EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS

2024 ESTORIL - PORTUGAL

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly financed by the European Union

EUROPEAN CONFERENCE ON
QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Enhancing the quality of the prediction of activities in Time Use Smart Survey using a microservice exploiting GPS data

Claudia De Vitiis - *ISTAT, Italy*

Author with:

Cappadozzi T., De Fausti F. , Inglese F., Michelini M., Terribili M.D. - *ISTAT, Italy*

Schouten B., Klingwort J. - *CBS, Netherlands*

Minnen J., Beyens P. - *hbits CV, Belgium*

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly
financed by the European Union

SMART SURVEYS in Time Use domain

- **Smart surveys**
  - combine data collection modes based on input from the data subjects (**active data**) with data collected **passively** by the **device sensors**
- **Geolocation data**
  - is one of type of sensor data very promising and useful for smart surveys, for survey topics cognitively burdensome or time-consuming such as diaries, already used for travel surveys, less exploited for Time Use Survey
- **Visualizing geolocation** would enable **respondents** to reconstruct their **daily activities** more easily by
  - distinguishing between **stops and travel times**, it helps remembering times and places visited
  - providing **tentative data** for the activities performed
- A crucial aspect is **obtaining respondent consent** to activate sensors, thereby allowing the tracking of smartphone movement
- **NSI** would greatly reduce the work involved in imputing missing trips, with respect to traditional TUS, improving the **quality** and the **timeliness** of the data produced

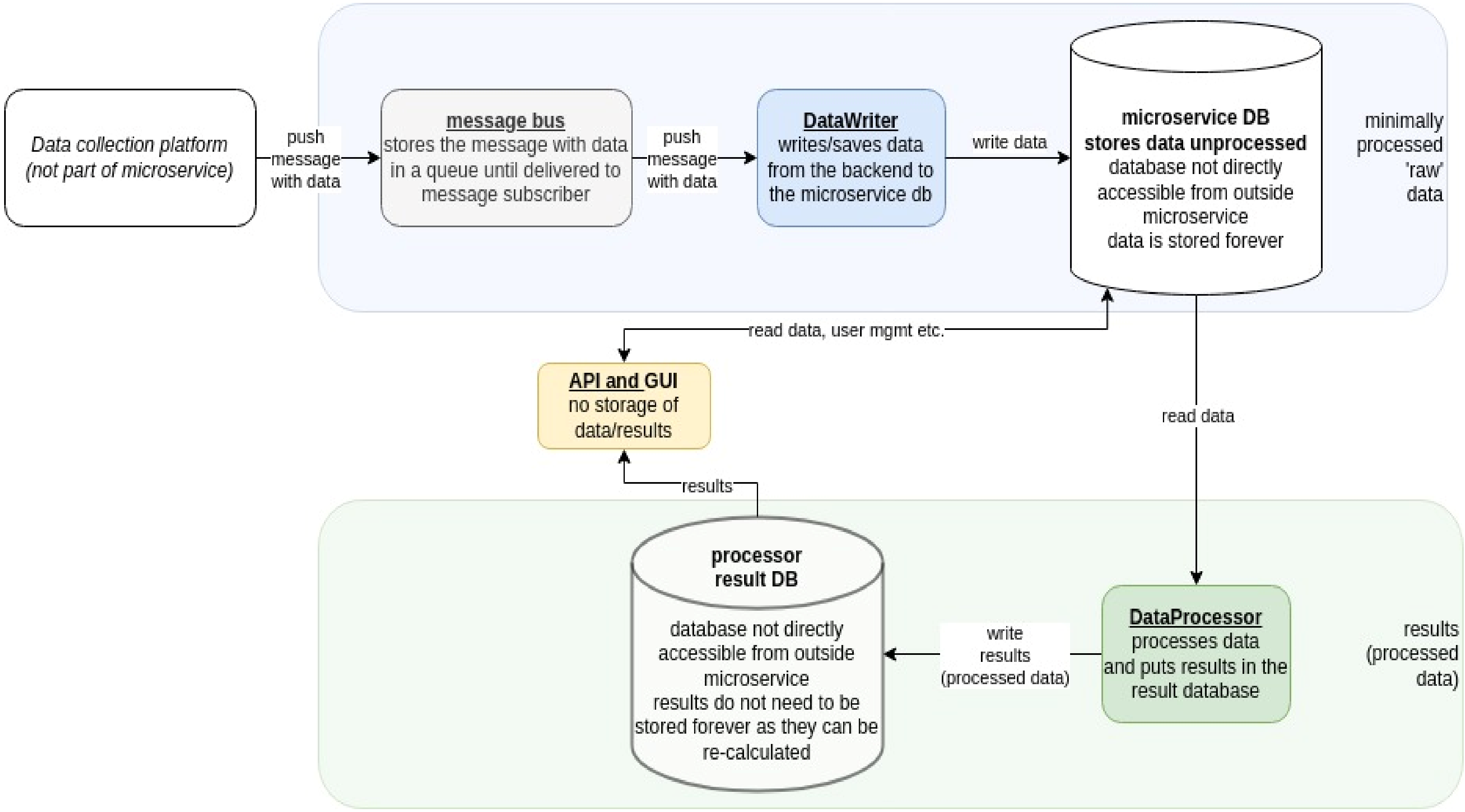
Smart Survey Implementation ESS-Net (**SSI**)

