Processing Outliers in GPS Trajectory Data Set

T. Tamizh Chelvam and V. Vijaya Samundeeswari

Abstract--- *The Global Positioning System (GPS) is a satellite- based navigation system that can be used to find your location quickly anywhere in the world. The determined location findings can be called as GPS trajectory data set and it could be displayed over a digital map, which is known as map matching. However, since the inaccuracy exist in GPS trajectory data set, it is not possible to display the data properly over a digital map. In this case, the positions may be seen to be outliers. To avoid this problem, it is necessary to analyze the received GPS trajectory data set and process them before map matching.*

Keywords--- *GPS, Outliers, Z Score, Lagrangian Interpolation, GPS Trajectory Data Set, Map Matching*

I. INTRODUCTION

IN past two decades, satellite navigation technology, specifically, the Global Positioning System(GPS) has been developed as an inevitable positioning technology for all kinds of navigation like land navigation, sea navigation and air navigation. It plays an important role due to round the clock, all seasoned, and free-of-cost service availability.

The satellites orbit the Earth every 12 hours at approximately 12,600 miles above the Earth. The satellites continuously transmit information towards the Earth. With 3 satellites, a GPS receiver can determine a 2D position which includes latitude and longitude. With four or more satellites, a GPS receiver can determine a 3D position which includes latitude, longitude, and altitude. While GPS provides an efficient way to collect latitude, longitude, and altitude, it is necessary to remember that there are errors inherent in any GPS trajectory data set. Map matching method will become less effective, if the erroneous GPS trajectory data set are not handled properly. This paper describes a processing method for outliers based on statistical and numerical methods. The input of the system comes from a etrex, Garmin GPS receiver. The results show an improvement in matching the GPS trajectory data set over a digital map.

II. GPS OUTLIERS SOURCES

There are many sources of possible errors that will lead to outliers causing the degradation of the accuracy in map matching. The following are the main possible sources listed and those can be considered when configuring GPS parameters to reduce outliers.

A. Satellite Geometry

Satellite geometry describes the position of the satellite to each other from the view of the receiver. If a receiver sees satellites and all are arranged for example in the north west, this leads to a "'bad"' geometry. If a position is determined in this case, the error of the positions may be up to 100 to 150 meters. If the satellites are well distributed over the whole firmament, the determined positions will be much more accurate.

B. Satellite Elevations

When a satellite is low on the horizon, the satellite signals must travel a greater distance through the atmosphere, resulting in lower signal strength and delayed reception by the GPS receiver. Position data should be collected using only satellites that are at least 15 degree above the horizon.

C. Multi Path

Satellite signals can reflect off larger nearby objects, such as buildings or cars, causing the GPS antenna to receive an erroneous signal (longer path length = false location). This phenomenon is known as multi path, which can induce errors of dozens of meters. Optimal accuracy can be obtained by collecting data in an environment devoid of large reflective surfaces and that has a clear view of the sky.

D. Signal-to-Noise Ratio (SNR)

This is a measure of the information content of a signal relative to the signal's noise. The satellite signals do not penetrate metal surfaces, buildings, tree trunks, or similar objects. When recording position data, the antenna of the receiver has to be placed away from these obstructions.

E. Ionosphere - Change in the Travel Time of the Signal

Before GPS signals reach the antenna on the earth, they pass through a zone of charged particles called the ionosphere, which changes the speed of the signal. If reference station and receivers are relatively close together, the effect of ionosphere tends to be minimal.

F. Troposphere - Change in the Travel Time of the Signal

Troposphere is essentially the weather zone of our atmosphere, and droplets of water vapor in it can affect the speed of the signals. The vertical component of the GPS position (elevation) is particularly sensitive to the troposphere.

G. Ephemeris Data Transmission from Base Stations

If the ephemeris data are transmitted with lower accuracy, meaning that the transmitted satellite positions do not comply with the actual positions, the inaccuracy of the position will be achieved for several hours.

T. Tamizh Chelvam Department of Mathematics Manonmaniam Sundaranar University Tirunelveli 627 012, Tamil Nadu, India. E-mail: tamche59@gmail.com

V. Vijaya Samundeeswari, Research Scholar Manonmaniam Sundaranar University Tirunelveli 627 012, Tamil Nadu, India. E-mail: vijivicky@yahoo.co.in

ISBN 978-93-82338-22-2 | © 2012 Bonfring

III. PROCESSING OUTLIERS IN GPS TRAJECTORY DATA SET

A data set is a set of measurements taken from some environments or process. If we have a collection of objects, and for each object, we have a set of same 'p' measurements. We can represent the collection of the measurements on 'n' objects as a form of 'n x p' data matrix. The 'n' rows represent 'n' objects on which measurements were taken. The 'p' columns of the data matrix may be referred to as variables, features, attributes, or fields[4].

The input of the system comes from a etrex, Garmin GPS receiver. Sometimes raw data are not convenient to process and it can be advantageous to modify them prior to process. Hence the XML based GPS Trajectory data are converted to a data set. The sample GPS Trajectory data set obtained from GPS receiver at various times is represented as 'n x p' matrix. Here 'n' represents the number of positions and 'p' represents the number of attributes known as DataId, Latitude, Longitude, Altitude and GPSDateTime.

Table I: Sample GPS Trajectory Data Set

Data Id	Latitude	Longitude	Altitude	GPS Date Time
520	13.0077155	80.2070173	105.40	15/4/2012 2:29:26
521	13.0076915	80.2068574	105.40	15/4/2012 2:29:27
522	13.0076686	80.2067013	105.40	15/4/2012 2:29:28
523	13.0076489	80.2065485	105.40	15/4/2012 2:29:29
524	13.0076304	80.2063965	105.40	15/4/2012 2:29:30
525	13.0076131	80.2062465	105.40	15/4/2012 2:29:31

In GPS Trajectory data set, sometimes data may lie in the unacceptable range (outliers). Without handling these outliers, result may not be the expected one. For this reason, the outliers should be identified and reconciled, and the proposed system is developed in this case.

The mathematical approach introduces some theoretical ideas behind processing outliers, leading to the development of this system. The modules that generally lie within the areas of processing outliers in GPS trajectory data set are given below.

- Outliers Identification
- Outliers Reconciliation

A. