Discussion Paper

Data quality aspects for location-tracking in smart travel and mobility surveys

Yvonne Gootzen
Jonas Klingwort
Barry Schouten

June 2025

Abstract

Travel Surveys are considered promising candidates to go ‘smart’. Respondents must be motivated and competent to report all travel details correctly for a specified period. Location tracking offers options to decrease the response burden and to improve measurement quality. The collected location data and added contextual information may also be input to travel mode and purpose predictions. However, respondents may perceive location tracking as privacy-sensitive. Furthermore, location data are subject to various types of error that, in part, can only be adjusted for with the help of respondents. It is, therefore, not evident that the promise of smart features holds in practice. For this reason, Statistics Netherlands conducted a first field test in 2018 using a proof-of-concept travel app. Response rates clearly showed variation across relevant subgroups in the population but were sufficiently high to justify further development and experimentation.

In 2022, Statistics Netherlands again conducted a travel-app-assisted field experiment, including the regular online travel survey as a concurrent option. At different time points, a population sample was offered the online option, and the sample consisted of randomized groups. Simultaneously, the requested tracking period was randomized, one full day or one whole week.

This paper analyzes and reports data quality aspects for location tracking in smart travel and mobility surveys to support decision-making about improving location data quality technologically or methodologically. It highlights technological limitations and challenges regarding smartphone brands, battery consumption, and missing data. Results suggest that up to 20% of tracking records may be unusable or require substantial imputation, emphasizing the need for tailored tracking strategies and a balanced trade-off between respondent interaction and advanced post-survey data correction methods.

Keywords: GPS, smartphone, sensor data, data processing

Contents

1 Introduction

Smart surveys employ the features of smart devices in collecting and processing data. They are particularly promising for surveys that are (cognitively) burdensome, demand detailed knowledge or recall, or include topics for which questions provide weak proxies. Travel surveys are typical examples of surveys that satisfy these criteria. Travel surveys collect data on people's movement patterns, transportation modes, and trip purposes to help planners and policymakers improve transportation systems. They provide insights into travel behavior, supporting infrastructure planning, congestion management, and sustainable mobility strategies. Over the last decade, a wide range of studies into sensor-assisted travel surveys has been conducted (e.g., McCool et al. (2024a), Harding et al. (2021), Gillis et al. (2023), and Lawson et al. (2023)). Studies into using an app-assisted approach, including location tracking, started in 2017 at Statistics Netherlands. The aim is to supplement data collection options in surveys on the general population's travel and time use.

While promising from a measurement perspective, the potential of increased survey participation rates by a reduced burden has yet to be demonstrated. The reasons are clear, in part. Survey nonresponse is not just about burden but also about making contact, making respondents feel sufficiently competent, and motivating them. Offering a smart option does not necessarily remove these reasons. An app-assisted location tracking approach demands some minimal digital skills. Also, location tracking data are subject to error (Harding et al. 2021; McCool et al. 2024b; Klingwort et al. 2025c) that can only be adjusted with the help of respondents. Finally, location tracking leads to microdata with surplus information relative to survey output needs. Consequently, respondents may perceive the detailed data as too sensitive to privacy. Evidence for these hesitations and perceptions can be found in Assemi et al. (2018), Struminskaya et al. (2020), Klingwort and Schnell (2020), and Lunardelli et al. (2024). Willingness to go smart may be the Achilles heel of smart travel surveys, and effective, nudge-to-smart recruitment and motivation strategies are paramount. These considerations have been a major driver behind the follow-up field test in 2022.

Various experimental studies have evaluated the design features of potentially effective recruitment and motivation strategies for smart surveys (Maruyama et al. 2015; Safi et al. 2017; Faghih Imani et al. 2020; McCool et al. 2024a; Winkler et al. 2023). The 2018 Statistics Netherlands first proof-of-concept study for a smart survey travel app, based on a general population random sample and a cross-platform app, attained registration rates of around 25% and seven-day completion rates of around 20%. The study randomized different incentive strategies. The registration rates were deemed sufficiently high to justify further research into effective data collection strategies and trade-offs in the active-passive involvement of respondents (Smeets et al. 2019; McCool et al. 2021).

Statistics Netherlands conducted a second large-scale field test for a smart travel survey between November 2022 and February 2023. The cross-platform app had been completely redesigned. The new app ('CBS Onderweg in Nederland') included several options for respondents to edit automated stop-track segmentations shown during the diary reporting period. In the experiment, three design features were randomly varied: the length of the reporting period (one day or seven days), the amount of respondent editing (full editing or limited editing), and the offering of the web diary as an alternative (direct at invitation, at first reminder or second reminder). For further details about the 2022 field experiment, see Schouten et al. (2024).

This paper supports decisions about the improvement of location data quality either technologically or methodologically. Location data quality has two elements: missing location data and imprecise/noisy location data. The options we have to improve these elements are the revision of the technology behind location tracking, the implementation of advanced AI-ML (AI: artificial intelligence, ML: machine learning) and imputation methods, and the involvement of respondents. Given that prevention is better than cure, improved technology is preferable. However, this has its limits. More specifically, the goal is to see what technological improvements are imperative and what remains for methodology improvement without or with the help of respondents. This paper does not discuss technology or methodology (UI-UX choices (UI: user interface, UX: user experience), AI-ML predictions, and imputation methods). This is done in other papers (see for example, Zahroh et al. (2025), Klingwort et al. (2025c), McCool et al. (2024b), Klingwort et al. (2025a), and Klingwort et al. (2025b)), but this paper provides the stepping stone. Specifically, data quality monitoring will likely remain paramount, given the ever-changing landscape of devices and sensor technology. This paper's scope is thus evaluating location data quality and identifying necessary improvements by answering the following research questions: 1) What is the quality of the location data? 2) What improvement is needed through technology? Moreover, 3) What improvement remains for methodology?

In order to discuss location data quality and any potential improvement, we need to be concise about what we mean. We will explore various dimensions: the number of measured geo-locations, the number of observed minutes with geo-location data, the length of participation, or the length of the gaps in the data. These quality dimensions are evaluated separately relative to the different experimental conditions, the manufacturer of the smartphone, and the sensor type. In addition, the extent to which geo-location accuracy and the smartphone's battery level affect the data quality is studied.

This discussion paper belongs to a series of discussion papers on smart travel surveys. The project, fieldwork, and response process are described in the discussion paper by Schouten et al. (2024), and the effects of respondent interaction are described in the discussion paper by Remmerswaal et al. (2025). As the discussion paper by Schouten et al. (2024) contains all relevant information on the fieldwork, the current discussion paper will only mention the information about the fieldwork that is necessary.

The report is structured as follows: First, we briefly describe the data used for the analyses, introduce definitions and notations and provide descriptives in section 2. We discuss qata quality related to technical device information in section 3. The smartphone brand is addressed in section 3.1, the provider and sensor in section 3.2, the battery level in section 3.3, and the accuracy in section 3.4. We move to information on user profiles regarding data quality in section 4. As a last step, we consider different post-processing options in section 5. In section 6, we end with a discussion and conclusion.

2 Data

This chapter provides information about the data, statistics, and definitions used that are relevant to this discussion paper. For detailed descriptions of the data, we refer to the discussion papers by Schouten et al. (2024) and Remmerswaal et al. (2025).

2.1 The 2022 - 2023 app-assisted fieldwork variants

Between November 2022 and February 2023, Statistics Netherlands conducted a large-scale field test. The new app (‘CBS Onderweg in Nederland’) included several options for respondents to edit automated stop-track segmentation shown during the diary reporting period.

The field test involved two samples: a sample of 667 former respondents from the Dutch Travel Survey (’Onderweg in Nederland’ or ODiN for short) and a fresh sample of 2544 persons of 16 years and older from the Dutch population register. The sample of former respondents, we refer to as the ’follow-up sample’. The fresh sample was randomly split into two equally sized subsamples which we will refer to as the ’full-editing’ and the ’limited-editing’ samples.

In the experiment, three design features were varied: the length of the reporting period (one day or seven days), the amount of respondent editing (full editing or limited editing), and the offering of the web diary as an alternative (direct at invitation, at first reminder or second reminder). In the ‘follow-up sample’, respondents were asked to use the app for seven days and complete the online questionnaire for one day (in the same week). Here, the app version allowed a lot of user interaction. but was the same for all sampled persons. The ‘full editing sample’ and ‘limited editing sample’, however, used different app versions. The full editing sample could do all possible edits and imputations on the presented tracks and stops. The limited editing group could not add stops or tracks and could not change start and end times. Within both editing samples there were two further randomizations. The first was into a one-day tracking and a seven-day tracking period.The second was on the timing of a concurrent regular web survey option. The different experimental variants are listed briefly below. For more details, see Schouten et al. (2024) and Remmerswaal et al. (2025).

- 1 = 1 day, Both app and regular ODiN survey offered at invitation.
- 2 = 7 days, Both app and regular ODiN survey offered at invitation.
- 3 = 1 day, App offered at invitation and regular ODiN survey at first reminder (after two weeks).
- 4 = 7 days, App offered at invitation and regular ODiN survey at first reminder (after two weeks).
- 5 = 1 day, App offered at invitation and regular ODiN survey at second reminder (after four weeks).
- 6 = 7 days, App offered at invitation and regular ODiN survey at second reminder (after four weeks).

Given that in this paper we consider the raw location tracking data and not the edits respondents performed we pool the data from the full editing and limited editing samples. We have in total 505 respondents from all samples, 212 from the follow-up sample and 293 from the fresh sample. Out of the 293 fresh sample respondents, 133 were in the one-day tracking subsample and 160 in the seven-day tracking subsample.

2.2 Observations in time and space

In this section, we report on the observed data regarding time and space throughout the fieldwork period. Regarding time, we report the average number of observed minutes per hour; for space, we report the number of measurements (geo-location → longitude, latitude, timestamp). These results give a more general impression of the data collection process.

Figure 2.1 shows the average number of observed minutes per hour during the fieldwork period. The average is computed for each hour individually over those users which were participating during the hour. Each data point in the figure corresponds to one hour. The red and green lines indicate points in time (invitations and reminders) of the different samples. The 'follow-up sample' and the 'full editing sample' were fielded first and the 'limited editing sample' followed. The figure clearly shows the impact of the fieldwork strategy. After the invitation was sent, observed minutes began to increase. This number decreased until a reminder was sent. Then, a reminder caused an increase in the observed minutes again. For the 'limited editing sample', the trend shows a somewhat larger level of observed minutes but, at the same time, also a more considerable variation.

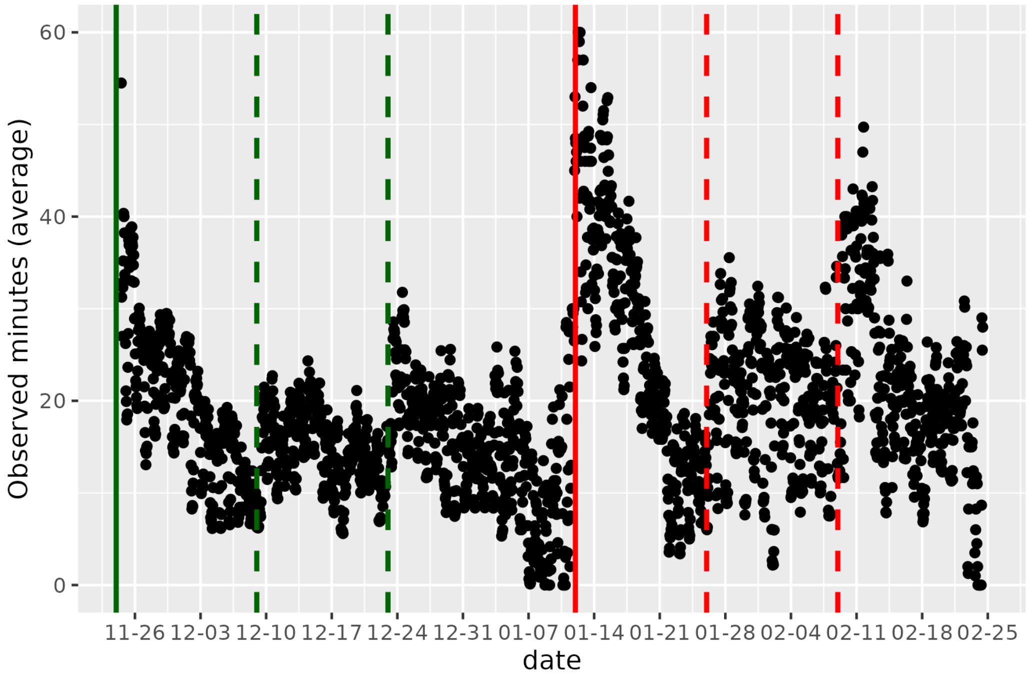


Figure 2.1 X-scale shows days, y-scale shows the average number of observed minutes per hour during the fieldwork period. Green lines indicate the invitations for the 'follow-up sample' and the 'full editing sample' (solid line) and two reminders (dashed lines), and the red lines indicate the invitation for the 'limited editing sample' (solid line) and two reminders (dashed lines).

Figure 2.2 shows the number of observed geo-locations. The red and green lines indicate the same information as in Figure 2.1. As in the previous figure, the fieldwork strategy is visible. Here, the effect of invitations and reminders is even more evident. A possible decrease in the impact of the reminders can be seen throughout the fieldwork period since the number of observed geo-locations increases after a reminder but never rises to the level of the invitation. It should be noted that the different patterns in Figures 2.2 and 2.1 cannot be directly compared because 2.1 shows averages and 2.2 totals.

