

Error Sources in the Analysis of Crowdsourced Spatial Tracking Data

Casper Van Gheluwe*[†], Angel J. Lopez*^{†‡} and Sidharta Gautama*[†]

* *Dept. of Industrial Systems Engineering and Product Design, Ghent University, Ghent, Belgium*

[†] *Industrial Systems Engineering (ISyE), Flanders Make www.FlandersMake.be*

[‡] *Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, ESPOL,*

Campus Gustavo Galindo Km 30.5 Vía Perimetral, 09-01-5863 Guayaquil, Ecuador

Email: {casper.vangheluwe, angel.lopez, sidharta.gautama}@ugent.be

Abstract—Governments are increasingly interested in the use of crowdsourced spatial tracking data to gain information on the travel behaviour of their citizens. To improve the reliability of reporting in such mobility studies, this paper systematically analyses the propagation of errors from low level operations to high level indicators, such as the modal split and travelled distances. We find that most existing metrics in literature are insufficient to fully quantify this evolution of data quality. The propagation channels are presented schematically and a new approach to quantify the spatial data quality at the end of each processing stage is proposed. This procedure, within the context of Smart Cities, ensures that the data analytics and resulting changes in policy are sufficiently substantiated by credible and reliable information.

Index Terms—data quality, geospatial data, crowdsensing, data processing, error propagation

I. INTRODUCTION

In the pursuit of more information on their citizens, governments are increasingly interested in crowdsourced data. This kind of data can be gathered relatively cheaply and quickly, but its analysis requires extra care as the measurement environment is uncontrolled. Often, governments are interested in this crowdsourced data to derive information that is necessary for their decision making. Governments look at spatial tracking data for mobility, quality of life or sustainability indicators. Crowdsourced mobility studies often use data gathered from smartphones [1, 2] or other Global Navigation Satellite Systems (GNSS) devices [3] to gain insight in the travel behaviour of citizens. However, this approach requires that the accuracy and the reliability of the data and transformation processes are clearly characterized. Studies have shown that errors that occur in early stages of the data processing can have drastic consequences on the accuracy of later stages and particularly on typical indicators that are reported at the end of mobility studies, such as the modal split and travelled distances.

II. EXISTING QUALITY METRICS

Over the past couple of years, several authors have defined the causes of data quality problems in GNSS mobility studies. These causes, along with potential metrics to measure their

This work is funded by the Flanders Agency for Innovation and Entrepreneurship through the FLAMENCO project (FLAnders Mobile ENacted Citizen Observatories).

impact, have been collected in this section. They are organized according to the stage where they occur.

A. Missing data and positioning errors

Positioning errors caused by inherent errors in the GNSS device or by signal reception issues have an influence on the trip lengths. Several authors have identified this phenomenon [4, 5] and Lopez has experimentally determined the distribution of the residual distance for several speeds [6, 7]. This overreporting has an influence of less than 1% at 30 km/h, and is therefore relatively minor for motorized traffic. However, it is more relevant for slower forms of transportation, such as pedestrians or cyclists. Ranacher *et al.* also determined that GNSS measurement errors cause a systematic overestimation of travel distance and offer a mathematical proof [8]. They offer a derivation of a formula for *OED*, the expected overestimation of distance. This equation can be reshaped to calculate the spatial autocorrelation of GNSS measurement errors *C*, which can be used as a quality metric to describe how accurately the GNSS sensor captured the movement of an object.

On the other hand, missing data can have a significant effect on underreporting of trip distance, especially for curves or other non-straight paths, because it can cut off the curvature of the trajectory in case one or more intermediate points are missing or if the sampling rate is too low. Based on the comparison of GNSS tracks and CAN-bus data recorded in the vehicle, Lopez reports that an average of 9% of the travelled distance is not captured as a result of missing data during the trip.

Biljecki *et al.* discovered that missing data can also affect the segmentation and transport mode classification processes [9]. If the signal shortage is long enough, it could mean that entire segments done with a different mode are not recorded. The exact location of a mode transition could also be lost if it occurs during a period of missing data.

