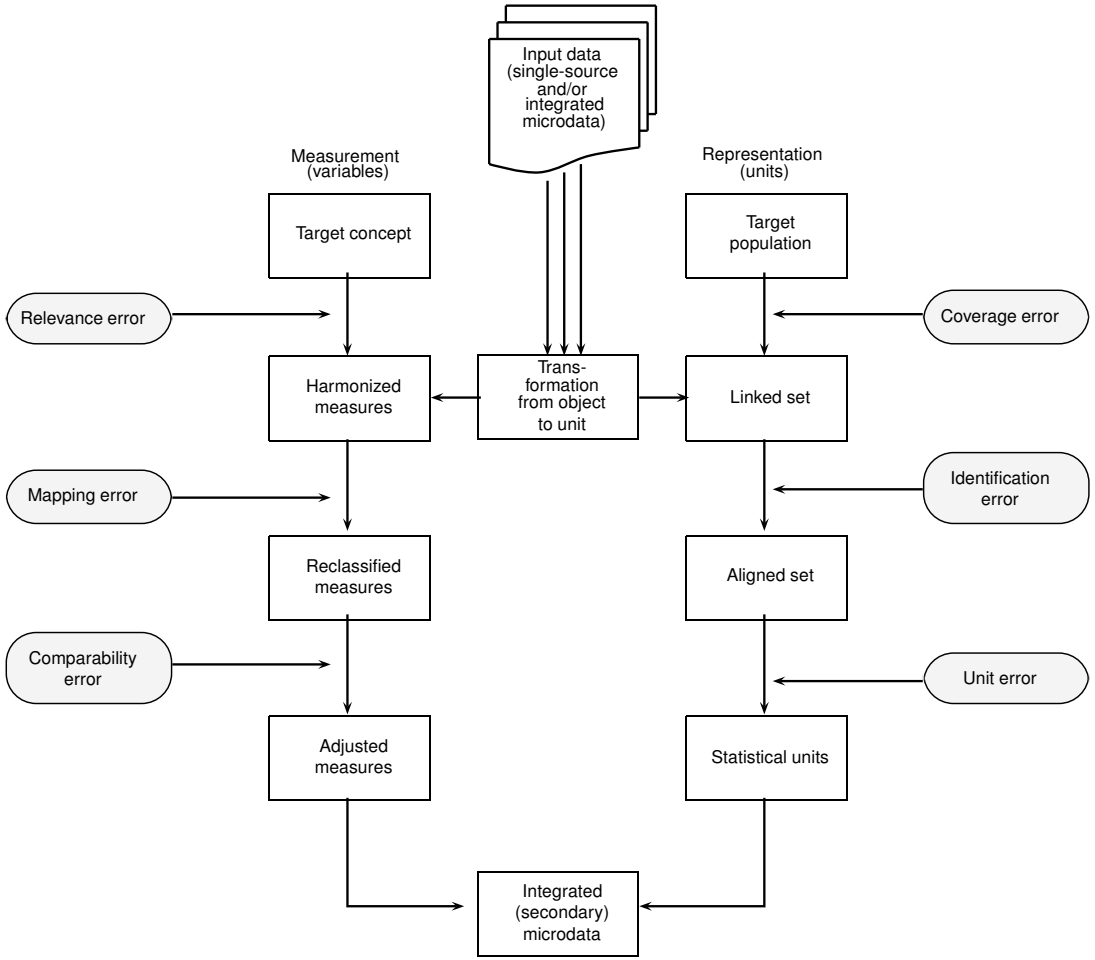
Data generation: Phase 1





Data generation: Phases 2 and 3



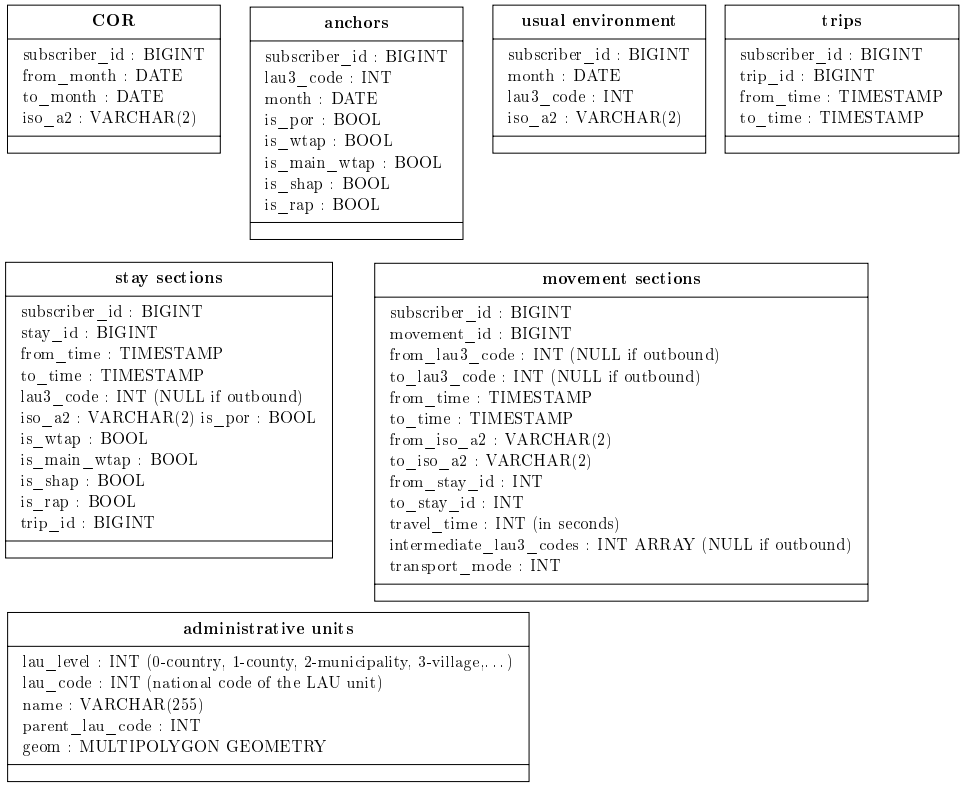


A data model

From

pseudonymised ID, time attributes, space attributes (. . .)

