# Validating a smart survey travel app: survey response versus algorithms

Yvonne Gootzen[1] Jonas Klingwort[2] Barry Schouten[3]

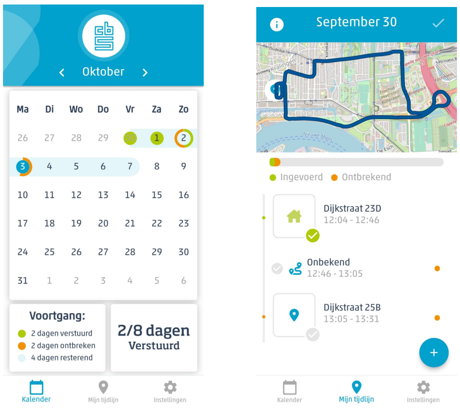
## 1 Introduction

The need for smart surveys in official statistics is increasing. This is caused by, for example, the fact that traditional diary surveys, such as time use, travel, or household budget surveys, are burdensome for respondents. Statistics Netherlands has taken the first step towards developing and potentially implementing a state-of-the-art smartphone app to measure travel and mobility behavior in the Netherlands. In 2022–2023, the app was tested in a large-scale field test in which different experimental conditions were used. The survey was sent to n=3,200 individuals from the general population and divided into three experimental study phases. This paper focuses on the first study phase (n=667) designed to validate the app data using survey responses. Therefore, respondents were asked to use the app for seven days and to report movements on one specified day during the same week in a web questionnaire.

## 2 The app and algorithms

The app collects GPS data on the respondents' travel and mobility behavior during a week in the Netherlands. The left panel of Figure 1 shows the digital app diary with the requested reporting period. The right panel shows the recorded GPS data, timestamp, and the classified periods to the respondent. These classified periods are algorithm-based clusters of the measured geo-locations classified into stationary (stop) or movement clusters (track). A relevant stop is when the respondent traveled/moved to a location with a purpose other than transport. A track separates two non-consecutive stops. A track can consist of multiple segments and transport mode switches separate segments. These classifications can be confirmed, edited, or deleted by the respondent. This paper compares this algorithm's performance (ALG 2022) to an algorithm without respondent interaction (ALG 2018). Both algorithms use time and radius as input parameters. The ALG18 uses the constant parameter set of 180s and 200m radius. If the measured geo-locations are for 180s in a 200m radius, these points are classified as stop, and consecutive stops are merged. If this condition is not met, the points are classified as track. The ALG22 follows the same concept but adjusts its parameters according to the accuracy of the measured GPS data. The output of both algorithms will be validated using the web questionnaire response.

[1] Statistics Netherlands & Eindhoven University of Technology, The Netherlands. yapm.gootzen@cbs.nl
[2] Statistics Netherlands, The Netherlands. j.klingwort@cbs.nl
[3] Statistics Netherlands & University Utrecht, The Netherlands. jg.schouten@cbs.nl



(a) App diary (b) GPS data and classifications

Figure 1: Example screenshots of the developed mobility app. The digital app diary (left) and the recorded GPS location data and state classifications (right) are shown.

