International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 3, Juni 2020, pp. 2742~2754
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i3.pp2742-2754

A comprehensive insight towards pre-processing methodologies applied on GPS data

R. Prabha[1], Mohan G. Kabadi[2]
[1]Visvesvaraya Technological University, India
[2]Department of Computer Science and Engineering, Sai Vidya Institute of Technology, India

Article Info

Article history:

Received Apr 30, 2019
Revised Nov 6, 2019
Accepted Nov 24, 2019

Keywords:

Data cleaning
Data pre-processing
Global positioning system
Receiver
Satellite

ABSTRACT

Reliability in the utilization of the Global Positioning System (GPS) data demands a higher degree of accuracy with respect to time and positional information required by the user. However, various extrinsic and intrinsic parameters disrupt the data transmission phenomenon from GPS satellite to GPS receiver which always questions the trustworthiness of such data. Therefore, this manuscript offers a comprehensive insight into the data preprocessing methodologies evolved and adopted by present-day researchers. The discussion is carried out with respect to standard methods of data cleaning as well as diversified existing research-based approaches. The review finds that irrespective of a good number of work carried out to address the problem of data cleaning, there are critical loopholes in almost all the existing studies. The paper extracts open end research problems as well as it also offers an evidential insight using use-cases where it is found that still there is a critical need to investigate data cleaning methods.

Corresponding Author:

R. Prabha,
Visvesvaraya Technological University,
Belagavi, Karnataka, India
Email: research.prabha.r@gmail.com

1. INTRODUCTION

The utilization of the Global Positioning System (GPS) has been increasing since the last decade as it is one of the most cost-effective navigational assistance [1]. With the proliferated usage of smartphone, various navigational applications and location services are directly dependent on the GPS data. The GPS system extracts the signal information from the satellites in order to obtain the location-specific information. On the basis of the usage of the different GPS receiver, the information of the location is generically provided in the form of Longitude, Latitude, and altitude [2]. The interesting factor about the GPS signal is its publicly availability and accessibility. From a technical viewpoint, the time factor and the spatial factor are the sole backbones of GPS satellites that bear an atomic clock with superior synchronization capability. They are also capable of rectifying and compensating any form of drift in the clock timing with the ground devices very spontaneously. A radio signal is being consistently transmitted by all the satellites of GPS that consists of updated positional data and time information of that position. It is also believed that latency between the GPS satellite is highly dependent on the distance from the earthly receiver and GPS satellite as it is free from any dependency of the speed of satellite and moreover the radio-waves have uniform velocity [3]. There is a typical computation carried out by the earthly receiver too which is responsible for computing the appropriate positional information after it obtained multiple data from multiple satellites. The computation will need to be carried out with higher accuracy. It is also believed that in order to compute the position information by the GPS receiver, there should be the presence of at least four GPS satellite within a line of sight. Although, this is a condition and this condition is quite hard to be satisfied in many real-time cases. The signal forwarded by the GPS satellite consists of much essential information. The first

form of content is a code which bears pseudorandom characteristics. Information of this code is only identified and inferred by the GPS receiver. The receiver can obtain the epoch on the basis of multiple parameters from this code. The second content of the GPS signal is the message that bears the information of the position of satellite and transmission of the particular epoch. The receiver then computes the time of flight on the basis of these two parameters, i.e. time of arrival and time of transmission and this computed information is required by any users [4]. It should be known that the offset of the clock maintained within a receiver as well as the position of the receiver is something that is required to be computed in parallel to each other. Finally, the obtained information is converted to longitude, latitude, altitude, speed, etc. which is forwarded to the navigational system of the user. The map-update, traffic monitoring, etc kinds of application uses GPS sensor to record the 3D coordinates X, Y, Z with a time stamp characterized by its value, that too have another characterized like volume. If the period of recording is very large as well as if any system failure happens then its will have veracity. The veracity as uncertain or missing data or redundant data plays a crucial role in the operating of the accurate traffic management system. The large GPS data traces consist of all three characteristics of volume, value, and veracity along with velocity and variety. This large volume of the spectral data poses huge critical challenges during the data processing most importantly their large volume of data with a low grading a mix blend of raw data and the uncertain aspects impacts largely the data analytics process in both viewpoint of data science and engaging to mine the useful insight. Therefore, this paper reviews the existing system of GPS data pre-processing. Section-2 discusses the essential of GPS data followed by essential of the pre-processing mechanism of GPS data in Section-3. Section-4 discusses existing research work followed by a briefing of open-end problems in Section-5. Finally, Section-6 briefs about the conclusion and future work direction.

2. ESSENTIALS OF GPS DATA

Before briefing about GPS data, it is essential to understand the fundamental structure of it. It consists of *a receiver, ground control station*, and *satellite*. At present, the GPS system transmits the signal using two significant frequency levels i.e. the first one is 1,575.42 MHz and second one is 1,227.60 MHz. The widely used commercial application used by common people uses GPS signals that are encoded by course/acquisition code, and this encoding system involves codes of identification of all the satellites. Special accessibility is also given to military application where the GPS signal is encoded with precise code [5]. Although, there is a sophisticated process to ensure that data offered by GPS is accurate, but still various external factors have an impact on the accuracy (Figure 1) viz. i) effect of troposphere that causes radio reflection inducing errors, ii) effect of ionosphere causing much slower speed of signal propagation causing error in transmission process, and iii) effect of multipath transmission causing adverse effect of reflection due to many physical structures on the ground [6].