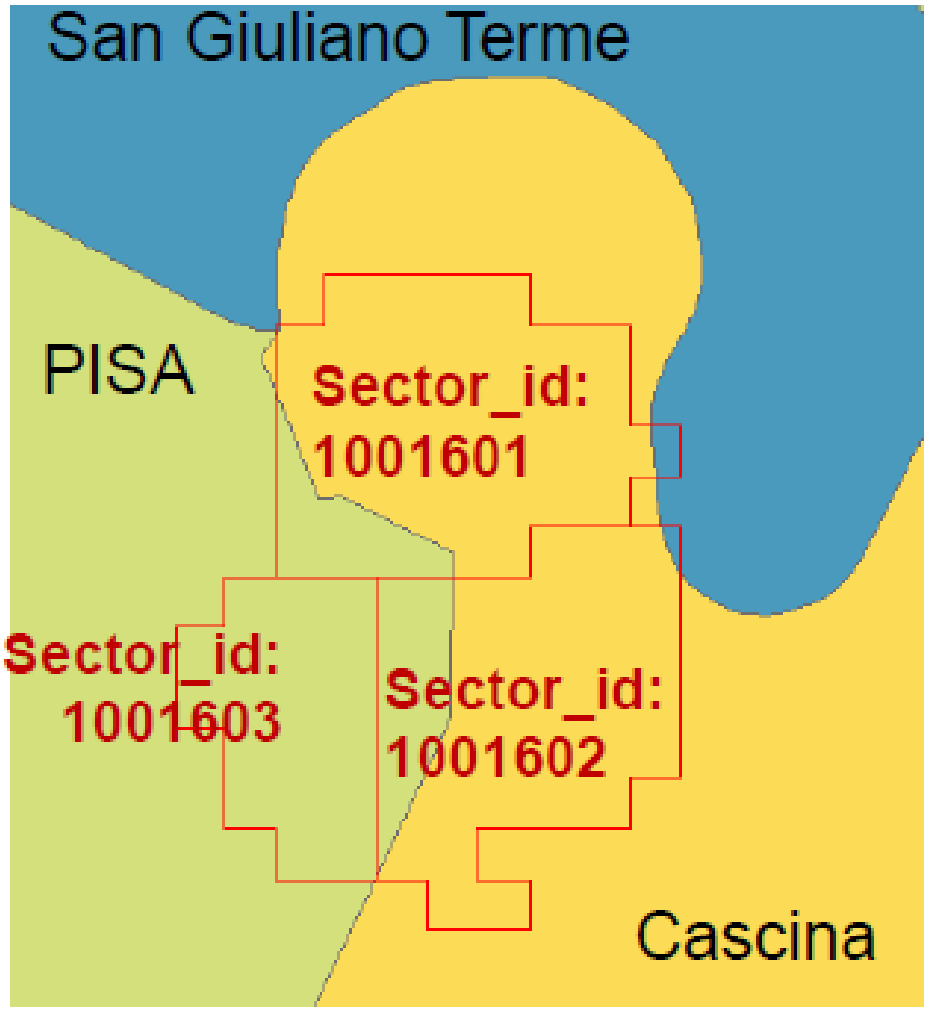
to



Geolocation of network events

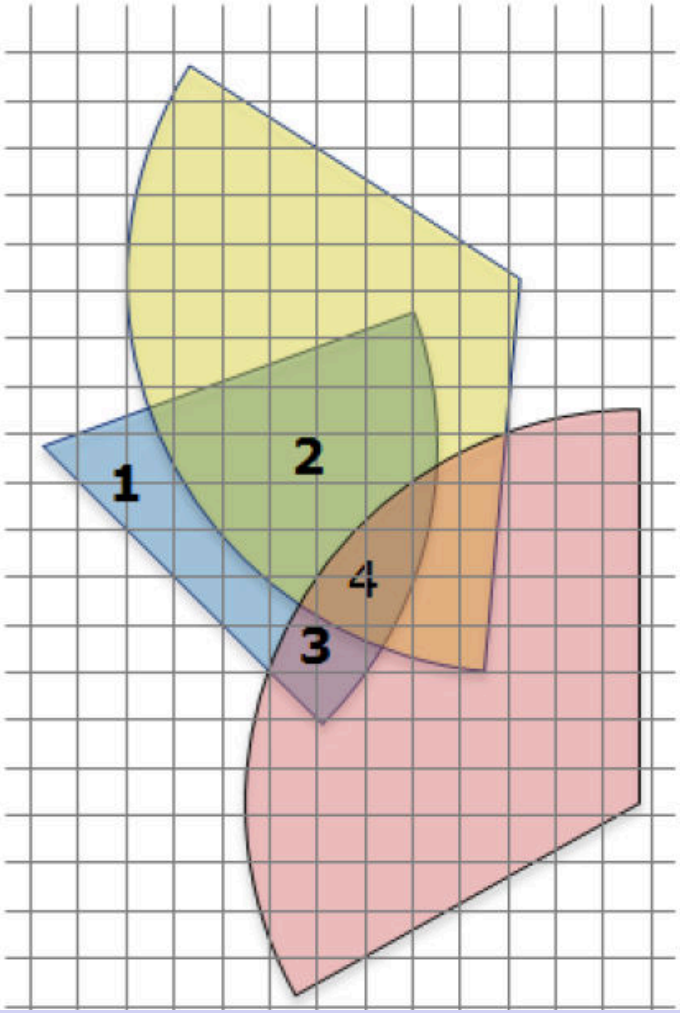
INSTITUTO NACIONAL DE ESTADISTICA

- **Best Service Area** Approach

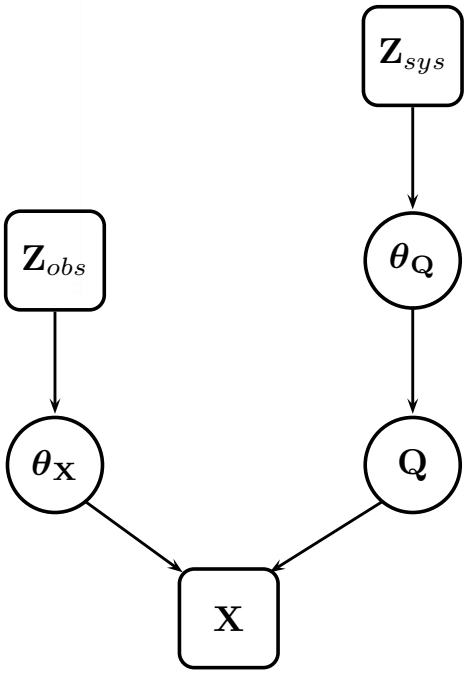


Geolocation of network events

- **Bayesian Approach**