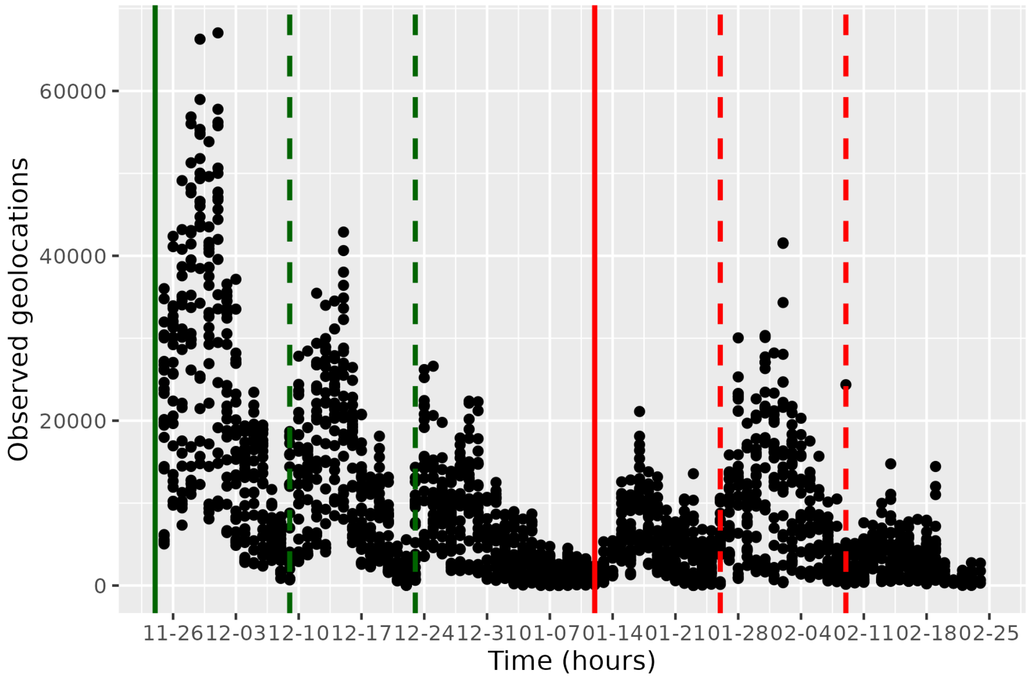


Figure 2.2 X-scale shows days, y-scale shows the total number of observed geo-locations during the fieldwork period. Green lines indicate the invitations for the 'follow-up sample' and the 'full editing sample' (solid line) and two reminders (dashed lines), and the the red lines indicate the invitation for the 'limited editing sample' (solid line) and two reminders (dashed lines).

2.3 Statistics and definitions

Here, the derived statistics for this report and their definitions are introduced.

1. **Number of measurements**: the number of measured geo-locations [0, ∞]
2. **Number of observed minutes**: the number of minutes within one hour for which at least one observation was present [0, 60]
3. **Hours range**: the number of hours between the first measurement and last measurement for which at least one observation was present [0, ∞]
4. **Hours largest gap**: the largest amount of consecutive hours without any measurements [0, *Hours range*]
5. **Hours missing**: the number of hours without any measurements [0, *Hours range*]

For each statistic, the minimum, maximum, quantiles (Q_{25}, Q_{50}, Q_{75}) and the average are calculated.[1] In addition, the following variables are considered for analysis (these variables are available per location data point):

1. **Samples**: the three samples (follow-up, full editing, limited editing)
2. **Brand**: brand of the smartphone (categorical variable)
3. **Provider/Sensor**: the sensor of the smartphone (categorical variable)
4. **Battery level**: the status of battery remaining (in %, [0, 100])
5. **Accuracy**: the accuracy of the measurements [0, ∞], low value = accurate measurement, high value = inaccurate measurement

[1] We do not report all metrics for each variable. The results are available and can be shared upon reasonable request.

It is important to note that some variables, such as accuracy and battery level, are not constant within a user but can vary per measurement/observation. A measurement/observation consists of a timestamp and geo-location. This variability is a key aspect of the data, reflecting the dynamic nature of smartphone usage. On the other hand, variables such as brand and sample are constant per user. Therefore, not all statistics and figures apply to all variables. This has to be considered when interpreting the results. The variable 'Accuracy' is a variable that is provided by the smartphone/sensor.

3 Technical device information

We start by considering the role of sensor technology. More specifically, we look at the smartphone brand, the different location sensor options, the battery usage of the app, the accuracy of the location points. Location tracking technology has been the topic of an large body of literature. Here, we limit ourselves to exploratory analyses. We refer to Safi et al. (2017) for an introduction to location tracking technology.

3.1 Smartphone brand

Table 3.1 shows the top-10 selection of brands using the number of users as selection criteria. The table is ordered by number of users (second column). The number of measured geo-locations (median) and the number of observed hours between the first and last measurements (median) are shown for each brand. The number and percentage of users with sufficient data quality (see section 4 how sufficient quality is defined) per brand are also shown. Most users use an iPhone or Samsung smartphone. Samsung phones measured more data than iPhones. Brands with smaller market shares (e.g., OPPO, OnePlus, Google, and Sony) measured more geo-locations than the two major brands. The variable 'Hours range' shows considerable variation between brands. About 75% of iPhone and Samsung data is of sufficient quality. Huawei, for example, only has 15% of users with sufficient data quality. All five users of Google phones have sufficient data quality.

Brand	Users	Observations (median)	Hours range (median)	Sufficient quality (n)	Sufficient quality (%)
iPhone	247	13654	157	185	74.90
Samsung	185	26822	142	145	78.38
OPPO	14	7513	46.5	8	57.14
HUAWEI	13	181	24	2	15.38
Motorola	9	210	10	4	44.44
OnePlus	7	28976	58	5	71.43
google	5	48269	192	5	100.00
iPad	5	7	0.00	0	0.00
Xiaomi	5	5042	194	4	80.00
Sony	4	70881.5	175.5	3	75.00

Table 3.1 Descriptive statistics of top-10 used smartphone brands.

Table 3.2 shows the number of days of the fieldwork variants split by major smartphone brands. The table shows the brand, the number of users per brand, the total number of users in the fieldwork variants, and the percentage of users/smartphone brands per fieldwork variant. Table A.1 is an extended version of this table and can be found in the Appendix.

The largest percentage of iPhone and Samsung brands is found in both the 1-day and 7-day variants. iPhone is the primary brand in both variants. None of the other brands make up more than 3% for either the 1-day or 7-day variants.

Number of days	brand	n_users	sum_users	percentage (per day)
1	iPhone	67	133	50
1	Samsung	53	133	40
1	other	13	133	10
7	iPhone	180	372	48
7	Samsung	132	372	35
7	other	60	372	17

Table 3.2 Cross-table of variants (one day and seven days) and smartphone brand. The percentage of brands is based on the number of users and is calculated for one and seven days.

Figure 3.1 shows data quality statistics split by the smartphone brand. The point size indicates the number of users (‘gebruikers’). All results are ordered increasingly (from top to bottom). In all figures, the mean and median are shown to contrast the outcome difference when choosing one over the other.

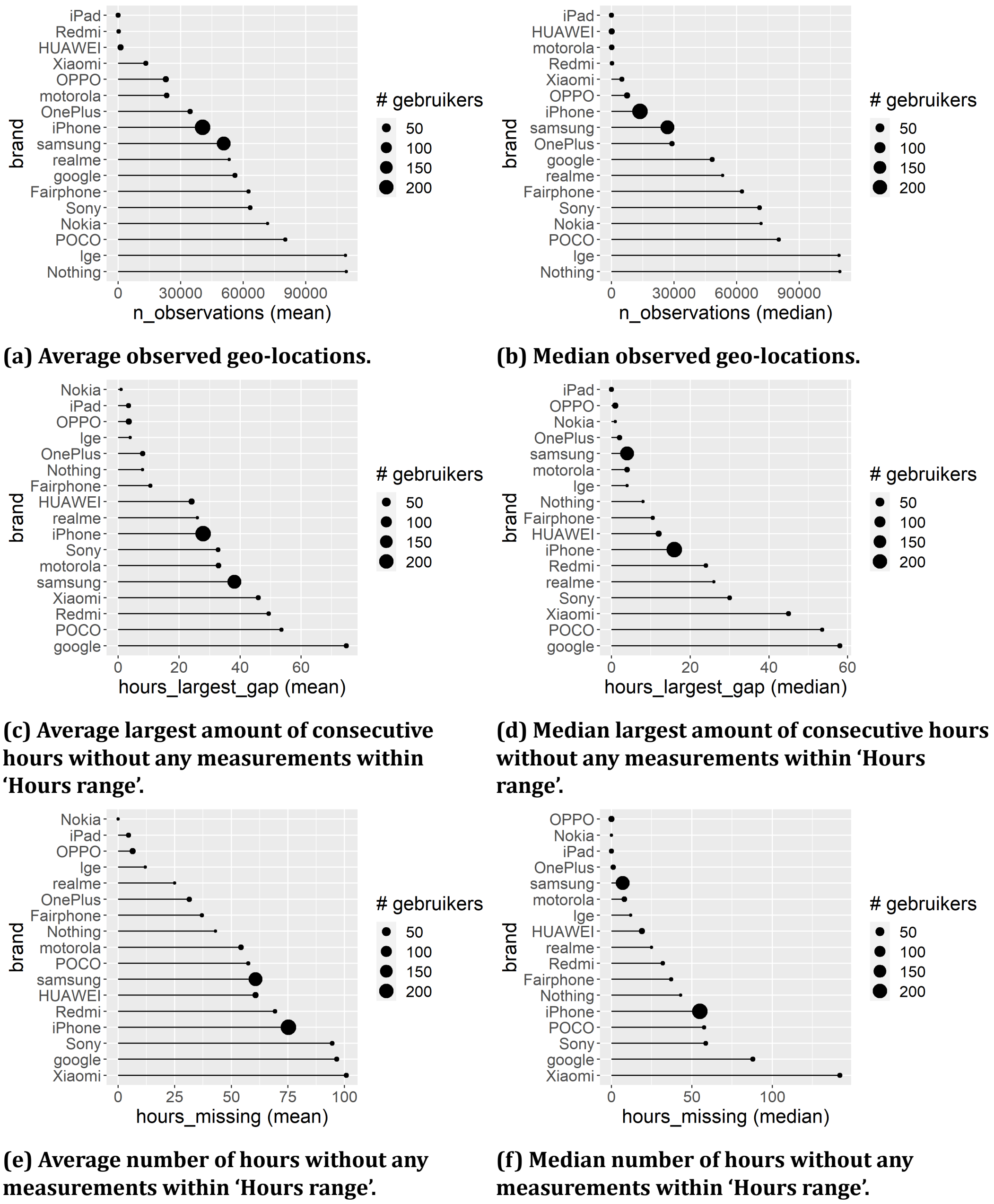
Figures 3.1a and 3.1b show the average and median number of observations per brand. Most important here is that the number of observations considerably decreases when considering the median. This finding is because some brands have more observations than the two major brands (iPhone and Samsung), but these numbers are based on fewer users (see Table 3.1).

Figures 3.1c and 3.1d show the mean and median for the largest amount of consecutive hours without any measurements within the ‘Hours range’. Again, numbers are larger when considering the mean. When considering the mean, Samsung has, on average, larger gaps than the iPhone. However, the iPhone has larger gaps when considering the median, and the number for Samsung decreases considerably.

A similar finding is shown in Figures 3.1e and 3.1f. Here, the number of missing hours is smaller when choosing the median over the mean. However, iPhone users have more missing hours in both cases than Samsung users. More details on the missing hours will be discussed in the context of Figure 3.2. When considering the median, the number of hours missing for Samsung users is small.

The mean and median for the statistics shown are considerably different. This finding suggests some large outliers in the data affecting the mean. When considering the two leading brands, Samsung users have more observations and fewer hours missing than iPhone users. For the largest gaps, the findings are mixed between the two brands.

Figure 3.2 shows the average observed minutes per hour during a day (24h). The color of the line indicates the brand (iPhone and Samsung) and the fieldwork variant (one day or seven days). The curves for iPhones are comparable between fieldwork variants, with slightly fewer observations for the 7-day variant. The Samsung phones show a greater difference between the 1- and 7-day variants. Most observations are made for the 1-day variant. As discussed and shown in Section 3.2, the two brands show a characteristic curve over the day, irrespective of the fieldwork variant.



(a) Average observed geo-locations.

(b) Median observed geo-locations.

(c) Average largest amount of consecutive hours without any measurements within 'Hours range'.

(d) Median largest amount of consecutive hours without any measurements within 'Hours range'.

(e) Average number of hours without any measurements within 'Hours range'.

(f) Median number of hours without any measurements within 'Hours range'.

Figure 3.1 Selected results of data quality statistics split by smartphone brand.

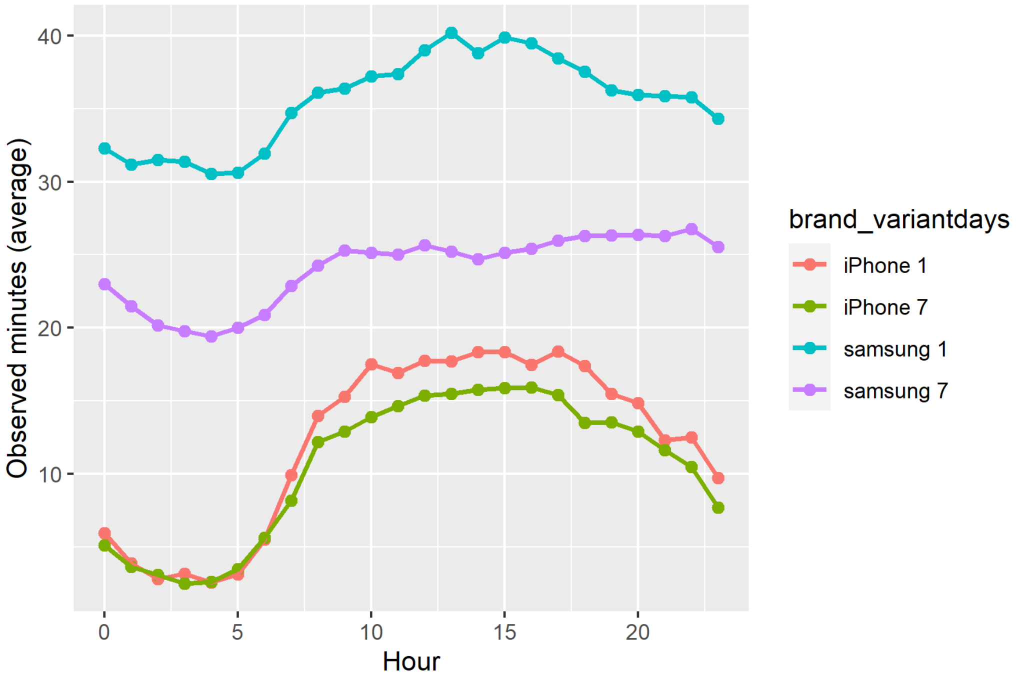


Figure 3.2 X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour split by smartphone brand and the number of days.