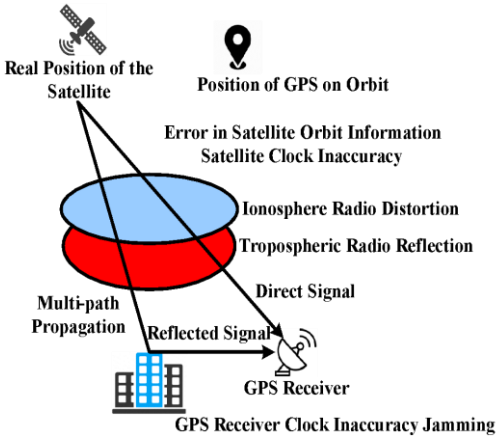

Figure 1. External factor for error in GPS data

In present times, a standard measure of *Dilution of Precision* or commonly known as DoP is used for checking how much is a degree of degradation has been invoked on GPS data in terms of accuracy [7]. Normally DoP value is smaller of there is a non-uniform position of GPS satellite or else DoP value is found higher. Another essential parameter is *signal* strength which signifies the level of signal stability too during

the reception state. Normally, GPS signal becomes unstable in the presence of obstacle or artifacts causing weaker strength of GPS signal. The third factor towards holding GPS data accuracy is aquantity of satellites with GPS capability. More the number of satellites, reliable are the positioning values. Discussion of such external factors causing degradation of GPS signal is publicly known; however, there are various internal factors too which is being investigated by the research community since the last decade. There are various forms and types of research-based solution to address this problem of pre-processing artifacts from GPS data. The standard process of performing this pre-processing the artifacts from GPS data are classified into two types viz. statistical-based approach and logical-based approach [8-11]. These methods were used traditionally to preprocess GPS data due to its behavior of less susceptible to the error streaming from sampling intervals. The density of the data point in GPS signal correlates itself with the probability factors of many notions of the vehicle moment on-track or off-track. Because of this, the low-density data point is considered an outlier in the case of GPS data. Majority of problems of artifacts in GPS data results due to missing data and following are the standard procedures to deal with the situation viz.

- *Outlier removal:* Traditionally, the data point of the GPS signal is initially sorted with an intelligent sense of either ascending or descending with the distance and medium, and the consistent data is taken as for further processing. The simplified method called Kerner density is used to get the density of the data points, and low-density data points are considered as redundant data. Other methods include adaptive density optimization, region growing clustering with knowledge. Most of the methods fail to handle outlier removal in the situation of the high-density data.
- *Trajectory Filtering:* In trajectory filtering the GPs data position accuracy is aims to be improvised. The approaches of adaptive Kalman filtering, particle filtering based methods are developed to smooth the noise than ensure reduction of error in the values of the data point. These filters interoperated the position and speed, but that is a computationally complex task.
- *System Model for GPS Data Collection:* The typical system model for the GPS data collection includes *N* user or custom devices equipped with the GPS sensor s.t D = {N_1, N_2, ...N_i}, where i =1 to N. The D_i record the data points of the GPS sensor and get logged with the local buffer, which is synchronized with a access point to cloud for the continuous stream and update of the data to the cloud and further for the numerical computing environment setup on the on-premise system A typical system architecture of the data generation, storage and the processing of the GPS data is shown in Figure 2.

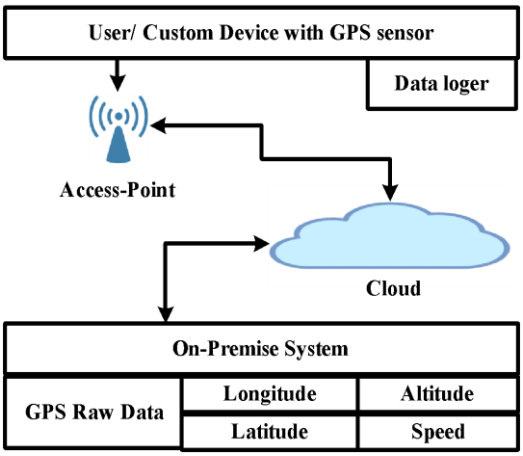

Figure 2. System model of the GPS data collection

3. PRE-PROCESSING GPS DATA

According to the existing research studies, all the process associated with the artifact removal of GPS data uses time-series analysis method and is broadly classified into two classes i) statistical based approach and ii) logical based approach. The research study on each approach is discussing as follows: i) Statistical based approach: Quantitative or statistical method is considered as one of the effective approaches to identify the best item sets and cleaned the datasets which are statistically closest to a user-specified data set [12]. Usually, the GPS data pre-processing method follows two significant phases; i) Error Detection and ii) Error Repairing [13].

3.1. Qualitative error/anomaly detection

This form of detection method deals with exploring statistical errors as follows;

- Error type: It relates to the search for the type of error and selecting the appropriate method to describe the patterns of legal data instance. Example- integrity constraints, first-order logic by the fractional method, functional dependencies, and denial constraints.
- Automation: This method clarifies how users are involved in the error detection method. Examples are the detection of functional dependencies andtracing all the replicated entries of data [14].
- Business-Intelligence: There are good possibilities of artifacts to occur on BI stack, likeerror-prone data are usually transmitted through certain communication channel with data processing capability. Meanwhile, majority of the strategies deals with tracing of the artifacts in data over actual database. Statistical Artifacts Tracing Taxonomy as shown in Figure 3.

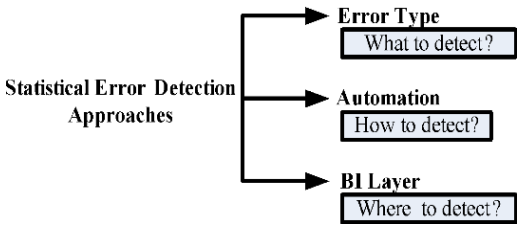

Figure 3. Statistical artifacts tracing taxonomy

3.2. Artifacts repairing method

Various instances of data are identified in this mechanism for ascertaining the essential quality demands of dataset. Similar as an error detection method, this method also addresses three significant questions like What, How and Where to repair. Error repairing method contains 3 phases as shown in Figure 4, viz.; i) Repair Target, ii) Automation and iii) Repair Model.

