Paradata as data source for census data collection monitoring: Brazilian census of agriculture case

Brazilian Institute of Geography and Statistics - IBGE

United Nations Statistical Commission

50th Session (2019) Side Events

Motivation

- Quality control in the census data collection phase is vital to produce quality statistics
- Inaccurate or fraudulent captured data may lead to undesired distortions
- Questionnaire microdata is the most common data source used for quality control

Motivation

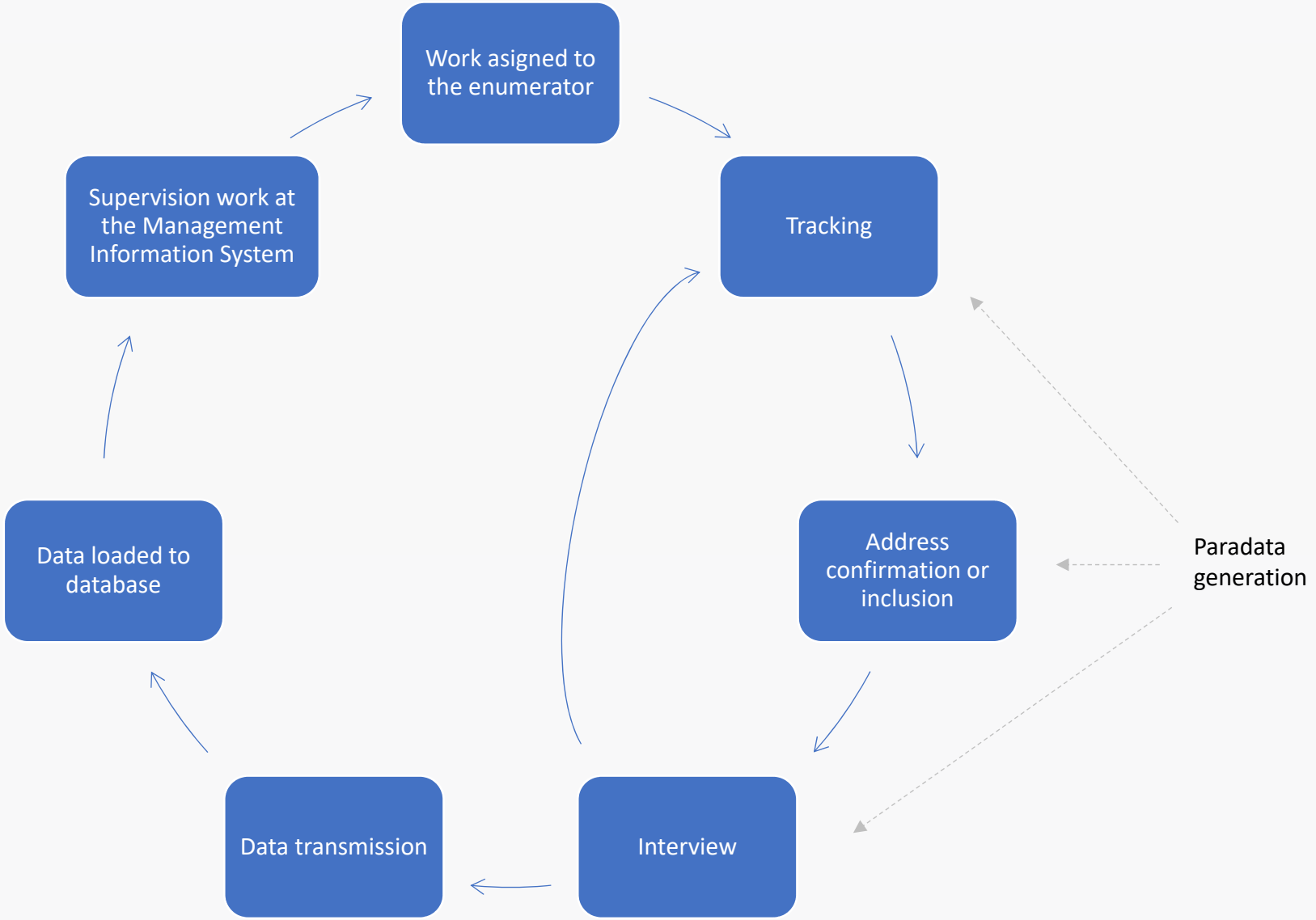
- **CAPI method generates a lot of “side” data (paradata) that may be used for quality purposes**
- **The use of GPS and hand held devices with high storage capacity enables the acquisition and storage of this data**
- **There´s no extra effort for the enumerator to collect paradata**
- **Proposal: use paradata as data source for data collection monitoring in the 2017 Census of agriculture**