- The **SSI project** aims at a general understanding and **elaboration of smart surveys at all design levels**, in the context of the European Statistical System (ESS)
- Building and **testing** on field the **smart services** in multiple countries, in **three survey domains** (**TUS**, **HBS**, **energy**), implementing **end-to-end data collection process**
  - For **TUS** domain, **location tracking** data requires **advanced and transparent ML**, establishing a trade-offs in privacy-by-design and post-survey processing

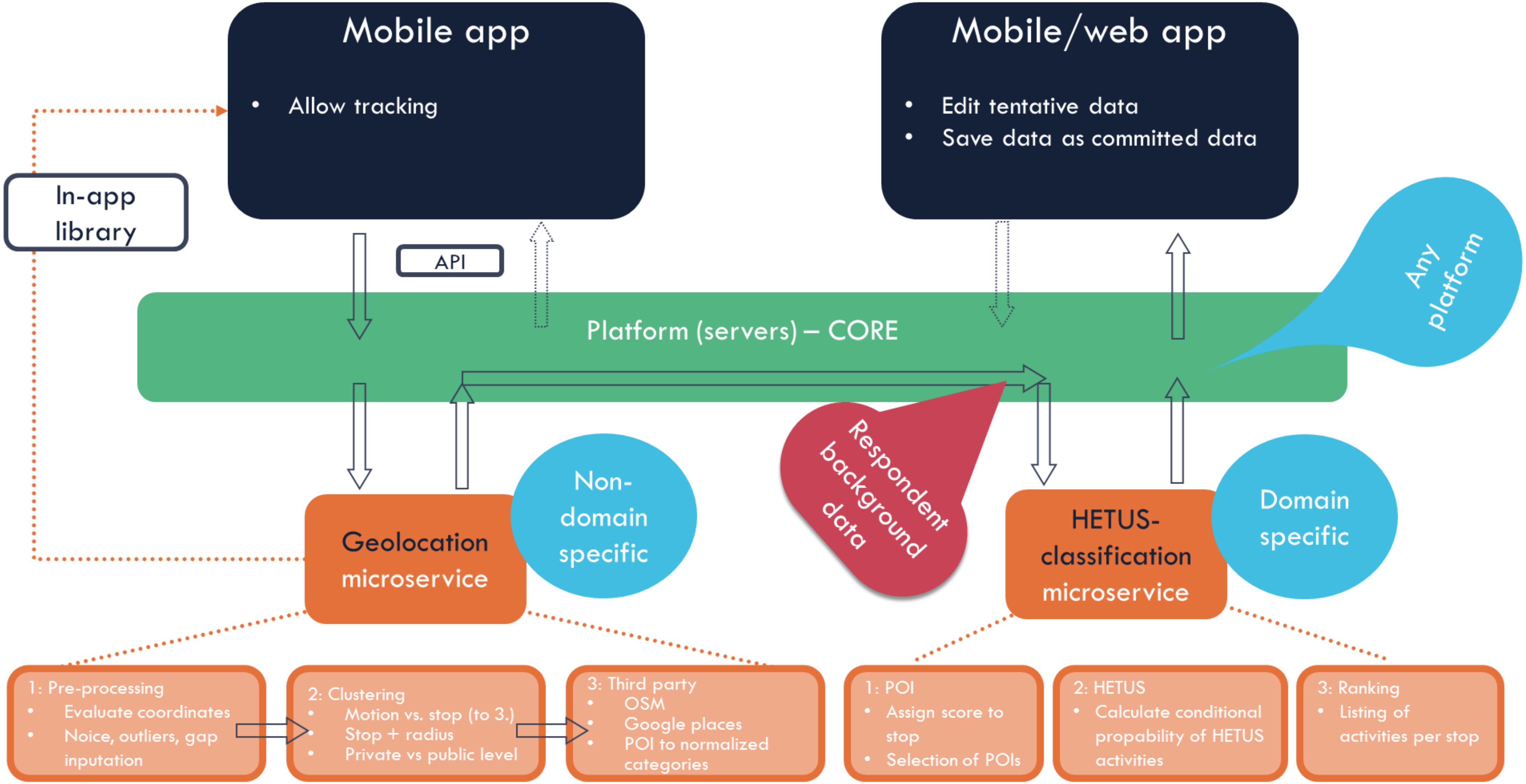
MICROSERVICES in SSI

- Central in SSI is the integration of **microservices**, which are **standalone services** with specific objectives, **detached from** a broader data collection **platform** architecture
- The microservice is seen here as **middle part software,** that is supportive to the respondents in reducing their burden to complete a time diary and is developed as an independent service to platforms
- The **data collection platform** governs the utilization of microservices, deciding who accesses them and when
  - There exists **no direct connection between respondents and microservices**
  - **No** direct **link** between **microservices** and the **primary database**, granting the data collection platform complete authority over the data provided to microservices
- The scope of the **Geotracking microservice** is to exploit **smart features** related to **position** and **movement** (GPS data) to detect **stop and track segmentation** and **predict travel mode** and **daily activities**, helping the respondent with **tentative data** for filling the diary of daily activities

Information view of the microservice in SSI



Geolocation microservice(s): development and binding



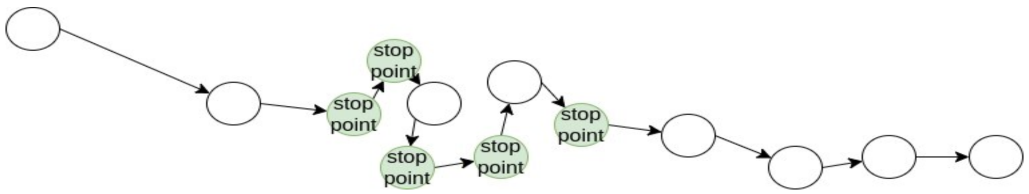
Segmentation of geolocation data

The stop detection algorithm is structured into four main components:

1. **Filtering**: GPS points undergo filtration based on their accuracy
2. **Identifying**: significant stop points are identified among the GPS data using a specific algorithm

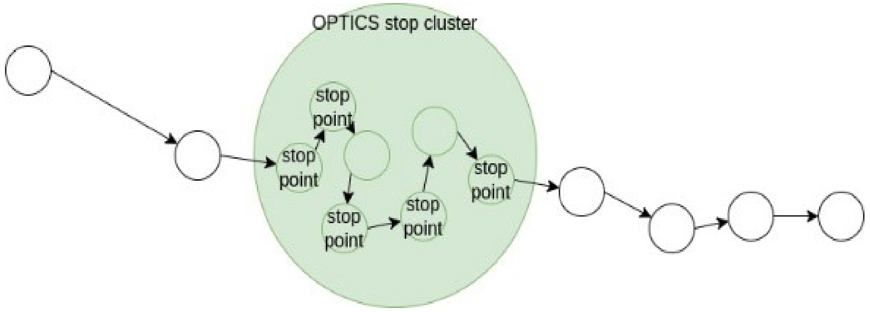
ATS algorithm: The first part, ATS, is engineered to perform **stop detection**, based on the research paper by *Bonavita et al (2022)*, focusing on the individual **stop-based adaptive trajectory segmentation** (ATS)

