Hindawi
Mobile Information Systems
Volume 2017, Article ID 4043237, 11 pages
https://doi.org/10.1155/2017/4043237

Research Article

Assessment of Smartphone Positioning Data Quality in the Scope of Citizen Science Contributions

Angel J. Lopez,[1,2] Ivana Semanjski,[1] Sidharta Gautama,[1] and Daniel Ochoa[2]

[1] *Department of Telecommunications and Information Processing, Ghent University, St-Pietersnieuwstraat 41, 9000 Ghent, Belgium*
[2] *Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral (ESPOL), Campus Gustavo Galindo, Km 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador*

Correspondence should be addressed to Angel J. Lopez; angel.lopez@ugent.be

Received 24 February 2017; Revised 8 May 2017; Accepted 24 May 2017; Published 21 June 2017

Academic Editor: Liang Chen

Human travel behaviour has been addressed in many transport studies, where travel survey methods have been widely used to collect self-reported insights of daily mobility patterns. However, since the introduction of Global Navigation Satellite Systems (GNSS) and more recently smartphones with built-in GNSS, researchers have adopted these ubiquitous devices as tools for collecting mobility behaviour data. Although most studies recognize the applicability of this technology, it still has limitations. These are rarely addressed in a quantified manner. Often the quality of the collected data tends to be overestimated and these errors propagate into the aggregated results providing incomplete knowledge of the levels of confidence of the results and conclusions. In this study, we focus on the completeness aspects of data quality using GNSS data from four campaigns in the Flanders region of Belgium. The empirical results are based on mobility behaviour data collected through smartphones and include more than 450 participants over a period of twenty-nine months. Our findings show which transport mode is affected the most and how land use affects the quality of the collected data. In addition, we provide insights into the time to first fix that can be used for a better estimation of travel patterns.

1. Introduction

Understanding human travel behaviour lies at the core of planning of transport services and ensuring development of sustainable communities. Traditionally, this data is collected based on self-reported insights into a person's daily mobility patterns. The self-reporting is usually done in the form of travel surveys or interviews. The drawbacks of these methods are well recognized in literature and include, among others, underreporting of short trips [1–3], overestimation of public transport travel times or underestimation of car travel times [4–6], obtaining incomplete and inconsistent information [7, 8], and rounding travel times and distances [9]. Recently, the availability of affordable Global Navigation Satellite Systems (GNSS) devices and mobile communication has presented a new way of acquiring data for mobility studies, including citizen science wherein volunteers participate in some aspects of mobility and environmental matters [10], among others.

Some examples of the GNSS data applicability to better understand individuals' daily mobility behaviour include implementation of GNSS data, together with geographic information system (GIS) technology and an interactive web-based validation application, to derive and validate trip purposes and transport modes [11]. Achieved results showed good match with the national travel survey findings indicating that the suggested approach might be promising alternative to paper diary based methods. Similarly, Zheng et al. [12] use GNSS data of 45 users, collected over six months' period, and apply supervised learning based approach to automatically infer transport mode and gain understanding on how to recognize transport mode exchange locations. Furthermore, Liu et al. [13] and Pan et al. [14] take a deeper look at spatial and temporal patterns of taxis' trajectories to

investigate interurban land use variations. They classify the study area into six traffic areas closely associated with various land use types (e.g., commercial, industrial, residential, institutional, and recreational) and find that human mobility data collected from location aware devices provides opportunity to derive urban land use information in a timely fashion. Using a similar dataset, Guo et al. [15] take a look at taxi tracks to derive trips' origin and destination locations. Hood et al. [16] develop a route choice model based on the GNSS data collected from smartphone users in San Francisco. They extract alternatives using repeated shortest path searches in which both link attributes and generalized cost coefficients were randomized. Their results show that bicycle lanes were preferred to other facility types, especially by infrequent cyclists. The obtained results are intended to be used for bike network infrastructure planning purposes. Duncan et al. [17] examine use of the GNSS data in objectively measuring and studying the relationship of environmental attributes to human behaviour in terms of physical activity and transport-related activity. Mavoa et al. [18] examine children's independent mobility data and implement sequence alignment to match GNSS and travel diary data. They successfully matched around 60% of all trips in between two datasets. Murakami and Wagner [19] explore the potential of validated GNSS tracks to improve self-reported trip data. By comparing the GNSS tracks and travel diaries, they found that, in general, self-reported trip distances were longer than those observed from the GNSS tracks.

Although most studies recognize the potential of GNSS based approaches for understanding human travel behaviours, it does have its limitations [7, 19]. These are rarely addressed in more detail or a quantified manner. Main reason for this is a general agreement that the GNSS data are more detailed than the self-reported insights, have higher spatial and temporal resolution, have high potential to replace traditional data collection approaches, and thus are a better basis for drawing conclusions on individuals' travel behaviour. Furthermore, the ground truth data that can be used to evaluate the quality of the GNSS based insights are rarely collected. There is a lack of awareness about the data quality of the raw sensor data and how their errors propagate into results. This gives an incomplete view on the levels of confidence of the results and conclusions. The aim of this paper is to deepen the understanding on the GNSS data applicability for mobility studies by providing systematic and quantified insights into collected data representativeness in describing individuals' travel behaviour. To do so, we will report on different GNSS based datasets that include trips made with various transport modes. For these datasets, we provide extensive description of the data collection process, as well as trips' related context, and report on how well they capture the actual travel behaviour. To describe the actual travel behaviour, we rely on data collected by independent sensors and/or GIS based validation procedure. By doing so, we hope to fill in some of the gaps recognized in the existing literature, raise the level of awareness about relevance of data quality reporting for GNSS based travel behaviour studies, and provide valuable reference for future research in this field.