Brazilian Agriculture Census - 2017

- 5 millions questionnaires => 70 millions
- 8.516.000 km² to be covered
- 127.000 Enumeration areas => aprox. 300.000
- 18.000 Enumerators and 4.000 field supervisors
  - => 210.000 15.000

censo agro 2017

Brazilian Agriculture Census - 2017



Paradata collected (1)

- **Geocordinates**:
  - **Enumerator tracking: each 16m of displacement**
  - **During the questionnaire completion: each 2 minutes**
- **Enumerator behaviour** (time record and action taken):
  - **Next/back button press**
  - **Begin/end of interview**
  - **Time taken to answer each question**
  - **Answer modification**
  - **Questionnaire reopen**

Paradata collected (2)

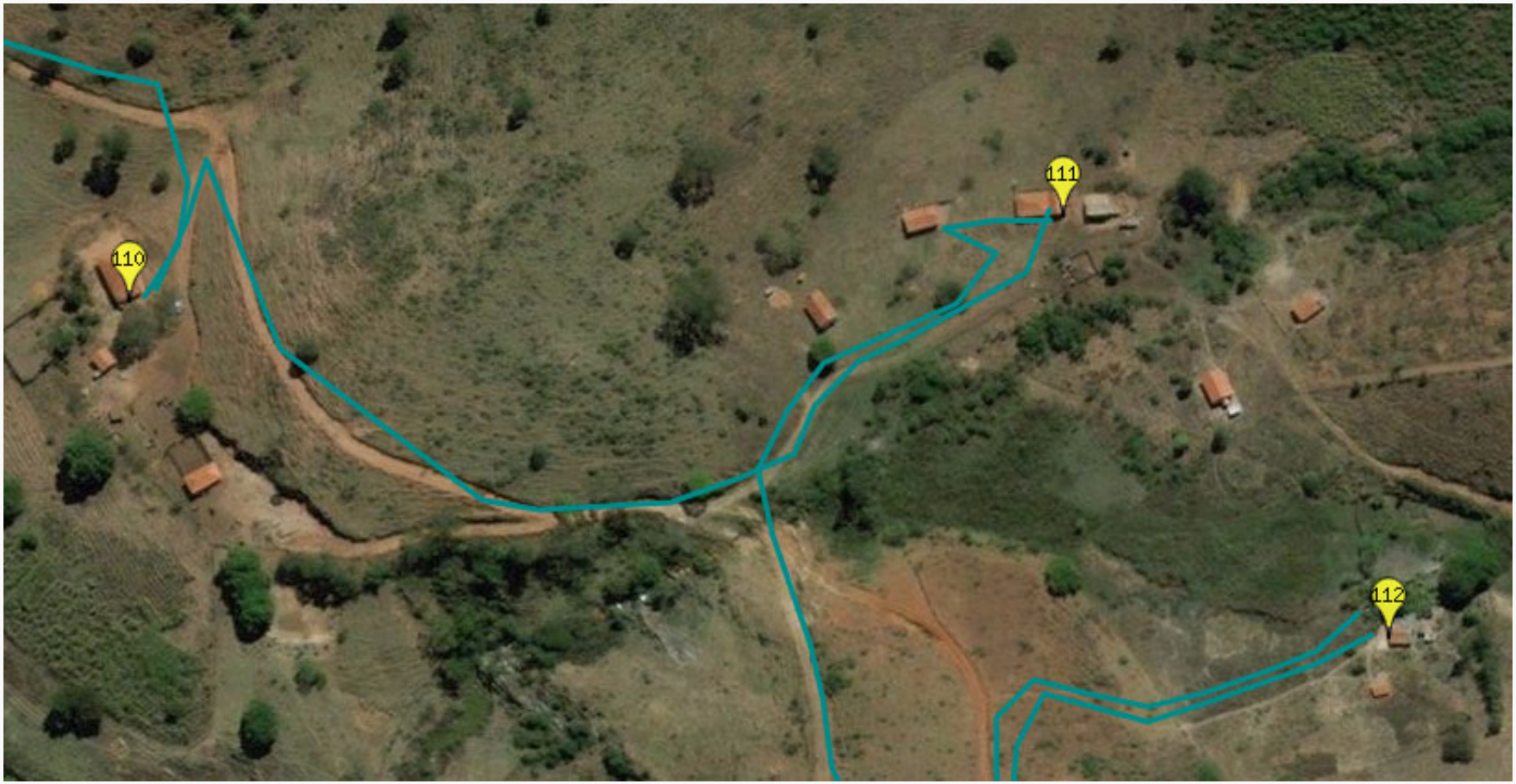
- 800.000.000 GPS tracking coordinates
- 90.000.000 GPS questionnaire coordinates
- 2.000.000.000 registers of enumerator actions