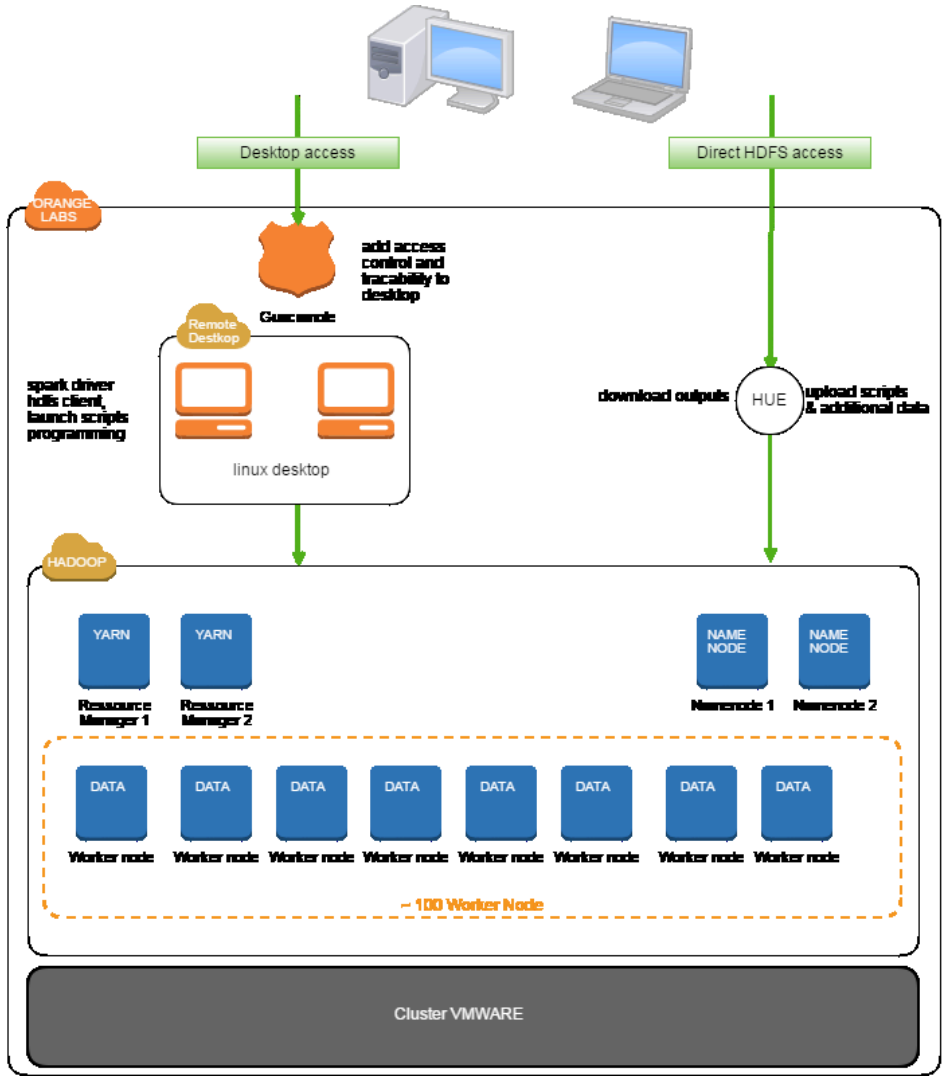


From data to target population: inference



- **Probabilistic sampling** not possible: the **curse of representativity**
- Adaptation of **species abundance problem** in ecological sampling:
  - **Hierarchical** model
  - Two working assumptions:
    - At t_0 individuals are assumed to be physically in the **territorial cell of auxiliary admin/survey data**.
    - Mobility patterns of individuals **do not depend on the concrete MNO** they are subscribed to.
  - **Bayesian** approach:
    - **Computational** scalability.
    - **Integration** with other data sources.

