

Detecting Outlier Segments in Uncertain Personal Trajectory Data

Sungsoon Hwang[1(✉)], Kushal Navghare[2], Umer Huzaifa[2], Ilyas Ustun[2], Alex Dzewaltowski[3], and Christopher Connaboy[3]

[1] Department of Geography, DePaul University, Chicago, IL 60614, USA
shwang9@depaul.edu
[2] School of Computing, DePaul University, Chicago, IL 60604, USA
{knavghar,mhuzaifa,iustun}@depaul.edu
[3] Scholl College of Podiatric Medicine, Rosalind Franklin University of Medicine and Science, North Chicago, IL 600064, USA
{alex.dzewaltowski,christopher.connaboy}@rosalindfranklin.edu

Abstract. Uncertainties such as missing data, noise, and outliers, are common in personal trajectory data due to extended indoor time. To improve the usability of such data, it is necessary to detect and repair errors. While smoothing-based methods can be used to deflate noise, such methods are not well suited to cleaning trajectory data in the presence of spatial outliers represented by abrupt change in location. We propose detecting spatial outliers at the level of segments using supervised machine learning algorithms. We partitioned a trajectory into segments based on time gaps (greater than recording interval) and used smoothed values of relevant features including movement state at the level of segments for detecting spatial outliers. Experimentation results indicate that Extreme Gradient Boosting outperforms Random Forest and Light Gradient Boosting Machine methods. The study indicates that gaps and movement state can be considered to partition a trajectory and improve outlier detection.

Keywords: Trajectory Data Preprocessing · Outlier Detection · Machine Learning · Trajectory Segmentation · GPS Errors

1 Introduction

A large volume of personal trajectory data has become readily available as Global Navigation Satellite Systems (GNSS) and location-aware devices (e.g., smartphones) have become more pervasive. Personal trajectory data track a person' whereabouts and movement at fine spatiotemporal resolution (e.g., every second to minutes). The data refers to a set of consecutive track points with a tuple of location and time (x, y, t) given sampling rate for any duration (e.g., hours to years) [1]. There are many use cases of personal trajectory data. Spatially explicit activity patterns that are inferred from personal trajectory data can be used to develop location-based recommendation systems. A large collection of historical trajectory data on road networks can inform real-time car navigation,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14814, pp. 418–426, 2024.
https://doi.org/10.1007/978-3-031-64608-9_28

traffic congestion mitigation, and transportation/land use planning. The environmental etiology of diseases has been largely unaccounted for due to challenges with measuring environmental exposure at a full scale in space and time. Now those challenges are being overcome by advances in the technologies mentioned above [2].

Uncertainties in personal trajectory data, however, have been the barrier to realizing the potential of the data. There are three major sources of uncertainty in personal trajectory data: a gap, noise and outliers. A gap refers to incomplete track points (records) caused by signal loss, and the latter two refer to inaccurate records caused by measurement error. GNSS signals do not typically penetrate thick structures like buildings, tunnels, and tree canopies, resulting in incomplete/irregular sampling in the trajectory data. Multiple sources like ionospheric delay (signal reception deterred by particles in ionosphere), multipath interference (signal reception deflected off by structures nearby), clock errors in GNSS receivers, and geometry among satellites, contribute to measurement errors in the trajectory data.

Noises are expressed as back-and-forth movement typically around an indoor location or slight misalignment with road networks outdoors. Figure 1 depicts track points that were logged for an hour when a person was staying inside a single building in downtown Chicago with many high rises [3]. Individual deviational points (or outliers) can occur and are typically represented by a sudden jump in location as observed in different geographic scales in Fig. 2.

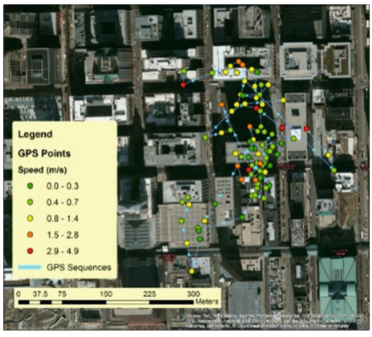

Fig. 1. Noise in personal trajectory data.