B. Preprocessing

Some mobility researchers apply interpolation, filtering or smoothing techniques to reduce the impact of positioning errors or missing data. Grochla and Połys applied Kalman filtering to GNSS trajectories recorded using a smartphone. They

propose two metrics to quantify the quality of the trajectory, and found that Kalman filtering actually introduces additional noise. While the resulting track will be more visually attractive due to smoothing, the average distance between the trajectory and the reference track can increase significantly [10]. Jun *et al.*, on the other hand, have found that a modified Kalman filter for GNSS data performs better than other common smoothing techniques at reducing the impact of GNSS random errors on estimations of speed, acceleration and travel distance [11].

C. Segmentation and transport mode classification

Segmentation and transport mode classification are closely related. Many techniques use point-based transport mode classification as a way to identify segments in a trip, while others perform segmentation first, followed by transport mode classification on each segment separately. Typically segmentation processes suffer from oversegmentation, leading to low precision in trip reporting, due to ambiguous situations where the users remain stationary for short periods, e.g. at traffic lights or in traffic jams. A segmentation algorithm or transport mode classifier that uses speed information could therefore have a high accuracy over time or distance, but yield significantly oversegmented trips nonetheless, if it contains many of these spurious short segments.

Prelicean *et al.* support the idea that comparing the accuracy of different segmentation and mode classification algorithms based on common metrics such as precision and recall is difficult [12]. To illustrate this, they offer the result of 5 techniques, which yield completely different segments but have identical precision and recall. Finally, they propose five new metrics that can more precisely evaluate the correspondence between the inferred and the true segmentation. These metrics still require a ground truth, which is typically not available in mobility studies.

An alternative way to quantify the performance of trip segmentation and mode classification algorithms, is to contrast certain indicators, such as the modal split or the distribution of trip distances, to known results from large scale surveys carried out by government agencies [13]. If the techniques that were used yield comparable values for these indicators, they are likely sufficient. Nevertheless, researchers have to consider that the population that participates in a mobility campaign may be biased compared to the general survey. For example, campaigns that study bicycle activity will likely attract a large portion of recreational cyclists, who will often travel longer distances at higher speeds.

D. Map matching

Map matching is an important step in the processing chain for spatial tracking data, as it links the trajectory to the road network. It is advantageous for two reasons. Primarily, it provides a way to reduce the impact of positioning errors by aligning the measured trajectory to a known road network. It can also be used to interpolate locations, by following the most likely route when data is missing. Finally, map matching allows analysts to build accessible visualisations to support the

conclusions of a mobility campaign, such as speed or intensity maps, by explicitly linking measurements to road segments.

Quddus, Noland and Ochieng have studied the influence of map matching on tracking data quality intensively. In 2005, they experimentally validated a map matching algorithm by comparing its output to a ground truth [14], and found that it exhibited a mean horizontal position error of 5.6 m, which can be reduced to 2.0 m if the analysis takes the distance between the traffic lanes and the road centreline into account. They also show that you can reduce the error if the vehicle speed is close to zero by using GNSS devices with dead reckoning support. Dead reckoning can augment the position accuracy by using information from additional sensors such as gyrometers, accelerometers and wheel speed. Next, the authors develop a metric that quantifies the integrity of map matching algorithms [15]. It considers several important factors that affect the uncertainty. The metric is quite robust for one test route, but further experiments are necessary to test its performance with other types of routes, such as in urban areas. Finally, the authors focused on the effect that the road network can have on the performance of map matching algorithms [16]. They particularly highlight five quality issues that may appear in the road network;

- Topological errors caused by features of the real world which were omitted or simplified
- Geometric errors due to the deviation of map features from their actual location in the real world
- Missing segments or the existence of old segments due to a lack of updates
- Incorrectly classified features (e.g. junction vs. roundabout)
- Timeliness of the data

They propose two additional quality metrics for map matching algorithms; the along-track ($|MC| = |MA| \times \cos(\phi)$) and cross-track ($|AC| = |MA| \times \sin(\phi)$) accuracy. For more detail we refer to their work.