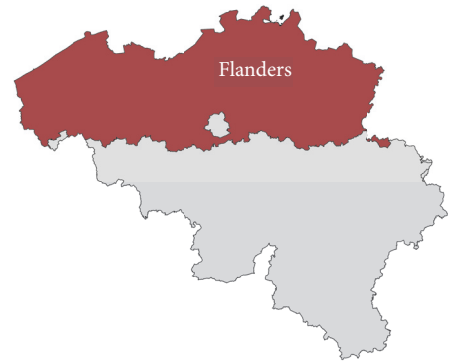

Figure 1: Flanders region in Belgium.

2. Method and Data

This paper presents an empirical analysis of the crowd sourcing data collected from four different campaigns that were launched in the Flanders, Belgium (Figure 1). The campaigns were part of different regional and European projects, each one with specific purposes, that are described further in Section 2.3. A common denominator across all projects is one of the data collection methods. This data collection method involved a smartphone application used to collect, among other information, GNSS traces from the participants.

The target population in this study is a generic population and the focus was on sustainable mobility and the transport mode that people employ for commuting and other daily trips. Furthermore, the aggregation level is at trip segment, where trip segment is a trajectory in which a single transport mode is used [20]. Hence, one multimodal trip may contain several trip segments, each travelled by different transport mode. Features of the study area and target population are shown in Table 1.

A description of the reported modes from the participants is depicted in Table 2. In this transport mode categorisation, passenger and driver utilise the same transport mode (e.g., car) but with different roles (i.e., one is driving the vehicle and one is not).

2.1. Data Representativeness. In order to have a representative set of mobility behaviour data for whole Flanders population (Figure 1), which is not biased for a single transport mode (i.e., only trips performed with a one specific transport mode), data from four campaigns are combined all together to create a more diverse dataset.

The resulting dataset follows a trip modal split (Figure 2 and Table 3) that is similar to the travel behaviour study conducted by the Department of Mobility and Public Works (Flemish Government). The Flemish research started in 1994 and it is known in Dutch as Onderzoek Verplaatsingsgedrag (OVG), which stands for "Travel Behaviour Survey." The study examines the mobility characteristics of families and individuals and focuses on the mobility behaviour of the Flemish [21]. OVG 4.5 is the last edition of that study and it covers the period from September 2012 to September

Table 1: Study area and target population.

Study area	Flanders
Area size (square kilometres)	13,522, source: National Committee of Geography of Belgium
Population in study area[1]	6,444,127
Target population	General
Transport modes	Foot, bike, drive, passenger, bus, tram, train, and moto
Trip activity	Commuting, business related, and recreational trips
Data collection period	Feb 2013–Jul 2015
Number of GNSS locations	10,048,552
Number of travelled kilometres	71,359
Trip segments	8,851
Devices (participants)[2]	457

[1]Eurostat. [2]A device or smartphone does not necessarily map a single user, since some devices were shared among the participants.

Table 2: Transport modes collected from the citizen science.

Transport mode	Description
Foot	Participant goes on foot
Bike	Bicycle/e-bike as a transport mean
Driver	Participant as a car driver
Passenger	Participant as a car passenger
Bus	Bus as a transport mean
Tram	Tram as a transport mean
Train	Rail as a transport mean
Moto	Two-wheeler vehicle such as scooter and moto

Table 3: Trip modal split comparison.

Transport mode	OVG 4.5	Dataset
Driver	51.9%	46.7%
Passenger	16.9%	3.1%
Bike	12.8%	30.1%
Train	1.7%	2.5%
Public transport[1]	3.5%	3.6%
Foot	10.8%	13.3%
Others	2.4%	0.7%

[1]Public transport merges data from modes: bus, tram, and metro.

2013 [22]. Therefore, we contrast our dataset with the OVG 4.5 and, apart from passenger and bike modes, all other modes are aligned with the official results. Differences in these two modes can be explained by the way that OVG figures are calculated. The figures are based on the main transport mode; thus subsequential modes are not considered (e.g., going to work may involve modes such as bike, train, and foot, although only train may be reported as a main mode). Nevertheless, our dataset captures all the modes present in the official results of the Flemish region, and, what is more, it reports similar average travelled distance with another study in the region [23], in which passenger and bike modes reported 17 and 4 km, respectively, and in our dataset those modes are 21 and 5 km. Considering this, combined dataset from all four projects is seen as good representation of overall population mobility behaviour.

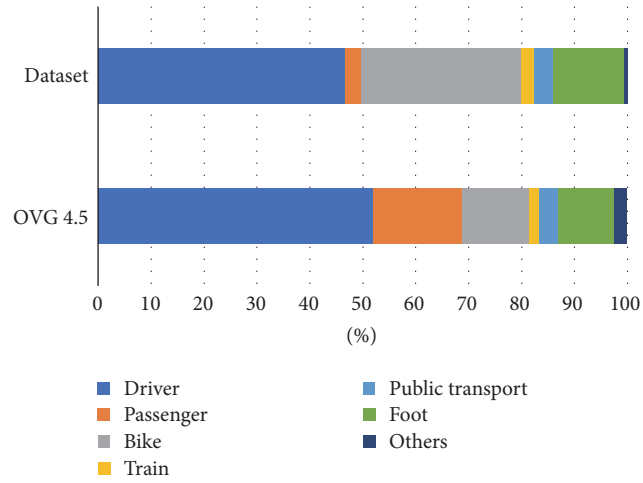

Figure 2: Modal split per trip segment of the crowd sourcing data (dataset) and the Flemish Travel Behaviour Survey (source: OVG 4.5 http://www.mobielvlaanderen.be/ovg/) (OVG 4.5).