Fig. 2. Outliers in personal trajectory data.

Incomplete and inaccurate track points account for a large proportion in 'personal' trajectory data in contrast to 'vehicle' trajectory data due to extensive time spent indoors in an increasingly urban setting. These uncertainties deter the development of accurate and reliable activity/place recognition algorithms. Further, most of these algorithms are developed based on relatively clean 'vehicle' trajectory data and thus those algorithms do not work well at uncertain personal trajectory data. Untreated uncertainties lower the usability of personal trajectory data in many real-world applications discussed above. Observations on uncertain personal trajectory data reveal that a gap, noise, and outliers are not independent of each other. That is, a gap, noise and outliers are interspersed with each other as they relate to structural interference.

Review of related work on cleaning trajectory data suggests several findings. First, smoothing feature values over sliding windows is most widely used to clean trajectory data. It is, however, not highly effective at cleaning the 'personal' trajectory data due to high level of uncertainties. Smoothing-based methods are more geared toward filtering out slight noise in outdoor trajectory data. Furthermore, despite the dependence of uncertain events (a gap, noise, and outliers) in trajectory data, existing methods rarely consider those events altogether. Relatedly, a recent study combines constraint-based methods for detecting outliers and smoothing-based methods for repairing noise [4]. This method, however, may not be effective at detecting outliers that are spatiotemporally clustered, as depicted at the lower right of the inset map of Fig. 2.

In this paper we will focus on detecting outlier segments from personal trajectory data. An *outlier segment* is defined as a set of one or more consecutive track points that are regarded as an outlier whose location deviates from the actual location beyond the acceptable level of error. The acceptable level of error will be formulated depending on intended use. There are several reasons why we believe detecting outlier segments is significant or takes precedence over other uncertain events discussed above. First, imputing unknown location during signal loss (gap) in trajectory data is sensitive to the presence of outliers [5] as much as smoothing noise is. In other words, detecting outlier segments should precede handling a gap and noise to improve the overall performance of trajectory cleaning. Second, outliers and slight noise should be treated differently due to different variations in errors. Smoothing can reliably clean trajectory data if outliers are singled out and repaired first. Third, outliers typically exist in segments rather than in an isolated manner. That is, detecting outlier segments is a more general case than detecting an individual outlier.

In this paper, we propose partitioning personal trajectory data to segments based on gaps and classifying segments into outlier vs. non-outlier category using supervised machine learning (ML) techniques. We consider the movement state of segments as an additional feature. ML algorithms are employed to account for dynamic, non-linear relationships among features associated with the likelihood of outlier segments. Segment-level outlier detection is considered less computationally intensive than individual-level analysis. We report on preliminary findings on using the proposed approach to detecting positional outliers in personal trajectory data.

The paper is organized as follows. Section 2 reviews existing methods for detecting errors in trajectory data. Section 3 describes the process of collecting, processing, and labeling personal trajectory data as well as the proposed method. Section 4 describes experimentation results related to the choice of parameters and machine learning algorithms. Section 5 concludes the study.

2 Related Work

This section focuses on existing methods related to detecting 'spatial' errors in 'offline' personal trajectory data. A spatial error can be quantified as the degree to which a measured location deviates from a true location. It is known that standard GNSS receivers available in the market today have a spatial (positional) error between 2 and 10 m in ideal outdoor conditions. However, horizontal accuracy varies greatly due to reasons as

discussed earlier. Due to difficulties with obtaining labeled data (i.e., track points with the actual location), error detection algorithms typically mark errors if a track point's measured location deviates from its 'expected' location. In general, data points with location deviation beyond acceptable level given intended use can be considered errors as a categorical variable.