- Repair Target: This process makes a different assumption about data and quality rules, e.g. trusting declared integrity constraints, trusting the complete data, allowing constraints relaxation, exploring the changing possibilities of data and constraints. However, most of theapproaches deals with rectification of data considering over a set of artifacts while there are also presence of approaches towards involving communication medium as a root cause of errors.
- Automation: Specifically, error repairing techniques are classified according to the user's involvement (i.e., Where and how humans involved). Some of the existing techniques are fully automated (e.g., database recovery). Other techniques involve human interaction during the repairing process which verifies the repaired work or incorporate training operation in order to carry out involuntary decision of repairing [15].
- Repair Model: The existing methods repair the database in situ and destruct the database. The queries answered by repair model, sampledconsidering various possibilities of rectification with parallel solution towards the probabilistic approach [16]. Some popular error repairing methods are discussed in [17-22].

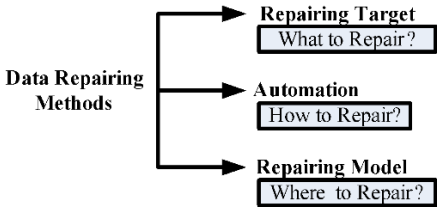

Figure 4. Standard data repairing method

Due to the increasing of analytical complexities, it is essential to know the statistical implication of data pre-processing. There are multiple techniques that exists which enhance the accuracy or efficiency of data pre-processing through statistical approach, e.g. Machine learning method. Some popular data pre-processing algorithms are discussed as follows.

Active-Learning for crowdsourcing is slowly increasing in popularity. Crowdsourcing is rapidly adopting in business fields for data pre-processing [23]. In the educational sector, there is an increasing necessity to address such complexity problemand multiple recent research studies employ an Active-Learning approach to solve the crowd queries [24-26]. The supervised learning methods (i.e., Support Vector Machine and Random forest) are a most important method to formulate the user input to data pre-processing, and Active-Learning is an algorithmic approach which elects the most informative datasets to acquire.

The several statistical data pre-processing approach has been presented in existing archives of research publication to more precisely and accurately clean the data repository e.g. The famous project "Eracer" was used for depicting the core process of data pre-processing over the noisy data can offer dual stages of learning operation. The famous graph-based methods are used for representing the message passing and relation algorithm which solve the inconsistencies [27]. Additionally, there are several recent approaches to represent the statistical outlier detection methods like [28, 29]. In [30], authors employ a machine learning approach to improve the pre-processing data reliability.

The extended work of performing sophisticated data preprocessing associated with clean the data and precedes machine learning training model is called Active clean [31]. This approach employs a selection method for most significant data and methods to rapidly update the machine learning model given new clean data. According to the study of [13], specific numbers of considered dataset are subjected to cleaning process while surplus data are further subjected to training. GPS trajectories or sensor data reading sequences are composed of imprecise or error-prone values. Even business database could be error-prone [32]. The existing approach of sequential data pre-processing considered the constraint associated with speed that is linked with consumption of fuel [33]. Determination of the errors associated with huge spikes can be carried out by constraint, while constraints based pre-processing repairs the dirty values with respect to mini/max speeds. However, the constraint associated with the speed is not successful for determining certain errors which is at par with the practical constraints of speed. For better investigation, it is essential to consider smaller version of errors. One small example to talk about is when there is a deviation of 1m over the readings of GPS. Apart from this, aggregating a massive number of errors, mining results can be seriously misled for example; not able to create clusters in inaccurate GPS readings with multiple small errors [34].

Furthermore, noise is usually associated with GPS raw data [35], and it increases an uncertainty signal on results that are undesirable to the authors and industrial engineers in general. However, the evaluation process defines how a dataset is reliable which include GPS error detections and missing data. Also, this evaluation includes sample size, rate, spatial coverage and existence of additional data type (i.e., weather). In the research study of Vitor et al. [36] investigated the limitations of prior work on the topic of data quality indicators (i.e., floating car data). Authors leveraged on the number of statistical indicators covers a number of statistical indicators including; reliability, accuracy and city spatial coverage and evaluate the specific data quality. The statistical indicators rely on a sequence of statistics, clustering and external data elements like road maps.

- Yuki-San Method: Such approach is used for settling various forms of statistical indicators which are basically of two type's viz. i) value: it represents the quality of the data, ii) veracity: it is mainly associated with data reliability from the source point. Such GPS based values are represented in the form of granularity and coverage factor. Micro-Temporal coverage (analyze the day time temporal coverage) and Spatial coverage (provide real-time spatial information). While, veracity is enumerated as; Missing data (compute any signal gaps from the dataset), reliability (measures the logical precision) and accuracy (spatial precision of GPS devices).
- Indicator of Spatial Coverage: This term is associated with the measurement of the distance based diversity of data of vehicle. Such values usually increase with the more density of traces of GPS. The entire process of spatial coverage is illustrated in algorithmic structure [36], where a set of traces associated with GPS over a defined Grid-Cell (Sgc) is weighted based on its relevance and formula of spatial coverage indicator can be represented as;

$$\text{Spatial coverage} \quad \frac{\sum_{gc \in Grid} \left(S_{gc} \mathrm{Y}_{gc} \right)}{\sum_{gc \in Grid} \mathrm{Y}_{gc}}$$

- *Missing Data:* Missing data refer as; the time signal gap between the two transmitted signals over a single cycle. It may be occurred by misuse of device or malfunctioning. The missing data represent a set of GPS trances which are missing, and it is formulated as;

Missing Data $(erfc(P)+G)^2$, Where RFC is a complementary Gaussian error function, P is a number of packets lost, and G is granularity.

– Reliability*:* The reliability covers the dataset objectivity, and it is computed as;

Reliability $(at+aT+rt+rT)^{-4}.k$, Where: (at) – awake trace ratio, (aT) – awake trip ratio, (rt) – reachable trace ratio, and (rT) – reachable trip ratio.

– *Accuracy:* Accuracy measured by inconsistency among the positions of GPS device and vehicle true location. Authors formalized the accuracy indicator by algorithmic form, and its resultant equation is defined as follow;

Accuracy Acc (median (eT), Where T represents Error of each trip. Yuki San method has experimented on data aggregated from four wheelers inSan Francisco and Nanjing. From the obtained results authors analyzed that proposed Yuki San method is very potential to uncover the value in floating car data sources in an automated manner.

