6/7 DECEMBER 2023

Second Workshop on Methodologies for Official Statistics

Smart Surveys: Methodological Issues and Challenges for Official Statistics

Claudia De Vitiis, Fabrizio De Fausti, Francesca Inglese, Monica Perez
ISTAT

Istat | Istituto Nazionale di Statistica

Outline

- Smart surveys – general aspects
- Methodological aspects
- ESSNet Smart Surveys projects
- Methodological tasks and involvement of Istat
- Conclusions

Istat

What are Smart Surveys?

- **Smart surveys:** surveys carried out using **respondent's smart devices**, combining data from **web questionnaire** with **sensor data**
  - The Smart Surveys can be designed for any device that has access to **passive data collection** capabilities through one or more **sensors** and provide access to their sensor data to other applications
  - The combination of sensor and app data with self-reports represents a **hybrid form of data collection** (in official social surveys data from diaries and sensors, i.g. for Household Budget Survey, HBS, and Time Use Survey, TUS).
    - ✓ **Traditional source of non-traditional data**
    - ✓ **Central role of respondent - Interaction (continuous and low intensity) of the individual with their personal devices**
- **Smart surveys produce Trusted Smart Statistics, Trusted Smart Surveys**:
  - Involvement of third parties (citizens, private companies and public bodies) with whom the NSI establishes a relationship of trust
  - Need to gain the respondent's trust through consent, engagement and use of Privacy Enhancing Technologies (PET) to protect personal data
- **Data are collected with reference to a specific sample of interest, a statistical sample, representative of a predefined target population**

Istat

Smart Surveys: Pros and cons for Official Statistics

❑ POTENTIALITIES

- ✓ The respondent can provide useful information for statistical production with **less disturbance** and **greater accuracy**
- ✓ The measurement capabilities of mobile devices can **supplement** or potentially even **replace** self-reports in surveys, providing **high frequency** and very **detailed data** for diary based surveys, where **passive data** can replace or support self-reports
- ✓ Used **in parallel with traditional** data collection modes, smart surveys can contribute to **non-response reduction** in social surveys

❑ ISSUES

- Need to acquire the **consent** of the respondent/data provider/citizen and involve him as an active subject
- **Coverage** problems, only a part of the population is reached and observed
- For NSIs, the **survey process** needs to be **redesigned**, several GSGPM steps are affected
- Data must be transmitted, stored and processed appropriately to **preserve privacy**
- **New methods** are needed to transform raw data into **statistical information**
- **Impact on data quality**, new types of errors, the estimates are produced by combining data from traditional and smart sources

Istat

Smart surveys: sensor data

- Sensor data collected passively (e.g. location, motion, activity trackers) and respondents' activities on smartphones (e.g. taking pictures, scanning receipts) increase available data sources
- It is necessary **to identify** sensor data that can provide the most **accurate measurement of the construct** to be investigated (from the raw data to the statistical variable)

Types of sensor data for social surveys in Official Statistics

- Image - Camera (to take photos, e.g. receipts for Household Budget Survey)
- Location - GPS (e.g. Time Use Survey)
- Acceleration - Accelerometer, Gyroscope (activity tracker for Health survey)
- External sensor, air quality (Living condition)

- ✓ Sensors can offer **objective data** to enrich or replace the **subjective answers** provided to questionnaire
- ✓ Sensors provide a **proxy** for the statistical variable
- ✓ Sensor data can also **reduce human error** by **shifting potential bias** in response (social desirability) to technology-based error

Istat

Smart surveys - Ethical and Legal aspects

- It is necessary to give **legal certainty** to users and citizens and to increase transparency and trust in those who manage the data
- The legal basis is the **GDPR** (Article 6.1 on the lawfulness of the processing)
- It is important to define a "co-ownership" **agreement between the parties** (the data controller determines the purposes and means of data processing)
- Management of **sensitive data**

It is necessary to refer to a privacy by design approach

Istat

Methodological issues/challenges

- **Literature** on **sensor data** is very wide and mainly technology oriented, but poorly developed for survey context, especially for official statistics
- **Multiple challenges** when collecting sensor and app data:
  - **participant selectivity**
  - **(non) willingness** to provide sensor data or perform additional tasks
  - **privacy** concerns and **ethical issues**
  - **quality** and **usefulness** of the data

❑ Crucial **methodological aspects** of the statistical process for a Smart Survey :