Several methods have been proposed to deal with spatial errors in trajectory data. Those methods can be classified into smoothing-based, constraint-based, and modeling-based approach. Smoothing techniques can be used to detect errors by comparing a track point's measured coordinate to its smoothed coordinate within temporal sliding windows. Constraint-based approach attempts to detect errors based on conditions imposed on data (e.g., errors if signal quality is poor, errors if the distance from the previous track point is beyond possible). A modeling-based approach builds a model that estimates the probability of track points' being errors as a function of related features, and mark errors if the probability is high. Below we review existing methods that fall into these categories.

Smoothing-Based Methods: this approach smooths geographic coordinates of a track point using filtering techniques of different kinds (e.g., mean, median, Kalman, particle) over sliding windows. A sliding window is typically made of consecutive track points before and after a track point to be processed for any error. This works to replace original coordinates with expected coordinates within sliding windows. The intent is to deflate noise that exists in raw trajectory data.

This approach is well suited to correcting (repairing) the location of slight noise, but the effectiveness of error correction is put into question in the presence of outliers. Smoothing in the segment of trajectory data containing outliers (Fig. 2) can result in more distorted coordinates than original coordinates [6]. Research also shows that the performance of filtering is sensitive to the size and partitioning of sliding windows [7]. Sliding windows of fixed size are typically used for convenience. Using sliding windows with disregard to uncertain events (such as gap) is expected to lower the performance of filtering. Little is known about effective partitioning strategies of sliding windows for smoothing and cleaning personal trajectory data.

Constraint-Based Methods: this approach constructs conditional rules that detect spatial errors based on threshold values. Threshold values can be determined globally (i.e., relative to the entire dataset) or locally (i.e., relative to temporal neighbors). Examples of global threshold values include extreme speed, acceleration, and altitude, where extreme values can be statistically determined. Spatial errors can be detected based on poor quality of GNSS signal indicated by a combination of high Dilution of Precision (DOP) (e.g., > 4), low number of satellites used (NSAT) (e.g., <4), and low signal-to-noise ratio (SNR) (e.g., <150). Examples of local threshold values include abrupt changes in speed or altitude. A track point beyond the maximum possible distance from the previous track point can be marked as an outlier (called range constraint), where the maximum possible distance is estimated based on the maximum possible vehicle speed [4].

While this approach is easy to implement, it is challenging to identify threshold values that are universally applicable. Moreover, in the case of rules based on locally unexpected values, the assumption should be made about the state (e.g., constantly moving) and

velocity of the previous track point or track points as a whole. If the previous data point is stationary, then the maximum possible distance from the previous data point is zero in theory, making it infeasible to detect outliers. Further, while a spatial outlier can occur individually in an isolated manner, the rule of 'range constraints' described above will fail to detect outliers if the previous track point is also an outlier. For this reason, we consider detecting *outlier segments* as a general case of an individual outlier. Lastly, it is difficult to construct rules that satisfactorily account for the dynamic, non-linear relations among different features; hence a modeling-based method is suggested as an alternative.

Modeling-Based Methods: this builds a model that classifies data points into a binary dependent variable y (error or not) based on a matrix of multiple features X. The relationship will be captured as coefficients (weights) of X in terms of mathematical models. Supervised machine learning (ML) algorithms for classification such as random forest and support vector machine (SVM) have been suggested as a method for detecting errors in trajectory data. A deep learning algorithm based on the Long Short-Term Memory (LSTM) was applied to detect outliers from outdoor data with about 86% accuracy [8].

3 Methods

We collected personal trajectory data around Chicago and its vicinity in the fall of 2023. A participant was asked to carry a GPS logger (QStarz BT-Q1000XT) all the time except for water activities for a total of 8 days. A GPS logger is configured to log a track point every second (i.e. with a 1-s recording interval) and stop logging it if a person is stationary for over 10 min. Total 170,488 track points were collected. A participant was asked to log the time and location at every start and end of trip made in a travel diary template provided to them during the tracking period. We ensure the accuracy of a travel diary that serves as ground truth before, during, and after data collection. That is, we directly supervise participants' completing a travel survey at the beginning of data collection, confirm the validity of a travel diary during data collection, and verify the accuracy of a travel diary against orthoimage of higher accuracy and independent sources (ESRI Imagery Hybrid) in ArcGIS Pro 3.0. We divided the collected trajectory data into training data (80%) and testing data (20%).