Paradata usage (1)

- **Management Information System made available reports to investigate the field operation by:**
  - **Interviews with short lenght**
  - **Interviews distant from the expected address**
  - **Places with many interviews**
  - **Interviews held in motion**

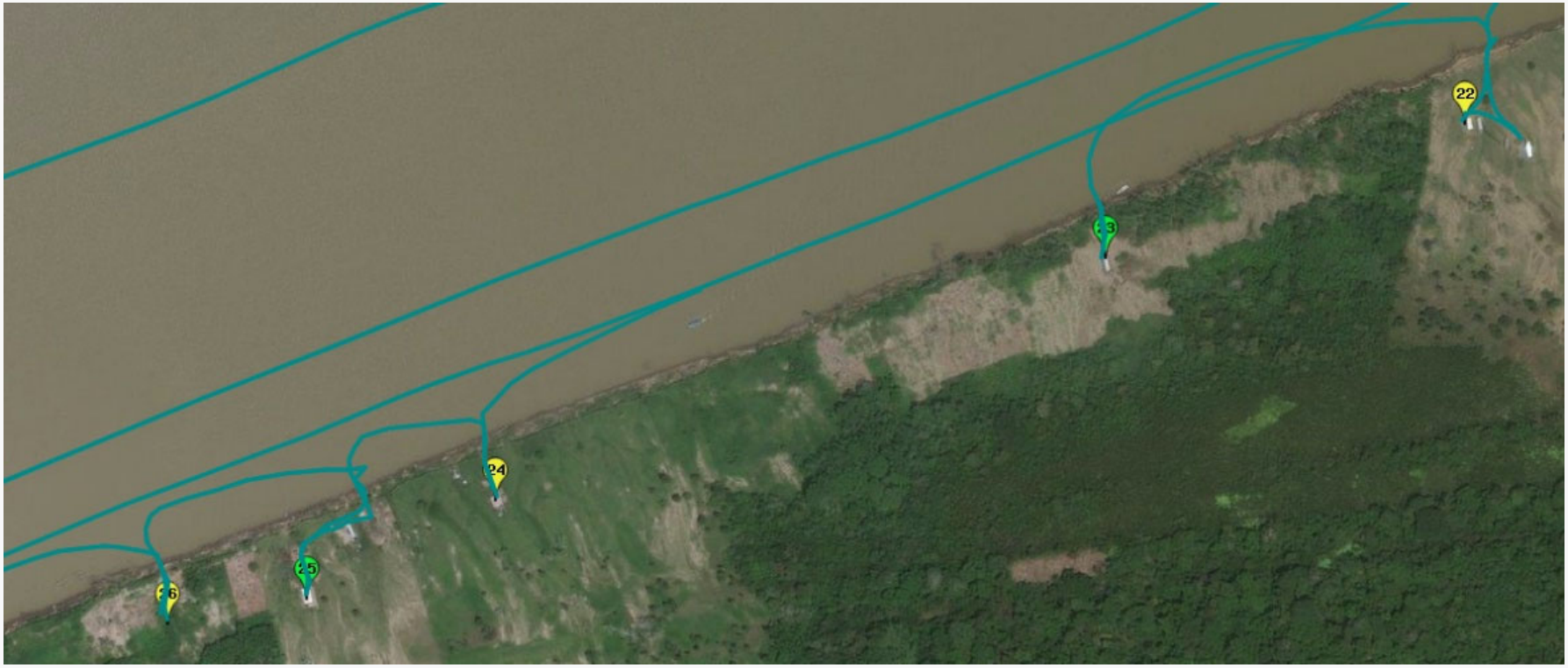
Paradata usage (2)