These metrics can be useful to compare several map matching algorithms if a ground truth is available, but are not suitable for realtime assessment of the map matching quality. Moreover, they fail to take into account some less obvious reasons why map matching can go wrong, such as:

- Network selection, for example if the transport mode of a segment was incorrectly classified
- Unauthorized manoeuvres, such as buses and taxis that can use bus lanes which are not available in the network. Some drivers also do not respect the transit regulations.
- Simplistic map matching methods may not take into account the driving direction, which can lead to inconsistent or even impossible paths.
- Some transport modes, such as walking and biking, have an inherently high freedom of movement which cannot always be constrained to a network.
- Missing data may lead the algorithm to assume a different trajectory than the true trajectory

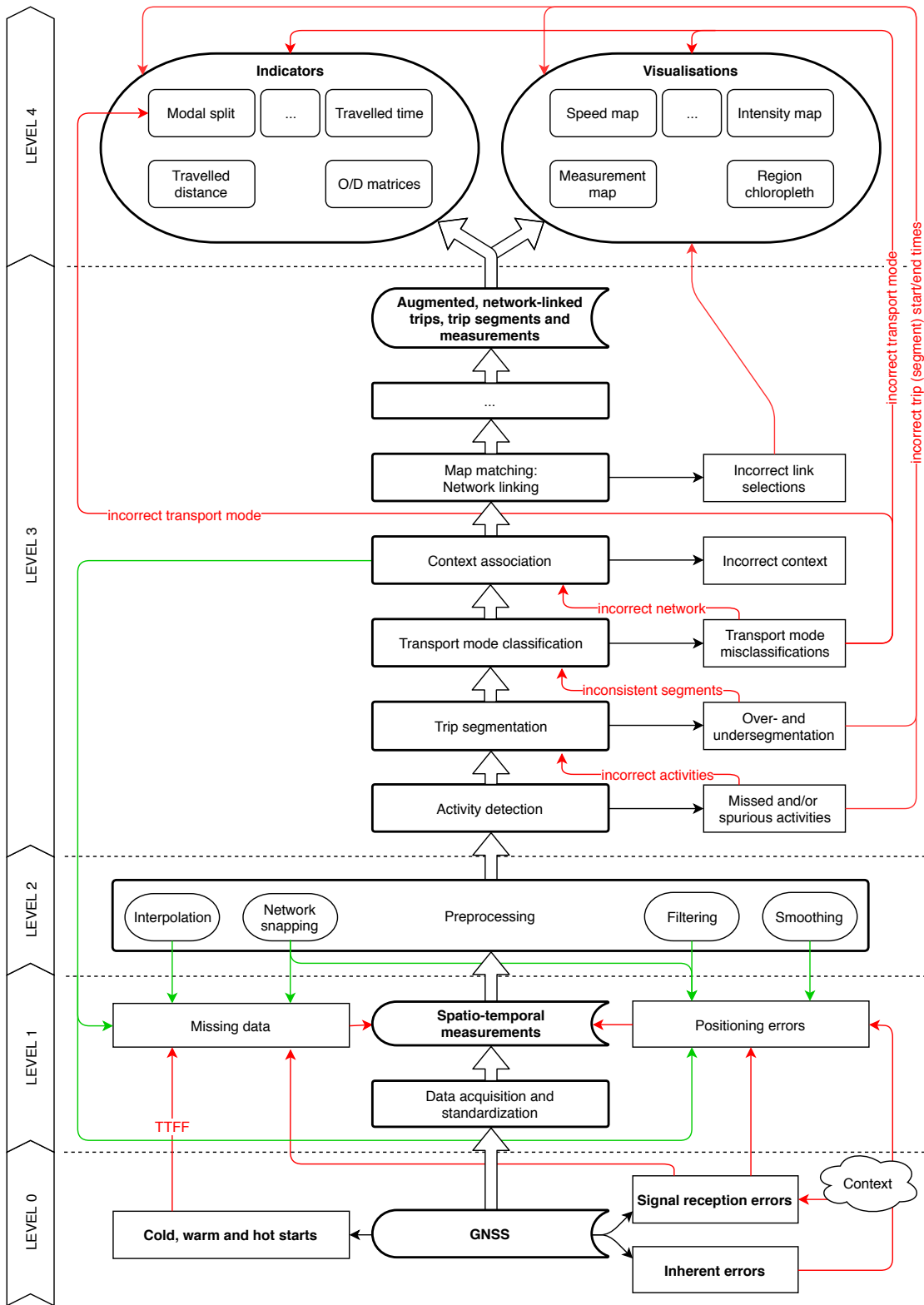


Fig. 1. Schematic view on the error propagation in mobility studies that use spatial tracking data. Data is transformed from raw coordinates to high level indicators through a series of consecutive processes. The red arrows indicate that errors originating from the source process have a negative influence on the quality of the destination process. Green arrows indicate that the source process corrects errors originating from the destination.