3.2 Provider and sensor

Table 3.3 shows the total number of measurements (geo-locations) split by provider. For each type, the number of measured geo-locations (median) and the number of observed hours between the first and last measurements (median). Nearly every user has observations based upon the 'GeolocatorPlatform'. For this provider, an observation is generated when the user starts or interacts with the app. This reason explains why this type has fewer observations than all other classes. While iPhones use one sensor (in the table 'ios'), different providers were used in Android smartphones. Most observations are made by 'ios' and 'GPS.' These two also contain the highest percentage of users with sufficient data quality. The statistics for the 'balanced' (well-calibrated and optimized) and 'fused' (combines multiple inputs) providers are comparable. For more details about the sensor configurations, see Schouten et al. (2024).

Provider	Users	Observations (median)	Hours range (median)	Sufficient quality (users)	Sufficient quality (%)
GeolocatorPlatform	503	15	113.00	0	0.00
ios	253	13032.00	155	184	72.73
fused	242	4603.5	123.00	157	64.88
balanced	241	4630	124.00	157	65.15
network	241	2234	125.00	129	53.53
gps	232	13701.5	124.00	171	73.71

Table 3.3 Number of measured geo-locations (median) split by 'Provider'.

Table 3.4 shows the total number of measurements (geo-locations) split by variables 'Provider' and 'Sensor'. The categories 'GPS' and 'Network' from the variable 'Provider' are summarized in 'Normal' in the variable 'Sensor'. Hence, using 'Provider' instead of 'Sensor' allows for more differentiated insights about the sensors.

Tables 3.5–3.7 show the fieldwork variants split by smartphone brand. The tables show the brand, the number of users per brand, the total number of users in the fieldwork variants, and the percentage of users/smartphone brands per fieldwork variants. Extended versions of these tables can be found in appendix Tables A.2–A.4. The percentages of iPhones and Samsung are the largest throughout all fieldwork variants. Only in variant 4 is the percentage of Samsung larger than that of iPhone. The percentages of all other brands are relatively small. These tables supplement other results in this paper, which aims to find differences in fieldwork variants. For example, Figures 3.2, 4.3, and 4.5.

Provider	Sensor				
	ios	balanced	fused	GeolocatorPlatform	normal
balanced	0	1879646	0	0	0
fused	0	0	1879627	0	0
GeolocatorPlatform	0	0	0	12101	0
gps	0	0	0	0	6644876
ios	9998410	0	0	0	0
network	0	0	0	0	983039

Table 3.4 Cross-table of variables ‘Provider’ and ‘Sensor’.

Variant	brand	n_users	sum_users	percentage
0	iPhone	106	212	50
0	Samsung	75	212	35
0	other	31	212	15

Table 3.5 ‘follow-up sample’ users and smartphone brand.

Variant	brand	n_users	sum_users	percentage
1	Samsung	16	42	38
1	iPhone	23	42	55
1	other	3	42	7
3	iPhone	25	50	50
3	Samsung	19	50	38
3	other	6	50	12
5	Samsung	18	41	44
5	iPhone	19	41	46
5	other	4	41	10

Table 3.6 ‘full editing sample’ and ‘limited editing sample’ users, fieldwork variants 1, 3, 5, and smartphone brand.

Variant	brand	n_users	sum_users	percentage
2	iPhone	23	48	48
2	Samsung	15	48	31
2	other	10	48	21
4	iPhone	23	54	43
4	Samsung	25	54	46
4	other	6	54	11
6	iPhone	28	58	48
6	Samsung	17	58	29
6	other	13	58	23

Table 3.7 ‘full editing sample’ and ‘limited editing sample’ users, fieldwork variants 2, 4, 6, and smartphone brand.

Figure 3.3 shows the median of the measured accuracy versus the unique minutes observed (natural limit of 60) split by the provider and sensor type. For visualization purposes, the maximum accuracy to be shown is set to 3000. Otherwise, large outliers would not allow for proper readability. In general, independent of the provider and sensor, most unique minutes observed have low values for accuracy (low value = very accurate, high value = inaccurate). The patterns of the 'Balanced' and 'Fused' sensors are similar: mostly low values, with larger values at the beginning and end of an hour (the tendency for a U-shape in the data points). 'GPS' constantly provides the lowest accuracy values, which show a larger variation within the first 30 minutes. 'iOS' is most comparable to 'GPS' but has larger values and more variation within the first 20 minutes. 'Network' shows the most different pattern with larger variation throughout the 60 minutes. 'GeolocatorPlatform' has only observed minutes during the first 15-20 minutes. Within these, a large variation is observed.

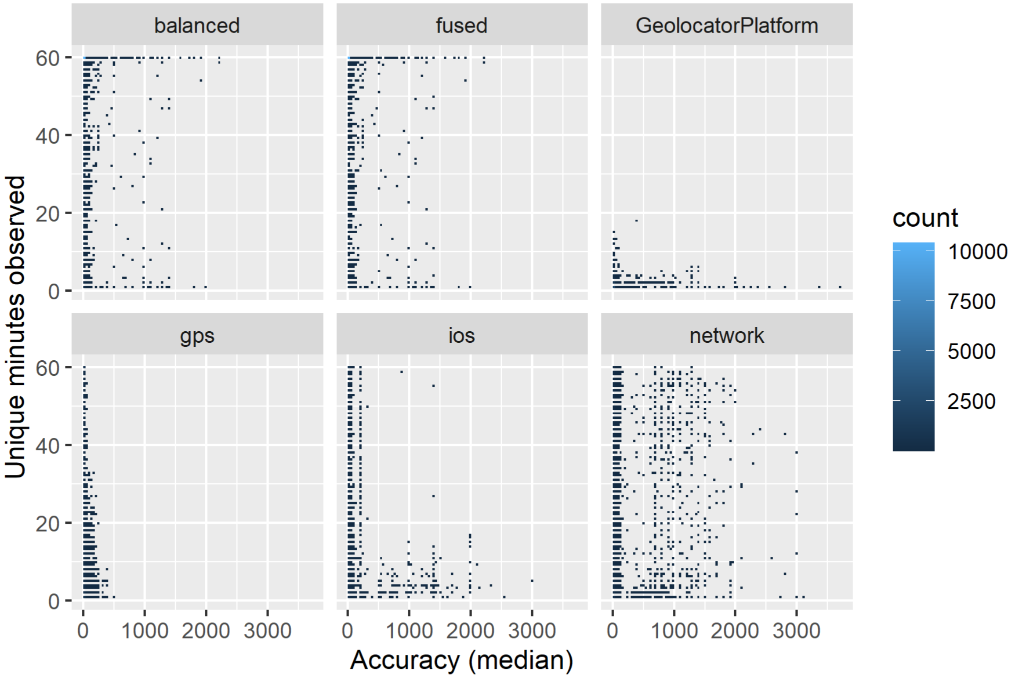


Figure 3.3 Measured accuracy and unique observed minutes split provider. The x-axis is restricted to 3000.

Figure 3.4 shows the average minutes observed (y-axis) per hour during the day (x-axis) split by provider and sensor type. The minutes observed with 'GeolocatorPlatform' are close to zero. The iPhone observations decrease steadily from hour 17 until hour 24, while the observations remain relatively stable for the normal, fused, and balanced sensors. Moreover, the number of observed unique minutes for the iPhone sensor is considerably smaller than the normal, fused, or balanced sensors used in Android phones. Most of the minutes are observed with the balanced and fused sensors. The normal sensor is in between the iPhone and these two sensors. The balanced and fused provide a similar number of unique observed minutes as the lines are stacked on each other. For all sensor types, there is a dip during nighttime in the observed minutes. This behavior can be due to inactivity in movement and sleep mode of devices.

Figures 3.5a and 3.5b show the average and maximum hours missing per provider and sensor. The point size indicates the number of users per sample. On average, 'Fused' and 'Balanced' have the lowest missing hours, followed by 'GPS' and 'Network' having a comparable amount. On average, most hours are missing for 'GeolocatorPlaform'. 'GPS' has the smallest maximum for missing hours. The other providers and sensors show comparable values. Figures 3.5c and 3.5d show the average and median for the largest gaps (in hours). Again, depending on the metric,

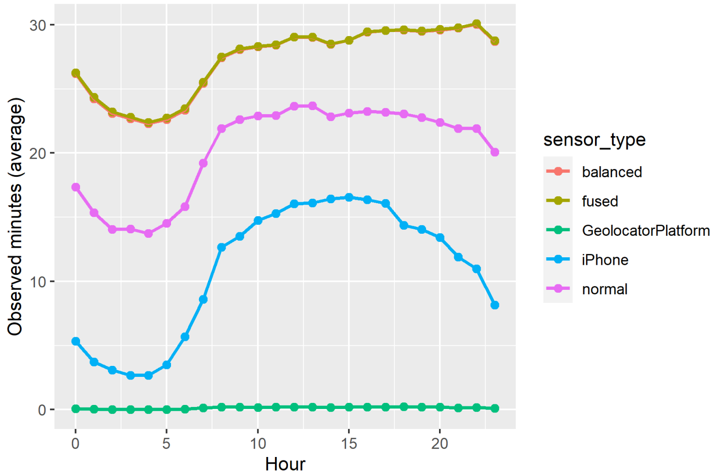
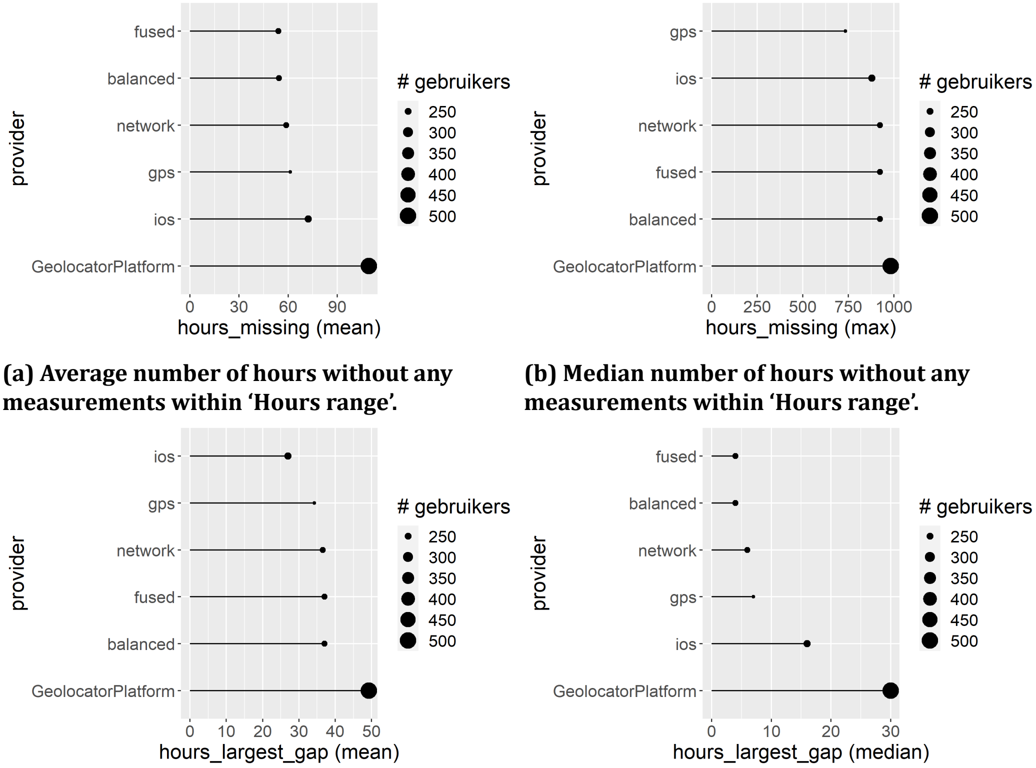


Figure 3.4 Average unique minutes observed (y-axis) during 24 (x-axis) split by provider and sensor type.

considerably different values are found, and the order of the categories changes.



(a) Average number of hours without any measurements within 'Hours range'.

(b) Median number of hours without any measurements within 'Hours range'.

(c) Mean largest amount of consecutive hours without any measurements within 'Hours range'.

(d) Median largest amount of consecutive hours without any measurements within 'Hours range'.

Figure 3.5 Selected results of data quality statistics split by provider/sensor type.

3.3 Battery level

Figure 3.6 shows the average observed minutes per hour during a day (24h). The line's color indicates the median battery level throughout the 24 hours. The battery level is highest from 4 am-8 am, decreases during the day, and goes below 60% in the evening hours before re-charging starts from about 11 pm to 3 am. The median battery level does not drop below 50%, suggesting no severe battery level problems for at least half of all users.

