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

Claudia De Vitiis[1], Pieter Beyens[2], Tania Cappadozzi[1], Fabrizio De Fausti[1], Francesca Inglese[1], Jonas Klingwort[3], Manuela Michelini[1], Joeri Minnen[2], Barry Schouten[3], Marco D. Terribili[1]

1 Italian National Statistical Institute (ISTAT), Roma, Italia

2 hbits CV, Bruxellex, Belgium

3 Statistics Netherlands (CBS), Den Haag, Netherlands

Abstract

Smart surveys offer new opportunities for developing social surveys, especially those based on burdensome compilation of diaries (Household Budget Survey, HBS and Time Use Survey, TUS), as they aim to exploit new data sources through personal devices (smartphones, tablets, wearables) that use sensors and provide information about themselves or their surroundings. On this topic, the European Statistical System (ESS) has financed two projects: presently the second project is ongoing, ESSNet Smart Surveys Implementation, starting in 2023, that aims to involve and engage households and citizens, to define and operationalize a new/modified end-to-end data collection process, and to test microservices, solution/component for Time Use and Household Budget surveys.

This paper focuses on the Time Use Survey (TUS) carried out making use of web and mobile applications and the inclusion of a microservice for geolocation. The microservice is seen here as the middle part of the software that is supportive to the household in reducing their burden to complete a time diary. The geolocation microservice is developed as an independent service to platforms. The collection of geolocation points (based on GPS sensor data), allows obtaining relevant information for TUS, predicting the HETUS activities as tentative data for the respondent during the filling of the diary. In this context, in fact, the role of respondents and the interaction with them during data collection is a crucial issue. Starting from the GPS coordinates and some additional information the geolocalized points are segmented into stop and motion; then, by adding geographical-spatial information, the places (Points of interest, POI) that can provide information on the activity are identified. One or more activities (with assigned probabilities) are associated with the POIs through an algorithm that exploits diverse information, such as place categories/activities taxonomy, the timing of the stop, country-specific indicators and user characteristics. The quality/accuracy of the predictions depends not only on the quality of raw data but also on the choice of the location/map features and the use of contextual data for the improvement of the classification algorithms.

Keywords: GPS data, HETUS Activities, Microservice, Smart Survey, Time Use

1. Introduction

Smart Surveys are surveys in which respondents are asked to employ smart devices (e.g. smartphones, tablets, activity trackers) to collect survey data through active and passive data collection of questionnaire and/or sensor data. Smart surveys involve dynamic and continuous interaction with the respondent and with her personal device(s). They combine data collection modes based on input from the data subjects (active data) with data collected passively by the device sensors (e.g. accelerometer, GPS, microphone, camera, etc.) (Struminskaya et al., 2020). A review of the literature can be found in (Smeets et al., 2019). In general, this innovative way of data collection offer new challenges to improve significantly the quality of the statistics produced by social surveys in the National Statistical Institutes (NSIs), while aiming at reducing the burden for the respondents.

This paper has a focus on the Time Use Survey (TUS) carried out making use of web and mobile applications and the inclusion of a microservice for geolocation, which centres around the collection of geolocation points. The microservice is seen here as the middle part of the software that is supportive to the household in reducing their burden to complete a time diary and is developed as an independent service to platforms. The scope of the microservice is to exploit smart features related to movement (GPS data) to detect stop and track segmentation (Bonavita et al., 2022) and predict the HETUS activities, helping the respondent with tentative data for filling the daily activity diary.

The paper is structured in four paragraphs: the context of the work, the architecture of the microservices, the details of the microservices for geolocation data (segmentation, travel mode prediction, activity prediction) and, final remarks on quality issues.