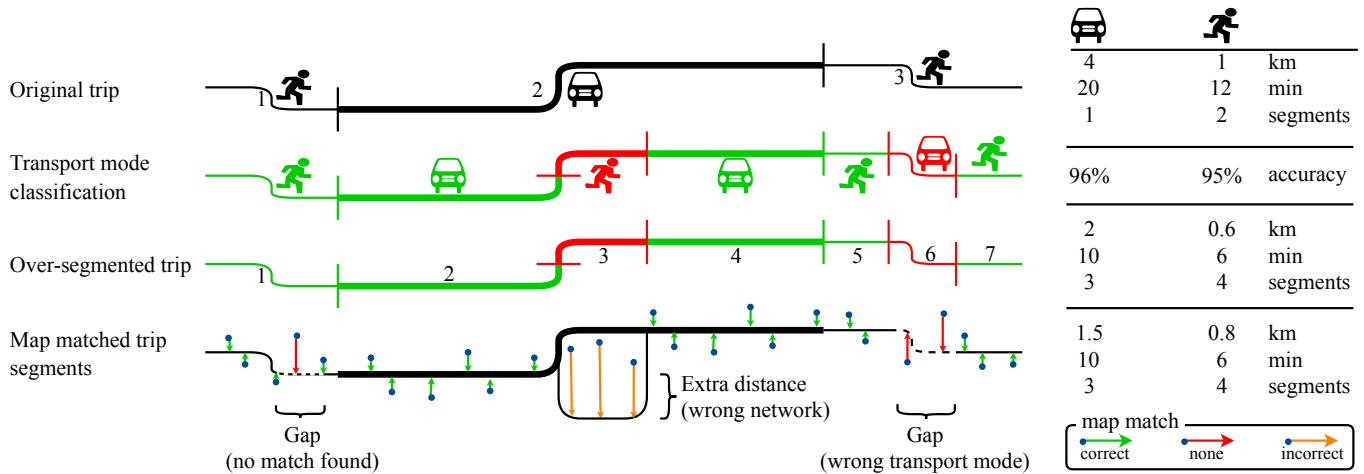


Fig. 2. An example showing the propagation of errors in Level 3 processing stages to errors in the Level 4 indicators. A trip consisting of 3 segments suffers from incorrect transport mode classification, leading to oversegmentation and finally to map matching errors. The accumulation of these errors has a significant impact on the calculated mobility indicators, which are listed on the right.

III. PROCESSING STAGES

The transformation from raw GNSS data points to mobility indicators and visualisations is a multi-stage procedure. Shen and Stopher subdivide this procedure into 5 interconnected stages: *preprocessing*, *trip identification*, *mode detection*, *purpose imputation* and *final result* [3]. The preprocessing stage downloads the data from the sensors, and does some preliminary validation. The validated GNSS data is then used as input for a trip identification algorithm. The resulting trips are issued to a transport mode classifier. Optionally, trip purpose imputation, i.e. the determination of the goal of the trip, can be performed using the results from the transport mode detection. Finally, the results are aggregated into a list of all trips, transport modes and purposes. In their conclusion, the authors mention that the outcome of the transport mode and purpose classification is likely manipulated in part by errors in the trip detection and segmentation processes that precede it.

Our model, which is presented schematically in Figure 1, expands upon this line of thought by explicitly marking which errors are generated by which processing stages and by indicating how these errors can propagate from the raw GNSS sensor data all the way to mobility indicators and visualisations, which are used by policy makers to make informed decisions.

In this section, we establish an architecture for the systematic preparation of mobility indicators and visualisations, organized in five levels [7] as shown in Table I. We further elaborate on the role of each component of the processing chain and how they are influenced by errors in previous processing steps.