2.2. Mobile Applications. Data on the campaigns is collected via two Android smartphone applications *Connect* [24] and *Routecoach* [25], which are developed at Ghent University in Belgium. The applications are part of MOVE, a smart city platform for supporting more sustainable mobility behaviour [26]. *Connect* is a mobile application to collect mobility behaviour data. It has two operating ways for gathering data: active and passive mode. Active mode requires the user's intervention to annotate the trip segment information such as purpose, transport mode, and starting/ending of the trip (Figure 3(a)), whereas, in passive mode, data is collected in background without requiring any action of the user. While logging data, trip segments are automatically detected based on stayed points (zones with no movement) and transport modes are classified using sensor data (GNSS and accelerometer sensor); however, the trip purpose is not inferred, but users can annotate their trips later on. In either active mode or passive mode, users can review their travel diary (Figure 3(b)) and correct whether it is needed (e.g., add a trip purpose). By default, *Connect* operates in passive mode but switches to active mode when a trip is started

FIGURE 3: Mobile applications to collect mobility behaviour data: (a) *Connect* trip annotation; (b) *Connect* travel diary; (c) *Routecoach* route suggestion; (d) *Routecoach* leader board.

manually, which means launching the application, filling in the data, and pressing the play button. Thus, there are no visual changes in the graphical user interface (GUI) in either active or passive mode.

Another feature of this application is that it sends survey questionnaires to test the user; such questionnaires are triggered based on some events like the use of certain transport mode [27]. For instance, questions regarding driving behaviour (frequency, accidents, and traffic) can be triggered after validating a certain amount of car mode trips, but such questions are not asked to nonfrequent car users who may under/overestimate some situations; another example of using the survey questionnaires is to collect user's demographic information, and, in this case, a survey is launched once at the beginning of the campaign. The questionnaires are configurable and their content is in function of mobility campaign (i.e., type of behaviour to be captured). Once the survey is filled in by the user, it is no longer triggered. Besides, *Connect* is developed under strict privacy guidelines; consequently, no registration is required to use the application.

The second application, *Routecoach*, shares similar features with *Connect*, but it adds route coaching and gaming features. A main goal of this application is to provide routing information between a defined pair origin/destination points (Figure 3(c)). To do that, *Routecoach* fetches routing information from multiples sources (e.g., the national rail company and the public transport company) to get schedules and routes, which are combined with other transport modes (walk, bike, and car) to suggest a set of feasible routes. Moreover, gaming features were used in the campaign

Table 4: A summary of the datasets per campaign.

	(i)	(ii)	(iii)	(iv)
Number of GNSS locations	3533752	3549218	1365198	1600384
Number of multimodal trips	2251	2463	1315	1803
Number of trip segments	2394	2738	1814	1905
Number of devices	40	18	19	380
Travelled kilometres	24737	22695	11063	12864
Data collection period	Jan–Sep 2014	Feb 2013–Jul 2014	Mar–Jun 2014	Jul 2014–Oct 2015
Application name	Connect	Connect	Connect	Routecoach
Type of logging	active	active	active	active
Sampling frequency	1 Hz	1 Hz	1 Hz	1 Hz
Sensor	GNSS	GNSS	GNSS	GNSS/FUSED[1]

[1] FUSED locations are not included in this study; a FUSED location is an estimation of the location based on the combination of various sensors such as WIFI, cell networks, GNSS, and Bluetooth.

to reward the most active users and encourage others to participate actively, for example, leader boards for the most biked kilometres (Figure 3(d)) and sustainability challenge board where friends could challenge each other to walk more kilometres during a week. More detailed description of the app and the campaign can be found in [27, 28].

It is worth mentioning that an active mode collects more precise and annotated data than passive mode, yet users might omit short trips (ATM, post office, and bakery) [29]. In contrast, more data is collected in passive mode but such data needs a processing chain (filtering, segmentation, points of interest, map-matching, activity, and mode detection) to be useful [30]. The aforementioned mobile applications work with a sampling frequency of 1 Hz in active mode and variable frequencies in passive mode to increase the time span of the device's battery; such frequencies depend on the battery level (a high battery level has a higher sampling frequency than a low level one).

2.3. Campaigns. Three projects made use of the mobile application *Connect* in their campaigns. For data collecting purpose, *Connect* was installed in a set of smartphones that were shared among participants in shifted periods (2-3 weeks). As part of the campaigns, the participants were asked to report their activity using the application in active mode. The following is a brief description of the projects:

(i) Multimodal electric mobility for commuter and business trips (Elmo) investigates whether electric two-wheelers, potentially combined with other types of durable mobility (classic public transport, taxi), could be a valuable alternative for work-related trips, such as commuting and business trip [31].

(ii) Electric vehicles in action (EVA) are a large-scale living lab platform with various types of electric vehicles, charging infrastructure, and data loggers. The purpose of this platform is to check which geographical placement of public charging stations is most fit [27]. The project also aims to assess the impact of electric vehicles on user behaviour, to further support the definition of standards, recommendations, scenarios, and roadmaps for the sustainable deployment of electric vehicles.