2. Context of the work

2.1 Smart Survey Implementation ESSNet (SSI)

The study of this paper is one of three case studies developed, tested and evaluated within the Eurostat-funded project Smart Survey Implementation (SSI). SSI aims at a general understanding and elaboration of smart surveys at all design levels and in the context of the European Statistical System (ESS). The four design levels considered and expanded to smart surveys are methodology, IT architecture and infrastructure, logistics and legal. Methodology concerns push-to-smart recruitment and motivation strategies, AI-ML, UI-UX and trade-offs in active and passive data collection. IT is about implementing smart data collection into services and embedding the resulting services in backend systems. Logistics is about the extension of the back-office and all human operations linked to processing smart data collection and updating AI-ML. The legal design legal comprises privacy risk assessments and mitigation measures and trade-offs in privacy-by-design choices. Smart surveys are especially promising for survey topics that are cognitively burdensome or time-consuming, that are non-central to the average respondent, or that lend themselves poorly to questions as proxy measures. The three ESS studies considered in SSI all satisfy one or more of these criteria. They are household expenditures, energy consumption and general time use. SSI attempts to learn how to set up smart surveys based on extensive studies into these topics. It does so by building and field testing the smart services in multiple countries. Perhaps the most challenging of the three is the smart Time Use Survey employing location tracking data. While potentially very useful for respondent as a framework to construct their diaries, location tracking data require advanced AI-ML and transparent, but complex, trade-offs in privacy-by-design and post-survey processing. A crucial aspect is obtaining user or respondent consent to activate sensors, thereby allowing the tracking of smartphone movement.

2.2 Survey domain need

The Time Use Survey requires respondents to fill in a diary for one or more days, in which they document all activities (named HETUS activities, such as *eating*, *working*, etc) carried out in specific locations, with whom they were carried out, and whether or not ICT facilities were used. In the event that an activity involves travelling away from home, respondents are required to provide a separate description of this activity, including details of the mode of transport used. In paper diaries, still utilised in many European countries, respondents often find it challenging to distinguish between the times of activities and travel to carry them out. This often results in incomplete or inconsistent information regarding travel activities and the mode of transport used and requires a time-consuming imputation process during the checking and correction phases. It is so evident why the utilisation of a micro-service application that employs GPS data would be a significant advantage for both the participants and the NSI. The application would enable respondents to reconstruct their day more easily by clearly distinguishing between stop and travel times, thereby reducing the burden of remembering times and places visited. Furthermore, the NSI would greatly reduce the work involved in imputing missing trips, thereby improving the quality of the data collected.

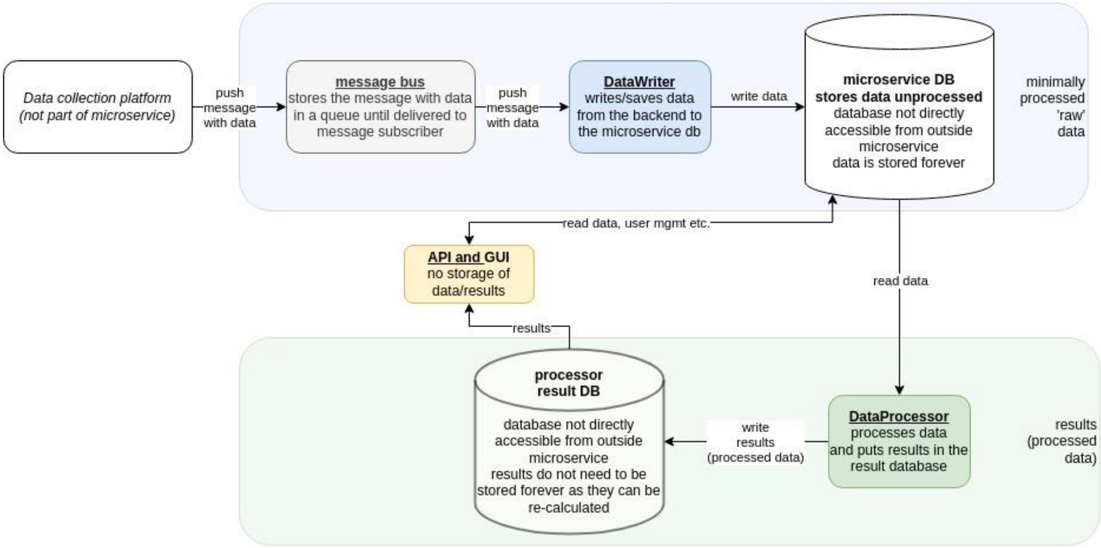
3. Architecture of the microservices in SSI Project

