Discussion paper

Eight years of developing a general population smart travel survey. *Lessons learned from two large field tests*

Jonas Klingwort
Barry Schouten
Daniëlle Remmerswaal
Danielle McCool
Yvonne Gootzen
Jelmer de Groot
Peter Lugtig
Marleen Schulte

February 2026

Summary

Statistics Netherlands (CBS) started investigating the potential of location tracking combined with machine learning in a general-population travel survey in 2017. CBS built and revised a cross-platform smartphone app with flexibility and transparency in methodology in mind. Two large-scale field tests were conducted in 2018 and in 2022–2023, randomly varying recruitment and user interface options. In this paper, we look back and make up the score. We look at the opportunities and challenges that we anticipated to exist in 2017. We also evaluate whether we overlooked opportunities or challenges. Moreover, we look at open (research) questions. Our focus is primarily on the methodology behind 'smart' travel surveys, but we also touch on technological constraints and boundaries.

Keywords

location tracking, mobility, sensor data, UI-UX, GPS, smartphone, machine learning

Content

1. Introduction

In this paper, we look back at eight years of research and development of a travel survey app, including location tracking and machine learning predictions. We confront the alleged opportunities of respondent burden reduction and improved data quality with the anticipated technological and methodological challenges. We summarize the analyses and conclusions from a series of papers that have been written within this time frame.

Surveys employing features of smart devices have been referred to as 'smart' surveys. In the European Statistical System (ESS) context, several smaller and larger projects have been oriented at testing and implementing smart device features[1]. The main motivations for embedding such features are improving data measurement quality, reducing respondent burden and modernizing the user interface and user experience. Survey topics that are time-consuming or cognitively burdensome, that require expert knowledge and/or that are hard to measure through questions, are considered the most promising to be revised (Schouten et al. 2025). Travel surveys satisfy these criteria. They are burdensome and require a detailed recall of daily travel and activities.

For travel surveys, three smart features can be employed to reduce respondent burden and/or improve measurement data quality: location tracking through smartphone sensors, real-time linkage to point-of-interest data and real-time machine learning predictions. The first may allow for complete and full recording of travel trajectories. The second and third may support segmentation of trajectories into stops and tracks, prediction of the modes of transport and prediction of the purposes of stops. In 2017, the literature on location tracking and transport mode prediction was rapidly growing. See the literature reviews included in the work by Smeets et al. (2019), Fourie et al. (2025), and Boer et al. (2026). The literature on stop-track segmentation and on stop-purpose prediction was, however, very thin. An early example of stop-purpose prediction is Xiao et al. (2016). The most recent research in this field was conducted by Zahroh et al. (2025a,2025b).

In 2017, CBS embarked on a program on how to develop smart surveys for use in official statistics. Travel surveys were considered strong candidates to be supplemented by the location-tracking and real-time computing features of smart devices. Statistics Netherlands recognized the potential for its general population repeated travel survey[2] and initiated a research project in collaboration with the Ministry of Infrastructure. In this paper, we adopt a retrospective look and evaluate how much of the promise made it to practice. The central question is whether the benefits to travel statistics accuracy still outweigh barriers to

[1] For an overview of the projects and a link to the project deliverables see https://cros.ec.europa.eu/dashboard/trusted-smart-surveys.

[2] See https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen for background to the survey in Dutch and https://www.cbs.nl/en-gb/our-services/methods/surveys in English.

implementation. Schouten et al. (2025) also review the "business case" of smart surveys. They consider the application of smart surveys to travel, physical activity and time-use official statistics. They use general criteria applicable to multiple settings. They conclude that the business case for travel statistics is modestly positive. Here, we look in much more detail at the case of travel statistics.

The traditional travel survey asked for one diary day only, partly because of the response burden. In this diary, respondents had to keep track of every trip they took, including start- and end-time, start- and end-location, mode of transport, and details on whether the trip was taken jointly with others. The basic idea of the smart travel survey was that GPS-tracking would be used so that the start and end locations, times, and route would be recorded. Respondents still had to provide travel modes and travel purposes. The labeled stops and tracks were the starting point for training of machine learning models.

From the outset, we envisaged there would be two main challenges. The first challenge was technical: stable location tracking technology was known to be a complex task. Enforcing relatively high frequency comparable tracking across the most prevalent platforms, Android and iOS, and across the wide range of brands and models was a known technological challenge. It demanded a trade-off between battery depletion and tracking accuracy. Another known technological challenge was the prevalence of gaps in location tracking. Gaps could be caused by a lack of connection to GPS, Wi-Fi, or cellular networks and by cold starts when smartphones are reactivated in resting mode. The second challenge was methodological: the recruitment and motivation of the general population to use location tracking would possibly be challenging. Next to willingness, we anticipated that insufficient motivation and skills to interact with a smartphone app would be potentially influential in the take-up of the app within the general population.

Two field tests were conducted between 2017 and 2025 with the goal of seeing how the technical and methodological challenges could be met. In both field tests, a cross-platform smartphone app was employed. The app was designed, developed, and revised by the Statistics Netherlands mobile app team. We had three reasons for in-house development. The first reason was the transparency of location-tracking technology and stop-track segmentation. Existing commercial vendors of apps were, and still are, not willing to reveal how they solve the methodological and technological challenges. The second reason was flexibility in training and updating machine learning models for travel mode and stop purpose predictions. Statistics Netherlands can link socio-economic and demographic auxiliary variables that were hypothesized to be important features in such models. The third reason was the flexibility in monitoring and evaluation of in-app paradata on respondent navigation and answering behavior. The evaluation of respondent interaction was considered very useful for understanding insufficient motivation and/or digital skills. Such insights may inform adaptive survey designs.

During this same period of eight years, technology progressed, and literature became richer. However, also influential societal events occurred, such as the COVID-19 pandemic and growing incidences of cybercrime and phishing. See, for

example, Gillis et al. (2023) and Lawson et al. (2023) for reviews in the context of travel surveys.

We outline this paper as follows: In Section 2, we describe the two field studies that form the basis of our evaluations. We then review the opportunities and challenges in Section 3. Next, in Section 4, we briefly discuss new challenges, (inter)national developments, and open questions. We close with a discussion in Section 5.

2. Field studies. Objectives and design

Statistics Netherlands fielded two large-scale experiments. Additionally, one small-scale study was conducted to collect training data. The first large-scale experiment in 2018 was mostly exploratory, aiming to understand the methodological and technological feasibility. It was successful and paved the way to further development and a more sophisticated second large-scale experiment in 2022–2023. In 2024, a separate study was conducted with eight University of Utrecht student assistants specifically designed to collect data to build and validate the algorithms upon.

Here, we explain the study designs. More details about the 2018 and 2022-2023 designs of the large-scale experiments can be found in McCool et al. (2021) and Schouten et al. (2024) , respectively.

2.1 The AVA18 field study

The first field study, called AVA18, was conducted between October 2018 and December 2018. It concerned two samples of equal size, each comprising 951 individuals. One sample was fielded in October, and another in November. Each sample consisted of a simple random sample from former Dutch Travel survey respondents and a fresh simple random sample from the Dutch population register. Both were equal in size in both months and thus in total consisted of 951 persons. Only persons 16 years of age and older were selected to avoid obtaining parental consent.

All samples were invited via postal mail, and these invitation letters included links and QR codes for the app stores, along with login usernames and passwords. Two reminder letters were sent. The reminder letters addressed those sampled persons who had not logged in to the app after two weeks and after four weeks, respectively. All respondents were asked to use the app for a full week. Because a week of reporting would always involve all days of the week, the starting day was not prescribed. Respondents could start anytime. The seven-day period was

deliberately chosen. Firstly, because it avoided detailed instructions on when to start, and secondly, because it did not create a time lag between reading the invitation and the start of the reporting period, as was the case in the diary of the traditional travel survey.

Apart from the split between former respondents and new sample units, there were two experimental conditions. The first experimental condition was the height and timing of incentives. Three strategies were randomly allocated across the samples: 1) a conditional incentive of 5 euros after registering the app and another conditional incentive of 5 euros after completing a week of reporting, 2) a conditional incentive of 10 euros after completing a week of reporting, and 3) a conditional incentive of 20 euros after completing a week of reporting. In all cases, sampled persons were given an unconditional incentive upon invitation. The second condition concerned parameters for stop-track segmentation. The app performed real-time stop-track segmentation based on a radius and a duration parameter. A stop was detected whenever location-tracking points remained within the specified radius for at least the specified duration. Both parameters were randomly varied across the samples. This was done to investigate the robustness of stop-track decompositions. As it was anticipated that it would be better to have too many stops rather than too few, the randomization was compared to the respondent's experience. Values were taken randomly from the interval [1 min, 5 min] for the duration and randomly from the interval [50 meters, 200 meters] for the radius. The random selection of the two parameters was independent for both the fresh sample and the sample of former respondents.

Table 1 gives the sample sizes for the incentive and target group conditions. The third condition on radius and duration parameters was randomly crossed with the two conditions.

Table 1: Samples in AVA18 for the two experimental conditions. The third condition on radius-duration parameters was randomly crossed with the other two conditions.

	5 + 5 + 5 euros	5 + 0 + 10 euros	5 + 0 + 20 euros	Total
Fresh sample	317	317	317	951
Former respondents	317	317	317	951
Total	634	634	634	1902

The app asked respondents to label the transport modes in the identified tracks and the purposes in the identified stops. Respondents could not add, delete, or edit stops and tracks. They could label a track or stop as spurious. Per day, they were asked to fill in a short questionnaire about peculiarities. The questionnaire included an open question about the type of errors respondents may have experienced in the stops and tracks.

An invitation to complete an online evaluation questionnaire was sent to all sampled persons who had registered for the app. Hence, the people who had

dropped out were also invited. The invitation was sent in the first week of January 2018. Results of the evaluation questionnaire have been published in an internal CBS report. Around 40% of the respondents reacted that the app was (very) user-friendly. 23% rated the app as (very) not user-friendly. The main concern was battery usage.

2.2 The AVA22 field study

Once the analyses for the AVA18 experiment were finished, we prepared plans for further app development and testing. These included a second large-scale experiment. The focus shifted to the app's user interface and the respondent's interaction with it. See Schouten et al. (2024) for a full description of objectives and a translation to the experimental design.

The AVA22 experiment was fielded between October 2022 and February 2023. Three samples were drawn. The first sample was a random sample from respondents to the traditional single-day travel diary survey who completed this in September 2022. The sample consisted of respondents who had reported only travel by public transport and/or at least three consecutive work-related travels. The sampled persons were asked to use the app for one week and to complete the regular one-day online diary for one of the seven days. Respondents received a conditional incentive of twenty euros. The second and third samples were fresh, simple random samples of persons aged 16 and older from the population register. The difference between the second and third sample was in the amount of editing respondents could make to the tentative stops and tracks shown to them in the app. There were two conditions: full editing (second sample) and limited editing (third sample). The limited editing sample could fill in gaps and could remove stops and tracks. The full editing sample could also add stops and tracks, and edit their start and end times. The samples received slightly different instruction materials with the invitation letters. Also, the in-app instructions were slightly different to accommodate the actual editing options that the respondents had.

Table 2: Sample sizes for the various conditions and target populations.

App & questionnaire	Location tracking	Timing of questionnaire	Full editing		Limited editing Fresh sample
			Follow up sample	Fresh sample	
YES	Seven days	Invitation	667		
NO	One day	Invitation		212	212
NO	Seven days	Invitation		212	212
NO	One day	1st reminder		212	212
NO	Seven days	1st reminder		212	212
NO	One day	2nd reminder		212	212
NO	Seven days	2nd reminder		212	212

Two experimental conditions were added for the fresh samples with limited and full editing options (see Table 2). The first condition concerned the length of the reporting period. One half was asked to participate for one day and received instructions to use the app on a specified day of the week. The other half received an invitation for seven days and could start whenever they wanted. The one-day sample, however, did not have to stop after the first day. In the invitation and reminder letters, it was stated that they could stay in the study for the full week. Both samples were offered a conditional incentive of 10 euros, given after completing the full study (either 1 or 7 days). The invited study duration condition was randomly crossed with the limited- and full-editing samples.

The second condition that was manipulated in the fresh samples concerned the mixed-mode design. We offered respondents in the second and third samples a mix of the regular online diary without location tracking and the app, including location tracking. However, the timing of the choice between the app and the online diary was randomized into three scenarios. In the first scenario, the choice was offered right at the invitation. In the second scenario, the choice was offered in the first reminder letter, sent after two weeks. In the third scenario, the choice was offered in the second reminder letter, sent after four weeks. In the second and third scenarios, the sampled persons were first asked to use the app. If they did not log in to the app, they were selected for the reminder letters. In case they opted for the online diary, they always had to fill in only one day. Hence, the seven-day sample had the choice between a full week of app reporting and one day of the online diary. Also, the incentive differed between using the app and using the online questionnaire; in the app, there was a conditional 10 Euro incentive, and in the online questionnaire a lottery after completion. The app's user interface was the same for the one-day and the seven-day samples. However, in the invitation and reminder letters, a weekday was specified for the one-day sample. Also, for the online diary, a weekday was specified. Consequently, the invitation and reminder letters were much longer than for the AVA18 field study.

2.3 The UU24 training data study

After AVA22, we concluded that the respondent-labelled travel modes and travel/stop purposes were not accurate enough as training data for machine learning models. Part of the location tracking data had gaps, part of the tracks comprised multiple travel modes, and part of the respondents did not (fully) label events. We decided to perform a small-scale, dedicated study to collect training data. Eight student assistants plus two Statistics Netherlands team members participated in the study.

The study addressed travel scenarios that were known to be harder to segment into tracks and stops. The study focused on travel mode prediction. The stops were mostly artificial.
The scenarios included:
– Bus and tram travel with many intermediate stops
– Public transport travels with multiple switches between modes
– Travels switching between walking and riding/driving

– Roundtrips by foot or by bicycle
– Travels close to waters (canals, rivers, lakes)

Detailed diaries were kept by the participants on all stops and travels. The diaries were aligned with the location tracking data.

3. An evaluation of opportunities and challenges

In the introduction, we mentioned the opportunities and the challenges that we saw in 2017. The two field tests, AVA18 and AVA22, and the small-scale study UU24 formed the basis for evaluating the promise of the opportunities and the solutions to the challenges. A series of research projects has been conducted to translate opportunities and challenges into manageable research questions. The projects concerned research master projects, PhD projects, internal Statistics Netherlands projects, and projects funded by Eurostat. The projects are pieces of the larger puzzle.

We first provide an overview of papers. Next, we discuss each opportunity and challenge individually. In doing so, we summarize the various projects. We conclude by listing the new challenges we identified over the past eight years.

3.1 The 2017 opportunities

The three smart features, location tracking, linkage to online point-of-interest (POI) data and real-time machine learning (ML) predictions were translated to five opportunities:

1. Obtain complete travel trajectories (location tracking)
2. Real-time stop-track segmentation (location tracking plus POI data)
3. Real-time travel mode prediction (POI data plus ML predictions)
4. Real-time stop purpose prediction (POI data plus ML predictions)
5. Longer reporting periods (location tracking)

In Table 3, we provide an overview of papers that investigated the opportunities. Per paper we also list the field experiments/study that were employed.

Table 3: Reports/discussion papers/journal papers for the five opportunities, including the field study, AVA18, AVA22, or UU24.

Opportunity	Paper	Data set
Trajectories	McCool et al. (2021) McCool et al. (2024)	AVA18 AVA18
Segmentation	Killaars et al. (2020) Gootzen et al. (2024) Klingwort et al. (2025b) Klingwort et al. (2025c)	AVA18 AVA22 AVA22 AVA22
Travel mode	Smeets et al. (2019) Fourie et al. (2025) Boer et al. (2026) Klingwort et al. (2025a) Klingwort et al. (2025b)	AVA18 AVA22, AVA18, UU24 AVA22, AVA18, UU24 AVA22 AVA22
Stop purpose	Zahroh et al. (2025a,2025b)	AVA22
Reporting period	Remmerswaal et al. (2024)	AVA22

3.1.1 Trajectories

The main motivation to consider the replacement of the traditional survey with an app with location-tracking is the availability of accurate travel trajectories. In traditional diaries, trajectories can only be constructed based on reported postal codes, start and end times, and travel modes. Especially in urbanized areas, there would be a multitude of potential trajectories. Given that respondents may react to traffic congestion and public transport delays, favorable trajectories could also vary from one day to another. Furthermore, it has long been conjectured that shorter travel, in distance or duration, may be overlooked by respondents. If respondents have their smartphones with them, which, to date, is the rule rather than the exception, these will not go unnoticed. Also, the segmentation of tracks into trips, which is tedious for respondents and requires detailed recall, may be facilitated by adding contextual points of interest based on their vicinity to observed trajectories. Finally, even if respondents would recall all travels and trips, they may not be motivated enough to go through the burden of providing all details. The location tracking trajectories remove part of the burden.

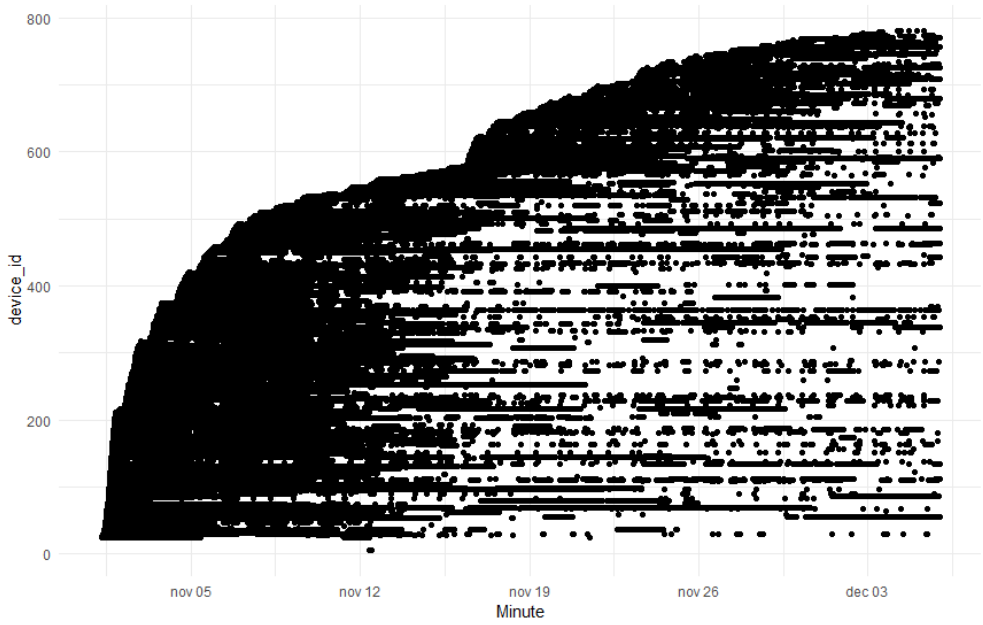

Figure 1: Participation patterns of the AVA18 app. Black dots represent the location data were provided in a particular hour. The app study started on Oct 30, and a reminder was sent on November 15. Black dots represent 10-minute intervals in which at least one observation was made. The device IDs are plotted on the Y-axis. The X-axis runs from Oct 30 up to Dec 10.

There are two prerequisites to this promise: Trajectories should indeed be accurate, and trajectories should be complete. McCool et al. (2021, 2024) investigated the properties of location data. The first prerequisite seems to hold true. There is noise in location data, but it can be reduced to acceptable levels with relatively high-frequency tracking. Smoothing techniques such as Kalman smoothing may be added to remove outlier points and further improve the accuracy of trajectories. The second prerequisite holds only conditionally: If location tracking is not interrupted, then trajectories appear to be mostly complete. It happens only very occasionally, according to respondents, that travels are missed because they forgot to take their device or because they are not seen by the stop-track decision rules. However, travels are frequently interrupted by failing technology. Figure 1 shows the available data for the AVA18 field test respondents. There are but a few respondents who had no interruption of location tracking for the entire reporting period. We will not discuss this issue here, as we suspected this to be one of the main challenges. We will return to the missing data problem in Section 3.2.2.

3.1.2 Stop-track segmentation

Segmenting smartphone-collected trajectory data into distinct “stop” and “track” episodes is critical for generating meaningful mobility diaries to be presented to the respondent. Further, these segmentations serve as input for travel mode and stop purpose predictions.

Killaars et al. (2020) investigated the sensitivity of stop-track segmentation to the radius and duration parameters. They used the AVA18 field study that had

randomized the two parameters per sampled person who logged in to the travel app. Figure 2 shows the number of stops detected in AVA18, dependent on the radius and duration parameters. It was concluded that the number of stops is relatively robust for a radius larger than 50 m. The duration parameter is more influential. Given that we anticipated that respondents would find it harder to add stops and tracks than to delete them, we decided to go for relatively short duration thresholds in segmentation for AVA22.