A. Level 0 – Raw (unprocessed) data

Like most sensing and measurement devices, GNSS sensors exhibit a fundamental uncertainty in their measurements. For GNSS sensors in particular, the sources of this uncertainty can be categorized into three clusters. A first group includes inherent errors, such as measurement errors due to deviations in satellite clocks, signal delays in the tropo- and ionosphere

TABLE I
DATA PROCESSING LEVELS

Level	Description
Level 0 (L_0)	Raw (unprocessed) data, it is data gathered from data sources (sensors, smartphones, etc.) at full resolution.
Level 1 (L_1)	Annotated data, it is the original data at full resolution but annotated with ancillary information, time referenced and transformed to a standardized format.
Level 2 (L_2)	Derived data from automatic processes, including quality improvements.
Level 3 (L_3)	Augmented data, L_2 data enriched by means of inference, data mining techniques and external data sources.
Level 4 (L_4)	Aggregated data as insights and analytics.

and the unpredictable effects of receiver noise and multipath fading [17]. These errors are typically of the order of 1 meter. The second cluster groups errors caused primarily by the spatial context in which the GNSS device is being utilized. The inability to accurately determine a position in urban canyons or underground has been established repeatedly in literature [5, 18]. Similarly, land use and – in the case of mobility studies – the transportation mode can significantly influence the accuracy of coordinates gathered using GNSS. Thus, these issues can lead to positioning errors if the position fix is inaccurate, or missing data in case a fix could not be determined at all. Finally the difference between cold, warm and hot starts of the GNSS sensor device can also lead to missing data [7].

B. Level 1 - Annotated data

In this stage the raw GNSS data is annotated, time referenced and converted to a standardized format for further processing. The data itself is not modified, therefore no additional errors are introduced unless the transformation process is flawed, for instance as a result of software bugs.

C. Level 2 - Derived data

Processes at this level may execute some operations to prepare the data for further analysis and to potentially improve the quality of the data. Several such processes can be applied. Interpolation is used to replace missing data with calculated information derived from adjacent points [19]. Positioning errors can also be reduced by attempting to snap coordinates to a known network, by filtering outliers based on one or more properties of the measurements [20] or their context [21], or by applying smoothing techniques [11].

D. Level 3 - Augmented data

After the preprocessing stage, the spatio-temporal measurements are passed through an activity detection algorithm, which is responsible for discovering potential activity of the user (i.e. movement). If an activity is detected, it continues to the trip segmentation process. There, the activity is further analyzed to detect intermediate stops. These stops might indicate a change of transport mode. The transport mode for each segment is determined. Finally each segment is map matched to align the GNSS track to the road network. This step can help with reducing missing data, by imputing intermediate locations, and can also reduce positioning errors, by aligning the geometry to the network.

Once the trip geometries are aligned to the network, it is fairly trivial to link specific measurements to the most appropriate edge in the network graph. This enables the aggregation of speeds, traffic intensities, etc. for specific network edges. Other processes, such as trip purpose imputation, may also run at this level. The data can also be augmented with, for example, weather information or personal information about the people undertaking the trips (e.g. age, social status, ...).

E. Level 4 - Aggregated data as insights and analytics

The trips and trip segments, along with their purposes, transport modes and other annotations can now be aggregated into specific indicators that are useful for researchers, urban planners and policy makers, such as the modal split, distributions of the travelled distance and origin-destination matrices. Additionally, the network linked measurements can be used to construct speed and intensity maps.

IV. ERROR PROPAGATION

As geospatial data is usually the subject of a large number of transformations and Geographic Information System (GIS) operations, the propagation of errors present in the data is of utmost importance [22, 23]. Errors the input data propagate to the outputs of each individual process. As some of these processes may run sequentially, the output of one process is likely used as input for other processes, and therefore the errors continue to propagate. Heuvelink proposes the use of Taylor series expansion to quantify the output error $U(\cdot)$ if the GIS operation $g(\cdot)$ is non-linear. Alternatively, one can use the Monte Carlo method, where one computes statistics of the output distribution by executing $g(a_1, \dots, a_m)$ repeatedly with randomly sampled values a_i from the input distribution.

The Monte Carlo method is easier to apply than the Taylor series expansion if $g(\cdot)$ is a complex operation, but only yields numerical results, and care has to be taken to properly condition the input values in case their distributions A_i are correlated.

The processing required for GNSS mobility studies requires complex transformations that cannot easily be described in terms of mathematical formulas or in- and output distributions. Often the input values will also be heavily correlated. As shown in Section II, a number of metrics to define the output quality of some of these individual processes have been described in literature, but the error propagation within a process chain formed by these individual modules has not yet been sufficiently studied.