# 3 Data

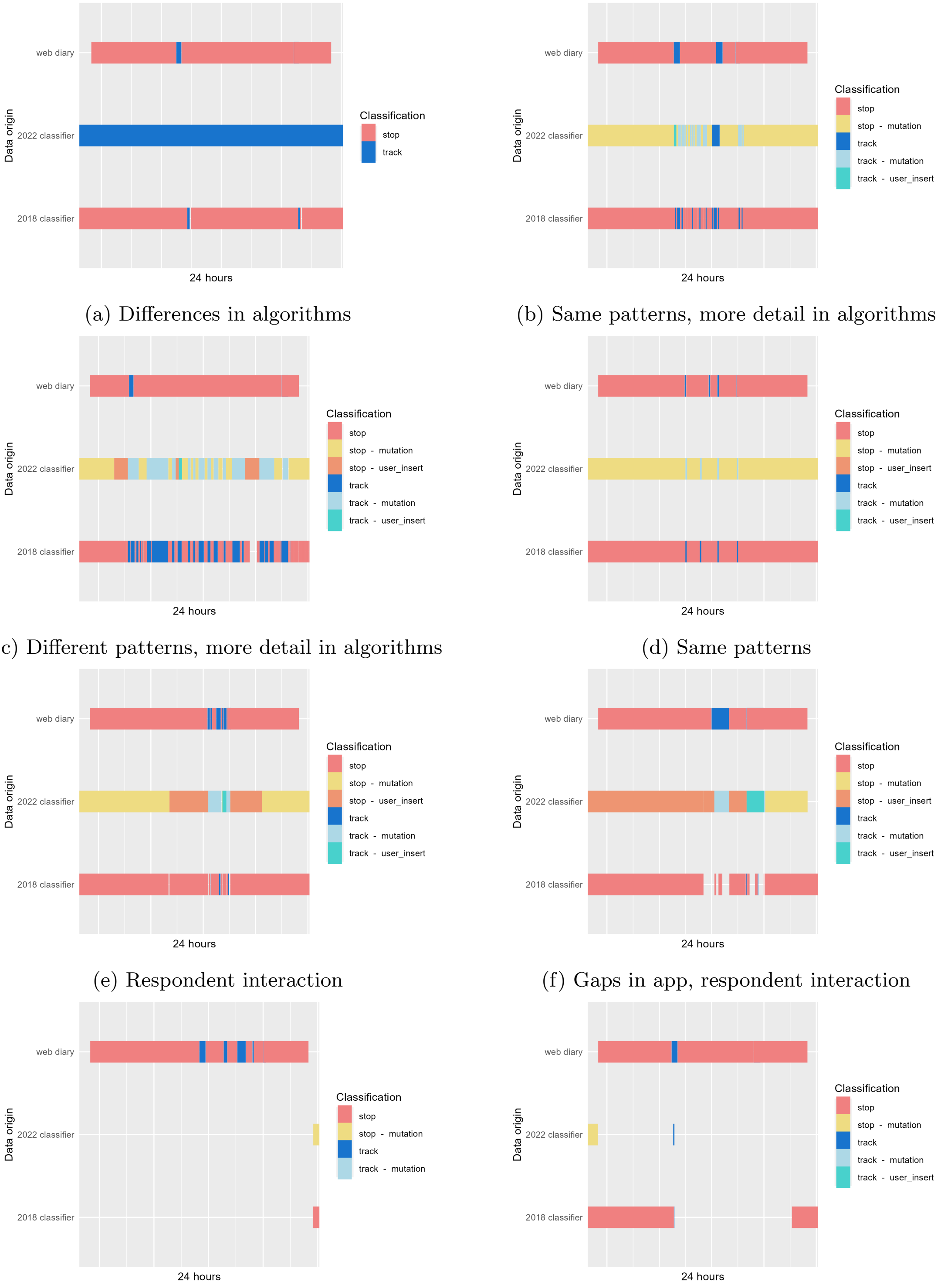
The data collection took place in Q4 2022 – Q1 2023. The sample consists of volunteers recruited through the regular CAWI survey (study phase 1) and a fresh random sample from the general population (study phases 2 and 3). The Phase 1 users were instructed to use the app for seven days and fill out a CAWI questionnaire on one predetermined day in the same week. Users were asked to complete the questionnaire at the end of the predetermined day or in the morning the day after. The users were informed that this setting is required to evaluate whether the app and the questionnaire result in the same information. The users were instructed not to use the app to complete the questionnaire. From the $n = 667$ users of the first study phase, 32% ($n = 212$) used the app. Of those, 23% ($n = 151$) have sufficient data quality regarding time and the number of measured geo-locations. 15% ($n = 101$) filled out the web questionnaire. 8% of the users ($n = 54$) were eligible for validation because they had app and diary data on the same day. The distribution of the days considered for validation is: 19% Monday, 20% Tuesday, 9% Wednesday, 13% Thursday, 11% Friday, 20% Saturday, and 8% Sunday. The task for the users in the web questionnaire was to report the movement periods. Periods when no movements were reported were imputed with stops. This is based on the assumption that all movements were reported and were not forgotten or concealed. The considered target variable for validation is a binary classification of whether a respondent is moving or stopped at a given time. The diary and app data were linked on the event and minute levels for the analyses. Events are in the form of periods classified into a stationary or movement activity.

# 4 Results

First, the event-level results are reported, and subsequently, the minute-level results.

## 4.1 Event-level comparison

Figure 2 shows examples of observed events. Each panel is based on a different respondent. The x-scale shows the 24 hours of the day on which web diary responses and app data are available.



(a) Differences in algorithms

(b) Same patterns, more detail in algorithms

(c) Different patterns, more detail in algorithms

(d) Same patterns

(e) Respondent interaction

(f) Gaps in app, respondent interaction

(g) Gaps in app

(h) Large differences

Figure 2: Examples of events in web diary and algorithms.

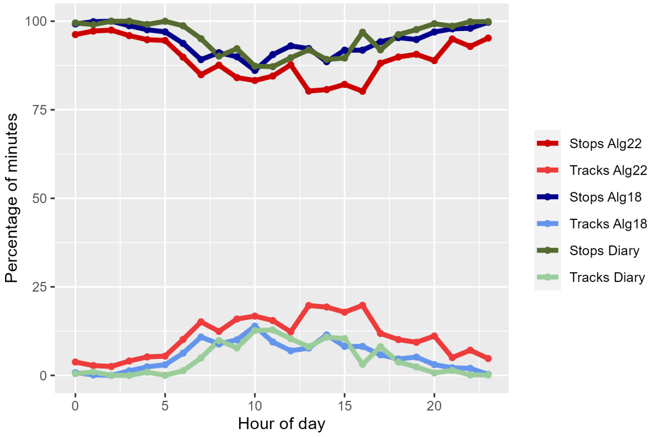


Figure 3: Percentage of minutes split by event and classification.