– *You-Sense Tool:* It is a monitoring tool which collects the GPS raw data via a mobile application. It tracks the position with GPS, Wi-Fi, and accelerometers. The advantage of YouSense is data pre-processing and data analysis. In [37] authors investigate multiple filter criteria for YouSense GPS data-pre-processing by statistical analysis of different person's dataset. YouSense collected the GPS data records and displayed according to the time stamp of GPS chip, and corresponding parameters are; Time millis, Longitude, Latitude, Accuracy, Altitude, Speed and Bearing. However, collected GPS data records provide high accuracy position data, but this data contains gaps (i.e., missing data errors). This data gaps may be planned gaps (i.e., the phone is not in operational mode, GPS device is switched off mode) or unplanned gaps (i.e., phone battery is dead, GPS device unable to receive signals). Hence, to resolve this kind of data gaps the dataset need to clean by i) filtering the wrong location information, and ii) fill-up the gaps during GPS device is switched off mode.

To understand the raw GPS data (Figure 5), authors developed a "Quantum Geographic Information System," i.e., QGIS tool that visualizes the GPS data (i.e., GPS viewing, editing as well as analysis). Also, this supports web map services. To repair GPS sequential trajectory data with the considering the variable as x = x [1], x [2]. In this case, x[i] is considered as ith point of data over a domain of finite structure. There is a specific timestamp ti linked with xi as well as artifacts with certain predefined range θ_i. There are various possibilities that the range of θ_i differs from each different forms of data which actually affects the accuracy score of the GPS readings. There are also good possibilities of directing a maximum value of θ_i for depicting highest possible artifacts for all forms of sequenced dataset.

Figure 5. Visualization of raw GPS data with multiple gaps in the GPS trace

The above (1) displays past-probability (P(x)) or also known as a likelihood of sequences x with respect to speed changes. Q (u_i) exhibits the future probability of speed changes ui, and P(u_i) represents the corresponding (log)past-probability where empirical distribution of probability Q carried over the speed factor that alters and can be determined using simplified statistical feature over the same sequence. Authors have formulated an issue associated with the rectification of the sequential data over a vast probability of computational complex problem [38] for the purpose of evaluating practical GPS data aggregated over using smartphone while the subject is mobile over the observation area. The presented study has considered comprehensive test environment with inclusion of errors. However, the only parameter to be identified in δ connected with cost associated with rectification of data as shown in Table 1.

$$P(x)=\sum_{i-2}^{n-1}P(\nu_i)=\sum_{1-2}^{n-1}\log Q(\nu_i) \quad (1)$$

Table 1. Strategy adopted in [38]

Methodology Adopted	Analysis
DPC, constant -factor approximation	Large budget
DP, dynamic programming	Exact
DPL, linear time heuristics	Fast, High error
QP quadratic Programming	Approximate distribution
SG, Simple greedy	Fastest
SCREEN	Existing approach

GPS trajectory data analysis is the trending research topic mainly used for transportation mode detection via GPS data analysis. There are diversified properties associated with the determination of mode of transportation (e.g. speed, latitude, location, longitude, acceleration, etc). Unfortunately, there is no inclusion of any mode of transportation charecteristics over the aggregated GPS data. The study carried out by [33] has presented a discussion of entropy factor P_E considering the mobility factor. A classifier design is developed for using learning machine is used for minimizing the training time without compromising accuracy.

– Permutation Entropy: This mechanism is used for identifying all the dynamic alterations of the computationally complex aspects. The variable PE is associated with the original series of time basically represents a Shannon entropy for all K symbols. Its mathematical representations is (2),

$$Hp(m)=-\sum_{j-1}^{K}P_j InP_j \quad (2)$$

Where *m* represents the embedding dimension, P_j represents distribution of probability factor associated with all the series of diverse symbol.

– *Extreme Learning Machine (ELM):* It is form of machine learning approach that targets using single hidden layer while a conventional training mechanism of feed-forward approach. The speed of the training using this approach is quite faster as compared to any legacy machine learning of neural network. The experimental analysis of such an approach is as follows: The Authors considered "Microsoft GeoLife dataset" which includes 17621 moving trajectories of 182 users in 3 years. These trajectories were recorded by different GPS loggers and GPS phones. Authors extracted the features from each trajectory and categorized into basic features (Average velocity, velocity variance) and sophisticated properties e.g. sophisticated features and P_E of velocity). The outcomes of training and testing from the features are shown in Table 2.

Table 2. Outcomes of training and testing

Sample sizes		
Training	Testing	Features
10%	90%	AV
20%	80%	DV
30%	70%	HCR
40%	60%	SR
50%	50%	VCR

4. EXISTING RESEARCH TRENDS

Apart from the standard methodology of GPS data preprocessing, there is various research contribution towards addressing *data cleaning* problems. The existing studies are broadly reported to adopt 4 different approaches, e.g. i) statistical-based approach, ii) logical approach, iii) outlier-detection approach, and iv) trajectory-based approach. The statistical-based approach is developed emphasizing time-series, prediction, trip detection, quantitative patterns, machine learning [39-48]. The existing logical-based approaches are reported to consider velocity constraints, reduction of travel distance, and human navigational system [49-51]. Nearly, similar problems are also considered when working with outlier-detection based approach where the consideration of driving behavior, statistical process controls, partitioning is carried

out [52-55]. Trajectory-based approaches are found to use security factor, congestion analysis, clustering, mining, updating map, similarity assessment [56-63]. Table 3 summarizes the research contribution of present times with respect to different parameters to exhibit that all the problems are associated with advantage as well as significant limitation too.