Trajectory data was partitioned into temporally consecutive segments, where a new segment starts whenever the difference in time from the previous track point exceeds a recording interval (1 s). This yielded a total of 143 segments. To label outlier segments in the collected trajectory data, research staff link the trajectory data to a travel diary based on timestamp and plot geographic coordinates of the trajectory data in ArcGIS Pro. The actual location of a track point is manually checked against a travel diary (for stop location) and ESRI Imagery Hybrid (for the location of a trip route and stop). Location deviation for each track point is calculated. For each of segments, research staff labels whether the segment can be marked as an outlier. An outlier segment is a set of temporally consecutive track points whose location unexpectedly deviates from the actual location. In Fig. 3, an outlier segment is depicted in pink line that is comprised of 11 track points in temporal sequence (labeled 1 to 11). These 11 track points (or traces) are not aligned with a road network, not constituting a trip.

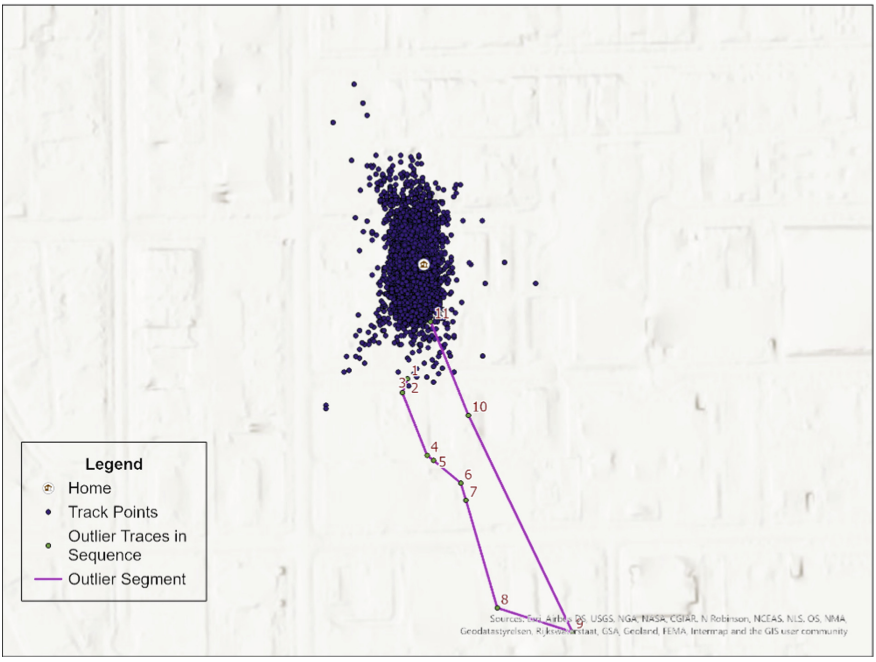

Fig. 3. An Outlier Segment in Personal Trajectory Data

The proposed method partitions the trajectory data into temporally consecutive segments as defined above and then flags outliers based on smoothed values of relevant features at the level of segments. For training machine learning (ML) algorithms, we consider descriptive statistics of the following features for each segment: speed, height, heading, PDOP, HDOP, VDOP, signal to noise ratio (snr_info), number of satellites used (sat_used), previous speed, previous heading, previous snr_info, previous sat_used, change in speed, change in heading, change in sat_used, distance from the previous track point (distance), the number of track points (num_points), and the number of track points falling into staypoints (total_staypoints).

To calculate total_staypoints, the individual track point is marked as staypoint or not using the staypoint-detection algorithm. The algorithm marks a staypoint if the distance among two or more consecutive track points exceeds a distance threshold and time span among those track points exceeds a time threshold [9]. In general, these threshold values should be chosen in relation to sampling rate, intended use, and data quality. These threshold values are chosen based on sensitivity analysis, to be discussed in the next section. For ML algorithms, we used random forest (RF), Light gradient boosting machine (LightGBM), and Extreme Gradient Boosting (XGBoost) for evaluation. Preliminary analysis indicates that these ML algorithms outperform k-nearest neighbor, support vector machine, and logistic regression.

4 Experimentation Results

To determine the optimal distance and time threshold values of a staypoint detection algorithm in relation to recording interval, we conducted sensitivity analysis. To this end we labeled outliers in a random sample of other trajectory datasets collected from the same GPS device with different recording intervals, described in [3]. Table 1 shows the optimal distance and time threshold (that maximizes precision and recall of detecting outliers in a balanced manner) for a staypoint detection algorithm. There is a relationship between distance threshold and recording interval as being a staypoint or not depends on speed. Therefore, distance threshold can be set to a product of recording interval (sec) and scale factor representing speed (in m/s). 3–6 m/s is recommended as scale factor considering walking speed and GNSS noise. It is, however, difficult to generalize the relationship between time threshold and recording interval given this analysis. Time threshold can be determined in relation to required granularity and intended use. Based on this analysis, we choose 3 m for distance threshold and 6 s for distance threshold.

Table 1. Optimal distance and time threshold by recording interval for staypoint detection

	Distance threshold	Time threshold
10-s interval	30 m	6 s
5-s interval	28 m	46 s
1-s interval	4 m	6 s

We evaluated classification accuracy of three ML algorithms (RF, LightGBM, XGBoost) on test data. Classification results with the best parameter values are shown in Table 2. Parameters are tuned to maximize F1-score. According to experimentation, XGBoost with precision 1 and recall 0.75 (F1-score 0.86) outperforms the other two ML algorithms (RF with F1-score 0.8 and LightGBM with F1-score 0.82).

Table 2. Precision, Recall, and F1-score for Detecting Outliers

	Precision	Recall	F1-score
Random Forest	1	0.67	0.8
Light GBM	0.9	0.75	0.82
XGBoost	1	0.75	0.86

Figure 4 plots relative importance of features used in classifying trajectory segments into outliers using XGBoost with 0.5 probability threshold. X axis in Fig. 4 shows the average gain of splits which use the feature. As expected, features indicating quality of GNSS signals (such as the number of satellites used and signal-to-noise ratio) contribute most significantly to this binary classifier. Other features like speed, distance, dilution

of precision (DOP), height, the number of staypoints, and change in heading are shown to be important.