(iii) Olympus focuses on networked mobility, aiming at integration between shared mobility (car sharing, carpool, and bike sharing) and private and public transports. It works at different levels of integration: end-users, mobility providers, and supporting services (e.g., charging infrastructure for electrical vehicles). Its campaign started in four Belgian cities (Antwerp, Ghent, Hasselt, and Leuven). In these cities and their stations, electrical shared cars and shared bikes are made available; therefore users can also opt for an electric variant [32].

On the other hand, the *Routecoach* application was a part of sustainable mobility campaign in province of Flemish Brabant. The application was freely available to download and most of the data was collected in passive mode.

(iv) The main aim of the campaign was to develop an evaluation and planning toolkit for mobility projects which is transferable and can be adopted by planners [33]. The data collection process lasted from January to April 2015. In total, 8303 users actively participated by downloading the freely available application and collecting the data on more than 30,000 trips, although, in this study, we only considered the users that manually reported their activity (380 users).

A summary of the collected data is presented in Table 4; some common features across the datasets allow us to merge them into one, for instance, the sampling rate, which is set to 1 Hz when application works in active mode (i.e., users manually report the starting/stop trip). Besides, this mode records a timestamp right after the user's action even if a GNSS location is not fixed.

The dataset (iv) incorporates data from the test period of the application but also postcampaign phase; thus it includes a longer period than the official campaign, though all data collected in passive mode (background data collection) are filtered out since the sampling frequency is not fixed; hence a theoretical estimation of the number of locations could yield an incorrect outcome.

Table 5: Aggregate features from the trip segments.

Name	Feature	Description
Collected GNSS	Number of GNSS locations	It is the actual number of collected measurements.
Expected GNSS	Theoretical number of GNSS locations	Having a sampling rate of 1 Hz, in theory, the number of GNSS locations to collect is equal to trip segment duration in seconds.
TTFF	Time to first fix	Time difference between the trip segment starting time and the first measurement timestamp.
Missing GNSS	Number of missing GNSS locations	Difference between the expected and collected GNSS.

Figure 4: Missing locations in a trip segment.

2.4. Data Quality. This study focuses on the completeness aspect of data quality. Consequently, the measurements (GNSS locations) are aggregated into a trip segments' level. Features such as number of collected locations, expected locations, time to first fix (TTFF), and missing locations are extracted from the trip segments (Table 5).

2.4.1. Missing Locations. Consider a missing location as a failed event of the GNSS device at getting a fix, so that such events occur during a trip segment. Therefore, a trip segment may include gaps (missing locations) where the location point is not determinate. We represent a trip segment as a list of consecutive locations points with annotations. When a trip segment is reported in active mode, it includes the user's annotations, such as purpose, transport mode, and start/end time. These temporal annotations, start/end time, are independent from the collected locations, since they are recorded right after pressing the play/stop button in the application (Figures 3(a) and 3(c)). Consequently, locations points might be present or not within a trip segment, particularly at the beginning, where a first location can be got after a while (Figure 4). In contrast, when a trip segment is collected in passive mode, the start time matches to the first location timestamp, as well as, the end time and the last location timestamp, because, in passive mode, those labels are inferred by the application using the timestamp of the locations. And yet, locations points may be not present between the first and last location points due to well-known issues of GNSS technology.

Issues like cold/warm start and signal reception can turn out in missing data within a trip segment; hence the resulting segment may include gaps along the trip. To make a distinction among gaps, the missing locations at the beginning of the trip, the gap before the first location, are associated with the cold/warm start issue of the GNSS device, which is the time that the GNSS device needs to acquire the first position after a period of inactivity. This issue often turns out in missing data at the beginning of the trip, especially when a device remains off for long periods [34, 35]. In contrast, the gaps after the first location point are linked to the signal reception issues such as signal loss [36] (e.g., underground travel, bridges, and tunnels) and multipath errors a.k.a. urban canyoning errors [37, 38].

Let S be trip segment represented by tuple $S_j = (P_j, A_j)$, where P_j is a list of location points $P_j = \{p_1, \ldots, p_m\}$ such that p is a location point with a timestamp, A_j is a list of users' annotations $A_j = \{start, end, mode, purpose\}$ with $start < end$, *start* is the starting time of the segment, *end* is the stopping time of the segment, *mode* is the transportation, *purpose* is the trip purpose, and j is a unique identifier of the segment. We assess the quality of the GNSS data using the following equations:

$$\text{Missing locations}_j = \frac{n\left(P_j\right)}{\left|A_{j,\text{end}} - A_{j,\text{start}}\right|} \tag{1}$$

$$\text{TTFF}_j = \frac{\left\lfloor t\left(P_{j,1}\right) - A_{j,\text{start}}\right\rfloor}{\left|A_{j,\text{end}} - A_{j,\text{start}}\right|} \tag{2}$$

$$\text{Missing within trip}_j = \text{Missing locations}_j - \text{TTFF}_j \tag{3}$$

$$\text{Collected locations}_j = 1 - \text{Missing locations}, \tag{4}$$

where $n(\,)$ and $t(\,)$ are functions to count the number of locations in P_j and to extract the timestamp from a location point. Having a sampling rate of 1 Hz, in active mode, the theoretical number of GNSS locations is equal to segment duration in seconds (i.e., difference between *end* and *start* times); thus the proportion of missing locations is equal to the number of collected locations over the theoretical ones (see (1)). The proportion of missing data linked to the time to first fix is the time difference between the first location timestamp

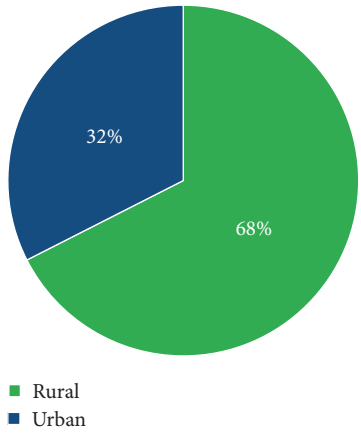

Figure 5: Land use of the trip segment based on the origin area.