Table 3. Summary of a different method for pre-processing GPS data

	Authors	Methodology	Capability	Advantage	limitation
Statistical-based Approach	Song [39]	Modeling with speed constraint factor	Determine the possibility of sequence	Higher accuracy	-Large spike error -Fail to identify small errors
	Jiang et al.[40]	Machine Learning	Prediction of favorable destination	Achieved satisfactory predication analysis, better performance	-Not applicable for other transportation services like taxis, private cars.
	Wang et al. [41]	Trip detection based on mobile data	Identify and eliminate false trip data	Accuracy between 95-97%	Narrowed study scope
	Zhang et al. [42]	The quadratic time constant factor approximation, Linear heuristics, greedy heuristic	1. Repairing a series of data with maximum likelihood.	Better performance w.r.t repairing and application accuracy	Induce computational complexity
	Equille et al. [43]	Identification of specific patterns of complex data	Detect and clean	Better detection performance. High data pre-processing accuracy in real-world datasets	Performance limited to synthetic data
	Cerqueira et al. [45]	Yuki San quality indicator	Automatically evaluate the quality	Automatically uncover the value in sources	Not applicable for real-world applications
	Peng et al. [46]	Machine learning	Can perform road safety analysis. Applicable to different types of data	Study applicable to VANET system	Lack benchmarking
	Sun et al. [47]	Uses existing tools for data filtering (YouSense)	Effective and simplified filtering	Applicable in real-time	High energy consumption. Need more memory space
	Granat [48]	Hidden Markov model, Enhanced Expectation Maximization	Supports faster assessment	Robust, reliable	Lack of extensive assessment
Logical Methods	Song et al. [49]	Polynomial time, the linear model constructed using time factor	can pre-processing stream data,	Better accuracy over the real dataset	Lack of extensive assessment
	Luo et al. [50]	Recommendation for optimal route data	Supports dynamic real-time travel planning	Effective route recommendation	Restricted computer power
	Weerakoon et al. [51]	Fuzzy logic	Multi-mode navigation system	Applicable for physically impaired pedestrians	Lack of extensive assessment
Outlier Detection Method	Hieu et al. [53]	Statistical, Shewhart control charts	Predictive performance	Offer granularity in outlier detection	No comparative analysis
	Lee et al. [55]	Data partitioning approach	Ability to identify outliers residing in sub-trajectory data	Better accuracy	Lacks any numerical analysis in an extensive manner
Trajectory Filtering Method	Patil et al. [56]	Identification and elimination of data anomaly	Maintains integrity, secrecy of cleaned data considers real-time parameters for anomaly detection.	Offers security and balance cleaning process at the same time	Lacks any numerical analysis in an extensive manner
	Wang et al. [57]	Congestion-based analysis of the trajectory	Identifies bottleneck and eliminates artifacts from trajectory data	Simpler visual analysis	Cannot support multiple task analysis

Table 3. Summary of a different method for pre-processing GPS data (*continue*)

	Authors	Methodology	Capability	Advantage	limitation
Trajectory Filtering Method	Idrissow [58]	Clustering approach	-Outlier detection, Stop detection, Interpolation, Map Matching	Improve the quality of the obtained clusters.	No benchmarking
	Yin et al. [59]	Recommendation-based filtering process	Offers the best performance on the recommended data	Achieved higher precision	Overhead not studied
	Peixoto [60]	Detection Stay points	Eliminate the noisy data	A good approach for location suggestion and detection of user experience	No benchmarking
	Shan et al. [61]	An experimental study, graphical-based filtering	Works better with a valid inference of map along with the proper update.	Applicable for different circumstances	The scope of map data not discussed with respect to computation
	Vementala et al. [62]	Geo-spatial similarity assessment	Better updating of the map	Faster processing	No benchmarking
	Tang et al. [63]	Simulation-based, diversified network data	Effective clustering performance	Construct the association with data semantics.	High computational cost

5. OPEN END PROBLEMS

From discussion made in the prior sections, it can be seen that there are various standard and unique approaches meant for addressing the data cleaning problems in GPS signals. However, it can also be seen that the majority of the researchers have not much considered about the problems associated with the *signal lapse* of the GPS data. The prime reason behind this is the usage of the standard dataset which misses these problems. Generally, information about such signal lapse can be obtained from the GPS device that obtain signaling from multiple GPS satellites. Such forms of dynamic data cannot be obtained from the standard dataset as they are a direct representation of any form of consistent interruption in GPS data with respect to time. Hence, there is a significant skip of problem consideration while attempting the GPS data cleaning process. It should also be known that consideration of such problem is of higher importance as they are highly practical and inevitable owing to the presence of different forms of infrastructure on the earth surface, e.g. trees, tall buildings, etc. A closer look into all the existing approaches exhibits that various methods indirectly attempts to solve this problem with the aid of time series analysis skipping the lapse factor. Recent works are not found to have any such consideration. However, a work carried out by Wheeler et al. [64], and Lachowycz et al. [65] have a unique approach where the authors have used the raw GPS data in order to check the lapse factor. This implementation permits various other forms of time-series data to be aggregated while investigating the lapse factor by retaining contextual spatial data as well as data obtained from accelerometers. However, this approach is only valid for outdoor applications and not indoor application resulting in missing data if the indoor application is considered. In the same year of 2010, there was a work carried out by Oliver and Badland [66] where the study ignored the participant-based information which fails to meet their critical factor. The next research methodology attempted for missing data was by using imputation technique by Troped et al. [67]. Irrespective of a slight difference in all these approaches, a common trait of usage of spatial data and temporal data is found to be used; however, all them serious misses any form of computational modeling for performing validation or benchmarking of the presented approaches of dealing with missing data from GPS signal. Eventually, the researchers working on standard dataset also ignored the fact that there is always a certain amount of error even in standard GPS data as such data are never claimed to consider any form of environmental factors. If such practical parameters are not considered in the dataset than there is always a fair chance of error degrading the accuracy of the analysis. There are various use cases to represent that missing data could significantly degrade the data quality of GPS Signal.