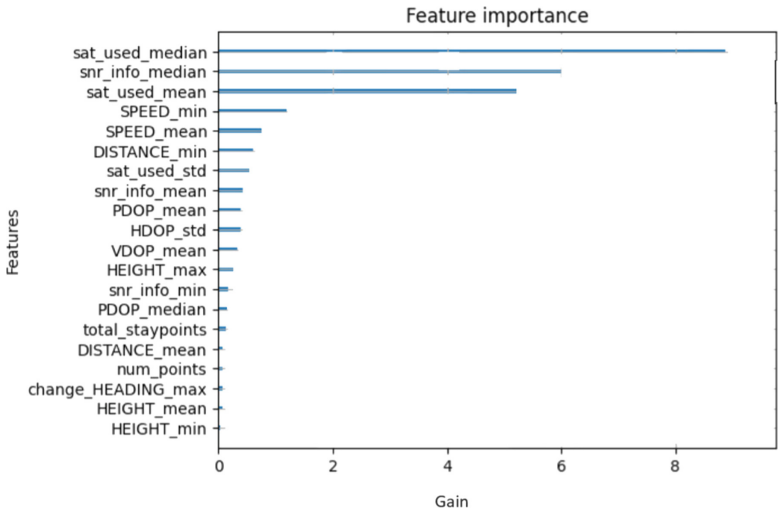

Fig. 4. Relative importance of features

5 Conclusions

In this paper, we consider methods for detecting positional outliers at the level of segments using supervised machine learning (ML) algorithms. We partition personal trajectory into segments based on time gaps and incorporate features pertaining to movement state using a staypoint detection algorithm, as well as other commonly available features. The study shows that segment-level classification with ML technique reaches an acceptable level of accuracy. The study also confirms that it is worthwhile considering movement states and uncertainties like gaps in detecting outliers.

The next step for the study is to devise and evaluate segmentation strategies that reflect movement state and detect spatial outliers at those segments. Such segmentation strategies are designed to incorporate movement state (whether a person was stationary or not) into feature space and aggregate track points into the spatiotemporally related unit of analysis that facilitates efficient error correction in the later step. Once outlier segments are identified, the location of outlier segments can be repaired in a different method than smoothing to make smoothing less sensitive to outliers. In the future research, other ML classifiers utilizing neural network will be considered. In general, the study indicates that it is necessary to consider interaction among uncertain events like gap, noise, and outliers in cleaning trajectory data.

Acknowledgments. This study was funded by the 2023 DePaul-RFUMS joint grant (project number 501763).

References

1. Liu, G., Gu, R., Wang, J., Yan, W.: Personal trajectory with ring structure network: algorithms and experiments. Secur. Commun. Netw. **2021**, e9974191 (2021). https://doi.org/10.1155/2021/9974191
2. Richardson, D.B., Volkow, N.D., Kwan, M.-P., Kaplan, R.M., Goodchild, M.F., Croyle, R.T.: Spatial turn in health research. Science **339**(6126), 1390–1392 (2013). https://doi.org/10.1126/science.1232257
3. Hwang, S., VanDeMark, C., Dhatt, N., Yalla, S.V., Crews, R.T.: Segmenting human trajectory data by movement states while addressing signal loss and signal noise. Int. J. Geogr. Inf. Sci. **32**(7), 1391–1412 (2018). https://doi.org/10.1080/13658816.2018.1423685
4. Fang, C., Wang, F., Yao, B., Xu, J.: GPSClean: a framework for cleaning and repairing GPS data. ACM Trans. Intell. Syst. Technol. **13**(3), 40:1–40:22 (2022). https://doi.org/10.1145/3469088
5. Hwang, S., Webber-Ritchey, K., Moxley, E.: Comparison of GPS imputation methods in environmental health research. Geospatial Health **17**(2) (2022). https://doi.org/10.4081/gh.2022.1081
6. Lee, W.-C., Krumm, J.: Trajectory preprocessing. In: Zheng, Y., Zhou, X. (eds.) Computing with Spatial Trajectories, pp. 3–33. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-1629-6_1
7. Thalmann, T., Abdalla, A.: Assessing the influence of preprocessing methods on raw GPS-data for automated change point detection. In: Huerta, J., Schade, S., Granell, C. (eds.) Connecting a Digital Europe Through Location and Place. LNGC, pp. 123–139. Springer, Cham (2014). http://link.springer.com/chapter/10.1007/978-3-319-03611-3_8. Accessed 31 July 2015
8. Sbetti, D., Tessaris, S.: Cleaning outdoor activity logs using deep learning. Presented at the SIGNAL 2022, The Seventh International Conference on Advances in Signal, Image and Video Processing, May 2022, pp. 1–6. https://personales.upv.es/thinkmind/SIGNAL/SIGNAL_2022/signal_2022_1_10_60003.html. Accessed 7 Feb 2024
9. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.-Y.: Mining user similarity based on location history. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2008, pp. 34:1–34:10. ACM, New York, NY, USA (2008). https://doi.org/10.1145/1463434.1463477