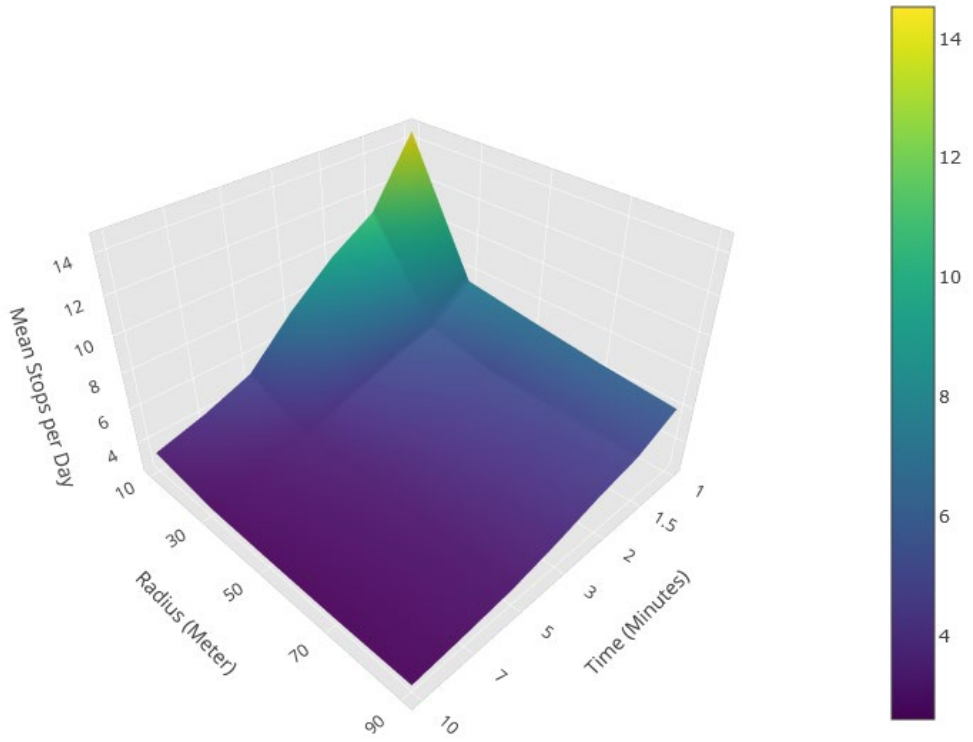

Figure 2: Mean stops per day for all stop-and-go parameters. Each data point contains the average of the mean number of stops per day for that specific stop-and-go parameter combination for all 674 respondents with data.

Three reports have been published that address the quality of the algorithm-based stop-track segmentation (Gootzen et al. (2024), Klingwort et al. (2025b), Klingwort et al. (2025c)). Summarizing these papers, moderate to high agreement was found (accuracy between 0.86 and 0.94) when comparing two automated segmentation algorithms against a response-based web diary that served as the ground truth. However, they also found that precision for stops was higher (0.96) than recall (0.89) in one algorithm, meaning that the algorithm was more successful at correctly identifying flagged stops than at finding all actual stops. Furthermore, it was emphasized that segmentation errors tend to arise from mis-detection of transitions (start/end of movement vs. stop), gaps due to sensor malfunction, and unrealistic fixed-threshold settings in the algorithm (e.g., minimum stop duration or stationary radius). In addition, it was found that minor changes in the ground-truth reference (label changes in the response-based web diary) shift accuracy by about seven percentage points, underscoring how sensitive segmentation quality is. Together, these studies suggest that while algorithmic stop-track segmentation

seems promising, the quality is still constrained by algorithmic assumptions, sensor inconsistencies, and ground-truth definition, meaning rigorous validation and algorithm tuning remain important.

3.1.3 Travel mode prediction

Automatic travel mode prediction (e.g., walking, biking, car, bus, or train) has been extensively analyzed in several studies by Smeets et al. (2019), Fourie et al. (2025), Klingwort et al. (2025a, 2025b), and Boer et al. (2026). The most straightforward algorithm, developed by Fourie et al. (2025), is based on a set of deterministic rules. This algorithm used GPS and OpenStreetMap (OSM) data, achieving an overall classification accuracy of 85% (balanced accuracy: 70%[3]). In total, 183 GPS and OSM features were considered in this study, and 13 were included in the final algorithm. This algorithm showed good generalizability to other data and another country (Italy). The work by Klingwort et al. (2025a) introduced a more complex algorithm based on a decision tree. The algorithm achieved a reasonable overall accuracy of 80% (balanced accuracy: 46%). About 200 features based on GPS and OSM were considered. The tree identified 39 features as important. However, the results also showed that the decision tree generalized less well and likely overfitted the data. The work by Smeets et al. (2019) and Boer et al. (2026) used more complex machine-learning algorithms (Random Forest and XGBoost). The study by Smeets et al. (2019) used a Random Forest with 127 features engineered from GPS and OSM data, achieving an accuracy of 85%. The study by Boer et al. (2026) found that XGBoost was the best-performing algorithm, achieving 91% accuracy (balanced accuracy: 92%). When tested on external data, the XGBoost still performed well. All studies conclude that speed-related GPS features have greater predictive power than OSM-based features, although OSM is required to achieve good model performance. Furthermore, the studies show that predicting public transport modes is most challenging. Overall, the studies conclude that a fully automated transport mode classification in the travel app may not yet be feasible; a hybrid approach with targeted prompts to the respondent to check predictions could be used to balance prediction accuracy and response burden.

3.1.4 Stop purpose prediction

Zahroh et al. (2025a, 2025b) investigated how well trip-stop purposes (e.g., home, shopping, work) can be automatically predicted. The AVA22 dataset was used, in which participants labelled stops with 12 distinct purpose labels. They developed and compared multiple machine-learning models, including artificial neural networks and extreme gradient boosting, using mainly spatio-temporal features (e.g., time of day, number of visits to a location) and administrative/contextual data. Results show that the best-performing models achieved a balanced accuracy of about 90 % for the “home” purpose class. However, performance for less frequent categories was much lower: the models tended to misclassify rare stop purposes into common ones. It was found that adding administrative data provided a slight improvement over models that use only place and time related features. It was also observed that extending the data-collection duration

[3] In balanced accuracy all travel modes are weighted equally, whereas in unbal;anced accuracy the travel modes are weighted by prevalence in the data.

improves predictive precision. The paper concludes that while smartphone-based tracking offers considerable potential for predicting trip purposes, it is not yet reliable enough to predict purpose without further refinement.

3.1.5 Longer reporting periods

Since we anticipated that both perceived and actual respondent burden would be strongly reduced by location tracking, longer reporting periods without increased drop-out seemed feasible. We started to explore a reporting period of a week. As an additional advantage, we saw the opportunity to avoid a prescribed starting day of the week.

Three research questions have been investigated: 1) What is the impact of longer reporting periods on response rates? 2) What is the impact of longer reporting periods on measurement data quality? 3) How does the increase in precision of statistics depend on the length of the reporting period? The third question distinguished within-respondent variation and between-respondent variation in travel. Apart from the methodological and statistical perspectives, the reporting period length has a legal perspective[4].

The first two questions have been studied in detail by Remmerswaal et al. (2024) based on the AVA22 field study. Table 4 is extracted from their paper and gives a summary of registration, participation and completion rates for the single-day and the seven-day samples. They found that the longer reporting period of seven days had higher resgistration rates than the one-day reporting period. Remmerswaal et al. (2024) did not draw strong conclusions about the impact of measurement data quality. In Table 5, they summarize five proxy indicators of data quality based on avaibaility of tracking data, stop-track segmentation, labeling of tracks and stops and editing of tracks and stops. The number of respondents was too small.

Table 4: Response rates per experimental treatment group for the AVA22 study

	Single-day		*Seven-day*		*p-value of difference*
	%	*95% CI*	*%*	*95% CI*	
Registation	*11*	*(9,13)*	*14*	*(12,16)*	*0.03*
Participation	*11*	*(9,12)*	*13*	*(11,14)*	*0.11*
Completion	*7*	*(6,9)*	*7*	*(6,9)*	*0.94*

Remmerswaal et al. (2025) investigated the balance between dropout and length of reporting period in three smart survey applications: travel, household budget, and physical activity tracking (Figure 3). They applied multi-level models to account for the clustering effect of events reported by the same respondents. For

[4] The legal perspective has been prominent in the AVA18 and AVA22 field studies, because location data is considered sensitive. The legal acceptance of longer reporting periods is closely connected to the distinction between person-level statistics and event-level statistics. The stronger the interest is in within-respondent variation, i.e., the person-level, the stronger the legal rationale becomes for longer reporting periods. The discussion is impacted by the traditional limitation to a single day of reporting. This tradition, at least in part, originates from the respondent burden in travel diaries.

travel, they then set an upper threshold on the total amount of in-app reporting time of all respondents. They then concluded that tracking multiple days increases precision. Hence, the actual burden in minutes remains the same, but standard errors are smaller when asking respondents to report longer. This is in part the result of differences in travel between working days and weekend days. The gain in precision is smaller with every extra day of reporting. The results, in theory, support an optimization of reporting period length. However, practical implications in implementation must obviously be accounted for. The reference/starting day discussion mentioned earlier may warrant a choice between one day or a week.

Table 5: Measurement quality of users with any geolocation data collected on the first full day

	1st full day of one-day group			**1st full day of seven-day group**			**Mean of first full six days of the seven-day group**		
		95% C.I.			95% C.I.			95% C.I.	
	Group mean	Low	Up	Group mean	Low	Up	Group mean	Low	Up
Mean minutes per day with a geolocation measurement.	661	565	763	608	520	700	502	459	540
Mean hours per day with at least one geolocation measurement	16	14	17	16	14	17	13	12	14
Percentage of days with compiled trips and stops.	88	84	93	84	79	89	89	86	92
Percentage of compiled diary days with labels.	78	71	85	73	66	80	69	65	73
Percentage of days with manually added stops or trips (participants with full editing only).	23	14	32	20	12	28	25	20	30
Total diary days	95			115			538		

95% Confidence Intervals (C.I.) are based on 1000 bootstrapped datasets.

Remmerswaal et al. (2025) investigated the balance between dropout and length of reporting period in three smart survey applications: travel, household budget, and physical activity tracking (Figure 3). They applied multi-level models to account for the clustering effect of events reported by the same respondents. For travel, they then set an upper threshold on the total amount of in-app reporting time of all respondents. They then concluded that tracking multiple days increases precision. Hence, the actual burden in minutes remains the same, but standard errors are smaller when asking respondents to report longer. This is in part the result of differences in travel between working days and weekend days. The gain in precision is smaller with every extra day of reporting. The results, in theory, support an optimization of reporting period length. However, practical implications in implementation must obviously be accounted for. The reference/starting day discussion mentioned earlier may warrant a choice between one day or a week.

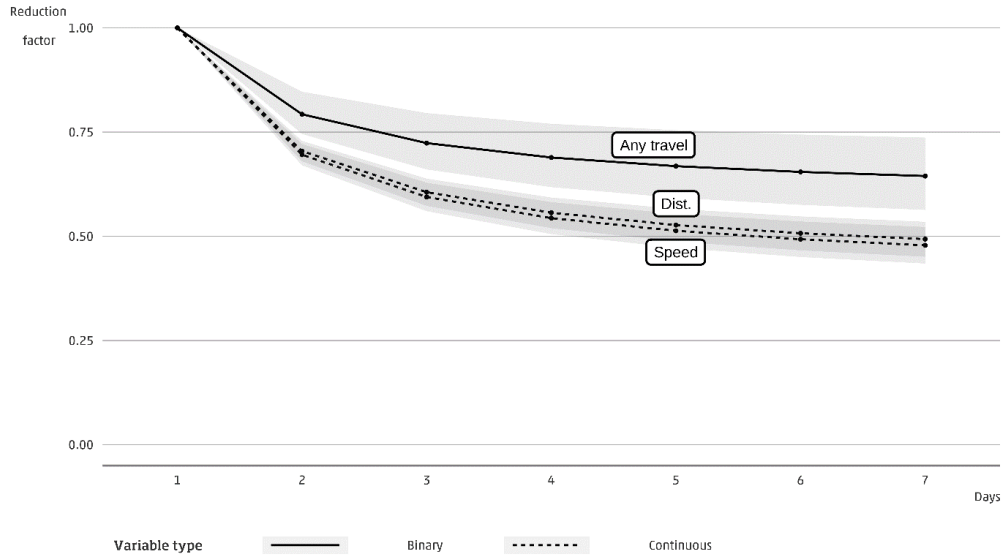

Figure 3: Theoretical approximation of the sample size reduction for three travel variables (any travel, distance, speed) when conducting multi-day diaries instead of a one-day diary and keeping the same precision. Shaded areas represent 95% confidence intervals. The X-axis shows the number of days in study, running from one to seven days.