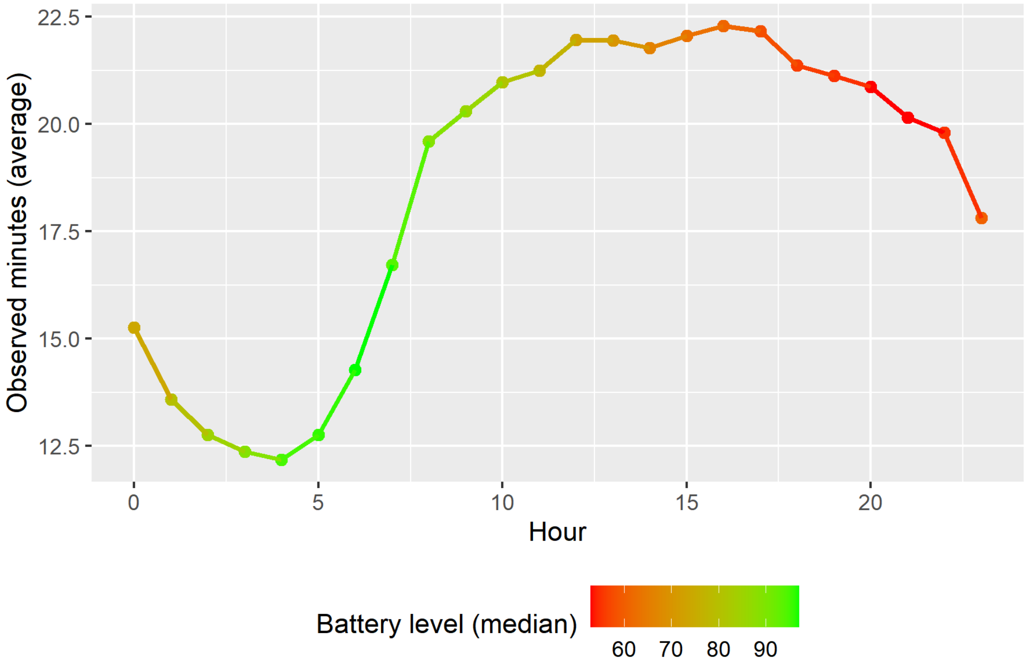


Figure 3.6 X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour. Color indicates battery level in %.

Figure 3.7 shows the average observed minutes per hour during a day (24h) split by 'Provider/Sensor'. The line's color indicates the median battery level throughout the 24 hours. Regarding the battery level, all panels show a comparable median battery level per hour. Again, it can be seen that a) the iPhone has, on average fewer observed minutes than 'Balanced', 'Fused', or 'Normal', and b) shows a different pattern throughout the day. Whether this pattern is due to different travel and movement behaviors or due to technicalities needs further research.

Figure 3.8 shows the average observed minutes per hour during a day (24h) split by the three samples (1 = 'follow-up sample', 2 = 'full editing sample', 3 = 'limited editing sample'). The line's color indicates the median battery level throughout the 24 hours. Regarding the battery level, all panels show a comparable median battery level per hour. However, the level of observed minutes differs between the panels. The 'full editing sample' and 'limited editing sample' are more alike than the level in in the 'follow-up sample', which is lower. This finding is interesting, considering the different reporting periods in the sample variants.

Finally, Figures 3.9 and 3.10 show the average observed minutes per hour during a day (24h) split by 'Brand'. For most brands, the same result as reported before is found: low battery levels in the early morning hours, high battery levels during the mid-morning, decreasing over the day, and low in the evening and night-times. Most brands show comparable patterns for the observed minutes during the day, while some have higher levels (OPPO or Nokia) than others (Xiaomi or Redmi). Figure 3.10 shows the same data but focuses on the two major brands: iPhone and Samsung. Similar patterns can be observed regarding battery level, though iPhone battery levels

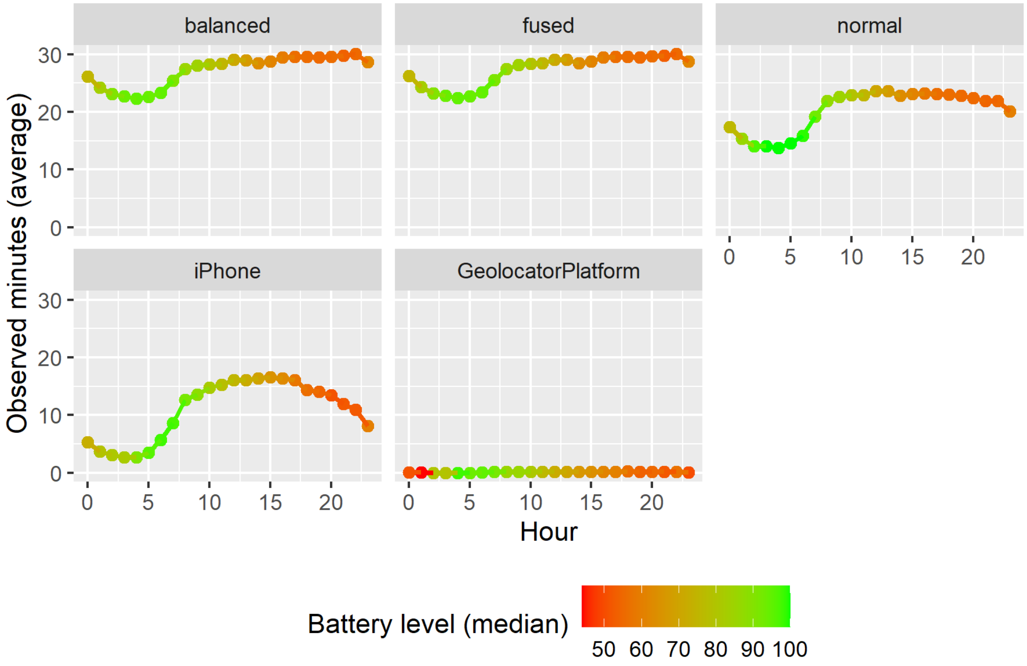


Figure 3.7 **X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour. Color indicates battery level in %. Panels show different sensor types.**

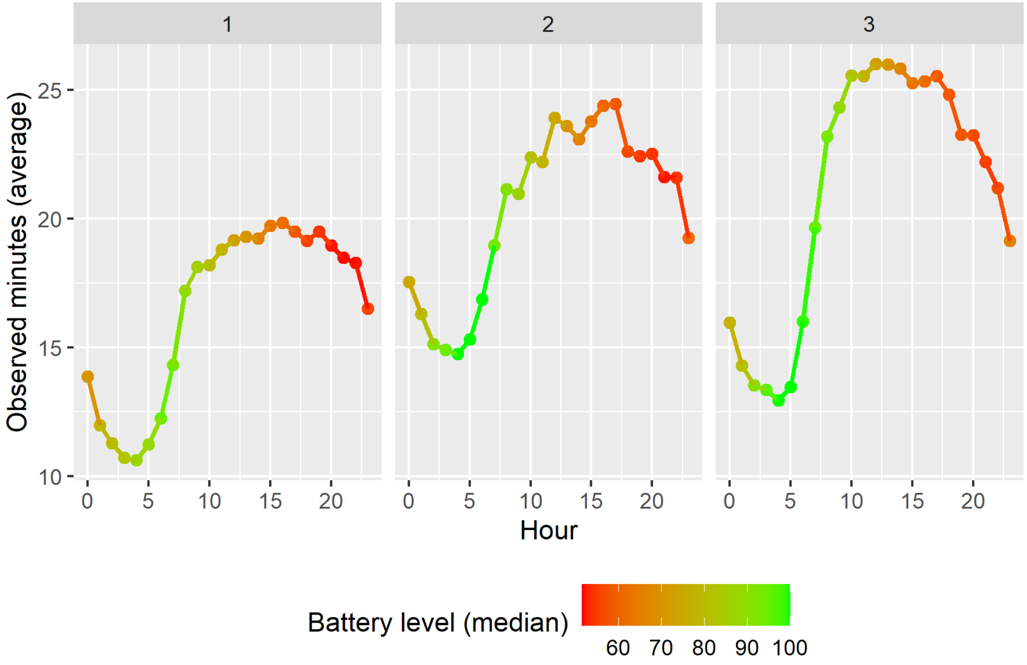


Figure 3.8 **X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour split by samples. Color indicates battery level in %. Panels show different samples.**

tend to be lower towards the end of the day. Different patterns/levels are observed in the minutes observed on average per hour during the day.

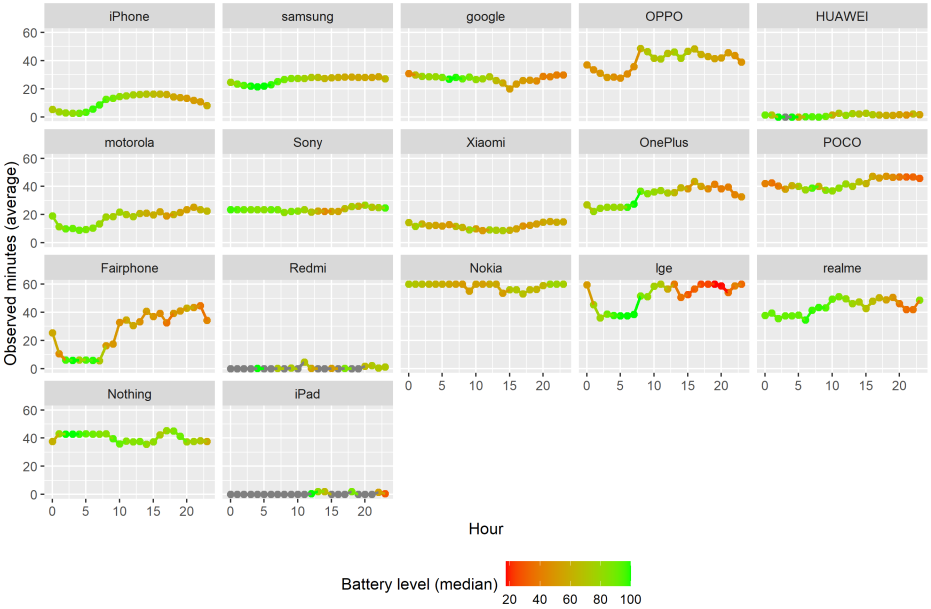


Figure 3.9 X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour split by smartphone brand. Color indicates battery level in %. Panels show different brands.

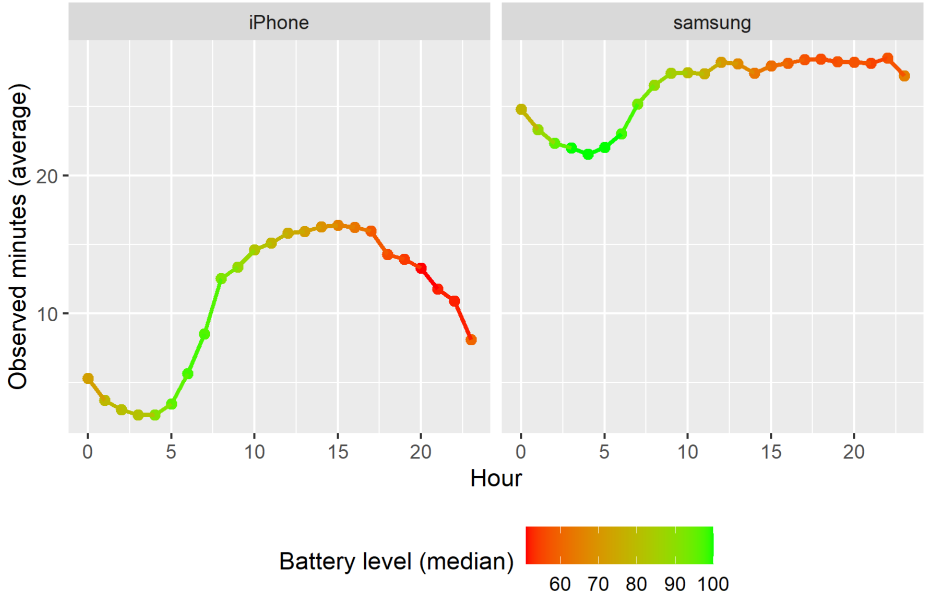


Figure 3.10 X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour split by smartphone brand (iPhone and Samsung only). Color indicates battery level in %. Panels show different brands.

3.4 Accuracy

Figure 3.11 shows the median accuracy per hour during the fieldwork period. The median is computed for each hour individually over those users for which the hour is within their hour range. Each data point in the figure corresponds to one hour. The red and green lines indicate points in time (invitations and reminders) of the different samples. It can be seen that some brands are present in each sample, while others are only present in one sample. In general, low values for accuracy are found for all brands. For only a few hours, some outlier values could be found.

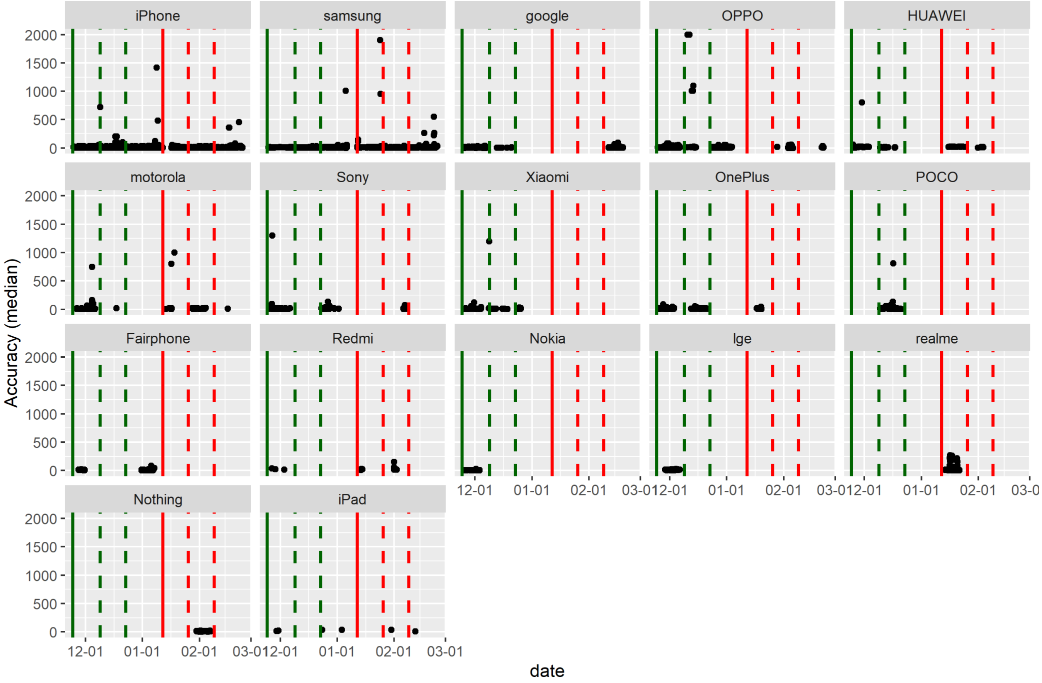


Figure 3.11 X-scale shows days, y-scale shows median accuracy per hour during the fieldwork period, split by smartphone brand. Green lines indicate the invitations for the ‘follow-up sample’ and the ‘full editing sample’ (solid line) and two reminders (dashed lines), and the the red lines indicate the invitation for the ‘limited editing sample’ (solid line) and two reminders (dashed lines).