- **Enumerator tracking on the field**



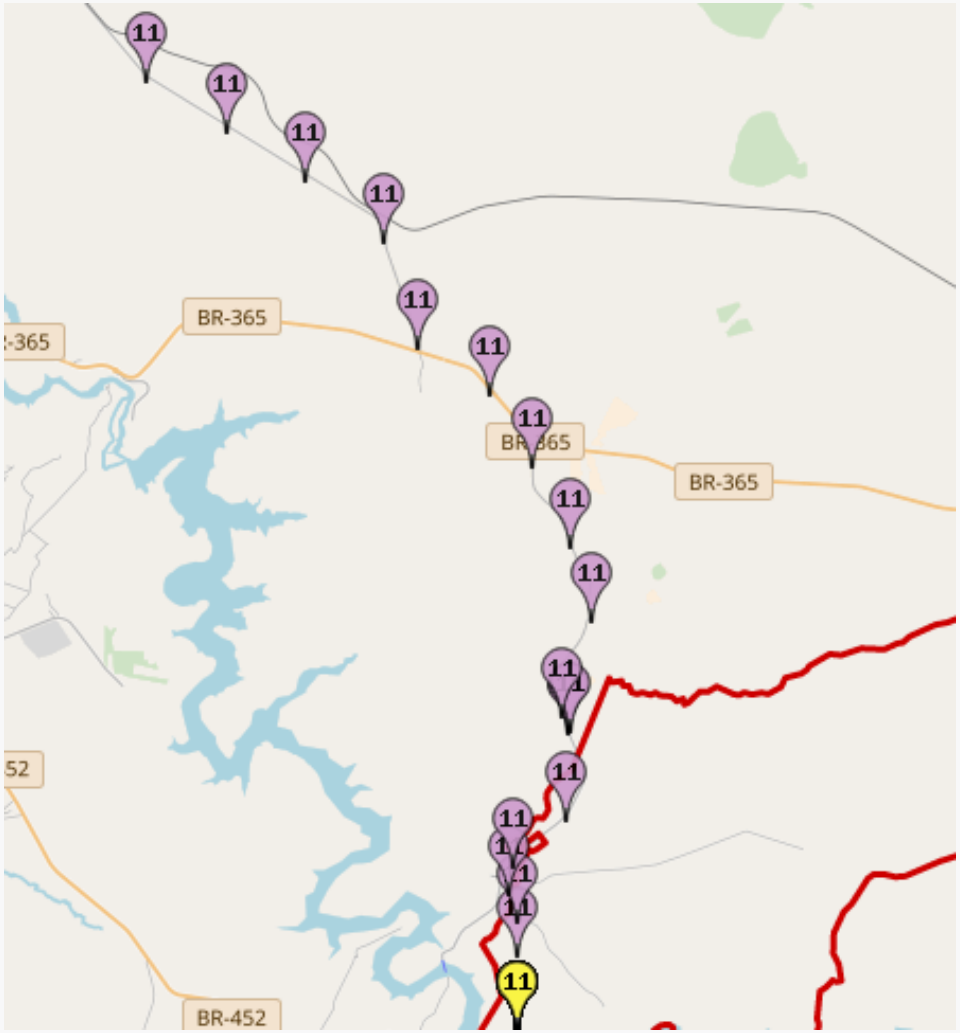
Paradata usage (3)