- *Use-Case-1*: The first use case is very common to everyone and is termed as a *drifting problem* that is highly inevitable and results in missing data. Figure 6 highlights the GPS traces of the dense forest area where it can be seen higher accuracy of tracking being maintained on the road area, but it starts showing random position when it enters the forest area. Hence, the positioning data in the forest area is missing, and there is no existing approach to address this missing data problem.
- *Use-Case-2*: This is another most encountered problem in GPS signal receiving characterized by *signal attenuation problem*. Figure 7 showcases a straight line in the circle which is a false route in the terrain

region. In such case, a linear line is drawn between the source and destination point which is highly inaccurate proving the complete loss of data. None of the existing research work has emphasized on this problem of missing data till date.

– *Use-Case-3:* This problem is usually more encountered in the urban area and very less in the rural area, and it results in *bouncing issue* of GPS signal. Figure 8 highlights three locations where the scattered GPS signal is received owing to the presence of tall buildings. The navigation system shows some separate tracks even on a straight road or vice-versa as they are incapable of tracing the original signals. Unfortunately, such problems also directly contribute to missing data where there is no effective solution found in the existing study.

From all this evidence, it is quite clear that there is a critical need for a reliable GPS service where the solution cannot be towards the external parameters but should be more focused on internal parameters.

Figure 6. Use case of drifting GPS signal

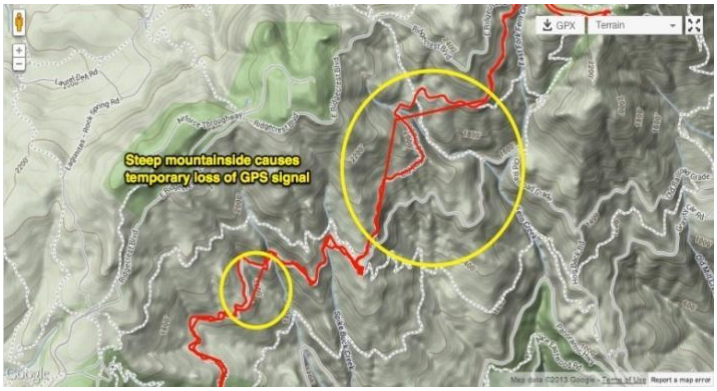

Figure 7. Use case of signal attenuation GPS signal

Figure 8. Use case of bouncing of GPS signal

6. CONCLUSION AND FUTURE WORK

This paper has presented a review of existing approaches towards GPS signal data and its quality. From the quality viewpoint, the paper attempts to highlight that data pre-processing of the GPS signal is one of the essential operations. Existing approaches towards data cleaning process are found to adopt many sophisticated and complicated approaches and have come up with the diversified result. However, their results are achieved under consideration of a specific error-free dataset and controlled research environment. As consideration of such forms of data itself should be re-thought of as without the presence of possible errors discussed in the form of use-cases in the prior section, it is not possible for addressing missing data problem in GPS signal.

Therefore, the future work direction should be towards considering the adoption of such dataset which has such characteristics of errors. Due to non-availability of such dataset, the future work will be initiated to perform computational modeling for error incorporation in such a way that there is a missing data mapping with the cases illustrated in prior section uses cases. As a part of the solution, the future work will be then focused on offering correlated data which higher probability of matching with the missing data. The work will be carried out using a combination of statistical approach, probability theory, and time-series analysis in order to evolve up with a new computational model. The performance of such a model will be assessed using certain benchmarked practices as well as comparative analysis with extensive numerical case studies.

REFERENCES

[1] A. K. Sen, A. B. Bhattacharya, "Radar Systems and Radio Aids to Navigation," *Stylus Publishing*, LLC, pp. 780, 2018.

[2] G.J. Sonneberg, "Radar and Electronic Navigation," *Elsevier, Technology & Engineering*, pp. 378, 2013.

[3] Guido Rizzi, Matteo Luca Ruggiero, "Relativity in Rotating Frames: Relativistic Physics in Rotating Reference Frames," *Springer*, pp. 456, 2003.

[4] G. Xu, "GPS: Theory, Algorithms and Applications," *Springer Science & Business Media*, pp. 340, 2007.

[5] Elliott D. Kaplan, Christopher Hegarty, "Understanding GPS/GNSS: Principles and Applications," Third Edition, Artech House, Technology & Engineering, pp. 1064, 2017.

[6] SabriCetinkunt, "Mechatronics with Experiments," John Wiley & Sons, pp. 880, 2015.

[7] Peter Teunissen, Oliver Montenbruck, "Springer Handbook of Global Navigation Satellite Systems," *Springer*, pp. 1327, 2017.

[8] Fan H., Zipf A., Fu Q., Neis P., "Quality assessment for building footprints data on OpenStreetMap," *Int. J. Geogr. Inf. Sci*., vol. 28, pp. 700–719, 2014.

[9] Berti-Equille L., Dasu T., Srivastava D., "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, Washington, DC, USA, pp. 733–744, Apr 2011.

[10] Hellerstein J. M., "Quantitative Data Cleaning for Large Databases," White Paper, United Nations Economic Commission for Europe (UNECE), 2008, [online], Available: http://db.cs.berkeley.edu/jmh/, [Accessed on 4 March 2018], [Retrieved on 05-03-2019].

[11] Bohannon P., Fan W., Geerts F., Jia X., Kementsietsidis A., "Conditional functional dependencies for data cleaning," *Proceedings of the 2007 IEEE 23th International Conference on Data Engineering*, Istanbul, Turkey. pp. 746–755, Apr. 2007.

[12] Berti-Equille, Laure, TamraparniDasu, and Divesh Srivastava, "Discovery of complex glitch patterns: A novel approach to quantitative data pre-processing," *In 2011 IEEE 27th International Conference on Data Engineering*, IEEE, pp. 733-744, 2011.

[13] Chu Xu, Ihab F. Ilyas, Sanjay Krishnan, Jiannan Wang, "Data pre-processing: Overview and emerging challenges," *In Proceedings of the 2016 International Conference on Management of Data*, ACM, pp. 2201-2206, 2016.