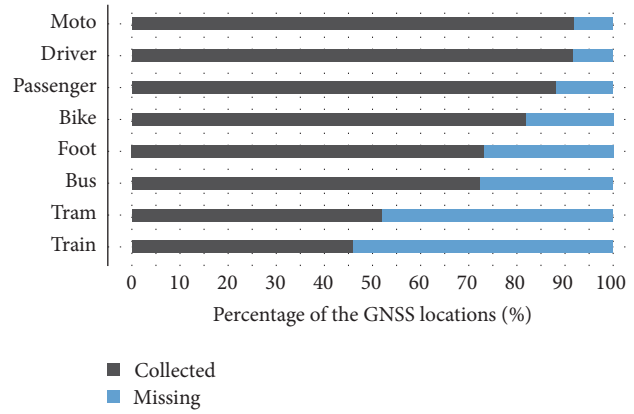

Figure 6: Percentage of missing GNSS locations per trip segment.

and the segment starting time over the theoretical number of locations (see (2)). The proportion of missing locations within a trip segment is the difference between the missing locations and the time to first fix (see (3)). The proportion of collected locations is equal to the missing locations subtracted from the unit (see (4))

2.4.2. Land Use. Land use plays an important role in this study; it provides an extra perspective to analyse the GNSS data quality and its relationship to the area where the data is collected; for instance, large structures that are extensively present across urban areas might affect the signal reception differently compared to that in rural areas, where such structures are hardly present. Consequently, a segment is classified as either rural or urban depending on its origin.

The segment origin contains relevant information to assess the missing location due to a cold/warm start effect. Since the time to first fix occurs at the beginning of the segment (Figure 4), the administrative area in which the segment started is used to classify it into rural or urban. Therefore, the administrative areas in Belgium are extracted from *OpenStreetMap contributors* (OSM). OSM is an open access platform for geospatial vector data and it is often considered complete and appropriate for planning studies in comparison to other commercial counterparts [39].

Using a geographic information system, we identify whether a segment lies in a rural or urban area [20], where a spatial operation (interception) is used between the administrative areas and the segments, which turns out in labelled segments based on the land use. A land use share is shown in Figure 5, where more trip segments start from rural areas.

3. Results and Discussion

In this paper, we present only results on the GNSS traces collected through smartphone and do not focus our analysis on any other type of data collected in the campaigns.

By comparing the collected GNSS locations and the theoretical number of locations, we calculated the missing data per trip and transport mode (Figure 6). In addition, we observed that *train* mode contains the highest percentage of missing data, collecting less than 50% of the data. To our understanding, a coach makes unfavourable conditions for signal reception because of its dense chassis, like passengers travelling in the low level of a double decker train and underground trajectories [40]. Besides, underground platforms and covered rail stations might affect getting the first fix.

For public transport, *tram* mode behaves slightly better than a train mode, gathering just above 50% of the data. Both transport modes share common things that can cause a signal loss, such as dense chassis, high voltage above them, and covered stations [41]. Instead, *bus* mode collects 72.4% of the data (Figure 6), a notorious improvement over tram, but perhaps not good enough for some applications.

Pedestrians are next on experiencing issues to gather the data; 26.8% of the data are missed on foot mode. This can be explained by user's behaviour for reaching certain destinations. For instance, walking under buildings to avoid rain and sun, taking shortcuts in between narrow corridors, and small stops in stores or covered areas.

Driver and passenger modes, both in car, perform good and collect around 90% of the data. Yet, we expected to hardly see dissimilarity among them, providing it is the same type of vehicle. But it turns out that a car passenger reports slightly less data than a car driver, 88.3% and 91.8, respectively. The differences can be in the interaction with the device, since a car driver must be focused on the road rather than manipulating the smartphone, whereas a passenger is free to interact with the smartphone all way long. These outcomes are comparable to a previous study [42], in which data quality was assessed using GNSS loggers instead. The mentioned study reported 9% of missing data in a car mode. That figure is similar to our findings, where 8.2% of the data are not captured by the smartphone.

Another interesting observation is the two-wheeler vehicles, bike and moto, that report 82% and 92%, respectively. It seems like the road restrictions push moped riders to share the same infrastructure as cars, having a clear view to the satellites most of the time and therefore gathering similar percentage of data. In contrast, bike mode makes use of other

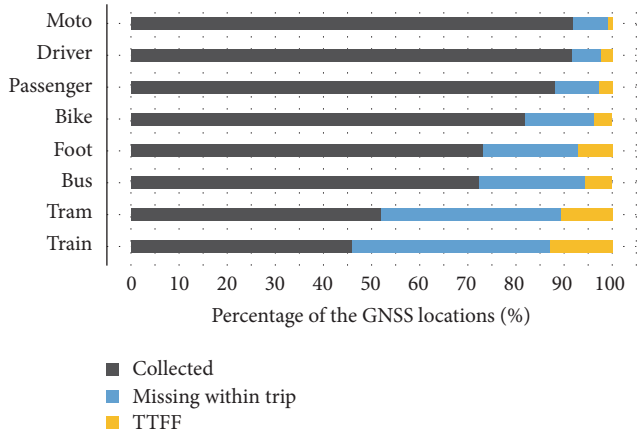

Figure 7: Percentage of missing GNSS locations at the beginning of the trip (TTFF).