Figure 3.12 shows the same data as Figure 3.11 but with a focus on the two major brands (iPhone and Samsung). In general, low values for accuracy are found for both brands. For only a few hours, some outlier values could be found. However, for iPhone and Samsung gaps are found during the fieldwork period. These gaps can be explained by known technical maintenance in preparation for the sampling for the ‘limited editing sample’ and do not indicate unknown technical problems in the app.

Figure 3.13 shows the average observed minutes per hour during a day (24h). The line’s color indicates the average measured accuracy throughout the 24 hours. The accuracy is lowest, especially during early morning or late evening/night time. During daytime, on average, larger values are observed. This finding might suggest that during the daytime, when moving/traveling, the accuracy tends to be lower than during the day when there is less moving/traveling.

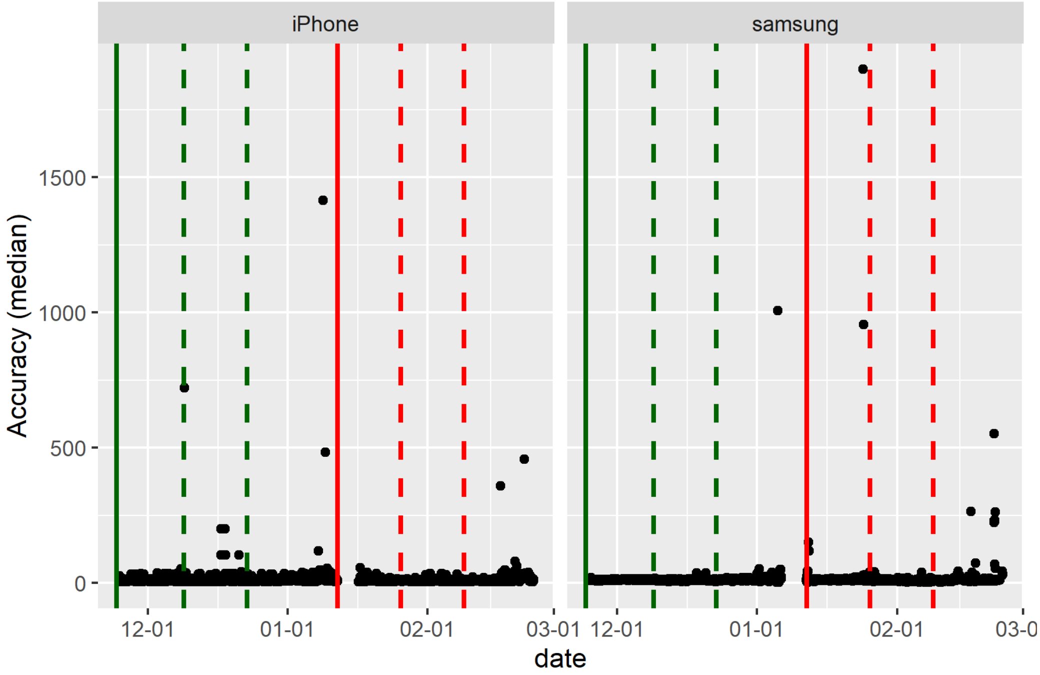


Figure 3.12 X-scale shows days, y-scale shows median accuracy per hour during the fieldwork period, split by smartphone brand (iPhone and Samsung only). Green lines indicate the invitations for the 'follow-up sample' and the 'full editing sample' (solid line) and two reminders (dashed lines), and the the red lines indicate the invitation for the 'limited editing sample' (solid line) and two reminders (dashed lines).

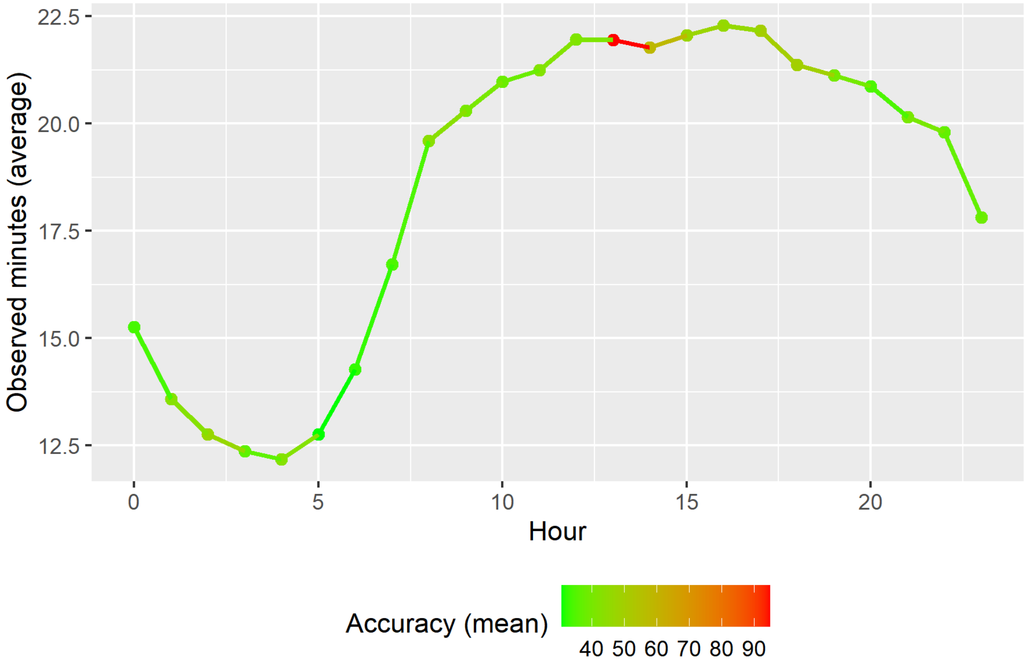


Figure 3.13 X-scale shows 24 hours, y-scale shows the average number of minutes observed per hour. The color indicates average accuracy.

4 User profiles

Next, we evaluate what is the impact of technical deficiencies and whether these relate to the different experimental conditions. Our ultimate goal is to set lower bounds for data quality. We investigate how to set such bounds. We will define two user profiles: insufficient data quality and sufficient data quality. The profiles are based on objective technical characteristics. We will separate the two types of respondents in the evaluation.

4.1 Lower bounds to data quality

Two parameters are chosen to categorize the users into data quality profiles. The number of observed hours (time) and the number of observed geo-locations (space). Table 4.1 shows the number of users classified as insufficient quality and sufficient quality for a given set of parameters. When a user has less than one hour of observed data or less than 2000 observations (measured geo-locations), the data is classified as of insufficient quality. With these parameters, 27% of the users are considered to have data of insufficient quality. When a user has one hour or more of observed data and 2000 or more observations (measured geo-locations), the user is considered to have data of sufficient quality. The effect of parameter choice on the percentage of usable and non-usable users is shown in Figure 4.1. The current parameter choice was based on a visual inspection of the data.

Table 4.1 Data quality user profiles

Hour	Observations	Number users	% Users	Classification
< 1	< 2000	136	27	insufficient quality
≥ 1	≥ 2000	369	73	sufficient quality

The parameter set of 0 hours and a minimum of 0 observations would result in 100% of users with sufficient data quality. However, there are more reasonable choices. As soon as the observations parameter increases, while the time value is held constant at 0, the number of users with sufficient data quality decreases. For example, with a minimum of 250 observations, this value decreases to about 80%. However, 250 observations is a small and likely too optimistic value. When considering the time aspect, this value would need to be increased to more than 10 hours for most parameter sets to lead to a substantial decrease in users with sufficient data quality. For example, when choosing 5 hours instead of 1 hour (see Figure 4.1), there will be no large impact on the number of classified users having insufficient or sufficient data quality. In the most extreme scenario chosen (more than 10000 observations or more than 48 hours of observed data), the number of users with sufficient data quality would drop to about 50%. Generally, there seems to be a tendency for two distinguishable groups: reasonable participation regarding time and space or nearly no participation. This finding might be due to an engagement problem, as a technical problem seems an unlikely cause and is, to the best of our knowledge, not present. For users with insufficient data quality, it needs to be considered whether they can be used in the analyses or whether they are treated, for example, as ‘nonresponse’. If treated as nonresponse, the weighting procedure might correct for this in the estimation procedure.

4.2 Samples and fieldwork variants

Table 4.2 shows the number of users per sample. The row order is based on the number of users. The number of measured geo-locations per user (median) and the number of observed hours

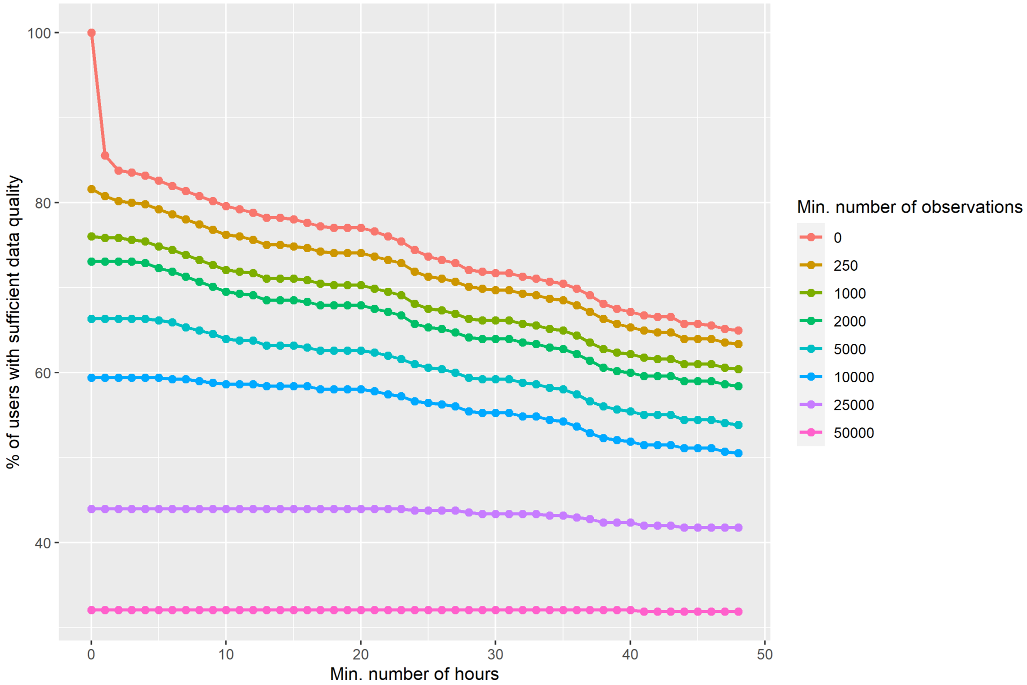


Figure 4.1 Percentage of users labeled as having data of sufficient quality for varying values of the minimum number of observations and the minimum number of hours required for a sufficient quality label.

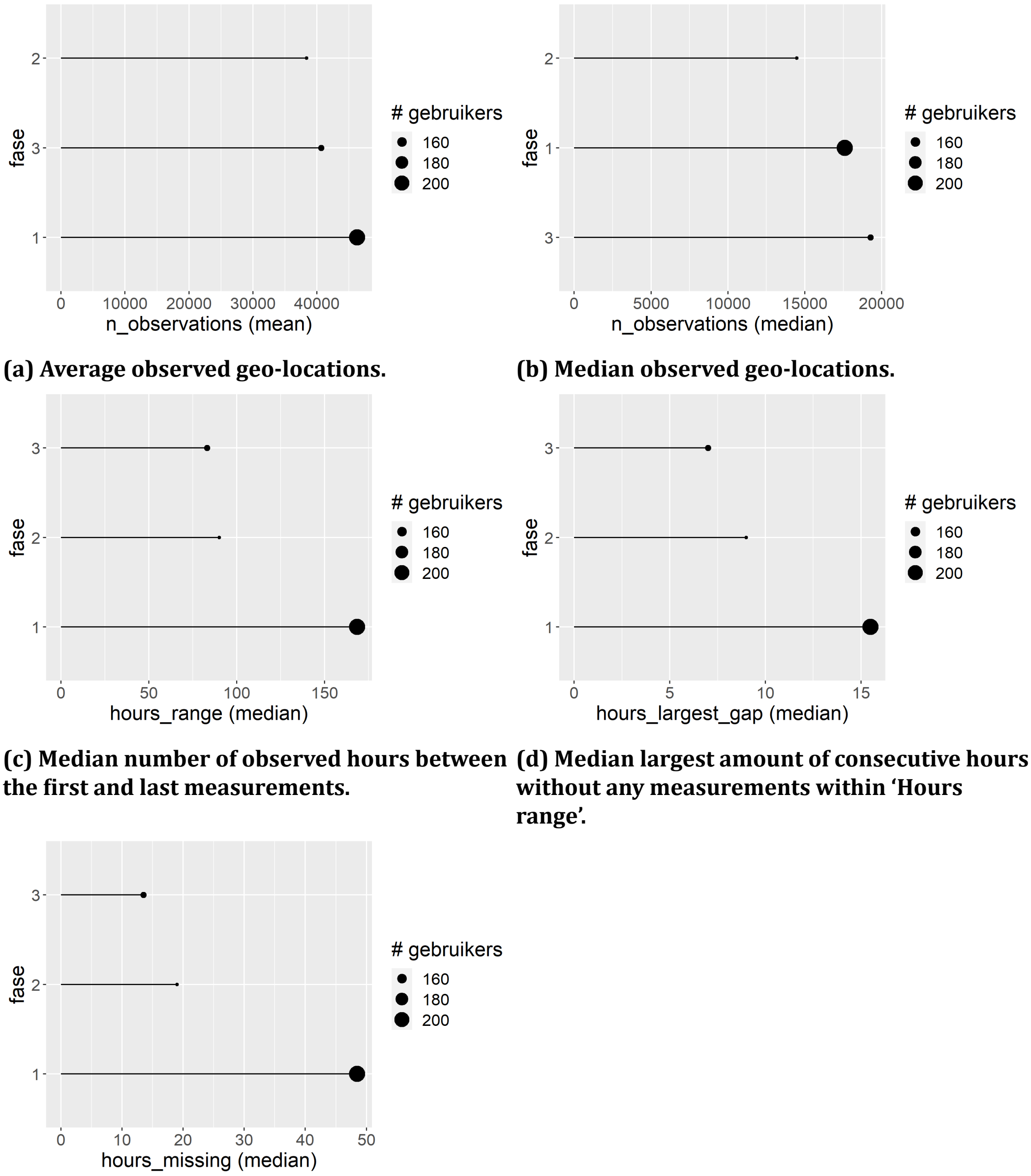
between the first and last measurements for each user (median) are shown. Most observations (median) are made in the 'limited editing sample', followed by the 'follow-up sample' and the 'full editing sample'. These are interesting results, given the different fieldwork lengths between the samples. Moreover, the number and percentage of users with sufficient data quality are shown, which is comparable between samples.

