Detection of outliers in crowdsourced GPS traces

Stefan S. Ivanović[*1], Ana-Maria Olteanu Raimond[1], Sébastien Mustière[1], Thomas Devogele[2]

[1] Université Paris-Est, IGN, COGIT Lab., France
[2] Université François Rabelais de Tours, Laboratoire d'informatique France

*Corresponding author: stefan.ivanovic@ign.fr

Abstract
Nowadays, crowdsourced GPS data are widely available in a huge amount. A number of people recording them has been increasing gradually, especially during sport and spare time activities. The traces are made openly available and popularized on social networks, blogs, sport and touristic associations' websites. However, their current use is limited to very basic metric analysis like total time of a trace, average speed, average elevation, etc. The main reasons for that are a high variation of spatial quality from a point to a point composing a trace and a need for referential data for evaluation of their quality. In this paper we present a novel approach for filtering and detection of outliers in crowdsourced GPS traces in order to assess their spatial quality intrinsically and make them more suitable for more advanced uses such as updating referential road network of French Mapping Agency – IGN. In addition, we propose a new definition of an outlier in GPS data, adapted to intrinsic assessment of spatial quality.

Keywords
Crowdsourced GPS trace, filtering, outlier detection, machine learning

I INTRODUCTION

Intensive sports activities of professionals and amateurs are very frequently recorded by using GPS devices. The traces obtained are then made openly available to the community through social networks, blogs, sport and touristic communities' web-sites. Currently, they are mostly used for visualization purposes and some basic metric data analysis (e.g. total time of the trace, distance, speed). On the other hand, they have a huge potential to be used for more advanced purposes. (Bergman and Oksanen, 2016). The aim of our research is to use crowdsourced GPS traces as a potential data source for highlighting updates in a referential road network of the French National Mapping Agency (IGN).

In our work, secondary road network in mountainous area such as hiking, bicycle and tractor paths are in our focus. Even if not always necessary, this network is very important for production of touristic maps and for other different applications such as defence and sport activities. These paths are very challenging for continuous update due to their intermittent nature (e.g. they appear and disappear very often) and various landscape (e.g. forest, high mountains, seashore) which make the update process time consuming and the traditionally used methods such as stereo-restitution, insufficient.

The crowdsourced GPS traces are collected without any protocol, with low and heterogeneous frequency sampling and mostly by low class GPS devices. In addition, they are made available with few or inexistent metadata. Moreover, various errors are introduced by different external factors such as topography, canopy, etc. These errors can cause a significant bias and may limit

the usability of the traces in different analysis. Thus, using GPS traces for updating authoritative data makes their quality issues very important. To assess the quality of GPS traces, a first step is to detect and filter outliers. This paper focuses on detection of outliers in crowdsourced GPS traces in order to improve their spatial quality.

II OUTLIERS DETECTION IN GPS DATA

Many studies have analysed factors that influence the quality of GPS data. Environmental factors have been found most influencing. First, topography that reflects on positional error and a number of fixed positions (Lewis et al. 2007; Cain et al. 2005; DeCesare et al. 2005). Second, canopy cover, especially the density of its coverage, type of species (Klimànek 2010; Tucek and Ligos 2002) and height (Janeau et al. 2004) affect positional error. Third, obstacles degrade the quality of GPS signal by causing multipath effect (Tucek and Ligos 2002). Since they all confirmed the influences of the environment on GPS data quality, it is necessary to take these influences into account when dealing with such quality anomalies like outliers are.

We define an outlier as a GPS point whose metrics and geometrics characteristics differ significantly from the characteristics of other points composing a GPS trace. Overall, most of works on outlier's detection in GPS data has been considered GPS measurements' errors as outliers and treated them by various filtering methods (Ordonez 2011; Duran 2012; Knight 2009). In addition, few works defined outliers from geometric point of view, as points that differ from the rest of the traces following the same path (Etienne 2013; Gil de la Vega et al. 2015). In our work, we only consider a single trace which does not require presence of other traces. That is important in situations where only one trace or few traces exist for the same path not allowing to define a pattern, like in some challenging mountainous areas.

Red arrows in Figure 1 point examples of outliers whose positions cause values of speed, distance and angle between them and their consecutive points significantly different compared to other points of the same trace.