3.2 The 2017 challenges

We expected four challenges, two technological and two methodological:

1. Technological
   a. Location tracking comparability
   b. Gaps in location data
2. Methodological
   c. Respondent willingness

d. Respondent user interface (UI) and user experience (UX)

In Table 6, we show the papers that evaluated the challenges.

Table 6: Publications for four challenges including the field study, AVA18 or AVA22.

Challenge	Paper	Data set
Tracking	Van Houwelingen et al. (2023) Gootzen et al. (2024)	AVA22 AVA22
Gaps	McCool et al. (2024) McCool et al. (2026) Gootzen et al. (2024)	AVA18 AVA18 AVA22
Willingness	Lugtig et al. (2022) Remmerswaal et al. (2024) Schouten et al. (2024)	AVA18 AVA22 AVA22
UI-UX	Remmerswaal et al. (2024) Giacobbe et al. (2024)	AVA22 AVA22

3.2.1 Location tracking comparability

In the paper by Gootzen et al. (2024), it was analyzed how smartphone-based GPS/sensor apps can both enhance and challenge data quality. On the one hand, passive location-tracking reduces respondent burden (no need to log every trip manually) and provides rich, time-stamped spatial trajectories that can support automated inferences about travel modes and purpose. However, a range of data-quality issues tied to the technology were identified: device brand, battery level, sensor accuracy, data gaps, and varying participation durations all affect whether the resulting GPS data is usable. The analysis suggests up to 20% of tracking records may be unusable or require substantial imputation due to missing or noisy data. The paper thus argues that while tracking technologies hold great promise for improving mobility survey data, achieving high quality requires carefully managing both technological (e.g., sensor settings, device heterogeneity) and methodological aspects (e.g., respondent interaction, editing capability, post-processing imputation) to offset errors inherent in passive tracking.

3.2.2 Gaps in location data

Gaps in location-tracking data have been extensively studied in AVA field studies. Despite attempts to improve tracking and assist respondents in avoiding gaps, they remain an influential source of error.

McCool (2021, 2024, 2026) discusses the causes of gaps and ways to handle them. They make a distinction between very short, short, and long gaps. Very short gaps are in the order of seconds and pose no problem. Short gaps are in the order of minutes. They may be caused by a cold start (the time needed for a mobile device to notice movement and change tracking frequency), underground travel, or buildings/infrastructure blocking GPS signals. The trajectories can often be reconstructed from the last seen and first newly observed locations. The longer gaps are in the order of hours. They mostly come from battery-saving and depletion. They also occur frequently during the night when the device is in

hibernation. The respondents in the AVA18 and AVA22 study were asked to 'white-list' the app to exempt the app from battery saving when other (battery-demanding) apps are running. Many respondents ignored the request or did not know how to exempt an app from battery saving. When ignored or applied incorrectly, the app can be stopped by the device operating system and only restarts when it is activated by the respondent. The resulting longer gaps occurred frequently in both AVA18 and AVA22 data.

McCool et al. (2024,2026) propose imputation methods for the short and long gaps, respectively. They evaluate their efficacy through simulations in which gaps are introduced in complete real-life location data. The methods are compared to naïve approaches (linear interpolation, case-wise deletion, day-wise deletion). They conclude that the advanced imputation methods can improve accuracy. However, they also argue that for long gaps, the improvement depends strongly on the mechanism underlying the missing data, which is, of course, unknown in the case of real missing data.

Figure 4 displays the impact of (simulated) short and long gaps on six travel statistics. In the simulations, the proportion of time periods without a measurement is evaluated. Sparsity levels of less than 20 to 30% have little impact on short gaps for all types of travel statistics. For long gaps, the picture is different. At these same sparsity levels, biases may already be 25% or more. The simulations, thus, show that missing data, when ignored, may give substantial shifts in travel statistics.

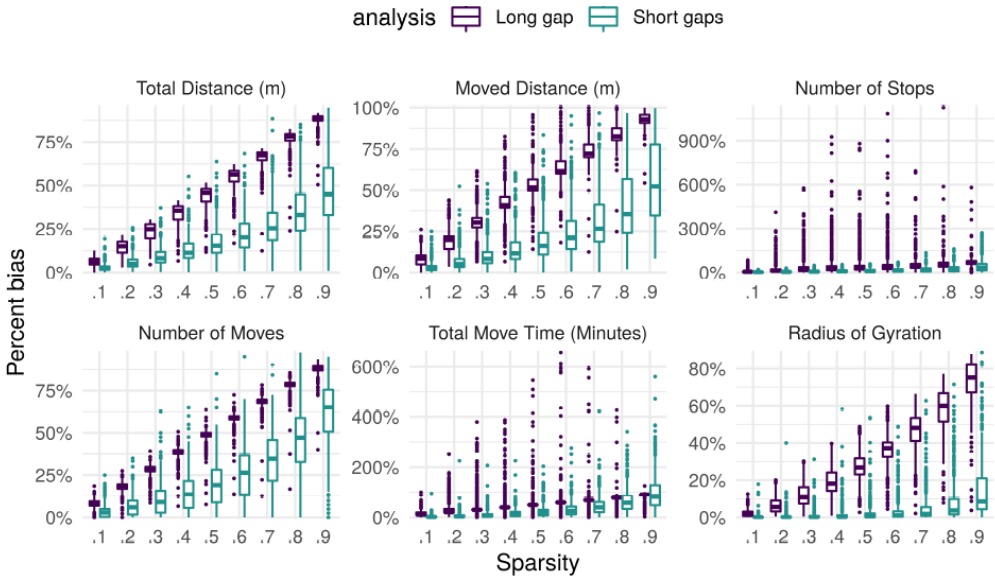

Figure 4: Impact of short and long gaps on bias in six travel statistics when applying naive interpolation. Sparsity reflects the proportion of time periods without a measurement.

These findings raise the question of how imputed data can be included in the analysis and estimation of statistics. The question applies most of all to in-depth analyses that employ microdata, such as road and rail infrastructure models. We

return to this question in Section 4.1, where we discuss data integration as one of the main new challenges. Table 7 shows the (simulated) performance of different imputation strategies as a function of gap length. Imputation methods using donor data by the same respondent or similar respondents can remove part of the bias. However, large amounts of donor data and longer tracking periods will be needed to further reduce biases. As expected, longer gap lengths lead to more and more bias, even after adjustment.

Table 7: Statistical properties of different strategies to handle gaps for different gap lengths. LI = linear interpolation, TWI = time window imputation, DTWBI = dynamic time warping imputation, DTWBMI-HI = dynamic time warping multiple imputation with high information, and DTWBMI-LO = dynamic time warping multiple imputation with low information. High information means at least 8 hours of location data is available before and after a gap. Low information means an hour of location data is available before and after a gap.

	Abs Bias	Trips	Med Bias	RMSE	TP Over	TP Under	TP Acc.
1 hr							
LI	0.000	1.03	0.000	0.10	6.5%	1.0%	93.2%
TWI	0.002	1.03	0.000	0.21	15.3%	11.1%	87.2%
DTWBI	0.000	1.03	0.000	0.13	4.5%	9.8%	92.5%
DTWBMI-HI	0.000	1.03	0.000	0.09	0.9%	8.4%	95.3%
DTWBMI-LO	0.003	1.03	0.000	0.11	4.7%	4.5%	95.1%
3 hrs							
LI	0.000	0.98	0.000	0.17	14.2%	3.0%	85.2%
TWI	0.025	0.98	0.000	0.21	23.5%	12.5%	91.5%
DTWBI	0.000	0.98	0.000	0.14	9.2%	8.7%	95.7%
DTWBMI-HI	0.007	0.98	0.000	0.13	6.7%	11.5%	95.5%
DTWBMI-LO	0.012	0.98	0.000	0.13	7.7%	8.4%	96.2%
6 hrs							
LI	0.000	0.95	0.000	0.32	26.0%	7.0%	72.8%
TWI	0.030	0.95	0.000	0.28	25.7%	21.9%	91.6%
DTWBI	0.010	0.95	0.000	0.23	12.0%	15.8%	94.2%
DTWBMI-HI	0.013	0.95	0.000	0.24	15.9%	20.0%	92.9%
DTWBMI-LO	0.015	0.95	0.000	0.24	14.3%	14.6%	94.1%
10 hrs							
LI	0.040	0.98	0.000	0.46	37.2%	16.0%	60.5%
TWI	0.029	0.98	0.000	0.32	33.4%	26.7%	92.5%
DTWBI	0.030	0.98	0.000	0.29	25.4%	17.6%	94.4%
DTWBMI-HI	0.003	0.98	0.000	0.28	20.2%	25.8%	93.5%
DTWBMI-LO	0.007	0.98	0.000	0.29	24.6%	19.3%	93.7%
12 hrs							
LI	0.060	0.98	0.000	0.44	35.1%	23.0%	62.6%
TWI	0.027	0.98	0.000	0.32	32.6%	27.6%	93.2%
DTWBI	0.030	0.98	0.000	0.27	18.7%	23.0%	95.4%
DTWBMI-HI	0.003	0.98	0.000	0.28	20.0%	31.5%	94.2%
DTWBMI-LO	0.000	0.98	0.000	0.28	20.7%	25.2%	94.5%

3.2.3 Respondent willingness

Independent of the impact of technological deficiencies on measurement and missing data, there is the potential impact of low and/or selective response. Location tracking and real-time predictions of stop and track characteristics may

be appealing to respondents because they remove the burden in terms of time and cognitive effort. But location-tracking data, when enriched with contextual information, also provides a big surplus of information relative to what is needed. Although location tracking is by now relatively standard in mobile device apps, respondents may object to using it in the context of surveys for privacy reasons. The two AVA field tests had various strategies to enhance willingness to participate (height and type of incentives, length of reporting period, and concurrent options to fill in a regular questionnaire) as the main target. They are described in Lugtig et al. (2022), Schouten et al. (2024), and Remmerswaal et al. (2025).