Figure 2 illustrates how low-level errors can propagate to higher levels and thereby affect derived information, such as modal split statistics, average distances or trip durations. In this example, a trip consisting of consecutive segments of foot, car and foot transportation is oversegmented, introducing short walking and driving segments, during periods which exclusively consist of car and walking, respectively. These oversegmentations and mode misclassifications lead to vastly different travel statistic and modal split outputs. Further steps in the processing chain can exacerbate this issue. If the map matching fails because the raw locations are too distant from the network, it can create a gap in the matched trajectory, which potentially decreases its length. Two additional problems may occur if there has been a transport mode misclassification. The points could either be linked to the wrong type of network (e.g. the pedestrian instead of the car network), which can cause needless detours and therefore increase the trip length, or the misclassifications could lead to gaps if there is no network for the incorrectly chosen transport mode.

V. ADDITIONAL METRICS

To be able to accurately characterize the quality and credibility of reported transport indicators, it is necessary to calculate quality measurements for each step in the processing chain separately. These measurements can then be aggregated into distributions if several measurements are combined to form a new, higher level entity, such as the segmentation of a set of spatio-temporal measurements into one or more trip segments. These calculations preferably occur for each entity that is produced by the process individually, but if no such calculations exist, one can instead fall back to using more general, pre-calculated metrics. We present such metrics for spatio-temporal measurement sequences and map matched trip segments.

A. Spatio-temporal measurement sequences

For spatio-temporal measurement sequences in GNSS tracking data, we calculate seven metrics. First, we determine the distribution of the time differences between subsequent points. Secondly, the heading changes between subsequent points are measured. Then the distance and time difference between the subsequent points are used to calculate the speeds. That

information is later used to determine the speed difference between successive measurements, along with the discrepancy between the calculated speed and the speed that is measured by the GNSS device, if that is available. For each of these metrics, a distribution is sketched using Bowley's seven-figure summary [24].

B. Map matched trip segments

Map matching services are typically not able to completely match a trajectory to a given network. This can lead to gaps in the matching, or to the non-alignment of a number of individual points. Thus, we determine the number of non-aligned points and the ratio of non-aligned points compared to the total number of points in the segment. For the gaps, we calculate the distance and duration between the start and the end of the gap (i.e. the final aligned point before the gap and the first aligned point after the gap, respectively). Additionally, we also consider the distribution of the distances between the original and network-aligned coordinates, and the duration and length of the micro-segments that are created by the matching process. Finally, if the specific map matching implementation supports this, we also keep track of the confidence values that are produced by the algorithm along with the actual matching.

VI. CONCLUSIONS

This paper presents a systematic analysis of error sources and error propagation in mobility studies. Previous research has focused on the validity of specific processes. These processes form a chain of operations that transform raw data to high level mobility indicators. Research has shown that errors that occur in the early processing stages can have profound effects on the later stages and the final results. Traditional metrics such as accuracy and recall are useful in many cases, but care must be taken to interpret those metrics correctly. Accuracies as high as 94% can be achieved for transport mode classification, but they may hide subtle issues, such as oversegmentation, which can have a serious impact on the validity of common mobility indicators, such as travelled time and distance with each transport mode. Our future research will focus on developing metrics that can be used to more clearly define the propagation of errors throughout the processing chain, along with applying and evaluating this technique on a number of data sets.