Sample	Users	Observations (median)	Hours range (median)	Sufficient quality (n)	Sufficient quality (%)
'follow-up sample'	212	17601.00	168.50	156	73.58
'limited editing sample'	148	19275.50	83.00	109	73.65
'full editing sample'	145	14483.00	90.00	104	71.72

Table 4.2 Descriptive statistics of geo-location data split by fieldwork sample.

Figure 4.2 shows data quality statistics split by sample. The point size indicates the number of users. All results are ordered increasingly (from top to bottom). The average and median of the observed geo-locations are shown in panels 4.2a and 4.2b of Figure 4.2. The reason for showing both statistics is that the order of samples changes depending on the statistic, and the number of observed geo-locations decreases substantially when the median is used. This finding indicates a skewed distribution of observed geo-locations with some large outliers. That is also why the median is shown in Figures 4.2c – 4.2e. In Figure 4.2c, the median of the range of hours for the 'full editing sample' the ´limited editing sample' is about half of the value for the 'follow-up sample'. A similar ratio is observed in Figure 4.2d for the median of the largest hours gap. Contrarily, in Figure 4.2e, the median of the total number of missing hours of the 'full editing sample' and the 'limited editing sample' is about a third of the value for the 'follow-up sample'.

Figure 4.3 shows the average minutes observed (y-axis) per hour during the day (x-axis) split by samples and variants. The variant '0' is the 'follow-up sample', while the codes '1–6' represent the 'full editing sample' and the ´limited editing sample' with their different variants, see Schouten et al. (2024) for details. Most of the minutes are observed in variants 1, 2, and 3.



(a) Average observed geo-locations.

(b) Median observed geo-locations.

(c) Median number of observed hours between the first and last measurements.

(d) Median largest amount of consecutive hours without any measurements within 'Hours range'.

(e) Median number of hours without any measurements within 'Hours range'

Figure 4.2 Selected results of data quality statistics split by samples (1 = 'follow-up sample', 2 = 'full editing sample', and 3 = 'limited editing sample').

Variants 1 and 3 use one day, and variant 2 seven days. In the other two variants, 4 and 6, which also used seven days, fewer minutes were observed. Variant 5 is the only variant with one day and few observations. It is mainly comparable to variant 0, which is the the 'follow-up sample'.

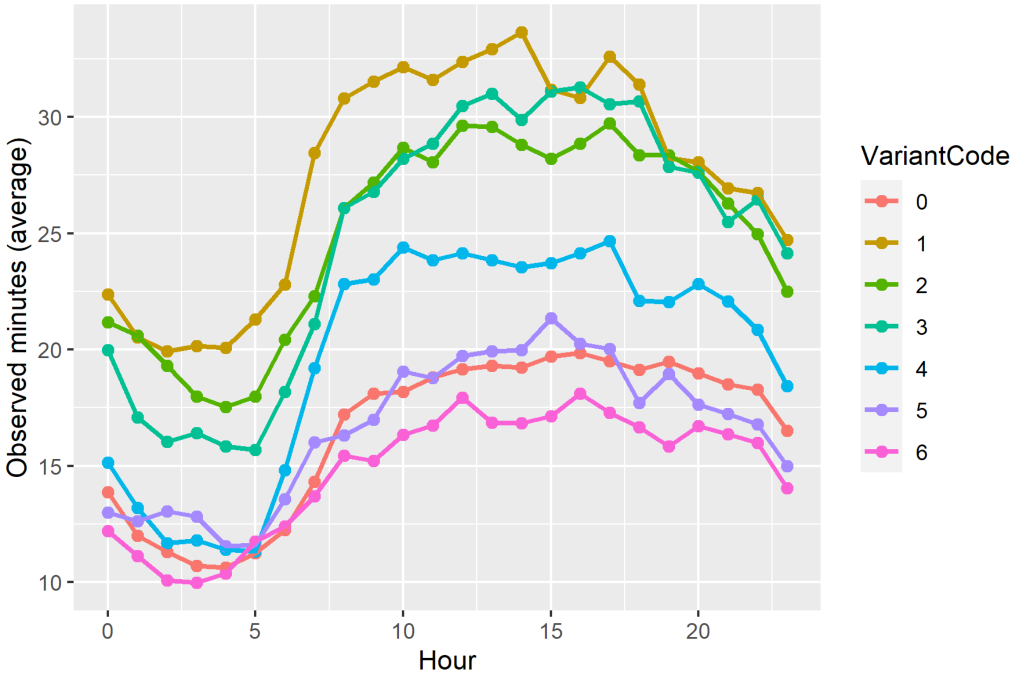


Figure 4.3 Average unique minutes observed (y-axis) during 24 (x-axis) split by fieldwork samples and variants.

The percentage of users with sufficient data quality was analyzed per sample and variant. The results are shown in Figure 4.4. Each panel shows a sample and fieldwork variant respectively. The panel '0' represents the 'full editing sample', while panels '1–6' indicate the 'full editing sample' and the ´limited editing sample' with their different variants. Results show that for variants with one day, an earlier and steeper decrease in users with sufficient quality compared to variants with seven days. Each panel shows a consistent percentage of users with > 25000 observations. Variant 5 shows the lowest percentage of users with > 50000 observations.

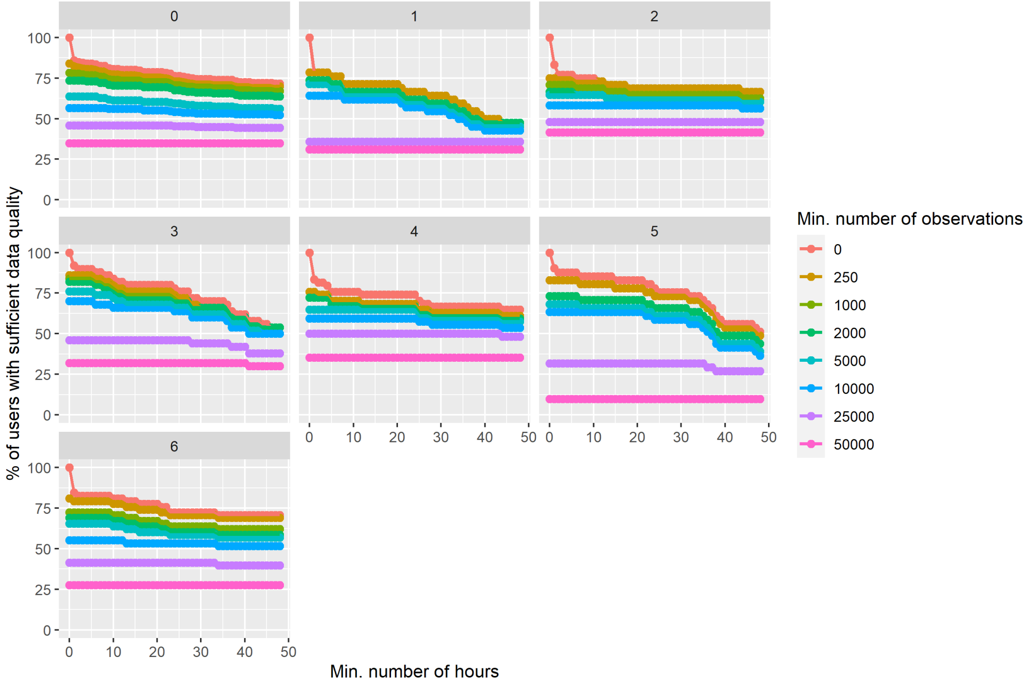


Figure 4.4 Percentage of users labeled as having data of sufficient quality for varying values of the minimum number of observations and the minimum number of hours required for a sufficient quality label, split by sample and variant.

Figure 4.5 shows the percentage of users classified as having sufficient data quality for a given set of parameters split by the two leading brands. For iPhone and Samsung, the 1-day variant shows a steeper and stepwise decrease over time compared to the 7-day variant. Other than that, there seems to be a tendency for the differences between the parameter sets to be larger for iPhone users than for Samsung users.

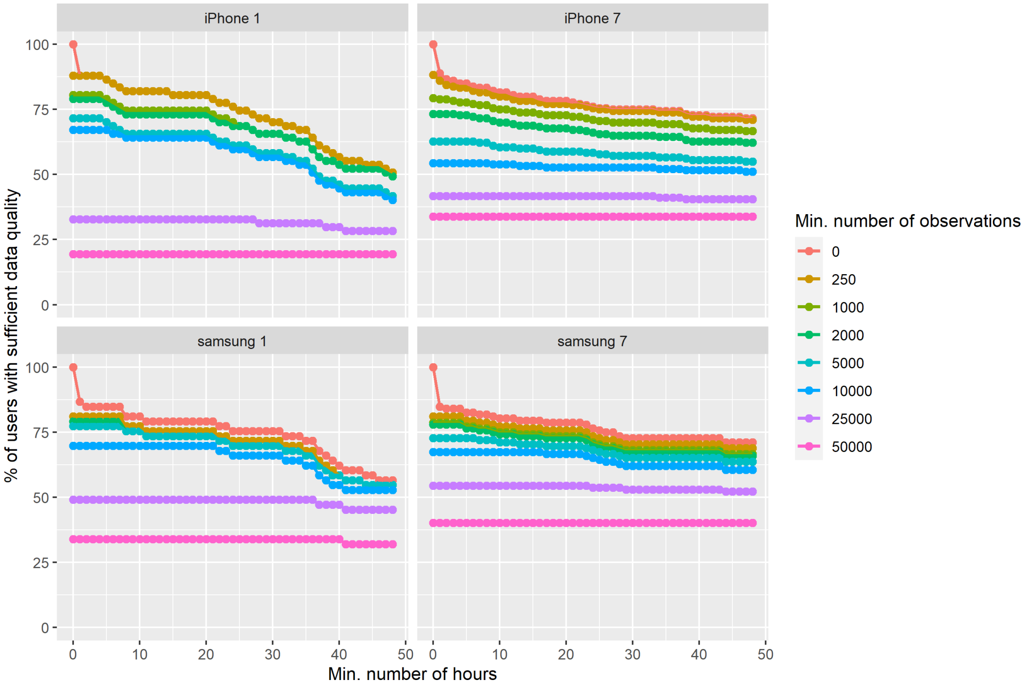


Figure 4.5 Percentage of users labeled as having data of sufficient quality for varying values of the minimum number of observations and the minimum number of hours required for a sufficient quality label, split by iPhone and Samsung phones and the number of days.

5 Data post-survey pre-processing

In the last evaluation, we consider the post-survey pre-processing of location data. So these concern adjustment methods that are applied after any editing by respondents and before making final stop-track segmentations and derivations of travel modes and travel purposes.

This section presents results of the post-processing routines applied to the data and their effects on the data quality. The following post-processing routines were applied: creating a reference period, cleaning manually detected noisy data points, cleaning automatically detected noisy data points, and applying data quality criteria to select eligible users with sufficient data quality.

All data collected were considered for post-processing (505 users, 21.397.699 observations). As a reminder, one observation is the combination of measured geo-location and timestamp. Different combinations of post-processing routines are applied in which the numbers of users (505) and observations (21.397.699) are always considered the starting point. Applying the routines might lead to a decrease in these numbers regarding users and observations.

5.1 Post processing routines

In the following, the applied post-processing routines are described. The order of the sections corresponds to the order of application of the routines. Table 5.1 is a brief overview the routines and their different versions.

Routine	Set of versions	Section
Reference period	{ period 1; period 2; period 3 }	5.2
Manual detection	{ apply; do not apply }	5.3
Automatic detection	{ apply median smoothing and keep points; apply median smoothing and delete points; do not apply smoothing and keep points }	5.3
Quality selection	{ apply; do not apply }	5.4

Table 5.1 Overview of applied post-processing routines and their different versions.

Hence, each scenario consists of four steps: the reference period step, the manual inspection step, the automatic detection step, and the quality selection step. This results in $3 \times 2 \times 3 \times 2 = 36$ scenarios. Not all scenarios are reported. As will be described in Section 5.3, the manual detection is applied or not in these scenarios, but will not be visualized in the results as the effects will hardly be visible. This decision reduces the presented scenarios to $3 \times 3 \times 2 = 18$.