The findings from the AVA18 and AVA22 field tests are very mixed. In AVA18, response rates were promising and comparable to the regular online survey. However, in AVA22, the response rates were lower. It is unclear whether the complex design of AVA22 affected response rates. The invitation letter that included both the smart and non-smart options, which determined the choice of reference period and the height of incentives, may have confused and discouraged respondents. AVA22 also followed the COVID-19 pandemic, which may have changed perceptions towards tracking and surveys in general. AVA22 did show that longer reporting periods of up to a week may not decrease willingness. In fact, response rates were slightly higher than for a single reporting day. This finding seems to confirm the assumption that respondents hesitate to download an app for one day. AVA22 also showed that the mix of smart and non-smart options may lead to more balanced response rates relative to either the smart or non-smart option. This turned out to be true across all background characteristics except for the migration background. Nonetheless, the general conclusion must be that optimal recruitment and motivation strategies have not yet been identified.

Table 8: Self-reported consent rates to the hypothetical participation in different types of smart surveys. The question was: Would you participate in a(n) CBS survey which asks you to? For each case the purpose was explained in the introduction.

Type of smart feature and survey	Yes	Maybe	No	DK
Share location	24.9%	24.9%	37.8%	4.1%
Share pictures of your house	11.8%	17.9%	58.9%	3.1%
Share data on energy use	40.9%	24.8%	28.0%	3.8%
Use an air quality monitor	47.4%	19.8%	24.7%	5.3%
Give the step counts on your mobile devices	39.0%	22.8%	32.4%	3.1%
Wear an activity tracker provided by CBS	20.2%	20.0%	48.3%	3.0%
Take pictures of receipts or upload digital receipts	13.8%	19.3%	56.0%	2.5%

Between 2019 and 2024, Statistics Netherlands conducted three experimental studies on the hypothetical willingness of the general population to participate in smart surveys (Struminskaya et al. 2020, 2021, Schouten et al. 2025). These studies were broader in scope, but in all of them, questions were asked after the

willingness to employ location tracking in the context of travel. The most extensive study was a perception survey conducted within the ESTAT-funded project Smart Survey Implementation. Table 8 shows the consent rates as measured. In the survey, respondents were asked for hypothetical willingness to participate in various surveys supported by smart features. In a second survey, respondents were invited to perform four smart data collections. Providing a GPS location was on the four. The consent rates allowed us to position location tracking within the wider range of smart features. Location-tracking consent rates turned out to be somewhere in the middle. Around 38% rejected the option. The relatively high proportion of 'maybe' answers, around a quarter of the respondents, marks the importance of context.

Based on the various consent studies, we concluded that location tracking is perceived as one of the most sensitive smart features. Furthermore, and perhaps related to sensitivity, the type of survey institute (academic, government, or market research) has a clear impact on willingness. Academic stakeholders seem to be trusted the most. The explicit mention of privacy-by-design data security measures does not increase willingness. Respondents do not seem to understand or believe the measures to be effective. However, respondents do understand and second the logic of using tracking in the context of travel. In the two hypothetical-actual willingness studies, respondents often remarked they did not believe it would be statistically useful to have just a few location measurements. Subsequently, they refused. All in all, it seems it is especially the logic of location tracking that should convince respondents. Explanations of how data are handled are mandatory, but do not help much in gaining trust.

AVA18 and AVA22 provided opportunities to study willingness directly. In Tables 9 and 10, we present the registration rates for AVA18 and AVA22, respectively. Larger incentives lead to higher registration rates, as may be expected. However, the increase already seems to converge at 20 euros. The concurrent design worked best when offering both options right from the start in terms of registration rates, but at the cost of the app options, and, hence, at the cost of location tracking data. We concluded that there is potential to reach response rates that resemble the regular travel survey, even when asking seven days instead of a single day. We also concluded that willingness is sensitive to how the study is introduced and presented.

Table 9: Registration percentages for AVA18

	Sample		Incentive experiment		
	previous respondents	Fresh sample	5 + 5 + 5 euros	5 + 0 + 10 euros	5 + 0 + 20 euros
Registrations	422	252	191	231	252
Percentage	44,4%	26,5%	30,1%	36,4%	39,7%

Table 10: Registration numbers and rates split against timing of the online alternative in AVA22. Standard errors are given in brackets.

	Direct		Reminder 1		Reminder 2	
	Count	Rate (SE)	Count	Rate (SE)	Count	Rate (SE)
App	90	10.8% (1.1%)	103	12.3% (1.2%)	99	12.0% (1.1%)
Questionnaire	83	10.0% (1.0%)	60	7.2% (0.7%)	35	4.2% (0.7%)
Combined	168	20.1% (1.4%)	157	18.8% (1.9%)	134	16.2% (1.3%)

3.2.4 Respondent user interface – user experience (UI-UX)

In a travel survey app, active participation and completion are key. In a travel survey app, active participation means checking and, if needed correcting, correcting stop-track segmentations and supplementing missing data. The main research question we asked ourselves was whether the average respondent is able and motivated to check, correct, and supplement segmentations and predictions. Given that tracking technology leads to missing data and segmentations and predictions are subject to error, the follow-up question is about the right balance in asking respondents to assist. Evaluations on respondent behavior could only be done for AVA22. AVA18 did not collect data on navigation and answering behavior in the app. Remmerswaal et al. (2024, 2025) investigated the AVA22 in-app behavior in detail. Giacobbe (2024) performed an analysis of the accuracy of respondent edits.

Before turning to in-app behaviour, we needed to evaluate the drop-out. Drop-out may be an indirect sign of insufficient motivation but may also be the result of a weak performance of the location tracking. The AVA22 study was subject to a considerable amount of drop-out. Two patterns could be observed: Drop-out shortly after registering the app and completing the short introduction questionnaire and gradual drop-out. The almost instantaneous drop-out without any follow-up activity has been observed as well in other types of smart surveys. The conjecture is that respondents scan the app and browse through the app main screens and decide not to continue. The gradual drop-out is conjectured to be mostly due to insufficient (perceived) technical performance. Figure 5 shows the drop-out for the single-day and the seven-day subsamples. The single-day respondents were encouraged to continue after the first day. We conclude that

the gradual drop-out may be reduced by improving the location tracking performance, but that direct drop-out is probably hard to battle.

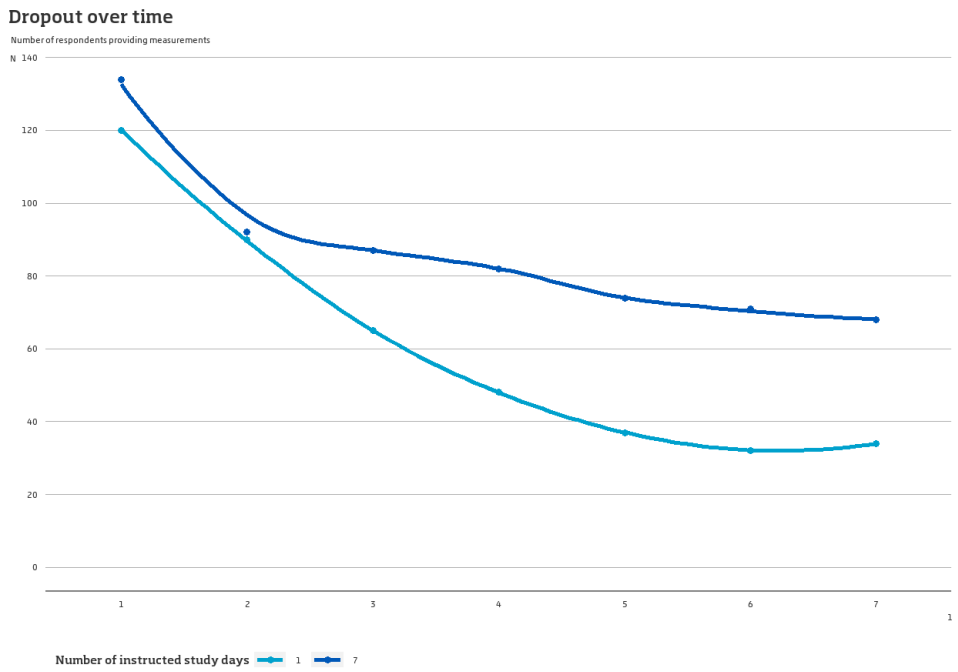

Figure 5: Drop-out rates as a function of time in-study for the single-day and seven-day samples. The X-axis shows the day in-study, running from one to seven days.

Next, we turned to the in-app behaviour of respondents who did not drop out directly after registration. We looked at various proxy indicators of respondent motivation and ability to interact with the app: time spent in the app, use of editing and labeling options, and in-app session characteristics throughout the reporting period.

Table 11 shows the amount of time respondents spent in the app at the first session and at later sessions. The first sessions are longer on average. Follow-up sessions are, on average, relatively short, indicating a low burden. Differences between conditions are too small to detect.

Table 11: Average time spent in-app in minutes per condition for the intro session and further sessions. Standard errors between brackets. AVA22

Condition	Intro session (in min)	Follow-up sessions (in min)
Full editing and one-day	3.9 (1.7)	1.1 (2.0)
Full editing and seven-days	4.3 (2.0)	1.6 (2.0)
Limited editing and one-day	3.7 (2.3)	1.1 (1.9)
Limited editing and seven-days	4.1 (2.3)	1.0 (1.6)

Table 12 provides insights into the proportion of respondents employing the different editing tasks. All tasks are performed by respondents, and they continue to do so throughout the week.

Table 12: Several indicators on in-app activity split by study duration and by editing conditions. AVA22

Indicators	One-day		Seven-days	
	Full editing (n= 61)	Limited editing (n= 72)	Full editing (n=84)	Limited editing (n=76)
Checked stop-tracks	100%	93%	90%	91%
Deleted >0 stop-tracks	79%	85%	76%	79%
Labeled stop-track	67%	63%	67%	60%
Added >0 stop-tracks	30%	Not applicable	50%	Not applicable

Finally, Table 13 compares in-app activity for the seven-day group for the first day and the later days. The first day shows more sessions and longer in-app time, which in part can be attributed to browsing options and scanning how the app needs to be used. The number of sessions and the in-app time stay at relatively high levels throughout the remainder of the reporting period. The main conclusion is that respondents employ all potential editing options they have available; they add and delete events, they change start and end times, and they label events. Furthermore, once they do, they tend to do so throughout the entire study period/ Similarly, respondents who are relatively inactive are so for the entire study period. Hence, respondents seem consistent, and relatively early on in data collection, it can be determined whether a respondent will be (very) active or not.

Table 13: Average number of sessions and in-app time per condition for first full day and all days. Standard errors between brackets. AVA22

	First full day of 7-day		Mean day of 7-day	
	Full	Limited	Full	Limited
Time in-app (in sec)	70 (48)	39 (27)	59 (51)	41 (25)
# In-app sessions	4.7 (2.7)	4.3 (2.3)	3.9 (2.6)	3.8 (2.2)
# hours location data	16.1 (7.5)	16.8 (6.6)	13.0 (8.0)	13.0 (8.0)
# tracks	3.6 (2.8)	5.1 (4.3)	3.0 (2.8)	4.5 (5.2)
# stops	3.3 (2.7)	5.1 (4.2)	2.8 (2.8)	(5.5)

4. New developments

Since 2017, technology and survey climate have changed. The literature on travel surveys employing location tracking and machine learning has grown substantially. We describe new challenges that emerged at Statistics Netherlands. We look briefly into field studies conducted internationally. Given the observed new challenges and experiences, we end with the open questions that we believe are most pressing.