REFERENCES

- [1] S. Vlassenroot, D. Gillis, R. Bellens, and S. Gautama, "Smartphones monitoren verplaatsingsgedrag: Universiteit Gent ontwikkelde MOVE-platform en CONNECT-app," *VERKEERSSPECIALIST (MECHELEN)*, no. 193, pp. 8–11, 2013, ISSN: 1379-4922.
- [2] I. Semanjski, A. J. Lopez, 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, p. 1035, Jul. 2016. DOI: 10.3390/s16071035.
- [3] L. Shen and P. R. Stopher, "Review of GPS Travel Survey and GPS Data-Processing Methods," *Transport Reviews*, vol. 34, no. 3, pp. 316–334, 2014. DOI: 10.1080/01441647.2014.903530.
- [4] "Can using global positioning system (GPS) improve trip reporting?" *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 2-3, pp. 149–165, Apr. 1999. DOI: 10.1016/S0968-090X(99)00017-0.
- [5] N. Schüssler and K. W. Axhausen, "Identifying trips and activities and their characteristics from GPS raw data without further information," *Arbeitsberichte Verkehrs- und Raumplanung*, vol. 502, 2008. DOI: 10.3929/ETHZ-A-005589980.
- [6] A. J. Lopez, I. Semanjski, D. Gillis, D. Ochoa, and S. Gautama, "Travelled Distance Estimation for GPS-Based Round Trips Car-Sharing Use Case," *Transactions on Maritime Science*, vol. 2, pp. 121–129, 2016. DOI: 10.7225/toms.v05.n02.003.
- [7] A. J. Lopez, "Processing crowdsourced data for the analysis of mobility behaviour," PhD thesis, Ghent University, 2018, pp. 21, 68–84, ISBN: 9789463551359.
- [8] P. Ranacher, R. Brunauer, W. Trutschnig, S. Van der Spek, and S. Reich, "Why GPS makes distances bigger than they are," *International Journal of Geographical Information Science*, vol. 30, no. 2, pp. 316–333, Feb. 2016. DOI: 10.1080/13658816.2015.1086924.
- [9] F. Biljecki, H. Ledoux, and P. van Oosterom, "Transportation mode-based segmentation and classification of movement trajectories," *International Journal of Geographical Information Science*, vol. 27, no. 2, pp. 385–407, Feb. 2013. DOI: 10.1080/13658816.2012.692791.
- [10] K. Grochla and K. Polys, "Using Kalman Filters on GPS Tracks," in *Man – Machine Interactions 4*, A. Gruca, A. Brachman, S. Kozielski, and T. Czachorski, Eds., Switzerland: Springer International Publishing, 2015, ch. XI.2, pp. 663–672, ISBN: 9783319234366.
- [11] 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. DOI: 10.3141/1972-19.
- [12] A. C. Prelipcean, G. Gidofalvi, and Y. O. Susilo, "Measures of transport mode segmentation of trajectories," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1763–1784, Sep. 2016. DOI: 10.1080/13658816.2015.1137297.
- [13] A. J. Lopez, I. Semanjski, S. Gautama, and D. Ochoa, "Assessment of Smartphone Positioning Data Quality in the Scope of Citizen Science Contributions," *Mobile Information Systems*, vol. 2017, pp. 1–11, Jun. 2017. DOI: 10.1155/2017/4043237.
- [14] M. A. Quddus, R. B. Noland, and W. Y. Ochieng, "Validation of map matching algorithms using high precision positioning with GPS," *Journal of Navigation*, vol. 58, no. 2, pp. 257–271, 2005. DOI: 10.1017/S0373463305003231.
- [15] —, "Integrity of map-matching algorithms," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 4, pp. 283–302, Aug. 2006. DOI: 10.1016/j.trc.2006.08.004.
- [16] —, "The effects of navigation sensors and spatial road network data quality on the performance of map matching algorithms," *Geoinformatica*, vol. 13, no. 1, pp. 85–108, Mar. 2009. DOI: 10.1007/s10707-007-0044-x.
- [17] M. A. Quddus, "High Integrity Map Matching Algorithms for Advanced Transport Telematics Applications," PhD thesis, Imperial College London, Jan. 2006, p. 270.
- [18] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 5, pp. 312–328, Oct. 2007. DOI: 10.1016/j.trc.2007.05.002.
- [19] S. E. Wiehe, A. E. Carroll, G. C. Liu, K. L. Haberkorn, S. C. Hoch, J. S. Wilson, and J. D. Dennis, "Using GPS-enabled cell phones to track the travel patterns of adolescents," *International Journal of Health Geographics*, vol. 7, no. 1, p. 22, May 2008. DOI: 10.1186/1476-072X-7-22.
- [20] J. Wolf, "Using GPS Data Loggers To Replace Travel Diaries In the Collection of Travel Data," PhD thesis, Georgia Institute of Technology, 2000.
- [21] N. Schüssler and K. W. Axhausen, "Processing Raw Data from Global Positioning Systems Without Additional Information," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2105, no. 1, pp. 28–36, Jan. 2009. DOI: 10.3141/2105-04.
- [22] G. B. M. Heuvelink, "Propagation of Error in Spatial Modeling with GIS," *Geographical Information Systems Volume 1 Principles and Technical Issues*, vol. 2, no. 14, pp. 207–217, 1999.
- [23] P. P. Siska and I.-K. Hung, "Propagation of Errors in Spatial Analysis," in *24th Applied Geography Conference*, Fort Worth, Texas, 2001.
- [24] A. L. Bowley, *An elementary manual of statistics*. PS King & son, Limited, 1915.