IT infrastructure



Package `mobloc`

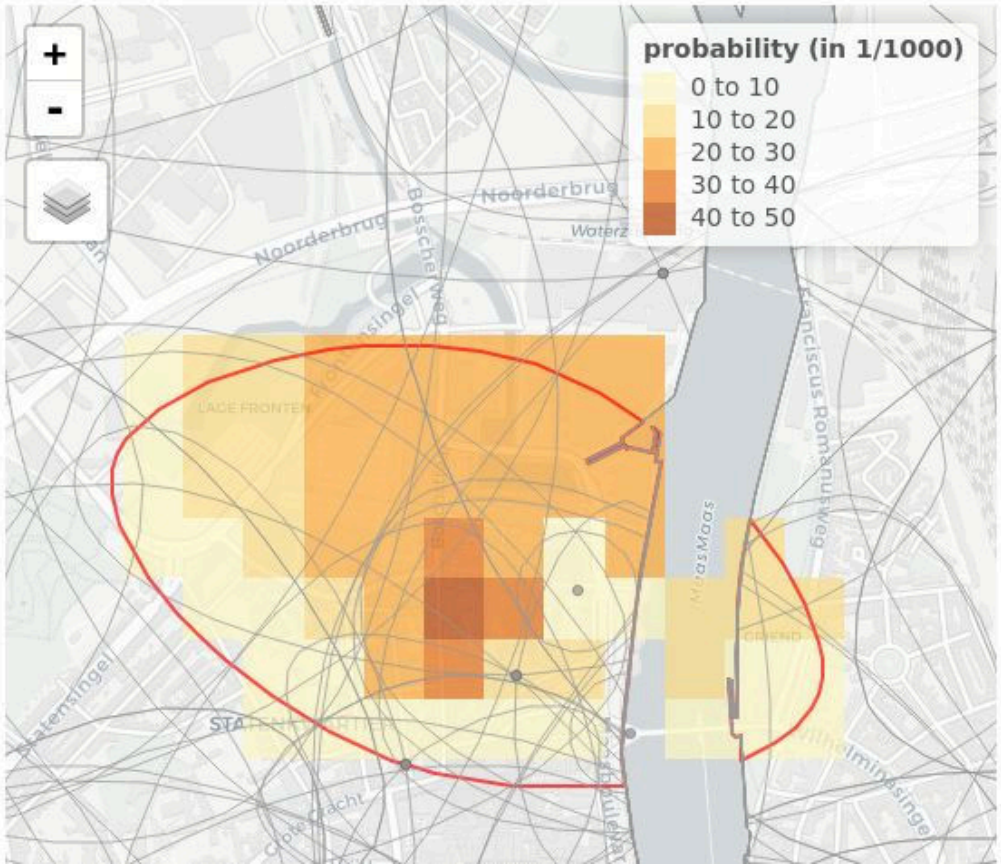
Cell Inspection Tool

Variable

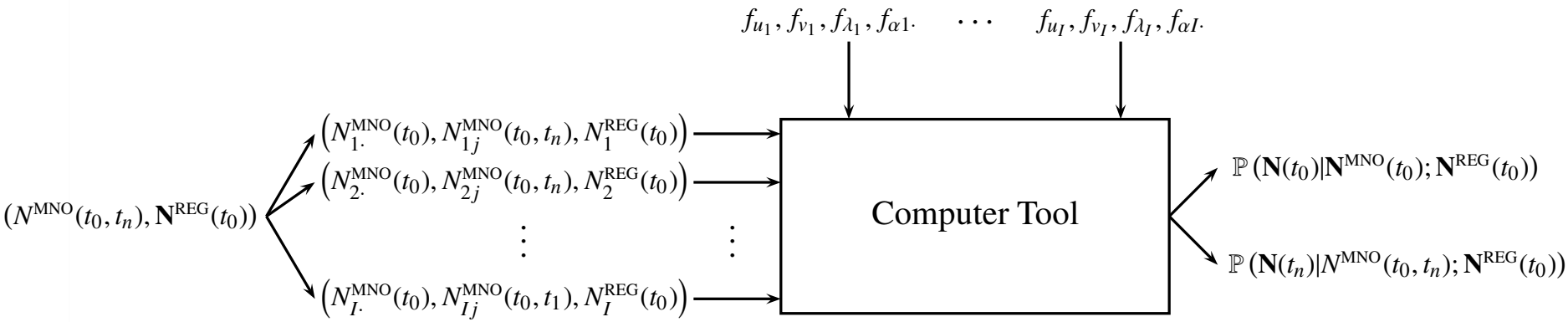
- (•) Probability
- ( ) Signal strength (dB)

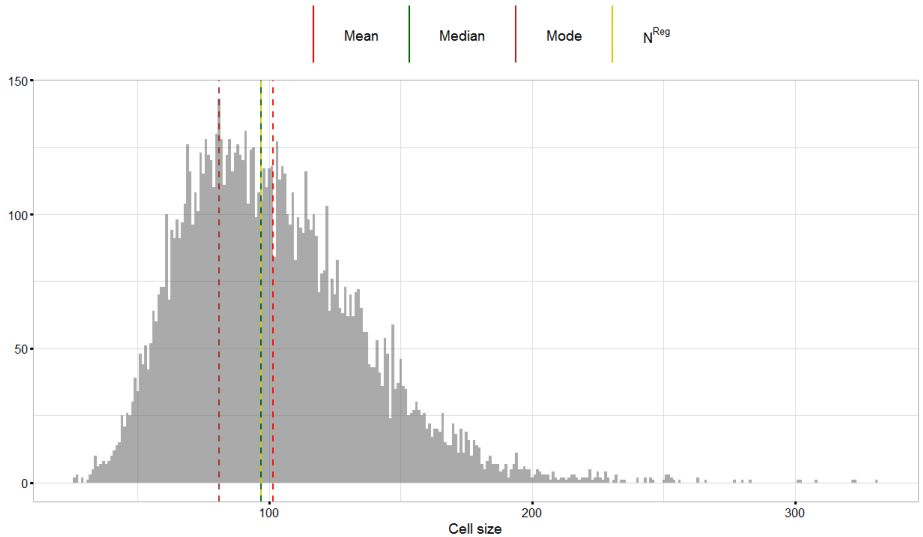
Selected cells

- [x] Cell 1
- [ ] Cell 2
- [ ] Cell 3
- [ ] Cell 4
- [ ] Cell 5
- [ ] Cell 6
- [ ] Cell 7
- [ ] Cell 8
- [ ] Cell 9
- [ ] Cell 10
- [ ] Cell 11
- [ ] Cell 12
- [ ] Cell 13
- [ ] Cell 14



Package `pestim`





Quality issues

- **CoP** affected by two generic facts:
  - MNOs **active part of the production process**.
  - Change of **inferential paradigm**.
  - **Higher** degree of **breakdown**.