4.1 New challenges

We identified four new challenges during the development and testing of the app:

- Maintenance and governance: We have learned that smartphone tracking technology requires constant maintenance. New versions of operating systems, new versions of mobile app development kits, new smartphone models and changing app store policies all imply a continuous revision of the underlying mobile-app tracking technology and the user interface of the app. Indirectly, not only the frontend of the app but also the backend of the app in-house may be affected. There are consequences as well for the preprocessing of location data and for machine learning models that provide predictions. On a less frequent basis, the app user interface must be modernized and aligned with general design principles. All in all, a governance strategy is needed that guarantees that changes in technology and changes in methodology can be incorporated in statistical processes without a severe impact on the comparability of statistics over time.
- Accurate training and validation data: When refining travel mode prediction and developing travel purpose prediction, it became apparent that accurate training and validation data are a requirement. We did not obtain such data in field tests. This made it hard for respondents to provide accurate labels. For this reason, we resorted to small-scale data collections. We directed participants to follow pre-specified travel scenarios and to keep a detailed diary. The resulting data (UU24) was too small to train models but large enough to evaluate models trained on complete field test data. We noted deficiencies in the trained models. To resolve the deficiencies, more data is needed. In AVA22, a sample of regular travel survey respondents was asked to use the app and fill in the regular online survey. However, only a relatively small proportion provided both forms of data. We speculate that instructions were not fully understood and/or the task was too burdensome. We conclude that dedicated data collections are needed to train and validate machine learning models.
- The sequence of segmentation and predictions: The natural sequence of steps is stop-track segmentation → travel mode prediction → purpose prediction. However, none of these steps is perfect. Even with complete location data, stop-track segmentation may miss shorter stops or tracks, making predictions hard or impossible. Given that gaps in the location data cannot be avoided, some time periods inherently will miss details on stops and tracks. Perhaps the most complicated factor is the distinction between stops that have a purpose

and stops that are transfers between travel modes. Transfers can be too short to detect, leading to multi-modal tracks. Multi-modal tracks are a complication that has not been studied extensively yet. Transfers can also be long and be confused with stops that have a purpose. Furthermore, no change of travel modes is possible without a minimal amount of walking. The potential to accurately segment and predict depends strongly on the length of the reporting period. The ideal sequence of steps, similar to the Expectation-Maximization (EM) algorithm, may be iterative. The three steps could be repeated until convergence is reached. Stop-track segmentation is revisited based on travel modes and stop purpose predictions. Travel mode and stop purpose predictions are performed again with the revised segmentation. Alternatively, stop detection may be implemented such that too many stops are found, and subsequent iterations are designed to cluster tracks. While methodologically feasible, such a cyclical approach severely complicates the mobile app code and the handling of data in the backend database.

- Data integration: Data integration of self-reports and edited and labelled location data is a challenge. The two types of data are very different in nature, each with its strengths and weaknesses. The natural tendency is to make the location-tracking-based statistics look like self-reports. The data set would be a series of tracks with begin and end times split into trips with begin and end times. The trips would get a travel mode and the tracks a purpose. However, all these variables may be the result of estimation/prediction. They must be flagged as predicted and accompanied by measures of uncertainty. Furthermore, in the transformation, all information about the exact trajectories is removed. It raises the question whether the location-tracking data becomes the 'victim' of a data model that was defined based on self-report. We could equally well map self-reports to trajectory data. Finally, multiple days of tracking imply that respondents appear multiple times. As the within-respondent variation in daily travel is smaller than the between-respondent variation, see Figure 3, the analysis needs to account for clustering. Such clustering effects obviously demand more advanced analysis strategies and reporting of standard errors.

4.2 International developments

A wide range of countries have tested app-based approaches for mobility research for over a decade now. National statistical agencies, other agencies, institutes, private companies, and academic teams in, for example, Switzerland, the Netherlands, Norway, Belgium, Australia, the USA, Israel, Singapore, New Zealand, and Canada have developed and trialed travel-survey apps. Table A.1 in Appendix A provides an overview of recent smartphone-based travel apps with passive location tracking (conducted in 2012–2022) that include at least 50 participants. This table is taken from the paper by Remmerswaal et al. (2025). It provides information on the project, study year, the country of data collection, the invitation mode, the type of sample, the incentive, the sample and response sizes, the study/tracking duration, whether the app is open source, which platforms are supported by the app, whether the app contains prediction elements, and whether a validation/annotation diary is included.

We summarize these characteristics. In addition, we aim to address the following questions: which other countries are conducting similar app-based mobility studies, how do their apps function and what are their key design features, whether these apps have been systematically tested, and what their main findings are. Ultimately, the goal is to synthesize these insights to understand what lessons can be learned and applied to future mobility research projects.

Different invitation methods were employed to recruit participants. The most common mode was postal mail, often used alone or in combination with other approaches such as crowdsourcing. Additional recruitment channels included email invitations to university populations, previous survey participants, and online convenience samples. Some studies also relied on interviewers or combined postal mail with traditional travel diary invitations.

Most studies relied on probability samples (PS), while several used non-probability samples (NPS) or a combination of both (PS + NPS). One study applied eligibility criteria in addition to probability sampling. Overall, probability-based recruitment predominated, though mixed and non-probability methods were also represented.

Incentives varied considerably in both type and value. Monetary rewards ranged from small conditional payments of about €5–€20 to larger fixed amounts exceeding €100. Common forms included gift cards, hourly payments, and lottery entries. Some studies provided no incentives, and others did not specify the reward type. On the whole, incentive strategies were diverse, reflecting differing budgets and participant engagement goals.

Sample sizes showed substantial variation, ranging from fewer than 200 participants to over 130,000. Most studies included between 600 and 20,000 participants, with a few exceptionally large-scale efforts. Several studies did not report exact sample sizes, and some combined volunteer and campaign-based recruitment, indicating diverse scales and participation frameworks.

The number of registered participants ranged from under 100 to over 15,000, with reported registration rates between 0.3% and 35.?%. Actual data submission rates were typically lower, ranging from 0.2% to 28.5%, and many studies did not clearly distinguish between registration and data contribution. Despite generally modest engagement levels, a few studies achieved notably high submission rates, suggesting successful participant retention strategies.

4.3 Open questions

Given the AVA18 and AVA22 results and given the international developments, the two most prominent methodological questions are:

- How to recruit and motivate respondents without overloading them with options and instructions?
- How to mix machine learning predictions and respondent interaction?

While the two studies, AVA18 and AVA22, had a strong focus on the evaluation of recruitment and motivation strategies, it is not yet clear what may work best. In part, the inconclusive findings come from the ambiguity in the number of reporting days. Whereas the current request is for one day of reporting, longer periods seem more natural for respondents when using an app. The concurrent mix of online and app-based surveying requires more thinking. On the other hand, findings are mixed because we have not yet been able to effectively instruct respondents on how to improve the technical performance of tracking. A considerable proportion of the respondents experienced gaps in location data. These gaps hampered stop-track segmentation. Effective communication strategies when tracking accuracy is insufficient require more UI-UX (user interface – user experience) consideration. Apart from these design choices, nonrespondents may simply not be willing to be tracked, despite guarantees about data security.

In both AVA18 and AVA2, respondents had to give the purpose of each stop and the travel mode(s) of each track. We have, thus, not been able to test and evaluate to what extent respondents are able and motivated to check predictions and to make changes if needed. The accuracy of prediction models is promising. But we have also learned in ESTAT-project Smart Survey Implementation[5] that respondent expectations are high. Respondents compare performance to other tools that they are familiar with. These tools do not have statistical objectives but seem to work flawlessly. Expressing uncertainty about predictions will be a challenge for the user interface.

5. The future of smart travel surveys

We set out to evaluate opportunities and challenges we saw at the start of the Statistics Netherlands smart travel survey project in 2017. The opportunities seemed greatest in improving measurement data quality by reducing respondent burden and incorporating contextual data through machine learning models. The reduced respondent burden and modernization of travel survey data collection could also be beneficial to the general population's willingness. However, we were aware of the large surplus of information in location data and the impact this may have on perceived intrusiveness. During the past eight years, we have learned that the opportunities still hold but do need more research. During these years, some challenges were resolved, but new challenges also presented themselves. The biggest challenge is perhaps the integration of smart and non-smart modes of data collection. We have found that smart data collection partly lives up to the promise in terms of the improvement in measurement quality, but we have also found that

[5] . We refer to https://cros.ec.europa.eu/dashboard/trusted-smart-surveys for the project deliverables

recruiting respondents for smart surveys is challenging. There may be a continued need for traditional diaries for those types of respondents who are not able or willing to participate in smart surveys.

In Table 14, we cross the five opportunities (full trajectories, stop-track segmentation, travel mode prediction, stop/travel purpose prediction and longer reporting periods) we identified against the four existing challenges (technical/location tracking comparability, missing data, general population willingness, and respondent UI-UX) and against the four new challenges (data integration, accurate training data, governance and maintenance, and prediction steps). We evaluate each combination as (still) promising, (still) challenging, and open/unresolved. We start by looking at the 2017 challenges, we then move to the new challenges and we end with the viewpoint of the opportunities.

Table 14: Global evaluation of opportunities and challenges. Green = promising, Orange = requires attention, Grey = open/unresolved.

	Opportunities				
	Trajectories	Segmentation	Travel mode	Stop purpose	Longer period
2017 challenges					
Location tracking					
Gaps					
Willingness					
UI-UX					
New challenges					
Data integration					
Training data					
Governance					
Prediction steps					

All 2017 challenges have been evaluated at least to some extent. However, open questions remain.

Location tracking comparability: Although there are subtle differences between Android and iOS devices, we see full technical comparability as feasible. Trajectories are similar, deviations in stop-track segmentation to 'truth data' exist, but do not depend strongly on the device, real-time prediction of travel modes and stop purpose is possible in all devices, and performances do not differ for longer tracking periods. The only exception to this is specific mobile device brands (HUAWEI and XIAOMI) for which the use of native mobile app libraries is not allowed. These brands give travel data that has very low data quality and are incomparable to other devices. Location data gaps, which we see as a challenge in itself, are in part a technical issue that is dependent on the type of platform and the type of device. So, including missing data does pose a threat to location tracking comparability.

Gaps: Location tracking data gaps pose a great challenge to the recording of trajectories. Consequently, they hamper adequate segmentation into stops, tracks,

and further into trips within tracks. Obviously, predictions cannot be made without data. However, smaller gaps of at most ten minutes can be overcome in mode and purpose prediction. The association between travel mode, stop type, and gap occurrence is also modest. A further positive finding is that we found no dependence between the prevalence of gaps and longer reporting periods. Location tracking is reset and re-initiated at the start of each reporting day. The frequency and size of gaps remain the same as time in-study progresses.

Respondent willingness: Respondent willingness remains a key issue. It is affected by the desire to record full and detailed trajectories. In AVA22, we have learned that the majority of respondents is willing to assist in stop-track segmentation and the labeling of stops and tracks, provided they are recruited in the app-study. Whether respondents are willing to assist in checking and, if needed, correcting travel mode and travel purpose predictions is as yet unclear. We conjecture that the majority of respondents will be sufficiently motivated to do so, given that, without predictions, they need to provide labels as well. However, it is likely that some of the respondents will 'satisfice' in labeling and do as little as is allowed by the app to complete the study. We have learned that longer reporting periods are not a barrier to initial willingness. By itself, respondents seem to find it natural to install and use an app for multiple days.

UI-UX: Apart from motivation, an effective UI-UX is a requirement for accurate measurement data. Respondents must be able and competent to interact with the app. Again, it is unclear whether respondents are sufficiently able to interpret predictions and revise if needed. The AVA22 study showed respondents can understand and interpret trajectories and the corresponding segmentations that are presented to them. They can also interact with the app for the decomposition into stops and tracks. We see that respondents have a steep learning curve. The in-app time drops after the first day and remains relatively stable throughout the remainder of the reporting period. This means that longer periods are not problematic for the UI-UX.

As may be expected, the new challenges also pose new questions.

Data integration: We see data integration as the biggest challenge in all aspects, from aligning trajectories to allowing for longer reporting periods. The yield from location tracking in terms of trajectories depends strongly on whether and how alternative manual data entry is organized. Manual data entry does not allow for a derivation of full trajectories. It is also prone to underreporting of shorter travels. Incomparability may result that in confounding with selection into tracking and manual data entry. Manual data entry does not imply segmentation by the respondent. While it does not exclude prediction of travel modes and purposes, some of the important features such as duration, distance, and time of the day are available, it will imply a different strategy towards training machine learning models. Furthermore, it must be questioned whether in manual data entry it is still natural to add prediction from a UI-UX perspective. On top of this , there is friction between the current single-day reporting and the more natural, longer reporting period in an app. Longer reporting periods are beneficial for imputation and

predictions. Hence, data integration requires attention to all opportunities that tracking offers.

Training data: Accurate training data is a must for stop-track segmentation decision rules and for building machine learning models that can be used for real-time predictions. In the UU24 small-scale study, we have tried to collect such data. It is possible to get full trajectories and segmentations both in diary data and location tracking data for extended reporting periods. However, the utility of the training data depends strongly on the location tracking performance of the app. This holds especially for the prevalence of gaps. Training data must be free of gaps. We expect, however, that accurate training data can be collected through separate dedicated data collections over longer periods of time.

Governance: Governance and maintenance of travel apps are demanding. The user interface requires frequent modernization, but we are confident this is feasible with the current expertise. However, the accuracy of travel trajectories deteriorates when in-app sensor technology is neglected. New operating versions and new smartphone models may require a full renewed look at how tracking is started and maintained. Inadequate maintenance may result in lower location data resolution and more gaps. It may also make longer reporting periods impossible because respondents need to check and supplement more and more when time-in study progresses. For the segmentation and predictions, we see relatively minor issues in maintenance. Machine models need updating in terms of libraries and dependencies. However, we expect that time dynamics in how and why people travel are modest and can be accounted for.

Prediction steps: The last new challenge, advanced, cyclical segmentation and prediction, is the least urgent and most unknown of challenges. In AVA18 and AVA22, we speculated on the use of open-source trip planners for public transport and perhaps also private transport. Embedding such facilities requires new infrastructure. It is unclear whether this can be done in real-time and what the role of respondents is. We conjecture that segmentation and predictions may benefit from more advanced and iterative decision rules, especially for longer reporting periods. However, also in the literature, this is still an open area.

Let us return to the opportunities. On the positive side, we conclude that longer reporting periods are feasible. The only barrier is data integration when respondents differ in how data is being collected. For the other opportunities, it starts with the trajectories. Segmentations depend on accurate trajectories. Stop and track prediction depends on accurate segmentations. Here, we see multiple barriers, both old and new. Gaps occur frequently in location tracking data. Willingness to perform location tracking is vulnerable. Consequently, the integration of questionnaire data and location tracking data is an issue. And making sure that location tracking data remains comparable in time requires constant maintenance and adequate governance protocols.
Perhaps the most pressing issue for the coming years is the effective recruitment and motivation of respondents. This issue is interlinked with data integration; concurrent manual and tracking options complicate both recruitment and data alignment. A decision needs to be made if and how to offer data collection

options. We advocate for full integration of options into the application frontend and backend. Ideally, the decision on how to provide data is not made at the onset of the study, but respondents can do this while participating to overcome communication challenges in recruitment. To the respondents, the application should be presented as one coherent user interface. A direct consequence of this recommendation is a fixed initial reporting period. We recommend a longer initial reporting period. During data collection, respondents may opt for a shorter reporting period. An integrated and coherent interface will not alter the perceived intrusiveness of location tracking. More reserach is needed here. Again, it seems favorable not to bother respondents at the onset. It seems better to let them get to know how and what is collected gradually so that they can make informed choices. A choice in how data are collected may have consequences for the comparability of statistics. This would occur when different subpopulations, e.g., younger versus older respondents, have different preferences. Options must be weighed carefully against the impact on the granularity and accuracy of travel data.

One way to embed smart surveys in data collection systems more generally is to move towards a longer relationship between the respondent and the official statistics agency. Some of our experiments have shown that respondents can be successfully recruited, provided sufficient resources are spent to convince respondents to install the app. In the future, such high up-front costs in fieldwork may be compensated by keeping respondents in a study, and for example, asking them to provide panel data for longer periods of time.

As a recommendation, we point out the importance of collaboration with other (national) statistical institutes. Much of the user interface, location tracking technology, and machine learning methodology is agnostic of country or region.

Acknowledgements

Several other colleagues at Statistics Netherlands and Utrecht University have been involved. In particular, we like to thank Rob Warmerdam, Victor Verstappen Ole Mussmann, Tom Oerlemans, Mike Vollebregt, Jesper van Thor and Coen van Heukelingen. The AVA18 and AVA22 field tests have been partially funded by Rijkswaterstaat. We like to thank Remko Smit for his support. We like to thank Maaike Kompier for her review comments and suggestions.

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands.

References

Allström, A., Kristoffersson, I., & Susilo, Y. (2017). Smartphone based travel diary collection: Experiences from a field trial in Stockholm. Transportation Research Procedia, 26, 32–38. https://doi.org/10.1016/j.trpro.2017.07.006.

Boer, Q., Y. Gootzen, J. Klingwort, D. Remmerswaal, P. Lugtig (2026): Machine-learning based transport mode prediction in a smartphone-based travel and mobility survey. Discussion paper, Statistics Netherlands.

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future Mobility Survey: Experience in Developing a Smartphone-Based Travel Survey in Singapore. Transportation Research Record: Journal of the Transportation Research Board, 2354(1), 59–67. https://doi.org/10.3141/2354-07.

Faghih Imani, A., Harding, C., Srikukenthiran, S., Miller, E. J., & Nurul Habib, K. (2020). Lessons from a Large-Scale Experiment on the Use of Smartphone Apps to Collect Travel Diary Data: The "City Logger" for the Greater Golden Horseshoe Area. Transportation Research Record: Journal of the Transportation Research Board, 2674(7), 299–311. https://doi.org/10.1177/0361198120921860.

Fourie, J., J. Klingwort & Y. Gootzen (2025): Rule-based transport mode classification in a smart travel and mobility survey. Discussion paper, Statistics Netherlands.

Flake, L., Lee, M., Hathaway, K., & Greene, E. (2017). Use of Smartphone Panels for Viable and Cost-Effective GPS Data Collection for Small and Medium Planning Agencies. Transportation Research Record: Journal of the Transportation Research Board, 2643(1), 160–165. https://doi.org/10.3141/2643-17.

Geurs, K. T., Thomas, T., Bijlsma, M., & Douhou, S. (2015). Automatic Trip and Mode Detection with Move Smarter: First Results from the Dutch Mobile Mobility Panel. Transportation Research Procedia, 11, 247–262. https://doi.org/10.1016/j.trpro.2015.12.022.

Giacobbe, G. (2024), Unveiling User Dynamics, Examining User-Initiated Changes and Socio-Demographic Influences in app-based Travel Survey Data, Master Thesis, Leiden University.

Gootzen, Y., J. Klingwort & B. Schouten (2024): Validating a smart survey travel app: Survey response versus algorithms. Mobile Apps and Sensors in Surveys (MASS), 6-7 March, Washington, DC, USA.

Gootzen, Y., Klingwort, J., Schouten, B. (2025), Data quality aspects for location-tracking in smart travel and mobility surveys, Discussion paper, Statistics Netherlands.

Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., & Crane, M. (2015). A Web-Based Diary and Companion Smartphone app for Travel/Activity

Surveys. Transportation Research Procedia, 11, 297–310. https://doi.org/10.1016/j.trpro.2015.12.026.

Gillis, D., Lopez, A. J., & Gautama, S. (2023). An Evaluation of Smartphone Tracking for Travel Behavior Studies. ISPRS International Journal of Geo-Information, 12(8), 335. https://doi.org/10.3390/ijgi12080335.

Hong, S., Zhao, F., Livshits, V., Gershenfeld, S., Santos, J., & Ben-Akiva, M. (2021). Insights on data quality from a large-scale application of smartphone-based travel survey technology in the Phoenix metropolitan area, Arizona, USA. Transportation Research Part A: Policy and Practice, 154, 413–429. https://doi.org/10.1016/j.tra.2021.10.002.

Killaars, L., Mussmann, O., Schouten, B. (2020), Stop and track detection in travel surveys, Discussion paper, Statistics Netherlands.

Klingwort, J., Y. Gootzen, and J. Fourie (2025a). Development and performance of a transport mode classification algorithm for smart surveys. Technical report. Smart Survey Implementation (SSI). Report number: WP3: Developing Smart Data Microservices. DOI: 10.13140/RG.2.2.30160.83203.

Klingwort, J., Y. Gootzen, M. Kompier & V. Toepoel (2025b): How smart are smart travel surveys? Evaluating trip segmentation, travel motive, and travel mode predictions. Mobile Apps and Sensors in Surveys (MASS), 4-5 June, London, UK.

Klingwort, J., Y. Gootzen, D. Remmerswaal & B. Schouten (2025c): Algorithms versus survey response: Comparing a smart travel and mobility survey with a web diary. Transportation Research Interdisciplinary Perspectives, 31: 101436.

Lawson, C.T., Krans, E., Rentz, E., Lynch, J. (2023), Emerging trends in household travel survey programs, Social Sciences and Humanities Open, 7, 100466.

Lugtig, P., Roth, K., Schouten, B. (2022), Nonresponse analysis in a longitudinal smartphone-based travel study, Survey Research Methods, 16 (1), 17 – 23.

Lynch, J., Dumont, J., Greene, E., & Ehrlich, J. (2019). Use of a Smartphone GPS Application for Recurrent Travel Behavior Data Collection. Transportation Research Record: Journal of the Transportation Research Board, 2673(7), 89–98. https://doi.org/10.1177/0361198119848708.

McCool, D., Lugtig, P., Mussmann, O., & Schouten, B. (2021). An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges. Journal of Official Statistics, 37(1), 149–170. https://doi.org/10.2478/jos-2021-0007.

McCool, D.M., Lugtig, P., Schouten, B. (2024), Maximum interpolable gap length in missing smartphone-based GPS mobility data, Transportation, 51, 297 – 327.

McCool, D., Schouten, B., Lugtig, P. (2026), Field Notes from the Travel App Frontier, PhD dissertation, Utrecht University, Faculty of Social Sciences and Behavioural Sciences.

Molloy, J., Castro, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., & Axhausen, K. W. (2022). The MOBIS dataset: A large GPS dataset of mobility behaviour in Switzerland. Transportation. https://doi.org/10.1007/s11116-022-10299-4.

Nahmias-Biran, B., Han, Y., Bekhor, S., Zhao, F., Zegras, C., & Ben-Akiva, M. (2018). Enriching Activity-Based Models using Smartphone-Based Travel Surveys. Transportation Research Record: Journal of the Transportation Research Board, 2672(42), 280–291. https://doi.org/10.1177/0361198118798475.

Patterson, Z., & Fitzsimmons, K. (2016). DataMobile: Smartphone Travel Survey Experiment. Transportation Research Record: Journal of the Transportation Research Board, 2594(1), 35–43. https://doi.org/10.3141/2594-07.

Remmerswaal, D., Schouten, B., Bakker, J., van den Heuvel, J., Klingwort, J. (2024), A smart Travel Survey. What is the role of the respondent? Discussion paper, Statistics Netherlands.

Remmerswaal, D., Lugtig, P., Schouten, B, Struminskaya, B. (2025). The effects of study duration on nonresponse and measurement quality in a smartphone app-based travel diary. Survey Research Methods, 19 (4), 389–407.

Roddis, S., Winter, S., Zhao, F., & Kutadinata, R. (2019). Respondent preferences in travel survey design: An initial comparison of narrative, structured and technology-based travel survey instruments. Travel Behaviour and Society, 16, 1–12. https://doi.org/10.1016/j.tbs.2019.03.003.

Safi, H., Assemi, B., Mesbah, M., Ferreira, L., & Hickman, M. (2015). Design and Implementation of a Smartphone-Based Travel Survey. Transportation Research Record: Journal of the Transportation Research Board, 2526(1), 99–107. https://doi.org/10.3141/2526-11.

Schouten, B., Lunardelli, I., Perez, M., D'Amen, B., Van den Heuvel, J., Nuccitelli, A., Lorè, B., Struminskaya, B., Zgonec, M. (submitted 2024), How does the general population think about smart surveys? To appera in the Journal of Official Statistics.

Schouten, B., Lugtig, P., Luiten, A. (2025). Can smart surveys have a positive business case? An evaluation based on three case studies, JOS special issue data integration, 41 (2), 547 – 568.

Schouten, B., Remmerswaal , D., Elevelt, A., Groot, J. de, Klingwort, J., Schijvenaars, T., Schulte, M., Vollebregt, M. (2024), A smart Travel Survey. Results of a push-to-smart field experiment in the Netherlands. Discussion paper, Statistics Netherlands.

Shankari, K., Bouzaghrane, M. A., Maurer, S. M., Waddell, P., Culler, D. E., & Katz, R. H. (2018). e-mission: An Open-Source, Smartphone Platform for Collecting Human Travel Data. Transportation Research Record: Journal of the Transportation Research Board, 2672(42), 1–12. https://doi.org/10.1177/0361198118770167

Smeets, L., Lugtig, P., Schouten, B. (2019), Automatic travel mode prediction in a national travel survey. Discussion paper, Statistics Netherlands.

Storesund Hesjevoll, I., Fyhri, A., & Ciccone, A. (2021). App-based automatic collection of travel behaviour: A field study comparison with self-reported behaviour. Transportation Research Interdisciplinary Perspectives, 12, 100501. https://doi.org/10.1016/j.trip.2021.100501.

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, a., Schouten, B. (2020), Mechanisms of willingness to collect smartphone sensor data and longitudinal consent: Evidence from the general population in the Netherlands, Public Opinion Quarterly 84 (3), 725 – 759.

Struminskaya, B., Lugtig, P., Schouten, B., Dolmans, R., Giesen, D. (2021), Sharing of smartphone sensor-collected data: Willingness, participation, and non-participation bias, Public Opinion Quarterly, 85 (1), 423-462.

Svaboe, G. B. A., Tørset, T., & Lohne, J. (2021). Recruitment Strategies in App-Based Travel Surveys: Methodological Explorations. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3988501.

Winkler, C., Meister, A., & Axhausen, K. W. (2024). The TimeUse+ data set: 4 weeks of time use and expenditure data based on GPS tracks. Transportation. https://doi.org/10.1007/s11116-024-10517-1.

Xiao, G., Z. Juan, and C. Zhang (2016). Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. In: Transportation Research Part C: Emerging Technologies 71, pp. 447–463. DOI: 10.1016/j.trc.2016.08.008.

Yazdizadeh, A., Patterson, Z., & Farooq, B. (2019). An automated approach from GPS traces to complete trip information. International Journal of Transportation Science and Technology, 8(1), 82–100. https://doi.org/10.1016/j.ijtst.2018.08.003.

Zahroh, S., Gootzen, Y., Klingwort, J., Lugtig, P., Schouten, B. (2025a), Predicting trip purpose in a smartphone-based travel survey. International Journal of the IAOS 41 (4), 985–995.

Zahroh, S., Lugtig, P., Gootzen, Y., Klingwort, J., Schouten, B. (2025b), Predicting trip purpose in a smartphone-based travel survey. Discussion paper, Statistics Netherlands.

Zhao, F., Pereira, F. C., Ball, R., Kim, Y., Han, Y., Zegras, C., & Ben-Akiva, M. (2015). Exploratory Analysis of a Smartphone-Based Travel Survey in Singapore. Transportation Research Record: Journal of the Transportation Research Board, 2494(1), 45–56.

Appendix A - Overview of (inter)national field studies

Table A.1. Smartphone-based travel apps with passive location tracking (conducted in 2012-2022) with at least 50 participants, from Remmerswaahl et al. (2025).

Authors year	Project and/or app name	Country, Year of data collection	Invitation mode	type of sample	Incentive	N invited	N (% for PS) app registrants	N (%) participants sending location data	Tracking days/study duration	Open source	Cross-platform (android and iOS)	Automatic mode prediction
(Molloy et al., 2022)	MOBIS, MotionTag	Switzerland, 2019	Postal mail	PS + eligibility criteria	Min. 100 CHF	90090	5375 (5.9%)	4218 (4.6%)	8 weeks	No	Yes	Yes
(Winkler et al., 2024) Remmerswaal et al 2024 (in review)	TimeUse+, MotionTag	Switzerland, 2022	Postal mail	PS	50 CHF	63081	3733 (5.9%)	2742 (4.4%)	28 days	No	Yes	Yes
	CBS ODIN	Netherlands, 2022	Postal mail	PS	€10 giftcard	2544	315 (12%)	292	1 or 7 days	Yes	Both	No
(McCool et al., 2021)	CBS TABI	Netherlands, 2018	Postal mail	PS	€5 unconditional + €10 or €20 conditional	1896	674 (35.4%)	541 (28.5%)	7 days	Yes	Both	No
(Geurs et al., 2015)	Dutch Mobility Panel, mobidot	Netherlands, 2013-2015	LISS panel	LISS panel	€15 per hour	800	655	600	3 times 2 weeks in 3 years	No	Both	Yes (+ Stop recognition)
(Storesund Hegglevoll et al., 2021)	Sense DAT, mobidot	Norway, 2018	Previous survey participants	NPS	Lottery	675	486	No distinction	4 weeks	No	Both	No
(Svabo et al., 2021)	TravelVu	Norway, 2019	Postal mail + crowdsourcing	PS + NPS	Lottery	Unknown	1821	No distinction	1 day or 1 week	No	Both	Yes
(Allström et al., 2017)	MEILI	Sweden, 2015	Alongside traditional travel diary invitation	PS	No	130000	495 (0.4%)	293 (0.2%)	7 days	Yes	Android only	Purpose prediction
(Gillis et al., 2023)	GPSWAL, CONNECT	Belgium, 2016-2017	Postal mail PS + crowdsourcing	PS + NPS	unknown	7000 + volunteers	237	No distinction	7 days	No	Android only	yes
(Roddis et al., 2019)	Future Mobility Sensing	Australia, 2015	Crowdsourcing	NPS	$50 giftcard	Unknown	400	235	2 weeks	No	Both	No
(Greaves et al., 2015)	?	Australia, 2013	Crowdsourcing	NPS	$50 AUD	unknown	276	No distinction	7 days	No	Both	No

(Roddis et al., 2019)	Future Mobility Sensing	Australia, 2015	Crowdsourcing	NPS	$50 giftcard	Unknown	400	235	2 weeks	No	Both	No
(Greaves et al., 2015)	?	Australia, 2013	Crowdsourcing	NPS	$50 AUD	unknown	276	No distinction	7 days	No	Both	No
(Shankari et al., 2018)	E-mission	USA, 2016	Email, University population	PS	No	>60000	>150 (0.3%)	No distinction	unspecified	Yes	Both	yes
(Lynch et al., 2019)	rMove	USA, 2018	Postal mail	PS	Giftcard	20500	655 (3.2%)	493 (2.4%)	7 days	No	Both	no
(Flake et al., 2017)	rMove	USA, 2016	Email, Previous travel survey participants	PS	$20 giftcard	2639	411 (27%)	No distinction	3 days	No	Both	no
(Hong et al., 2021)	Future Mobility Survey	USA, 2016-2017	Postal mail PS + convenience sample	PS + NPS	Unknown	Unknown	15097	No distinction	2 days	No	Both	Yes
(Nahmias-Biran et al., 2018)	Future Mobility Survey	Israel 2016-2017	interviewers	PS	Unknown	38500	11928 (31%)	No distinction	2 days	No	Both	No
(Cottrill et al., 2013; Zhao et al., 2015)	Future Mobility Survey	Singapore, 2012-2013	Previous survey participants	NPS	SG$30 (USD$25)	10500	>1500 (~14%)	No distinction	2 weeks	No	Both	Yes + stop recognitio
(Safi et al., 2015)	ATLAS II	New Zealand, 2014	Crowdsourcing	NPS	Unknown	186	77	73	3 days	No	iPhone only	No
(Patterson & Fitzsimmons, 2016)	DataMobile	Canada, 2014	Email, University population	PS	Lottery	47606	892 (1.9%)	No distinction	2 weeks	No	Both	No
(Yazdizadeh et al., 2019)	MTL-Trajet	Canada, 2016-2017	Online convenience sample	NPS	Lottery	unknown	11433	7773	7 days	Open data	Both	No
(Faghih Imani et al., 2020)	City Logger	Canada, 2017	Crowdsourcing	NPS	Lottery	22804 emails + online campaign	2041	1550	5 days	No	Both	Yes (android only)

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information