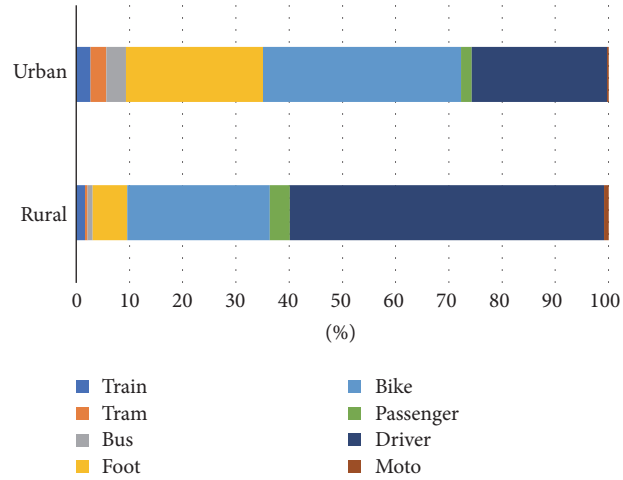

Figure 8: Modal split of the trip segments based on the land use.

road facilities (bicycle lanes and bicycle highway) and it has few restrictions (a biker could ride on a walking path or take same shortcuts as pedestrians). Hence, it faces similar issues for collecting data like pedestrians.

3.1. Time to First Fix. Considering the time to first fix as a part of the missing data (Figure 7), we calculated it using the trip start time and the timestamp of the first GNSS location. It turns out that it follows the same trend for all the modes except for on foot mode, in which we observed that not only does the overall data collection struggle in this mode (26.8% of missing data) but also the first fix represents 7.2% of the trip segment.

Both train and tram missed more than 10% of the data before getting the first fix. Having significant gaps at the beginning of the trip leads to travel reporting issues, in which the truth origin of a trip may be far from the first location point, thus underreporting the travelled distance, while the gaps within a segment can be corrected using GIS tools and map-matching techniques. Those techniques align location points with the road network and also interpolate missing locations, yielding a good estimation of the travelled distance.

Additionally, we noticed that travelling in a car as a passenger generates more missing data during the trip than in a car as a driver, since the missing data at the beginning are very close, 2.3% and 2.7% for car and passenger mode, respectively; we can tell the difference is on the interaction with the device.

3.2. Land Use Effects. To analyse further the missing data, we considered the land use as factor that could influence the GNSS signal reception. We relied on the information provided by the OSM to extract the administrative areas; thus trips segments were classified as rural and urban depending on trip origin. Results are shown in Figure 8, in which a modal split of the trip count is provided for both urban and rural areas.

Results in Figure 9 were expected for rural and urban areas; that is, urban areas are affected more than rural areas.

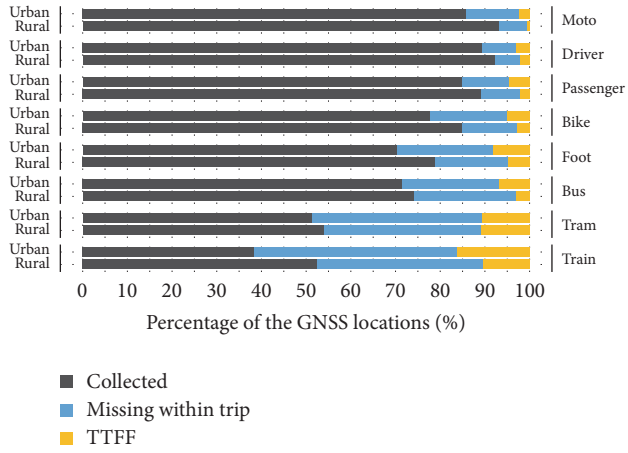

Figure 9: Percentage of missing locations based on the land use.

However, the effect on tram mode seems to be the same in both areas, particularly the time to get the first fix, where we notice similar figures of missing data, 10.8% and 10.7% for rural and urban. Hence, it could be that infrastructure facilities are similar; besides a public transport like tram is not that popular in rural areas, which turns out that those trips are from surrounding urban areas.

Considering the experiment setup, where the sampling frequency is 1 Hz, we can estimate the average delay at getting the first fix. Therefore, we calculate the average travel time per mode and combine it to the percentage of missing data due to the time to first fix (Table 6 and Figure 10), which turns out to be an average estimation of the cold/warm start problem in smartphones.

As was pointed out before, land use exhibits no effect on a tram mode to first fix; in both areas, it reports similar figures. In contrast, modes such as moto, passenger, bike, and bus are increased twofold when these are compared to rural areas (Figure 10).

As an overall outcome, the smartphones reported 84% of the GNSS data; 16% of the data was missed (Figure 11). The

Table 6: An estimation of the time to first fix based on an average travelled time (minutes).