The objective of the SSI project is to actively involve households and citizens while designing and implementing a novel or adjusted end-to-end data collection process through innovative data collection techniques. Central to this process is the integration of microservices, which are standalone services with specific objectives, detached from a broader data collection platform architecture. This setup ensures several key points:

- There exists no direct connection between respondents and microservices; instead, the data collection platform governs the utilization of microservices, deciding who accesses them and when.
- Similarly, there is no direct link between microservices and the primary database, granting the data collection platform complete authority over the data provided to microservices.

In terms of information flow, the diagram in Figure 1. illustrates how microservices store, manipulate, manage, and distribute data.

Figure 1: Information view of the microservice

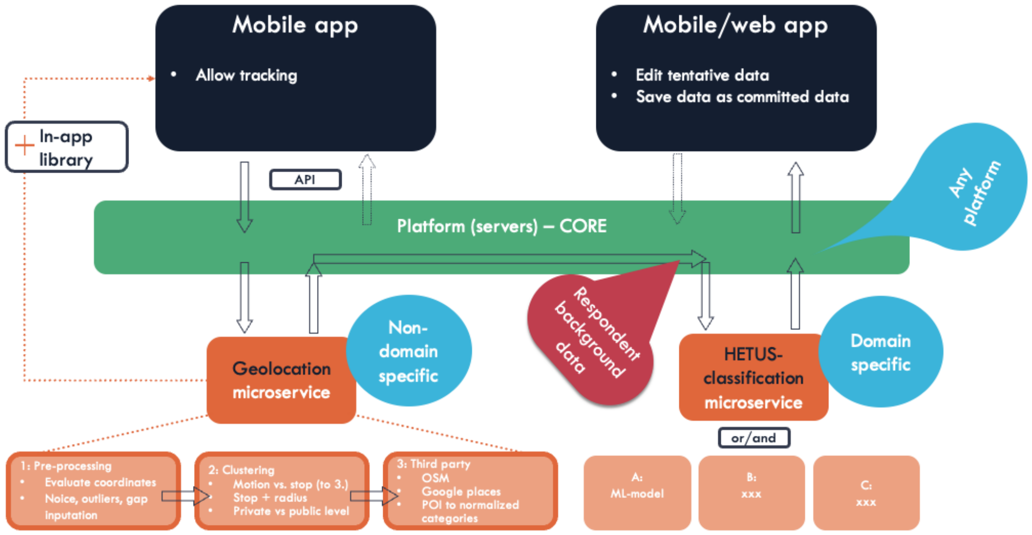

Data collection platform (not part of microservice)
push message with data
message bus
stores the message with data in a queue until delivered to message subscriber
push message with data
DataWriter
writes/saves data from the backend to the microservice db
write data
microservice DB
stores data unprocessed
database not directly accessible from outside microservice
data is stored forever
minimally processed 'raw' data
read data, user mgmt etc.
API and GUI
no storage of data/results
read data
results
processor result DB
database not directly accessible from outside microservice
results do not need to be stored forever as they can be re-calculated
write results (processed data)
DataProcessor
processes data and puts results in the result database
results (processed data)

A microservice is deployed as a collection of Docker containers, each runtime element is deployed as a container. To uphold data privacy, sensitive and personal information is confined to the database of the data collection platform. The microservice is strictly prohibited from extracting user or respondent data into its own databases. This restriction is enforced through a "microservice link" established between the microservice and the data collection platform, typically in the form of an identifier such as an ID, UUID, GUID, etc. Leveraging container technology erects barriers between different components utilized in the study setup, enhancing data security and isolation.

4. The microservices for Geolocation data

Integrated into the end-to-end process involving a data collection platform, the microservice is a pivotal component. Within the SSI project, these developed microservices will be incorporated into the MOTUS data collection platform (Minnen et al, 2023) to assess feasibility. This integration showcases how users or respondents can view microservice outputs as provisional data, which can then be incorporated into their timelines. The geolocation microservice itself is constructed as two distinct microservices. The first is a non-domain-specific microservice, while the second is more domain-specific. The diagram below illustrates the interaction between the components of the data collection platform and the microservice(s). The end-to-end process aims to assist users/respondents in completing their timelines through the microservice. With the user/respondent's consent, the mobile app tracks their geo-positions. These positions are then routed through the core component of the data collection platform and serve as input for the geolocation microservice.

Figure 2: Binding between data collection platform components and microservice(s)



The output of this microservice is subsequently processed through the platform to reach the second part of the microservice, which aims to associate both the travel mode and the activity prediction, corresponding to the HETUS classification of travel mode, places and activities, as it is illustrated in the following paragraphs. The way in which these parts of the microservice will be connected to each other in the architecture is still being studied within the SSI project.

4.1 Geolocation data, segmentation

The stop detection algorithm is structured into four main components, with an additional step to incorporate supplementary context from third-party databases (public or proprietary):

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.
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.

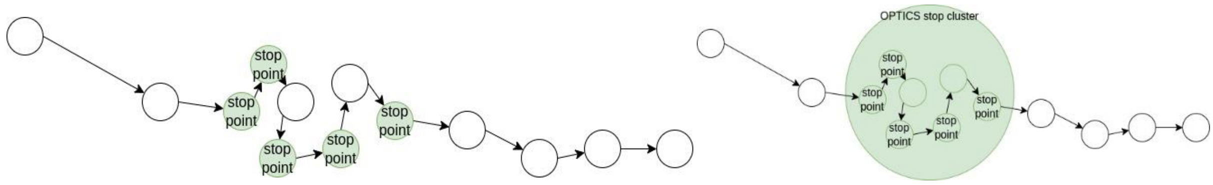
Additionally, an extra step involves attaching extra context by identifying points of interest or nearby places within the stop clusters (e.g., utilizing OpenStreetMap or Google Places).

The necessary input parameters include timestamp, longitude, latitude, and accuracy. The GPS points must be arranged in chronological order.

4.1.1 ATS-OPTICS

The first four steps are executed using the developed ATS-OPTICS algorithm, which comprises two parts. The initial part, known as (s)ATS, is engineered to perform stop detection and is based on the research paper by Bonavita *et al* (2022). Specifically, the implementation focuses on the individual stop-based adaptive trajectory segmentation (ATS) from this paper. In essence, the algorithm identifies a GPS point as a stopping point if there is a duration of more than "t" seconds between the current GPS point and the subsequent GPS point which is more than "x" meters away. The algorithm's parameters, both temporal and spatial, can be adjusted for tuning. By default, the implementation utilizes a spatial parameter of 50m (as recommended in the paper) and a temporal parameter of 180s. It's worth noting that the algorithm also supports deriving the temporal parameter from the GPS data using a Thompson tau statistic, as described in the paper. The second part of the process involves clustering the stop points to form stop clusters. This task is accomplished using the widely recognized OPTICS algorithm, which is based on DBSCAN and applies a density-based approach to spatial data (source: https://en.wikipedia.org/wiki/OPTICS_algorithm). Since OPTICS does not inherently consider the temporal dimension of the data, additional processing is performed to partition the spatial clusters based on time.

Figure 3: Stop identification and clustering



To ensure a polished output, the merging of stop clusters is conducted, followed by the addition of track clusters between the stops. 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. If users or respondents label stop clusters as private places (e.g., home, work, etc.), these locations will subsequently be recognized as private locations in future instances.

4.2. Travel mode prediction

Transport mode classification is possible after the geolocation data is segmented into stops and tracks. This approach requires a database with information about transport mode infrastructure, such as OSM. One option to determine which transport mode is used in a track cluster is to map the geolocation data of the segmented track clusters to the infrastructure data. Other approaches, such as machine learning, can be considered, but it was decided not to do so, as previous results were not promising (Smeets et al., 2019). 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. The transport modes available in OSM are motorized vehicles on roads, trains, trams, buses, subways, bicycles, and on foot. It can be calculated which transport mode has the largest proportion in the track cluster considered.

The transport mode with the largest proportion is then considered as the most plausible mode and can be assigned to this cluster. Depending on the infrastructure, it can be the case for certain track-clusters that no OSM is available for a transportation mode. It is also possible to generate multi-modal clusters when different transport modes have the same proportion in a cluster. In these cases, respondent interaction might be required to select the correct transport mode.

The quality of the infrastructure data is particularly important in this approach. If different transport modes have different numbers of data points, this can lead to biases towards or against certain transport modes. 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.

4.3. Activity prediction (WP2.2 - ISTAT)

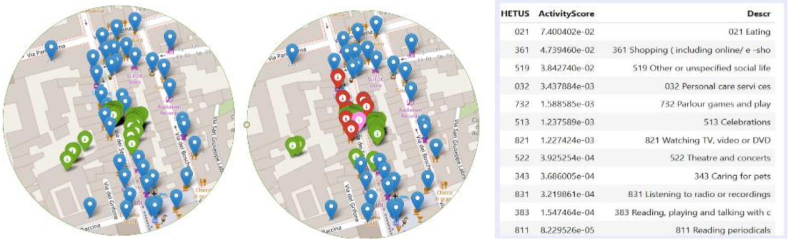
The last microservice associates HETUS activities distribution (with assigned probabilities or scores) to the stop identified in the first microservice, through an algorithm that exploits several information, such as place categories taxonomy, timing of the stop, country-specific indicators derived from previous HETUS survey data and user characteristics. Categories of place from the third party (Google Places GP or OSM) are mapped to the HETUS classification of places.

The procedure consists of the following steps:

- A score (POI-score) is assigned to each POI inside an adaptive radius around the stop centre location, based on the weighted median of the distances calculated between each POI and all GPS points of the stop, weighting by the accuracy of GPS points.
- A short list of POIs is identified using the elbow criterion on the POI scores.
- Through a Bayesian decomposition, for each POI of the short list the conditional probability of HETUS activities are calculated starting from the distribution observed in TUS data. 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.
- 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.

The main steps of the activity prediction procedure are displayed for an example stop in th e following figure.

Figure 4: Main steps for the activity prediction, input and output



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

5. Discussion and quality issues

The presented microservice is under development in the SSI project: the main pipeline steps have been implemented but the assessing and testing phase is ongoing. The quality assessment is a crucial part of the successive activity of the SSI project, for evaluating the impact of different choices and types of data on the goodness of predicted variables to be used as tentative data in the survey, through an interaction with the respondent.

A major concern in smart surveys and sensor data passively collected is the quality of the data itself. While it is true that sensor data acquired passively can lead less measurement errors than self-reports, it is also true that these data are not free from biases. In the case of location data, for instance, the GPS data themselves and the subsequent algorithms used to segment the tracks and to predict travel mode, trip purpose and activities carried out at a location, are not fail-proof: GPS tracks may be incomplete with missing data determined by people's behaviour or by the situation in which the GPS signal may be lost; GPS tracks may be complete but, for instance, the trip detection algorithms may be too sensitive (over-identifying trips) or not sensitive enough (under-identifying trips); heterogeneity in sensor quality between different types of smartphones influence the measurements and the derived statistical information, as well as the quality of contextual data sources (map services) used to add place details, can influence the accuracy of the prediction. The quality of map services may be different across European countries: this must be taken into account in this context, together with overcoming the stringent GDPR issues that can prevent the use of Google Places.

References

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

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*