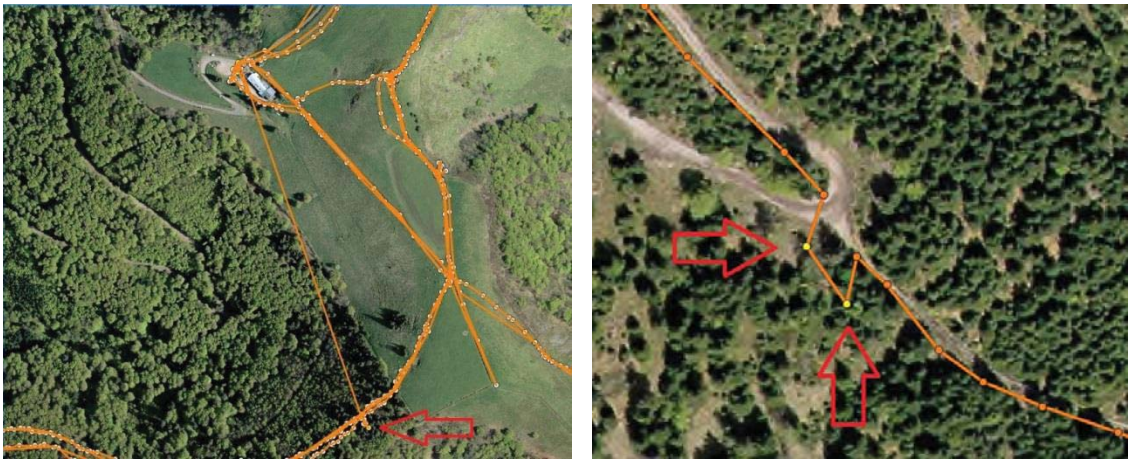

Figure 1: Examples of outlier points.

III TEST DATA

Crowdsourced GPS data recorded during sport activities such as running, cycling, walking and hiking in Vosges Mountains are collected from different websites and portals of French sport associations. We chose this area due to various landscapes with a plenty of secondary roads, as well as an important number of crowdsourced GPS traces.

In total, we selected 436 traces composed by 292337 points. GPS points are theoretically described by their position (WGS84), elevation and timestamps. Practically, some points have missing attributes, that is, missing elevation or timestamp or even both. Six indicators measuring the completeness of the attributes are computed. We found that 106206 points (36%) lacked timestamp, whereas the situation with elevation was far better – only 6580 points (2%) lacked elevation. Regarding the traces, we noticed that 157 (36%) had no timestamps at all, whereas 287 (66%) had at least one missing timestamp.

IV METHODOLOGY AND RESULTS

Due to the heterogeneous nature of crowdsourced GPS data and more random than systematic influences of environment and other known factors, the detection of outliers is a complex task. Many intrinsic and extrinsic indicators may be computed to detect outliers. However, determining relevant criteria and thresholds as well as how to combine them is also a challenging task. Thus, to manage these difficulties, we propose an approach based on supervised machine learning techniques in order to generate generic rules and thresholds.

Our approach is composed by four steps: i) the first one consists in filtering noise such as redundant points, negatives values, etc.; ii) the second is the computation of different intrinsic and extrinsic indicators; iii) the third one consists in using machine learning, which involves a supervised sampling followed by applying a classification algorithm; iv) finally, the generated rules and thresholds are applied on non-classified points of the test area.

First, within the step of noise filtering, redundant points (e.g. overlapping consecutive points) are filtered by means of zero distance values in a small time window. Negative speed values are used to detect structural errors in GPX files such as two traces irregularly merged into one or errors of GPS clock. This is an important step since it was spotted that in some cases negative speed values caused geometric anomalies of the traces. In total 4594 points (2%) were filtered thanks to this step.

To formalize outliers' detection, we propose intrinsic metrics and extrinsic indicators. The former are calculated from GPS measures only, such as distance, speed, direction and elevation between consecutive points. The later are calculated based on the analysis of the spatial context in which GPS points are recorded such as type and density of forest, slope and its curvature, proximity of obstacles (e.g. cliffs, buildings, forest) and other features (e.g. river, lake, building).

In total, 15 different indicators were computed. One representing direction change, three based on distance and speed respectively, taking into account their mean and variations between consecutive points, four derived from elevation (GPS and corresponding DTM) and four based on spatial context such as: proximity of obstacles (e.g. buffer of cliffs, buildings, forests), point in lake/river/building, point in the forest, type of forest. It is important to stress that due to missing attributes (e.g. lack of timestamps) for some points, some measures cannot be calculated (e.g. speed). The indicators are calculated for each point taking into account previous and next point in the trace.

In order to apply the machine learning techniques, first a sampling zone was chosen so that it represented faithfully the state and heterogeneities of the entire test area and the entire pattern of GPS points. Percentage of missing attributes (timestamps and elevation) of sampled points compared to the total percentage of missing attributes differed for only 2%. In total 2342 points were sampled and manually classified as outliers (3%) or regular points (97%). The learning process was conducted in WEKA software package using JRIP algorithm proposed by Cohen (1995). Evaluation of results by means of 30 fold cross validation is presented in a confusion matrix in Figure 2.