The y-scale shows the different sources of the event classifications. The legend shows the different events in the data. Figure 2a shows large differences between ALG18 and ALG22, and Figure 2b shows the same pattern in all three sources. With more details in the algorithms, Figure 2c shows different patterns between web response and algorithms. Again, with more details in the algorithms, Figure 2d shows the same pattern in all three sources. Figure 2e shows several respondent interactions (also visible in other panels), Figure 2f shows gaps in the GPS data which the respondent in the ALG22 fills, Figure 2g and Figure 2h shows different patterns and large gaps in the GPS data.

Figure 3 shows the percentages (y-axis) of the events per hour (x-axis) aggregated over all users. The percentages of available data per hour were calculated for the events to compare the imputed web response and the algorithm classifications that contain missing data. The algorithms start recording tracks earlier than reported in the web response. A large difference between the reported tracks and ALG22 is observed during the afternoon. With the current parameter setting, ALG18 is closer to the web response than ALG22 is. ALG22 classifies more tracks than ALG18 and the web response. This difference will change with a different parameter setting. The difference between ALG18 and ALG22 is due to differences in algorithms (fixed parameters vs constantly adjusted parameters), but it might also be that the respondents labeled more stops in ALG22 than they reported in the diary. However, this cannot be disentangled with the data at hand.

## 4.2 Minute-level comparison

In addition to the event-level comparison, the data was linked on the individual minute level. From this, the confusion matrix was derived, and the following quality metrics were derived: Accuracy (ACC), balanced accuracy (ACC$_b$), precision, recall, and F1. The metrics were calculated with each classification as ground truth. This will have effects on precision and recall. The results are shown in Table 1. Overall, good results are achieved. The ACC ranges from 0.86 to 0.94. This means the algorithms and survey responses are the same for 86% and 94% of all minutes, respectively. For ACC$_b$, even slightly higher values are achieved. The ALG18 yields slightly better results than the ALG22 compared to the survey response. Comparing ALG18 with ALG22 results in slightly higher values than comparing algorithms with survey responses. Switching the ground truth has the largest effect for ‘ALG22 – response’ and ‘response – ALG22’. The values change by about 7%. For example, in the ‘ALG22 – response’ comparison, the precision is 0.96, i.e., 96% of the minutes identified as stops were actually correct. Here, recall is 0.89, i.e., 89% of minutes correctly identified as stop were correctly identified. This is the reverse in the ‘response – ALG22’ comparison. The F1 score suggests a good performance for all comparisons.

| | ACC | Precision | Recall | $ACC_b$ | F1 |
|---|---|---|---|---|---|
| ALG22 – response | 0.86 | 0.96 | 0.89 | 0.93 | 0.93 |
| response – ALG22 | | 0.89 | 0.96 | | |
| ALG18 – response | 0.91 | 0.96 | 0.95 | 0.95 | 0.95 |
| response – ALG18 | | 0.95 | 0.96 | | |
| ALG22 – ALG18 | 0.94 | 1.00 | 0.94 | 0.97 | 0.97 |
| ALG18 – ALG22 | | 0.94 | 1.00 | | |

Table 1: Quality metrics of classifications in the app data and web response.

## 5 Discussion & Conclusion

Using smart surveys to complement or replace traditional diary-based mobility surveys is an important topic in official statistics. In this paper, we presented research on validating a smart survey travel app in which we compared algorithm-based classifications with survey responses about mobility behavior. The central results of this study are summarized below.

First, the results of the event-level comparison show a heterogeneous picture of patterns occurring in the different classifications (see Figure 2). It also shows that some respondents interact with the app, labeling periods or filling gaps in the data. When aggregating the events, there are indications that the algorithms record mobility earlier than reported and record more daily movements. However, the algorithms also differ slightly (see Figure 3). Second, the obtained quality metrics show promising results when comparing the linked data on individual minute levels. Overall, high accuracy, balanced accuracy, and F1 values are achieved. Evidence shows that the ALG18 matches the survey response more than the ALG22 (see Table 1). Whether this is due to the different parameter settings of the algorithms or the respondent interactions remains unclear to this moment.

These results sound promising, but the study and results are subject to a central limitation. The 55 respondents used for this analysis are 8% of the initial sample. Various selection mechanisms (non-response for app, non-response for web questionnaire, insufficient app data quality) cause the result to appear more positive than it actually might be. Furthermore, the measures of fit can only measure the similarity between the various classifications since there is no one true ground truth available. In addition, the data is constituted mainly by long stops (e.g., nighttime), causing much agreement between the response and the algorithms, resulting in more similarities.

Future work will focus on imputing the app data and repeating the analysis with ‘complete’ data to estimate the missing data’s effect better. In addition, the quality metrics will be calculated during ‘peak times’ (e.g., 6 am – 10 am and 4 pm – 7 pm), not considering long stops (e.g., nighttime or work time). Furthermore, the extent to which respondent interaction improves quality will be studied.

To sum up, this validation study has shown that the measurements with smart surveys go in a similar direction to those of the traditional diary survey, but the apps capture more signals. However, this study also shows how essential the assumptions and details that go into the algorithm are to capture these signals.