- **DATA COLLECTION STRATEGIES**
  - **Involvement and motivation** of respondents, recruitment materials and user interface
  - **Incentives**: Provide meaningful feedback to the respondent
- **DATA PROCESSING**
  - Use of **Machine learning** algorithms to **transform sensor data** (images, signals, voice) into statistical information
  - Use of strategies/appropriate ML methods to **improve** prediction **accuracy**
- **QUALITY OF DATA**
  - **New types of Errors: Coverage / Representation error - Measurement error - Mode effect implications**
  - Strategies to prevent/address errors: ✓ **paradata**
    ✓ **contextual data**

Istat

Data collection issues/challenges

- **Strategies for effective data collection**
  - How to **recruit and motivate people** to increase participation
    - ✓ **Interviewer support** at least for certain categories of the population
    - ✓ **Incentives**
      - ✓ Providing individual feedback (e.g., on household expending, time spent on activities)
      - ✓ Incentives - monetary
    - ✓ **Communication strategies**, change of cultural approach, exploiting the active role of respondent

  - **Open issues**
    - ✓ Do we need to involve interviewers?
    - ✓ Do we need respondents' feedback on data quality?

Istat

Machine learning algorithms: use/automation/accuracy

- **Use**: exploiting machine learning algorithms to handle and **process data** provided by smart devices such as **images**, **signals**, **voice**, and to **transform sensor data into statistical information**

- Crucial point in the use of ML: to what level **automation** can replace the direct acquisition of information or replace manual processes without affect data quality and/or increasing respondent burden

- Under what circumstances **results from ML models can be used directly as statistical data**, and under what circumstances data should be fed back to respondents

  - EXAMPLES
    - in **HBS OCR** is used for reading images of receipts and classification algorithms are necessary to trace to the COICOP classification the products declared or acquired from the images.
    - In **TUS** ML algorithms can be used to support the respondent in filling in the activity diary, providing suggestions based on the prediction of activities based on locations (data from **GPS** matched with contextual information from map services).

- Paramount is the **improvement of ML accuracy**:
  - Human-in-the-loop - respondent involvement through query to acquire missing data (label), respondent involvement in checking the quality of the data
  - Contextual data

Data quality - New types of errors

- **Representation/selection errors due to the use of a mobile devices**
  - **Coverage** problems due to unavailability of a smartphone or other mobile devices
  - **Non-response**
    - ✓ Unwillingness to participate, to download and install an app, to use the app actively or passively
    - ✓ Unwillingness to share sensor data due to smartphone inability and activities involved and/or privacy concerns
    - ✓ Unwillingness to provide consent

Total Survey Error framework

- **Measurement errors from sensor data**
  - ✓ Incorrect starting concepts
  - ✓ Sensor inaccuracy (imprecision, time inequivalence, device inequivalence)
  - ✓ Systematic and random measurement errors due to sensor quality (heterogeneity of devices)
  - ✓ Anomalies in measurements (outlier, noise, missing data, etc.)
  - ✓ Respondents' behaviour (incorrect initialization of the measurement or wrong use of the devices)
- **Need to define a DQ framework for sensor data**

(*Mussmann, Schouten, 2019; Keusch, Struminskaya, Antoun, Couper, Kreuter, 2019*)

Istat

Data quality framework and strategies to control errors

DATA QUALITY (DQ) and SENSOR DATA

- Need to define a **DQ framework for sensor data** to take into account :
  - quality of the raw data (characteristics and properties of sensors)
  - quality of processed data (measurement context - behaviour of the participants, errors related to the collection and processing phases)
- Need to implement complex **strategies to control errors** - monitoring dashboard and interaction with respondents involved into the quality control, **monitoring indicators**, **paradata and contextual data**
  - ✓ **Paradata** - to control data collection phase and to acquire information on app functionality, app usage, device information, sensor performance, etc.
  - ✓ **Contextual data -** to characterize users' day-to-day situations that have an influence on their smartphone and app usage, and consequently on data quality (respondent behaviour)

(Immonen, Pääkkönen and Ovaska, 2015)

Istat

ESSNet Projects on Smart Surveys (2020-2022 and 2023-2025)

In the context of the European Statistical System (ESS), two projects have been financed on this topic

- **The ESSNet on Smart Surveys**, which developed its activities in years **2020-2022**, delivered preparatory work to create an European wide methodological and architectural framework to share and re-use smart survey solutions and components, for supporting NSIs in doing smart surveys

- **The ESSNet on Smart Surveys Implementation (SSI) 2023-2025**, a new project aiming at implementing the defined framework for the domain of social surveys, mainly Time Use and Household Budget surveys

Istat

ESSNet Projects on Smart Surveys (2020-2022)

Two main activities / work-packages:

- **WP2** (Coordinated by CBS)
  - Evaluate the use of previously developed tools for European social surveys, such as Time Use Survey (TUS) or Household Budget Survey (HBS), through tests and field surveys to asses whether these tools can be used in different national contexts, tests of functionality and usability and pilot
- **WP3** (Coordinated by **ISTAT**– other partners: CBS, Destatis, Statistics Poland, INSEE)
  - Analyze the feasibility and the specifications for a European platform that supports the shared use of Smart Survey solutions, through the definition of a **conceptual framework** following a top-down design approach)

The most important results of the WP3

- ✓ **Description of a general framework**, addressing the smart surveys from different perspectives (**methodological, technological, architectural**), providing a new and **useful reference scheme** for the NSIs

Istat

ESSNet Smart Surveys Implementation (SSI) Project 2023-2025

- The goal of the SSI project is to **implement and demonstrate the concept of Trusted Smart Surveys**, realizing a **proof of concept** for the complete, end-to-end, data collection process and demonstrating a solution combining:

  1. **Involvement and engagement of citizens** as active contributors
  2. Acquiring, processing and combining **data collected from smart devices** and other appliances
  3. Contributing to the **trustworthiness** by guarantying **strong privacy safeguards**

- The SSI project is currently carrying out different tasks:

  - Definition of methodological standard for recruitment, ML, human-in-the-loop, mode-effect
  - Development of microservices, platform independent components, implementing specific functions (ML algorithms) for receipt scanning in HBS and geolocation data in TUS
  - Experimentation of smart survey process to Household Budget Survey - HBS (in view of the implementation of the new Regulation in 2026) and Time Use (HETUS) surveys
  - Legal and Privacy issues (Istat coordinates WP5)

Istat

Methodological tasks and involvement of Istat

- **Development of standards** for **machine-learning models** within smart survey process (data collection and processing):

  In most ML classification problems, it takes little effort to achieve close to 80% accuracy, but it is increasingly difficult to push for the last 20%. This is a significant challenge for official statistics that require high precision and accuracy.

  - What to do when the quality of the machine learning outcome is too low
  - When should respondents be asked to provide new input (a picture or open-text) because no meaningful information could be extracted
  - How and when should the training datasets used in the ML be updated/improved
  - Case studies are the ML methods used in HBS and TUS

    ✓ For **TUS** ISTAT is defining the methodological procedures and requirements for the microservice using **geolocation** data to **predict the HETUS activities** (selection of features from GPS data and OSM to infer location, ML model to predict activities from location)

- Two **pilot surveys** are carried out in SSI, to investigate **representation and measurement error**:
  - **Perception survey**, national survey on attitude towards smart surveys (in Italy, Netherlands and Slovenia)
  - **Time Use smart survey** in a **mixed-mode** perspective

Istat

Cross-national perceptions survey

Pilot survey on a national representative sample

To evaluate the opinions and attitudes of different target populations with respect to surveys using new technologies (mobile device, app)

The goal is to answer questions such as:

- ✓ *Do people understand the benefits of smart features (burden reduction, additional knowledge, better proxies of concepts of interest)?*
- ✓ *What objections do people have against smart features (still too expensive, not interesting/relevant, don't see logic/benefits, privacy issue, etc.)?*
- ✓ *To what extent are trade-offs between benefits and objections spontaneous?*
- ✓ *Do people understand what they are consenting to, are their expectations in line with practice?*
- ✓ *What is informed consent according to the interviewees?*
- ✓ *Do perceptions depend on the amount of control over the data/process?*
- ✓ *Are there differences between countries in all of this?*
- ✓ *How do perceptions depend on the sponsor/stakeholder of the survey?*

Cross-national perceptions survey

"New methods of data collection for statistical surveys"

General survey – representative sample

(4000 individuals, two- stage stratified sample)

Self-reported **paper questionnaire**, embedded login credentials to go into online smart part

+

On line smart survey

Web responsive questionnaire,
4 blocks embedding smart services

- **Communication** and **recruitment**: preference on communications, motivations to partecipate motivation discouraging partecipation
- **Attitude** to use smart device
- **Experience** using app and internet
- **Opinions** on advantages on using app to make research related to the topic
- **Willingness** to share app data with NSI
- **Motivations** for unwillingness to share app data with NSI (secutity, privacy, etc)
- Importance of **informed consent** (the need to know in advance which data will be collected/shared, need to control collected/shared data)

- **Consumption** →*scanning* receipts
- **Living condition** → scanning images of energy meters (electricity/gas/water)
- **Phisical activity** → providing *step counts from pedometers*
- **Mobility** → sharing *location* by GPS