[14] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "Crowder: Crowdsourcing entity resolution," *PVLDB*, vol. 5(11), 2012.

[15] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided data repair," *PVLDB*, vol. 4(5), pp. 279–289, 2011.

[16] G. Beskales, I. F. Ilyas, and L. Golab, "Sampling the repairs of functional dependency violations under hard constraints," *PVLDB*, vol. 3(1-2), pp. 197–207, 2010.

[17] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi, "A cost-based model and effective heuristic for repairing constraints by value modification," In *SIGMOD*. ACM, pp. 143–154, 2005.

[18] F. Chiang and R. J. Miller, "A unified model for data and constraint repair, In *ICDE*, pp. 446–457, 2011.

[19] X. Chu, I. F. Ilyas, and P. Papotti, "Holistic data pre-processing: Putting violations into context," In *ICDE*, pp. 458–469, 2013.

[20] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," In *PVLDB*, VLDB Endowment, pp. 315–326, 2007.

[21] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu, "Interaction between record matching and data repairing," In *SIGMOD*, ACM, pages 469–480, 2011.

[22] M. Volkovs, F. Chiang, J. Szlichta, R. J. Miller, "Continuous data pre-processing," In *ICDE*, pp. 244–255, 2014.
[23] A. Marcus and A. Parameswaran, "Crowdsourced data management: Industry and academic perspectives," *Foundations and Trends in Databases*, vol. 6(1-2), pp. 1–161, 2013.
[24] D. Haas, J. Wang, E. Wu, and M. J. Franklin, "*Clamshell: Speeding up crowds for low-latency data labeling*," *PVLDB*, vol. 9(4), pp. 372–383, Dec. 2015.
[25] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu, "Corleone: Hands-off crowdsourcing for entity matching," In *SIGMOD*, 2014.
[26] B. Mozafari, P. Sarkar, M. J. Franklin, M. I. Jordan, and S. Madden, "Scaling up crowd-sourcing to very large datasets: A case for active learning," *PVLDB*, vol. 8(2), 2014.
[27] C. Mayfield, J. Neville, and S. Prabhakar, "*ERACER: a database approach for statistical inference and data pre-processing*," In *SIGMOD*, 2010.
[28] M. Balazinska, A. Deshpande, M. J. Franklin, P. B. Gibbons, J. Gray, M. H. Hansen, M. Liebhold, S. Nath, A. S. Szalay, and V. Tao, "Data management in the worldwide sensor web," *IEEE Pervasive Computing*, vol. 6(2), pp. 30–40, 2007.
[29] S. Madden, "Database abstractions for managing sensor network data," *Proceedings of the IEEE*, vol. 98(11), pp. 1879–1886, 2010.
[30] M. Yakout, L. Berti-Equille, and A. K. Elmagarmid, "Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes," In *SIGMOD*, 2013.
[31] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, "Activeclean: Interactive data pre-processing while learning convex loss models," *In Arxiv*, 2015, [online], Available: http:// arxiv.org/ pdf/1601.03797.pdf.
[32] Abdulla, Raed, Aden Abdillahi, and Maythem K. Abbas, "Electronic Toll Collection System based on Radio Frequency Identification System," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8(3), pp. 1602-1610, 2018.
[33] S. Song, A. Zhang, J. Wang, and P. S. Yu, "SCREEN: stream data pre-processing under speed constraints," In *SIGMOD*, pp. 827-841, 2015.
[34] S. Song, C. Li, and X. Zhang, "Turn waste into wealth: On simultaneous clustering and pre-processing over dirty data," In *IGKDD*, pp. 1115-1124, 2015.
[35] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Mapmatching for low-sampling-rate GPS rajectories," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, New York, NY, USA, pp. 352–361, 2009.
[36] Cerqueira, Vitor, *et al.*, "On Evaluating Floating Car Data Quality for Knowledge Discovery," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19(11), pp. 3749-3760, 2018.
[37] Sun, Qiyun, Rein Ahas, AntoAasa, Wanggen Wan, and Chi Yuan, "GPS data pre-processing and analysis based on YouSense mobile application," pp. 11-8, 2017.
[38] Zhang, Aoqian, Shaoxu Song, and Jianmin Wang, "Sequential data pre-processing: a statistical approach," In *Proceedings of the 2016 International Conference on Management of Data*, ACM, pp. 909-924, 2016.
[39] S. Song, "Time Series Data Cleaning," 2013, [online], Available: http://ise.thss.tsinghua.edu.cn/sxsong/.
[40] Jiang, Jian, Fei Lin, Jin Fan, Hang Lv, and Jia Wu, "A Destination Prediction Network Based on Spatiotemporal Data for Bike-Sharing," Complexity, 2019.
[41] Wang, Bao, Linjie Gao, and Zhicai Juan, "A trip detection model for individual smartphone-based GPS records with a novel evaluation method," *Advances in Mechanical Engineering*, vol. 9, no. 6, 2017.
[42] Zhang, Aoqian, Shaoxu Song, and Jianmin Wang, "Sequential data cleaning: a statistical approach," In *Proceedings of the 2016 International Conference on Management of Data*, ACM, pp. 909-924, 2016.
[43] Berti-Equille, Laure, TamraparniDasu, and Divesh Srivastava, "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," *In 2011 IEEE 27th International Conference on Data Engineering*, IEEE, pp. 733-744, 2011.
[44] Zaytar Mohamed Akram and Chaker El Amrani, "MetOp Satellites Data Processing for Air Pollution Monitoring in Morocco," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8(6), pp. 4584-4592 2018.
[45] Cerqueira, Vitor, *et al.*, "On Evaluating Floating Car Data Quality for Knowledge Discovery," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19(11), pp. 3749-3760, 2018.
[46] Peng, Zhe, Shang Gao, Zecheng Li, Bin Xiao, and Yi Qian, "Vehicle safety improvement through deep learning and mobile sensing," *IEEE network*, vol. 32(4), pp. 28-33, 2018.
[47] Sun, Qiyun, Rein Ahas, AntoAasa, Wanggen Wan, and Chi Yuan, "GPS data cleaning and analysis based on YouSense mobile application," *4th International Conference on Smart and Sustainable City (ICSSC 2017)*, pp. 11-8, 2017.
[48] Granat, Robert, "A Time Series Analysis Method for Geophysical Sensor Networks," Retrieved on 05-03-2019.
[49] Song, Shaoxu, Aoqian Zhang, Jianmin Wang, and Philip S. Yu, "Screen: Stream data cleaning under speed constraints," *In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ACM, pp. 827-841, 2015.
[50] Luo, Zhongwen, HuiminLv, Fang Fang, Yishi Zhao, Yuanyuan Liu, Xiuqiao Xiang, and Xiaohui Yuan, "Dynamic Taxi Service Planning by Minimizing Cruising Distance Without Passengers," *IEEE Access*, vol. 6, pp. 70005-70016, 2018.
[51] Weerakoon, K. M. K., K. S. Rupasinghe, T. P. Withanaarachchi, G. M. R. I. Godaliyadda, and M. P. B. Ekanayake, "Outdoor human navigation with gps and sensor systems," *In 2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, IEEE, pp. 1-6, 2017.