5.2 Reference period

This routine is motivated by the fact that once a respondent installed the app and participated, data was collected continuously until either the end of fieldwork period or until the app was uninstalled by the user. This means that although only one or seven days of participation were asked, it is possible that data was collected beyond this period. Considering the results by the results presented in this discussion paper, there is evidence for large variations in the length of the collected data per user and the amount of data collected per user. For example, the minimum observed hours is 0 (the app was used for less than 60 minutes), and the maximum observed hours is 983 (the app was used for about 41 days). Creating a reference period makes sense from the perspective that these data are to be used to produce an official statistic. Any data collected outside the reference period will not be considered for the statistic. Without such a period, no statements about the proportion of missing data could be made, or it would be impossible to calculate, for example, the average number of trips per *time unit* or the total number of trips by car per *time unit*, as no *total time unit* is defined.

The following rules were used to generate different reference periods.

1. Consider the timestamp of the first observation and add six days (7-day variant), or consider the timestamp of the first observation and consider only this single day (1-day variant). Observations after one or seven days are not considered, depending on the variant.
2. Consider the period of seven days (7-day variant) or one day (1-day variant) in which most observations with classified app data are available.
3. Consider the period of seven days (7-day variant) or one day (1-day variant) in which most observations are available.

By *classified app data* in the second reference period option, we mean the result of the classifications from the in-app algorithm in combination with user annotations (mutations, deletions, and inserts of events). Other versions of the reference period will be addressed in this chapter as well.

5.3 Noisy data

As shown in all previous internal reports (Gootzen and Klingwort 2023; Klingwort and Gootzen 2023b; Klingwort and Gootzen 2023a), there are noisy data points collected that might have a (negative) impact on the start/stop classification of the algorithm (for details about the start start/stop classifications algorithm, see McCool et al. (2021), Gootzen and Klingwort (2022), Oerlemans et al. (2023), and McCool et al. (2024b). Here, we apply two routines that consider noisy data points. First, based on manual data inspection, some data points were identified as noisy. Second, using a smoothing algorithm, noisy data points were identified. Next, we will explain these two routines.

Manual detection

Based upon manual data inspection, noisy points were identified in the data. This routine identified a total of 2.410 noisy data points. This is 0,013% of the data. As shown in Table 5.1, we consider scenarios where these points are deleted and kept. However, given the small amount of manually detected noisy data points, the effect of deleting/keeping the data points is considered negligible. In case the points are kept, it is still possible that these will be identified through the subsequently applied smoothing algorithm (and adjusted or deleted). This post-processing routine is not shown in the results (Section 5.5). Given that this step only concerns about 0,013% of the data, it should be reconsidered whether such a manually routine should be applied.

Automatic detection

Different versions of automatic detection of noisy data points are available: mean smoothing, median smoothing, and Kalman smoothing. The different versions have been studied in Gootzen and Klingwort (2022). Based on previous research and results, it is opted for the median smoothing. As currently implemented, the version adjusts the location (longitude and latitude) of a noisy data point to align with the non-noisy data points. However, the timestamp has not been adjusted at this point. In extreme cases, this might lead to strange results regarding the calculated speed. Due to this potential problem, we implement different versions in the post-processing routines. In one version, we apply the median smoothing and keep the adjusted points; in a second version, we apply the median smoothing and delete the adjusted points; and in a third version, we apply no smoothing.

5.4 Quality selection: user profiles

In Section 4, data quality profiles were introduced. These are based on two parameters: the number of observed hours (time) and the number of observed geo-locations (space). A simulation study determined that a user with less than one hour of observed data or less than 2000 observations is classified as having insufficient/low-quality data. When a user has one hour or more of observed data and 2000 or more observations, the user is considered to have data of sufficient quality and is eligible for analysis. With this parameter set, about 75% of the users are classified as sufficient quality and 25% as insufficient quality. This routine is applied as the last step. We consider scenarios in which this routine is applied and not applied.

5.5 Results

The effects of applying the post-processing routines to the data are shown in Figures 5.1 and 5.2. In Figure 5.1, the effects of the number of observations are shown, and in Figure 5.2, the effects on the number of users. This split is relevant because the different routines are targeted at both these aspects.

In both figures, the x-axis shows the percentage of removed observations and users, respectively. In both figures, the y-axis shows the labels of the post-processing routines. In total, 18 processing scenarios are shown on the y-axis. The bars represent the percentages of data removed, and the coloring of the bars indicates the different processing routines. The reference period of each scenario is indicated by colored dots in front of the bar (the reference period is also included in the label) to improve readability.

In Figure 5.1, the percentage of removed observations by each step is always relative to the total number of observations (21.397.699). First, we focus on the reference period routine. The reference period is a constant in each scenario; therefore, the number of observations removed due to this routine does not differ between scenarios. For reference period 1, about 30% of the observations are removed. For reference period 2, about 20% and for reference period 3, about 8% are removed. In the scenarios where the automatically detected data points are not smoothed but removed, about an additional 27% of the observations are removed. Accordingly, for reference period 1, about 47% of the observations are removed; for reference period 2, about 34% of the observations are removed; and for reference period 3, about 43% of the observations are removed. The quality selection criteria have only a small effect on the number of removed observations. In each scenario where this routine is applied, between 0,2% and 0,5% of the observations are removed.

To sum up, choosing reference period 1 would result in the largest removal of observations from the three considered reference periods. Reference period 3 would result in the second largest removal, and the fewest observations are removed when considering reference period 2. If it is considered to remove suspicious data points and not smooth them, about 27% of observations would be removed. Applying quality selection criteria does not cause substantial removal of observations. This might change if other threshold values are chosen (see section 5.4 for the chosen parameters), although the chosen parameter set seems to be reasonable. When not applying the quality selection routine, no users are removed.

In Figure 5.2, the percentage of removed observations by each routine is always relative to the total number of users (505). First, we focus on the reference period routine. Only in reference period 2 are users removed (about 10%). This is because this period considers the period in which most observations with classified app data are available. However, there are users without classified app data who have been removed. The routine of smoothing data points does not cause any removal of users. Finally, the quality selection routine strongly affects the percentage of removed users. If this routine is applied, up to 35% (reference period 1) of the users are removed, and at least about 21% (reference period 2). Of course, no users are removed when the quality selection routine is not applied.

To sum up, only reference period 2 would cause the removal of users. Variations of smoothing/dealing with noisy data points do not directly cause the removal of users but might cause this indirectly. This means that some users might fall under the threshold used for the quality selection routine due to removing these points. Choosing reference period 1 would result in the largest removal of users (35%). Reference periods 2 and 3 would cause between 20% and 30% removal of users.

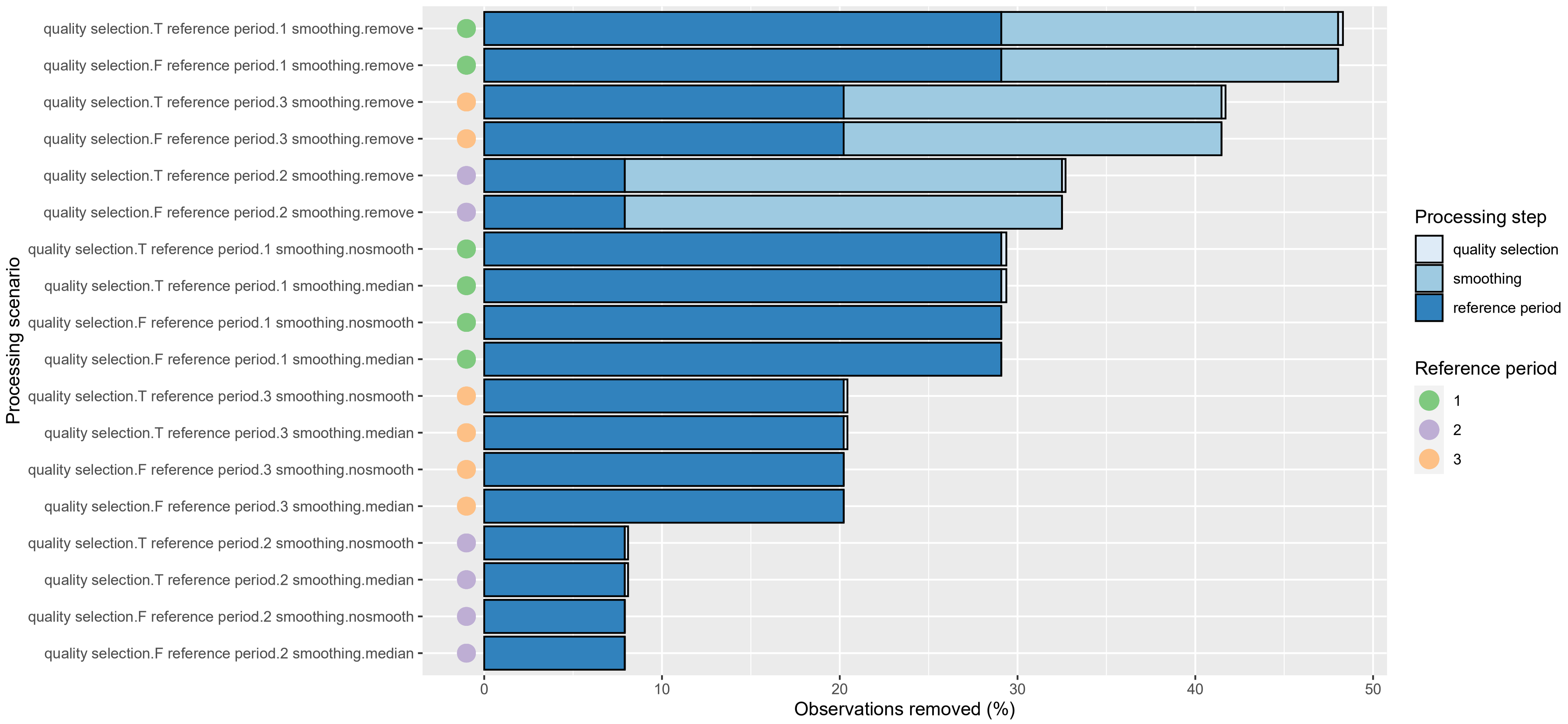


Figure 5.1 **The x-scale shows the percentage of removed observations, and the y-scale shows the processing scenarios. The colored bars indicate the different processing routines, and the colored dots indicate the different reference periods.**

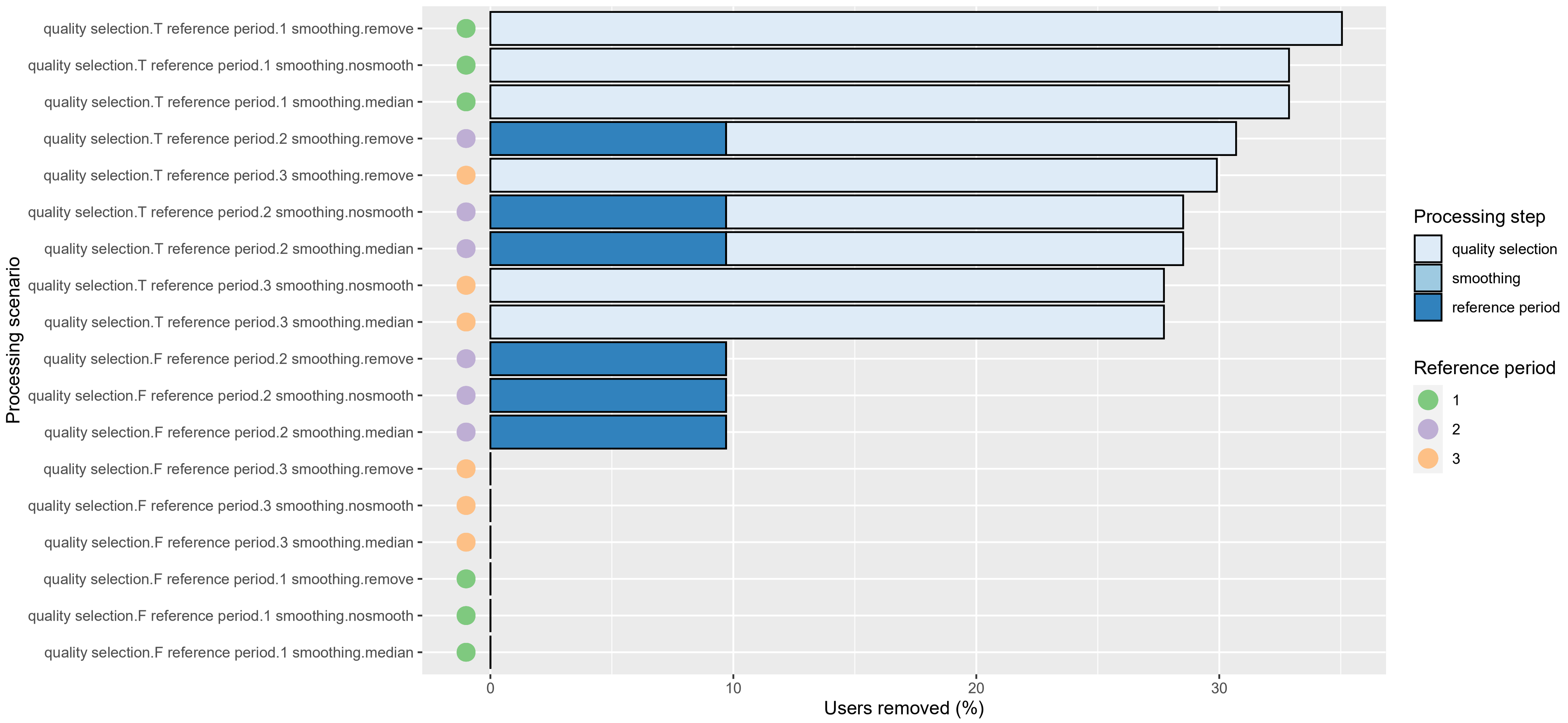


Figure 5.2 **The x-scale shows the percentage of removed users, and the y-scale shows the processing scenarios. The colored bars indicate the different processing routines, and the colored dots indicate the different reference periods.**

In summary, this analysis focused on evaluating different post-processing routines and their effects on the data in terms of observations and users. We considered four post-processing routines: defining a reference period, manually detecting noisy data, automatically detecting noisy data, and quality selection. Results show that some of the routines affect the removal of observations and users stronger than others. The reference period and the quality selection routines have a large impact. When choosing a reference period based on day or week patterns (reference period 1 in this report), most observations and most users were deleted. However, fewer observations and users were removed when opting for a somewhat more user-oriented reference period, such as reference period 2 or 3. When setting up the study design, these findings indicate that flexibility and user orientation could be considered instead of specifying a fixed start. However, when solely leaving the choice to the users, there is the risk of introducing bias because users might use the app in periods with more than the usual travel/moving behavior. In Table 5.2, we show the distribution of the starting day. That is the day on which the first observation was made. Users tend to start using the app around the end of the week. Most started on a Friday, and Friday to Sunday comprise about 53% of all starting days. Table 5.3 shows the distribution of the starting day split by the different day variants (one day and seven days). The 1-day variant shows less variation per day compared to the 7-day variant. In the 7-days variant, most users started on a Friday (nearly 4-times more than on a Wednesday or about 3-times more than on a Tuesday). This finding suggests that assigning a starting day leads to a more equal spread of days over the week. Table 5.4 shows the distribution of the starting day split by the data quality indicator (insufficient and sufficient data quality). There seems to be no apparent relation between the starting day and the data quality.