	Predicted		
Actual class		Outlier	Not an outlier
	Outlier	61	16
	Not an outlier	16	2249

Figure 2: Confusion matrix of 30 fold cross validation

The number of correctly classified outliers is almost four times bigger than the number of misclassified, while overestimation and underestimation are balanced. In addition, a global precision of the approach calculated from the matrix is 98%. Both results confirm high performance of the approach and generated rules, particularly taking into account presence and random distribution of missing attributes.

In total, five rules as well as thresholds were generated and presented in Figure 3.

1. AngleMean>=87°.54 and DistDiffMed>=1.05 => outlier
2. AngleMean>=71°.23 and SpeedRate>=1.5 => outlier
3. AngleMean>=74°.80 and DistDiffN<=0.21 => outlier
4. AngleMean>=83°.15 and SpeedRate <=0.85 => outlier
5. AngleMean>=56°.43 and DistMean>=8847.31m => outlier

Figure 3: Rules generated in machine learning

Where: i) AngleMean represents an average value of 3 direction changes; ii) DistDiffMed a relation between two consecutive distances (before and after a point that is evaluated) and a median distance of a trace; iii) DistDiffN a normalized value of two consecutive distances; iv) DistMean defines a mean distance of two consecutive distances and v) SpeedRate represents the velocity change rate being proposed by Winden et al (2016).

Finally, the generated rules were applied on the entire test area. As a result, 9309 points (3%) were recognized as outliers. In Figure 4, an example of a successfully detected outlier is illustrated.

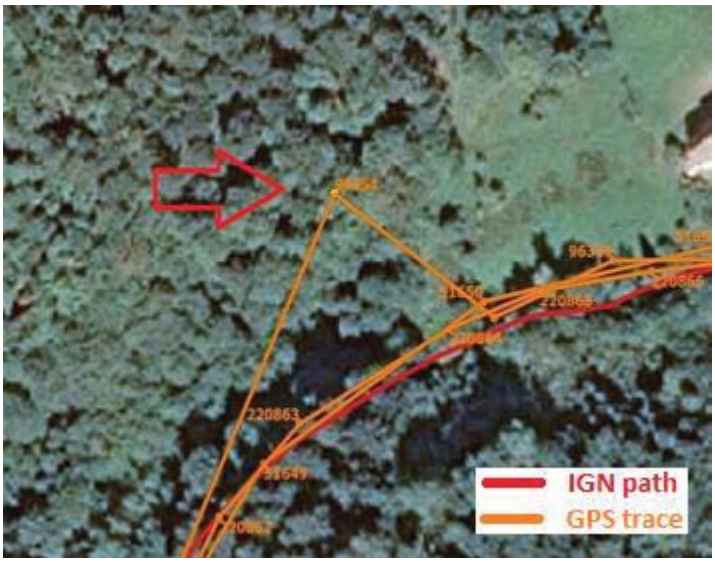

Figure 4: An outlier detected within non-classified points

Although the successfulness of the approach is high, there is a specific situation where it does not perform properly causing overestimations. This happens when treating traces with very high sinuosity and low spatial resolution such as illustrated in Figure 5.

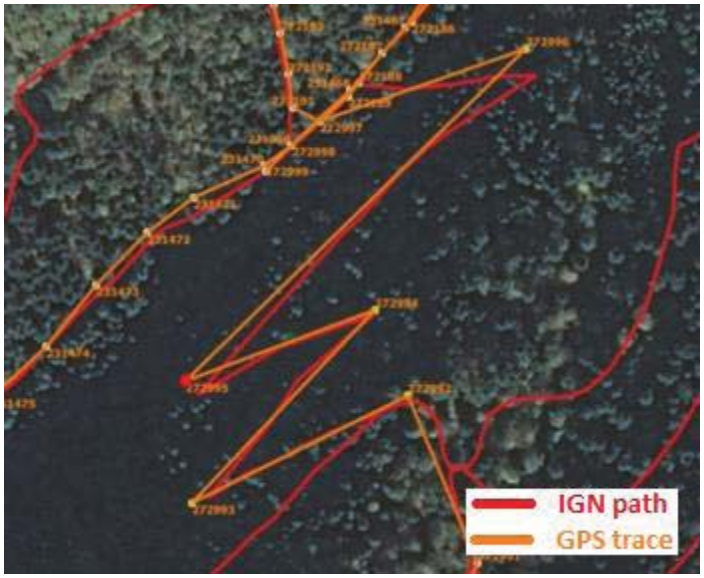

Figure 5: Misclassified outliers

All five points composing a part of the trace that fits referential road are identified as outliers. Such geometries of referential roads are not frequent (in our test area only 11), thus the overestimation produced are not numerous. However this should be treated by taking into account the sinuosity of referential roads.