Transport mode	Travelled time		Average TTFF	
	Average	Std. Dev.	Rural	Urban
Bike	21.9	11.0	0.6	1.1
Bus	27.0	16.1	0.8	1.9
Driver	18.6	11.1	0.4	0.5
Foot	13.7	10.5	0.7	1.1
Moto	32.9	9.6	0.2	0.8
Passenger	27.5	14.9	0.6	1.3
Train	33.2	12.7	3.4	5.4
Tram	18.6	9.7	2.0	2.0

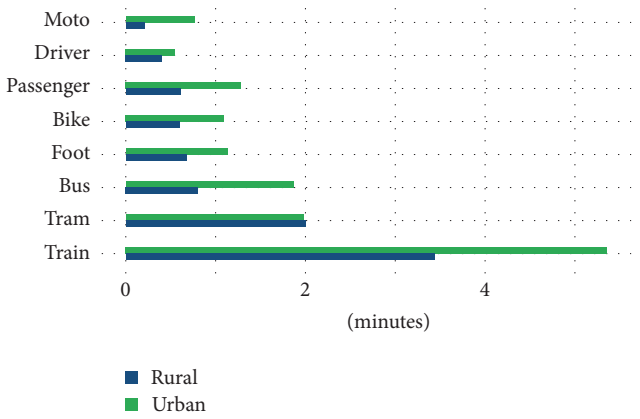

Figure 10: Average time to first fix in smartphones (minutes).

missing data is split in 4% as a part of the cold/warm start problem and 12% is missed within the trip segment (between the first location point and the end of the trip) that can be linked to signal reception issues.

4. Conclusions

This paper presents outcomes from four campaigns combined all together, in which the quality of the mobility behaviour data was analysed, more precisely, the completeness aspect of the data. These results can be used as a reference for current and futures mobility studies that use smartphones as a collecting tool, where these figures for missing data can help to put measurements on the time travelled and the distance on multiple transport modes into context. In addition, the results can help to set up parameters on the smartphones applications and to assess the influence on the reporting.

Travelling by train has the most impact on the GNSS reception of the smartphones. However, it may be not that difficult to overcome the lack of data, since that transport mean follows a fix pathway and timetables as well. Therefore, GNSS traces can be aligned with the railway and even dummy-locations can be extrapolated to fill in the gaps. It will require extra steps in the processing chain (map-matching and interpolation) to do this processing, which can lead

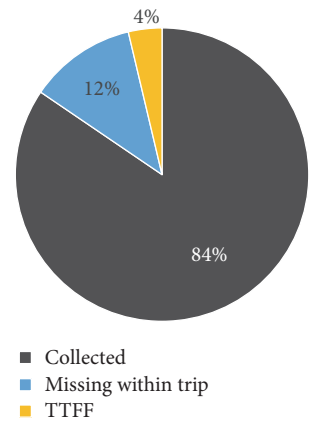

Figure 11: Overall missing data in smartphones as a data collecting tool.

to other types of errors (oversegmentation and misleading trajectories).

It turns out that people travelling in similar vehicles can produce different outcomes based on their role within the vehicle. For example, a person in a car can be either passenger or driver, where drivers reported less missing data than passengers. We have seen that the issue is not related to gathering a first fix during the trip, which leads to the conclusion that the human-device interaction (angle and position) while collecting data has an impact on the quality of the data collection.

Finally, our findings show consistency with a previous study [42], in which a driver mode reported comparable figures (9% of missing data). However, that study made use of GNSS loggers installed in a car instead of smartphones, where the GNSS loggers were mainly affected by the cold/warm start because of the long periods of being turned off (e.g., a car being on a parking lot). In contrast, smartphones are more likely to be working a whole day and, besides, users might be gathering GNSS locations through any other mobile application installed on their devices, which gives a fast response at getting a first fix.

Disclosure

The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; and in the decision to publish the results.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research is funded by INTERREG North-West Europe Project New Integrated Smart Transport Options (NISTO),

the Flemish Government Agency for Innovation by Science and Technology, and the Flemish Institute for Mobility.

References

[1] S. Bricka and C. R. Bhat, "A comparative analysis of GPS-based and travel survey-based data," *Transportation Research Record*, Article ID 1972, pp. 9–20, 2006.

[2] J. Wolf, M. Oliveira, and M. Thompson, "Impact of underreporting on mileage and travel time estimates: results from global positioning system-enhanced household travel survey," *Transportation Research Record*, vol. 1854, no. 3, pp. 189–198, 2003.

[3] S. G. Bricka, S. Sen, R. Paleti, and C. R. Bhat, "An analysis of the factors influencing differences in survey-reported and GPS-recorded trips," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 67–88, 2012.

[4] C. Carrion and D. Levinson, "Value of travel time reliability: a review of current evidence," *Transportation Research Part A: Policy and Practice*, vol. 46, no. 4, pp. 720–741, 2012.

[5] J. Bates, J. Polak, P. Jones, and A. Cook, "The valuation of reliability for personal travel," *Transportation Research Part E: Logistics and Transportation Review*, vol. 37, no. 2-3, pp. 191–229, 2001.

[6] I. Semanjski and S. Gautama, "Sensing human activity for smart cities' mobility management," in *Smart Cities Technologies*, I. N. Da Silva, Ed., InTech, 2016.

[7] J. Du and L. Aultman-Hall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 3, pp. 220–232, 2007.

[8] P. Stopher, C. FitzGerald, and M. Xu, "Assessing the accuracy of the Sydney Household Travel Survey with GPS," *Transportation*, vol. 34, no. 6, pp. 723–741, 2007.

[9] F. Witlox, "Evaluating the reliability of reported distance data in urban travel behaviour analysis," *Journal of Transport Geography*, vol. 15, no. 3, pp. 172–183, 2007.

[10] Q. Jiang, F. Kresin, A. K. Bregt et al., "Citizen Sensing for Improved Urban Environmental Monitoring," *Journal of Sensors*, vol. 2016, Article ID 5656245, 2016.

[11] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 3, pp. 285–297, 2009.