Istat

Survey to evaluate mode effect

Pilot survey – Field test of a Time Use Smart Survey – mode effect assesment – Autumn 2024

- The goal is to **assess differences** between **data collected through smart survey** and **traditional** web or paper diary survey, separate sample selection differences from mode measurement effects
- The field test aims to
  - ✓ **Compare** estimates from the surveys **using smart features** with estimates obtained **without** the involvement of smart features (using or not the **geolocation data** to foster compilation of diary of daily activity)
  - ✓ **Estimation of pure mode effects** (measurement effects) and the size of a break in time series caused by a shift from a traditional survey to a smart survey considering data collected from the surveys using paper-based diary and web/app-based data (Using data from traditional TUS 2023)

- ISTAT is implementing a smart survey on general population for the first time
- A sample of 3000 individuals will be selected for a set of municipalities, interviewers will be involved to support respondents
- The research platform MOTUS (Hbits, VUB) will be used for the implementation of the questionnaires and data collection

Istat

Final remarks

- ✓ **Smart surveys** have great potential and represent **a major challenge for social surveys** on a methodological, technological and organizational level
- ✓ Smart surveys can **encourage citizen participation**, **reducing the statistical burden** and **financial costs** in statistical production, but require a **greater initial effort**
- ✓ Smart surveys have **methodological implications on data quality**: mode effect and the quality of the collected data are two crucial issues to address
- ✓ A big methodological challenge in smart surveys is **data accuracy**, differently from big-data context, in smart surveys errors can be investigated and even corrected
- ✓ Regarding **smart data acquisition**, issues are the **active-passive trade-off** and the **engagement** and involvement of the respondents.
- ✓ The trade-off among **quality**, **burden** and **privacy preserving** of data collected is crucial.
- ✓ Combining different techniques (**multi-technique surveys**) remains, at the moment, the most effective strategy both to reach different population targets and to ensure the quality of the data as a whole

Main references

- Bähr, S., Haas, Georg-C., Keusch, F., Kreuter, F., & Trappmann M., 2020. “Missing Data and Other Measurement Quality Issues in Mobile Geolocation Sensor Data”. *Social Science Computer Review*. DOI: 10.1177/0894439320944118
- Benedikt, L., Joshi, C., Nolan, L., de Wolf, Nick. & Schouten, B., 2020. Optical Character Recognition and Machine Learning Classification of Shopping Receipts - @HBS>An app-assisted approach for the Household Budget Survey
- Biemer, P. P., de Leeuw, E., Eckman S., Edwards B., Kreuter T., Lyberg L. E., Tucker N. C. 2017. *Total Survey Error in Practice*. West B. T. (Editors). John Wiley & Sons, Inc., Hoboken, New Jersey
- ESSnet on Smart Surveys (2020-2021). https://ec.europa.eu/eurostat/cros/content/essnet-smart
- Immonen A., Pääkkönen P., and Ovaska E. 2015. “Evaluating the Quality of Social Media Data in Big Data Architecture”. *Digital Object Identifier* 10.1109/ACCESS.2015.2490723
- Mussmann, Ole; Schouten, Barry (2019): Final methodological report discussing the use of mobile device sensors in ESS surveys. MIXED MODE DESIGNS FOR SOCIAL SURVEYS - MIMOD, WP5 deliverable. Edited by EUROSTAT.
- Keusch, F., Struminskaya, B., Antoun, C., Couper, M. P.; Kreuter, F. 2019. “Willingness to Participate in Passive Mobile Data Collection”. In *Public Opinion Quarterly* 83, pp. 210–235.
- Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M. 2019a. “Trusted smart statistics: motivations and principles”. *Statistical Journal of the IAOS*, 35. https://ec.europa.eu/eurostat/cros/system/files/sji190584.pdf 2)
- Ricciato, F., Giannakouris, K,. Wirthmann, A., Hahn, M. 2020. “Trusted Smart Surveys: a possible application of Privacy Enhancing Technologies in Official Statistics”. *SIS Conference*.https://it.pearson.com/content/dam/region-core/italy/pearson italy/pdf/Docenti/Universit%C3%A0/Pearson-SIS-2020-atti-convegno.pdf
- Struminskaya, B., Lugtig, P., Keusch, F., Hohne, J.K. 2020. “Augmenting Surveys with Data fromSensors and Apps; Opportunities and Challenges”, Social Science Computer Review, 089443932097995

Istat

Thank you for your attention!

Istat | Istituto Nazionale di Statistica