Mo.	Tue.	Wed.	Thu	Fr.	Sa.	Sun.
70 (14%)	46 (9%)	47 (9%)	73 (14%)	122 (24%)	75 (15%)	72 (14%)

Table 5.2 Distribution of weekdays for all 505 users on which the first observation was recorded. Percentages are row percentages.

Variant	Mo.	Tue.	Wed.	Thu	Fr.	Sa.	Sun.
1 day	18 (14%)	16 (12%)	22 (17%)	18 (14%)	24 (18%)	12 (9%)	23 (17%)
7 days	52 (14%)	30 (8%)	25 (7%)	55 (15%)	98 (26%)	63 (17%)	49 (13%)

Table 5.3 Distribution of weekdays for all 505 users on which the first observation was recorded, split by 1 vs day variants. Percentages are row percentages.

Variant	Mo.	Tue.	Wed.	Thu	Fr.	Sa.	Sun.
Insufficient quality	18 (13%)	11 (8%)	12 (9%)	18 (13%)	31 (23%)	21 (15%)	25 (18%)
Sufficient quality	52 (14%)	35 (9%)	35 (9%)	55 (15%)	91 (25%)	54 (15%)	47 (13%)

Table 5.4 Distribution of weekdays for all 505 users on which the first observation was recorded, split by data quality indicator. Percentages are row percentages.

We list some points to stimulate potential discussions. First, other reference periods can be defined, which we list below.

1. Invitation calendar day: The baseline reference period is equal to the starting calendar day as requested in the invitation letter.
2. Invitation calendar day plus 24h: Like 1, but extended to the first 24 hours following registration.
3. Actual starting day: The first calendar day confirmed/validated by the respondents in the app.
4. Actual starting day plus 24h: Like 3, but extended to the first 24 hours after registration.
5. Optimal overlap: The period of maximal overlap between the app and the diary data.

Second, we also mentioned the issue of the smoothing algorithms that are considered only to adjust space but not time. This problem has yet to be studied within the project. If smoothing algorithms are applied in the future, we recommend investigating this issue. Third, the relevance of using some of the routines, such as the automatical smoothing, needs to be discussed, given that a new app version is in development.

6 Conclusion and future work

Location tracking is a promising feature for travel surveys, passenger mobility surveys, and other time-use surveys where travel is relevant. However, location tracking comes with technological and, consequently, methodological challenges. These may be relaxed by continuous monitoring of technological advances and sophisticated use of sensor options. However, we expect both respondent involvement and advanced post-survey adjustment routines to be imperative.

In order to get a sense of what bounds of data quality to expect and what follow-up actions are needed, we assessed location data quality in this report. We evaluated relative to the technical specifications and various experimental conditions embedded in a large field study.

We conclude that location data tracking technology still showed various deficiencies, especially for specific brands. This may, in part, be caused by the decision to simultaneously record three types of sensor measurements during the field study. This choice affected battery usage and, consequently, missing data. However, also without this choice, we suspect that battery management demands further tailoring and continuous re-evaluation when new models/operating systems are launched.

We evaluated a range of lower bounds on data quality and their consequences for the proportion of eligible respondent tracking days. In our study, the most natural choice would imply that around a fifth of records are discarded or, at best, require large-scale imputations.

Finally, we considered several post-survey pre-processing options. This assessment must be considered preliminary since we have not yet confronted the options with respondent edit scenarios. This trade-off between respondent involvement and advanced AI-ML and imputation routines remains an area for further research.

We see further tailoring of tracking routines and motivated trade-offs between respondent editing and adjustment methods as the main future steps. With respect to the latter, research is ongoing on the potential of missing data imputation and respondent editing.

Acknowledgments

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

We like to thank Chris Lam and Coen van Heukelingen for useful comments on an earlier version of this manuscript.

References

Assemi, B., H. Jafarzadeh, M. Mesbah, and M. Hickman (2018). “Participants’ perceptions of smartphone travel surveys.” In: *Transportation Research Part F: Traffic Psychology and Behaviour* 54, pp. 338–348. DOI: 10.1016/j.trf.2018.02.005.

Faghih Imani, A., C. Harding, S. Srikukenthiran, E. J. Miller, and K. Nurul Habib (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.” In: *Transportation Research Record* 2674.7, pp. 299–311. DOI: 10.1177/0361198120921860.

Gillis, D., A. Lopez, and S. Gautama (2023). “An Evaluation of Smartphone Tracking for Travel Behavior Studies.” In: *ISPRS International Journal of Geo-Information* 12.8, p. 335.

Gootzen, Y. and J. Klingwort (2022). *Optimal parameter choice and sensor selection for ODiN app-algorithm: a simulation study. Internal CBS Memorandum*.

Gootzen, Y. and J. Klingwort (2023). *Data quality exploration AVA22: Phase 1. Internal CBS Report. 26-01-2023*.

Harding, C., A. Faghih Imani, S. Srikukenthiran, E. Miller, and K. Nurul Habib (2021). “Are we there yet? Assessing smartphone apps as full-fedged tools for activity-travel surveys.” In: *Transportation* 48, pp. 2433–2460.

Klingwort, J. and Y. Gootzen (2023a). *Data quality report: AVA22 phases 1–3. Extended Analysis. Internal CBS Report. 16-06-2023*.

Klingwort, J. and Y. Gootzen (2023b). *Data quality report: AVA22 phases 1–3. Internal CBS Report. 25-05-2023*.

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, and V. Toepoel (2025b). *How smart are smart travel surveys? Evaluating trip segmentation, travel motive, and travel mode predictions*. Conference paper. Mobile Apps and Sensors in Surveys (MASS).

Klingwort, J., Y. Gootzen, D. Remmerswaal, and B. Schouten (2025c). “Algorithms versus survey response: Comparing a smart travel and mobility survey with a web diary.” In: *Transportation Research Interdisciplinary Perspectives* 31, p. 101436. ISSN: 2590-1982. DOI: https://doi.org/10.1016/j.trip.2025.101436.

Klingwort, J. and R. Schnell (2020). “Critical Limitations of Digital Epidemiology: Why COVID-19 Apps Are Useless.” In: *Survey Research Methods* 14.2, pp. 95–101. DOI: 10.18148/srm/2020.v14i2.7726.

Lawson, C., E. Krans, E. Rentz, and J. Lynch (2023). “Emerging trends in household travel survey programs.” In: *Social Sciences and Humanities Open* 7, p. 100466.

Lunardelli, I., J. van den and Heuvel, B. Schouten, B. D’Amen, B. Loré, A. Nuccitella, M. Perez, and M. Zgonec (2024). “How do respondents think about surveys with smart features?” In: *Deliverable 1.2 of project Smart Survey Implementation, Statistics Netherlands*.

Maruyama, T., Y. Sato, K. Nohara, and S. Imura (2015). “Increasing Smartphone-based Travel Survey Participants.” In: *Transportation Research Procedia* 11, pp. 280–288. DOI: 10.1016/j.trpro.2015.12.024.

McCool, D., P. Lugtig, O. Mussmann, and B. Schouten (2021). “An App-Assisted Travel Survey in Official Statistics: Possibilities and Challenges.” In: *Journal of Official Statistics* 37.1, pp. 149–170. DOI: 10.2478/jos-2021-0007.

McCool, D., P. Lugtig, and B. Schouten (2024a). “Longitudinal smartphone data for general population mobility studies.” In: *Journal of Official Statistics* 37.1.

McCool, D., P. Lugtig, and B. Schouten (2024b). “Maximum interpolable gap length in missing smartphone-based GPS mobility data.” In: *Transportation* 51.1, pp. 297–327. DOI: 10.1007/s11116-022-10328-2.

Oerlemans, T., T. Schijvenaars, and M. Vollebregt (2023). *Documentatie ODiN App-variant 27-02-2023*.

Remmerswaal, D., B. Schouten, J. Bakker, J. van den Heuvel, and J. Klingwort (2025). “A smart travel survey – What is the role of the respondent?” In: *Discussion paper, Statistics Netherlands*.

Safi, H., B. Assemi, M. Mesbah, and L. Ferreira (2017). “An empirical comparison of four technology-mediated travel survey methods.” In: *Journal of Traffic and Transportation Engineering* 4.1, pp. 80–87. DOI: 10.1016/j.jtte.2015.12.003.

Schouten, B., D. Remmerswaal, A. Elevelt, J. de Groot, J. Klingwort, T. Schijvenaars, M. Schulte, and M. Vollebregt (2024). “A smart travel survey: Results of a push-to-smart field experiment in the Netherlands.” In: *Discussion paper, Statistics Netherlands*.

Smeets, L., P. Lugtig, and B. Schouten (2019). “Automatic Travel Mode Prediction in a National Travel Survey.” In: *Discussion paper, Statistics Netherlands*.

Struminskaya, B., V. Toepoel, P. Lugtig, M. Haan, A. Luiten, and B. Schouten (2020). “Understanding Willingness to Share Smartphone-Sensor Data.” In: *Public Opinion Quarterly* 84.3, pp. 725–759. DOI: 10.1093/poq/nfaa044.

Winkler, C., A. Meister, U. Isenschmid, K. Lerdo de Tejada Acosta, B. A. Le, and K. W. Axhausen (2023). “Time Use+. Field Report and Lessons Learned.” In: *ETH Zürich, Working paper 2023-10*. DOI: https://doi.org/10.3929/ethz-b-000634863.

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

Appendices

A Extended tables

Number of days	brand	n_users	sum_users	percentage (per day)
1	OPPO	3	133	2
1	motorola	1	133	1
1	HUAWEI	3	133	2
1	Redmi	1	133	1
1	google	1	133	1
1	OnePlus	2	133	2
1	iPhone	67	133	50
1	Sony	1	133	1
1	iPad	1	133	1
1	samsung	53	133	40
7	Fairphone	2	372	1
7	lge	1	372	0
7	google	4	372	1
7	motorola	8	372	2
7	HUAWEI	10	372	3
7	Nothing	1	372	0
7	iPad	4	372	1
7	OnePlus	5	372	1
7	iPhone	180	372	48
7	OPPO	11	372	3
7	POCO	2	372	1
7	realme	1	372	0
7	Nokia	1	372	0
7	Redmi	2	372	1
7	Xiaomi	5	372	1
7	samsung	132	372	35
7	Sony	3	372	1

Table A.1 Extended version of Table 3.2. Cross-table of variants days (1 day and 7 days) and smartphone brand. The percentage of brands is based on number of users and calculated for both 1 and 7 days.

Variant	brand	n_users	sum_users	percentage
0	Fairphone	2	212	1
0	google	3	212	1
0	iPhone	106	212	50
0	OPPO	4	212	2
0	HUAWEI	4	212	2
0	samsung	75	212	35
0	iPad	1	212	0
0	Redmi	1	212	0
0	POCO	1	212	0
0	lge	1	212	0
0	motorola	3	212	1
0	Xiaomi	4	212	2
0	OnePlus	5	212	2
0	Sony	2	212	1

Table A.2 Extended version of Table 3.5. 'follow-up sample' users and smartphone brand.

Variant	brand	n_users	sum_users	percentage
1	samsung	16	42	38
1	OnePlus	1	42	2
1	iPhone	23	42	55
1	iPad	1	42	2
1	google	1	42	2
3	HUAWEI	1	50	2
3	OnePlus	1	50	2
3	OPPO	3	50	6
3	iPhone	25	50	50
3	samsung	19	50	38
3	Redmi	1	50	2
5	motorola	1	41	2
5	Sony	1	41	2
5	samsung	18	41	44
5	iPhone	19	41	46
5	HUAWEI	2	41	5

Table A.3 Extended version of Table 3.6. 'full editing sample' and 'limited editing sample' users, fieldwork variants 1, 3, 5, and smartphone brand.

Variant	brand	n_users	sum_users	percentage
2	HUAWEI	3	48	6
2	iPad	1	48	2
2	iPhone	23	48	48
2	samsung	15	48	31
2	Nothing	1	48	2
2	motorola	2	48	4
2	OPPO	3	48	6
4	OPPO	2	54	4
4	HUAWEI	1	54	2
4	iPad	1	54	2
4	motorola	1	54	2
4	samsung	25	54	46
4	iPhone	23	54	43
4	Nokia	1	54	2
6	iPhone	28	58	48
6	iPad	1	58	2
6	google	1	58	2
6	HUAWEI	2	58	3
6	motorola	2	58	3
6	OPPO	2	58	3
6	POCO	1	58	2
6	Redmi	1	58	2
6	realme	1	58	2
6	Xiaomi	1	58	2
6	samsung	17	58	29
6	Sony	1	58	2

Table A.4 Extended version of Table 3.7. ‘full editing sample’ and ‘limited editing sample’ users, fieldwork variants 2, 4, 6, and smartphone brand.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague, Netherlands
www.cbs.nl

Prepress
Statistics Netherlands Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70
Via contact form: www.cbs.nl/information