[12] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw GPS data for geographic applications on the web," in *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*, New York, NY, USA, April 2008.

[13] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai," *Landscape and Urban Planning*, vol. 106, no. 1, pp. 73–87, 2012.

[14] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi GPS traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 113–123, 2013.

[15] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris, "Discovering Spatial Patterns in Origin-Destination Mobility Data," *Transactions in GIS*, vol. 16, no. 3, pp. 411–429, 2012.

[16] J. Hood, E. Sall, and B. Charlton, "A GPS-based bicycle route choice model for San Francisco, California," *Transportation Letters*, vol. 3, no. 1, pp. 63–75, 2011.

[17] M. J. Duncan, H. M. Badland, and W. K. Mummery, "Applying GPS to enhance understanding of transport-related physical activity," *Journal of Science and Medicine in Sport*, vol. 12, no. 5, pp. 549–556, 2009.

[18] S. Mavoa, M. Oliver, K. Witten, and H. M. Badland, "Linking GPS and travel diary data using sequence alignment in a study of children's independent mobility," *International Journal of Health Geographics*, vol. 10, article no. 64, 2011.

[19] E. Murakami and D. P. Wagner, "Can using global positioning system (GPS) improve trip reporting?" *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 2-3, pp. 149–165, 1999.

[20] P. Stopher, E. Clifford, J. Zhang, and C. FitzGerald, *Deducing Mode and Purpose from GPS Data*, Working Paper of the Austrian Key Centre in Transport and Logistics, Institute of Transport and Logistics Studies, University of Sydney, Sydney, Australia, 2008.

[21] OVG, "Flemish Travel Survey," Department of Mobility and Public Works, 2013, Available: http://www.mobielvlaanderen.be/ovg/.

[22] OVG, OVG Flanders 4.5: Travel Survery, 2014.

[23] L. H. Immers and J. E. Stada, Transportation Systems, 2004.

[24] Move., "Connect mobile application," Ghent University. Google Play store, 2013, Available: https://play.google.com/store/apps/details?id=com.move.tripdiary.

[25] Move, "Routecoach mobile application," Ghent University. Google Play store, 2015, Available: https://play.google.com/store/apps/details?id=com.move.routecoach.

[26] A. J. Lopez, I. Semanjski, D. Gillis, J. De Mol, R. Bellens, and S. Gautama, "Development of smart city platform as a tool for supporting more sustainable mobility behaviour," in *Serious Health Games and Apps Conference*, vol. 32, 2015.

[27] S. Vlassenroot, D. Gillis, R. Bellens, and S. Gautama, "The use of smartphone applications in the collection of travel behaviour data," *International Journal of Intelligent Transportation Systems Research*, vol. 13, no. 1, pp. 17–27, 2015.

[28] I. Semanjski, A. J. L. Aguirre, J. De Mol, and S. Gautama, "Policy 2.0 platform for mobile sensing and incentivized targeted shifts in mobility behavior," *Sensors (Switzerland)*, vol. 16, no. 7, article no. 1035, 2016.

[29] D. P. Wagner, "Lexington area travel data collection test: GPS for personal travel surveys," Final Report, Off. Highw. Policy Inf. Off. Technol. Appl. Fed. Highw. Adm. Battelle Transp. Div. Columbus, 1997.

[30] J. Wolf, D. Dr, and R. Guensler, *Using GPS data loggers to replace travel diaries in the collection of travel data*, Georgia Institute of Technology, 2000.

[31] VIM, "Multimodal electric mobility for commuter and business trips," Flanders Institute for Mobility, 2015, Available: http://www.vim.be/projects/elmowork.

[32] Olympus, "Networked mobility solutions, 2015," Available: http://www.olympus-mobility.com.

[33] Nisto, "New Integrated Smart Transport Options," 2015, Available: http://www.nisto-project.eu.

[34] P. Stopher, Q. Jiang, and C. FitzGerald, "Processing gps data from travel surveys," in the 2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications, 2005.

[35] N. Schuessler and K. Axhausen, "Processing raw data from global positioning systems without additional information," *Transportation Research Record*, no. 2105, pp. 28–36, 2009.

[36] G. Draijer, N. Kalfs, and J. Perdok, "Global Positioning System as Data Collection Method for Travel Research," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1719, pp. 147–153, 2000.

[37] N. Schuessler and K. Axhausen, "Identifying trips and activities and their characteristics from GPS raw data without further information," in *Proceedings of the 8th International Conference on Survey Methods in Transport*, Annecy, France, May 2008.

[38] J. Jun, R. Guensler, and J. Ogle, "Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1972, pp. 141–150, 2006.

[39] H. H. Hochmair, D. Zielstra, and P. Neis, "Assessing the completeness of bicycle trail and lane features in OpenStreetMap for the United States," *Transactions in GIS*, vol. 19, no. 1, pp. 63–81, 2015.

[40] E. Bertran and J. A. Delgado-Penín, "On the use of GPS receivers in railway environments," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1452–1460, 2004.

[41] R. Mázl and L. Přeučil, "Sensor data fusion for inertial navigation of trains in GPS-dark areas," in *Proceedings of the 2003 IEEE Intelligent Vehicles Symposium, IV 2003*, pp. 345–350, usa, June 2003.

[42] A. Lopez, I. Semanjski, D. Gillis, D. Ochoa, and S. Gautama, "Travelled distance estimation for GPS-based round trips: car-sharing use case," in *Proceedings of the Fifth International Conference on Data Analytics*, pp. 87–92, DATA ANALYTICS 2016.