Concerning the influence of external (environmental) factors, it is obvious that based on the generated rules, the impact and importance of the external factors on appearance of outliers were not discovered. It is likely that a link between them exists, however it is difficult to be detected and modelled in crowdsourced traces due to their extreme heterogeneity, and a lack of two very important information such as precision of GPS device and the quality of signal.

For example, the same obstacle produces a multipath and subsequently outlier while collecting data by low precision GPS, whereas it is not a case for high precision GPS under the same conditions. The same situation is with canopy cover. More precise device using better quality signal can produce a trace without outliers even in closed coniferous forest, while lower precision device supplied with low quality signal would cause outliers even in an open area.

V CONCLUSION

GPS traces are widely spread as crowdsourced data due to the significant number of sportsmen's and amateurs recording them during sport or leisure activities. Our research is aiming to use them for highlighting updates in referential secondary road network. To do that, it is necessary to assess the traces quality, particularly to deal with significant anomalies and of their geometry that we defined as outliers. In this paper, we presented a novel approach for formalization and detection of the outliers despite the high heterogeneity of crowdourced GPS traces and dominantly random influences of external factors on their quality. The outliers were modelled and detected using 15 intrinsic metric and spatial context indicators calculated for each point in a trace. The successfulness of the approach is generally high, except in the case where a sinuosity of a referential road is very high. By analysing generated rules and the nature of crowsourced data, we conclude that causes of the outliers are obviously highly heterogeneous. Thus, the most efficient way to model and detect them is by means of their intrinsic indicators, particularly taking into account that links between spatial context and crowdsouced traces' quality proved to be weak.

References

Bergman C., Oksanen J. (2016). Conflation of OpenStreetMap and Mobile Sports Tracking Data for Automatic Bicycle Routing. *Transactions in GIS* 00(00): 00-00

Cain III J.W., Krausman P.R., Jansen B.D., and Morgart J.R. (2005). Influence of topography and GPS fix interval on GPS collar performance. *Wildlife Society Bulletin 33*(3), 926-934

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Twelfth International Conference on Machine Learning*, Tahoe City, California

DeCesare N.J., Squires J.R., and Kolbe Y.A. Effect of forest canopy on GPS-based movement data. *Wildlife Society Bulletin 33*(3), 935-941

Duran, A., Earleywine, M. (2012). GPS Data Filtration Method for Drive Cycle Analysis Application. In *SAE 2012 World Congress,* Detroit, Michigan

Etienne, L. (2011). Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction àla détection de comportements inhabituels - Application au trafic maritime. PhD thesis. Institut de Recherche de l'Ecole Navale.

Gil de la Vega, P., Ariza-Lopez, F.J., Mozas-Calvache, A.,T. (2015). Detection of outliers in sets of GNSS tracks from volunteered geographic information. In *Agile Conference 2015,* Lisbon

Janeau G., Adrados C., Joachim J., Gendner J.P., Pépin D. 2005. Performance of differential GPS collars in temperate mountain forest. *C. R. Biologies 327*, 1143–1149

Klimànek M. 2010. Analysis of the accuracy of GPS Trimble JUNO ST measurement in the conditions of forest canopy. *Journal of Forest Science 56*(2), 84–91

Knight, N. L., Wang, J. (2009). A comparison of outlier detection procedures and robust estimation methods in GPS positioning. *J. Geodesy 62*(4), 699-709

Lewis J.S., Rachlow J.L., Garton E.O. and Vierling L.A. (2007). Effects of habitat on GPS collar performance: using data screening to reduce location error. *Journal of Applied Ecology 44*, 663–671

Ordoñez, C., Martínez, J., Rodríguez-Pérez, J., and Reyes, A. (2011). Detection of Outliers in GPS Measurements by Using Functional-Data Analysis. *Journal of Surveying Engineering137*(4), 150-155

Tuček, J. Ligoš J. 2002. Forest canopy influence on the precision of location with GPS receivers. *Journal of Forest Science 48*(9), 399–407

Van Winden, K., Biljecki, F. and van der Spek, S. (2016). Automatic Update of Road Attributes by Mining GPS Tracks. *Transactions in GIS* 00(00): 00-00