GPS point is a **stopping point** if there is a **duration** of more than "t" seconds between the current GPS point and the subsequent GPS point with more than "x" meter **distance** (**t=180s**, **x=50m**)



3. **Clustering**: Grouping points together to form clusters resembling stops
4. **Postprocessing**: Refining the clusters by merging them to reduce their number and ensuring alternation between stops and tracks

OPTICS algorithm based on DBSCAN, density based approach to spatial data



Segmentation of geolocation data (2)

- As a **final step**, the stop locations are **cross-referenced** with a **Points of Interest (POI) database** (such as **Google Places** or **OpenStreetMap**) to augment them with additional information

  - **Travel infrastructure** and **POIs** from the map crucial for predicting travel mode and activity
  - Comparison between POIs in **Google Places** and **OpenStreetMap:**
    - **GP** has **greater coverage** and is more regularly **updated than OSM**
    - POI **coverage varies** significantly **across countries**

Transport mode classification

- **Transport mode classification** is performed **after the geolocation data is segmented** into stops and tracks
- This approach **requires** a **database with information about transport mode infrastructure**, such as **OSM**
  1. After mapping the geolocation data to the OSM data, the **number of OSM geolocation points per transport mode** within a track cluster needs to be determined
  2. The transport modes available in OSM are motorized vehicles on roads, trains, trams, buses, subways, bicycles, and on foot
  3. The transport mode having the **largest proportion in the track cluster** is considered the **most plausible** and assigned to the cluster

- The **quality of the infrastructure** data is particularly important in this approach
- The **quality** and **density** of the various transportation modes can **vary depending on the country**
- There are still a number of open questions at this step, such as the **data quality and comparability of the infrastructure** data, how to deal with **multi-modal track clusters**, but also how different segmentation algorithms affect this method