- Example: **accuracy** dimension

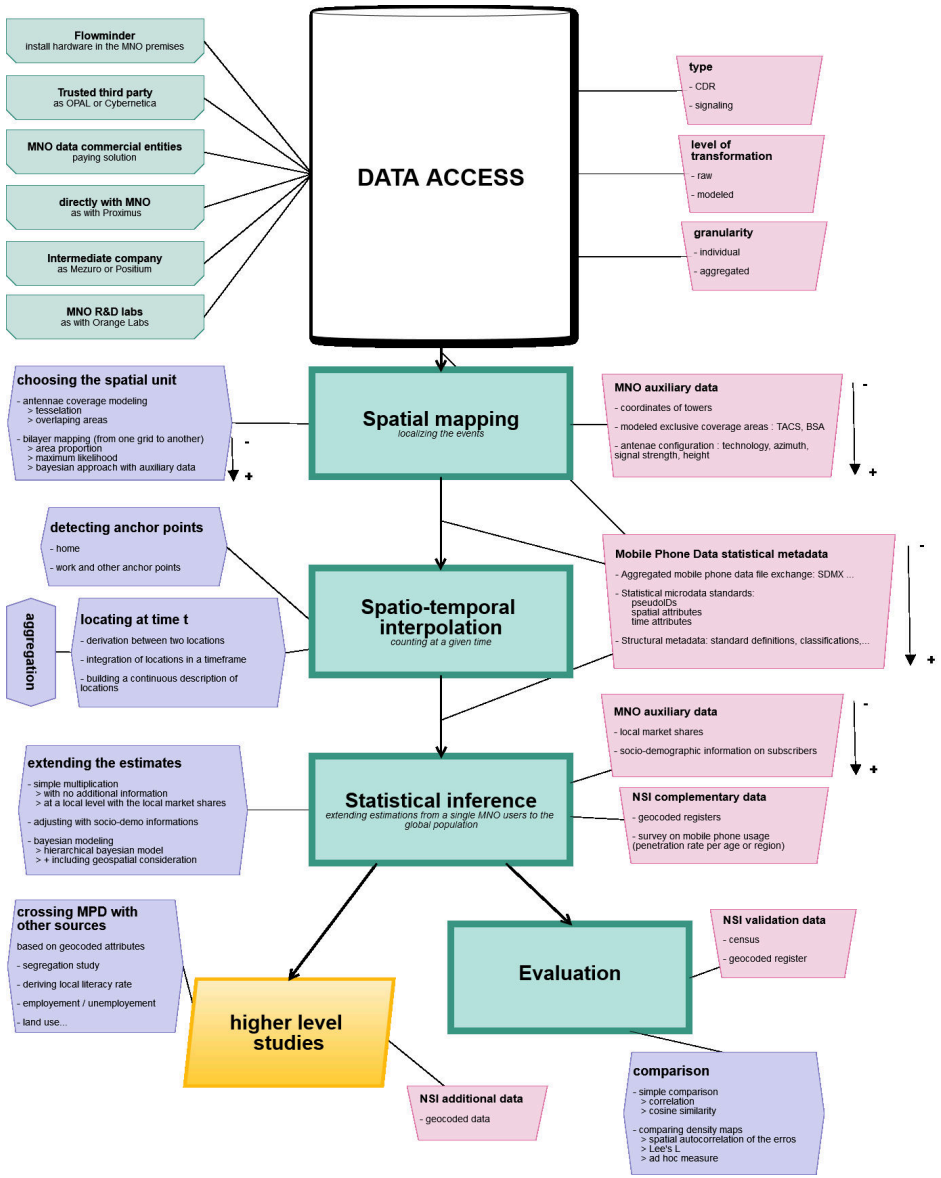
  From **confidence** intervals to **credible** intervals

  Model **checking** and model **assessment**

Main conclusions

- **Access blocked**: further work on **perceived risks** and **collaboration**.
- **Total Survey Error** paradigm still **valid**.
- **Geolocation**.
- **No probability sampling**: hierarchical **models**?
- **Computational complexity** for storing, accessing, and computing **in situ**.
- **Quality Assurance Framework** needs revision: **active role** of data holders and change of **inferential paradigm**.

The future research



The future research