[52] Sun, Chuan, Jianyu Wang, Lian Xie, Duanfeng Chu, and Liqun Liu, "Data Cleaning of Speed Monitoring Based on Driving Behavior Characteristics for Commercial Vehicle," *In IOP Conference Series: Materials Science and Engineering*, IOP Publishing, vol. 392(6), pp. 062156. 2018.
[53] H. T. Hieu, T. Y. Chou, Y. M. Fang, T. V. Hoang, "Statistical Process Control Methods for Detecting Outliers in Gps Time Series Data," *International Refereed Journal of Engineering and Science (IRJES)*, vol. 7(5), ver. I, pp. 08-15, May 2018.
[54] Gupta, Manish, Jing Gao, Charu C. Aggarwal, and Jiawei Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26(9), pp. 2250-2267, 2014.
[55] Lee, Jae-Gil, Jiawei Han, and Xiaolei Li, "Trajectory outlier detection: A partition-and-detect framework," *In 2008 IEEE 24th International Conference on Data Engineering*, IEEE, pp. 140-149, 2008.
[56] Patil, Vikram, Priyanka Singh, Shivam Parikh, and Pradeep K. Atrey, "Geosclean: Secure cleaning of gps trajectory data using anomaly detection," *In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, pp. 166-169, 2018.
[57] Wang, Zuchao, Min Lu, Xiaoru Yuan, Junping Zhang, Huub Van De Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE transactions on visualization and computer graphics*, vol. 19(12), pp. 2159-2168, 2013.
[58] Idrissov, Agzam Y., "A data cleaning framework for trajectory clustering," 2012.
[59] Yin, Peifeng, Mao Ye, Wang-Chien Lee, and Zhenhui Li, "Mining GPS data for trajectory recommendation," In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Cham, pp. 50-61, 2014.
[60] Douglas Alves Peixoto, "Mining Trajectory Data," *Technical Report*, Nov. 2013.
[61] Shan, Zhangqing, Hao Wu, Weiwei Sun, and Baihua Zheng, "COBWEB: a robust map update system using GPS trajectories," *In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, pp. 927-937, 2015.
[62] Vementala, Nikhil, Paolo Papotti, and Mohamed Sarwat, "A Framework for Interactive Geospatial Map Cleaning using GPS Trajectories," *In Proceedings of the 10th ACM SIGSPATIAL Workshop on Computational Transportation Science*, ACM, pp. 19-23, 2017.
[63] Tang, Lei, Yaling Zhao, Zongtao Duan, and Jun Chen, "Efficient similarity search for travel behavior," *IEEE Access*, vol. 6, pp. 68760-68772, 2018.
[64] Nihad, El Ghouch, En-Naimi El Mokhtar, Zouhair Abdelhamid, and Al Achhab Mohammed, "Hybrid approach of the fuzzy C-means and the K-nearest neighbors methods during the retrieve phase of dynamic case based reasoning for personalized follow-up of learners in real time," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9(6), pp. 4939-4950, 2019.
[65] Lachowycz K, Jones AP, Page AS, Wheeler BW, Cooper AR., "What can global positioning systems tell us about the contribution of different types of urban greenspace to children's physical activity?" *Health Place*., vol. 18, pp. 586–94, 2012.
[66] Oliver M, Badland H, Mavoa S, Duncan MJ, Duncan S., "Combining GPS, GIS, and accelerometry: methodological issues in the assessment of location and intensity of travel behaviors," *J Phys Act Health.*, vol. 7(1), pp. 102-108, 2010.
[67] Troped PJ, Wilson JS, Matthews CE, Cromley EK, Melly SJ., "The built environment and location-based physical activity," *Am J Prev Med*, vol. 38, pp. 429–438, 2010.

BIOGRAPHIES OF AUTHORS

Prabha R. pursued Bachelor of Computer Science and Engineering from Madurai Kamarajar University, 1997. Completed Master of Technology in Computer Science and Engineering from Visvesvaraya Technological University, Belagavi in 2010. She is currently working at Nettur Technical Training Centre, Bangalore as a Assistant Manager and pursuing Ph.D from Visvesvaraya Technological University. Her Research focuses on IoT, Machine Learning and Deep Learning. She has 17 Years of Teaching experience and 4 Years of Research of Research experience.

K. G. Mohan received Bachelor Degree in Electrical Engineering from University of Mysore in 1984 and Master of Technology in Power and Energy System from KREC (Mangalore University) during 1995. Received Ph.D with the specialization of Computer Architecture from Anna University in the year 2007. He has published and presented more than 30 research papers in reputed International Journals and Conferences. His area of research includes Low Power Architecture design, Wireless Sensor Networks, IoT, Cloud Computing, Network Security and Cryptography. He has 30 years of teaching experience and 14 years of Research Experience.