HETUS Activity prediction algorithm

- The **activity prediction algorithm** processes the **stops** identified by Geotracking MS
  - **Input** data for each stop are the **GPS points**, the **timestamp** and the **associated POIs**

- **Steps of algorithm for activity prediction**

1. POIs identification and selection of a short list of POIs in each stop.
   - A score (**POI-score**) is assigned to each POI, based on the **weighted median of the distances** between each **POI** and all **GPS points of the stop**, weighted by the **accuracy** of GPS points.
   - A **short list of POIs** is identified using the **elbow criterion** on the POI scores.
2. Determination of (conditional) **probability of HETUS activities** for each POI selected in the short list.
   - Through a **Bayesian decomposition** (taking the idea from Cheng *et al* 2022), for each POI of the short list the conditional probability of HETUS activities are calculated starting from the **distribution observed in TUS data** (Italian data for the moment).
   - The **variables considered** (duration and time of the day, HETUS place category, occupational status, age classes) in the decomposition **are linked with** the corresponding variables observed in the stop and for the specific respondent.
3. Assignment to a stop of a **rank** of the HETUS **activities** based on a final score.
   - Finally, a **rank of the HETUS activities** is assigned to the stop, based on a final score calculated **aggregating the probabilities** of the activity **weighted** by the POI-score associated with the activity for each POI in the shortlist.

HETUS Activity prediction

GPS points and POIs of the stop

In red the short list POIs of the stop

HETUS	ActivityScore	Descr
021	7.400402e-02	021 Eating
361	4.739460e-02	361 Shopping (including online/ e -sho
519	3.842740e-02	519 Other or unspecified social life
032	3.437884e-03	032 Personal care servi ces
732	1.588585e-03	732 Parlour games and play
513	1.237589e-03	513 Celebrations
821	1.227424e-03	821 Watching TV, video or DVD
522	3.925254e-04	522 Theatre and concerts
343	3.686005e-04	343 Caring for pets
831	3.219861e-04	831 Listening to radio or recordings
383	1.547464e-04	383 Reading, playing and talking with c
811	8.229526e-05	811 Reading periodicals

The resulting list of activities with final scores

Quality assessment

The microservice will produce three outputs to be used in smart surveys through UI:

- stop / track segmnetation, to support the respondent in remembering trips and stops
- transport mode prediction, to be used as survey variables or tentative data to be confirmed in travel or time use survey
- activity (stop purpose) prediction, tentative data to be confirmed in travel or time use survey

- The **quality of the prediction** is influenced by various factors:
  - **heterogeneity** in **sensor quality** between different types of smartphones – influencing the **input data**
  - **heterogeneity** in **contextual data sources** - map quality – influencing the **output data**, **predictions** of travel mode and activities
  - performance of the **algorithms** implemented, tuning of **parameters**

Next steps

- **Microservice development is still ongoing**, stop/track segmentation is defined, trip and activity prediction still needs improvement but is very promising
- **Impact analysis**
  - ✓ Assessment on the quality of prediction depending on **sensor and map quality**
  - ✓ How much the prediction of activity depends on the **use of TUS data** and the **user profile**?
  - ✓ **Test** and **output evaluation** on different **data sets**
- **Implementation** of the microservice to be connected to different platforms
- **Test of the microservice** in large and small test in different **countries**, within ESSNet SSI

References

Bonavita A., Guidotti R., & Nanni M. (2022). Individual and collective stop-based adaptive trajectory segmentation. *Geoinformation* 26, 451-477

Cheng J., Zhang X., Luo P., Huang J., Huang J. (2022). An unsupervised approach for semantic place annotation of trajectories based on the prior probability. Information Sciences, Volume 607, August 2022, Pages 1311-1327. Elsevier ed. https://doi.org/10.1016/j.ins.2022.06.034

Sadeghian P., Håkansson J., Zhao Xiaoyun. (2021). Review and evaluation of methods in transport mode detection based on GPS tracking data. Journal of Traffic and Transportation Engineering (English Edition) · July 2021. DOI: https://doi.org/10.1016/j.jtte.2021.04.004

Smeets, L., Lugtig, P. & Schouten, B. (2019). Automatic Travel Mode Prediction in a National Travel Survey. *Statistics Netherlands, Discussion Paper*

Struminskaya, B., Lugtig, P., Keusch, F., & Hohne, J.K. (2020). Augmenting Surveys with Data from Sensors and Apps; Opportunities and Challenges, *Social Science Computer Review, 089443932097995*

Thank you for your attention!

devitiis@istat.it

Instituto Nacional de Estatística
Statistics Portugal

eurostat

The conference is partly
financed by the European Union