- **Enumerator tracking on the field**



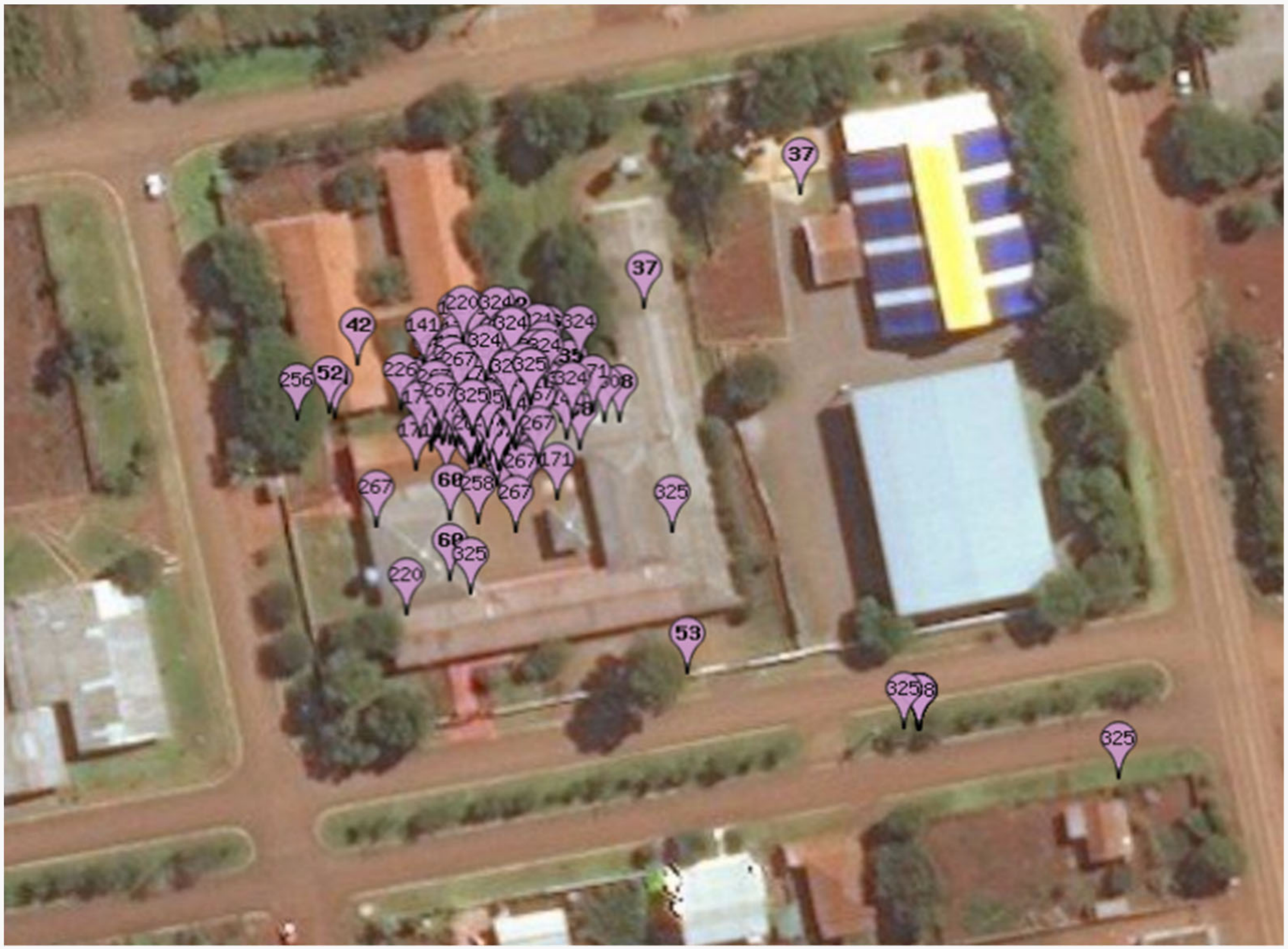
Paradata usage (4)

- **Interview held in motion**



Paradata usage (5)

- **Multiple interviews conducted in same place**



Results

- 12.7% (139.176) of the collected questionnaires suggested as suspicious by the management system were reviewed by the field staff

- Many fraud attempts were detected by the field staff by checking the available reports

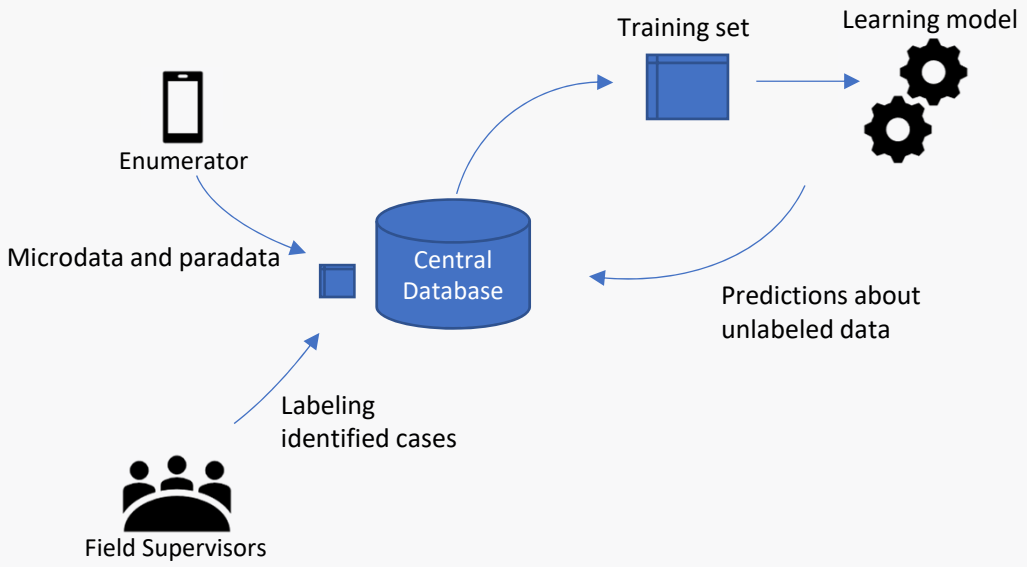
- Many enumerators mistakes were detected and corrections were made in time, improving this census efficiency and results

Conclusions

- **Relevance of use of paradata monitoring the field work of a census operation**
- **Paradata complemented the analysis of the microdata and became a powerful tool for measuring the quality of the data collection phase**
- **Made possible to identify suspicious or fraudulent work**
- **However, paradata related to questionnaire navigation was not totally explored. Further analysis may disclose valuable information regarding fraudulent patterns in this data**

Future research

- **Automate the identification of suspicious cases by designing a supervised machine learning model to check the quality of the collected data**
- **Use labelled data registered by the field supervisors to automatically classify questionnaires as suspicious based on the paradata generated by the enumerator**



Thank you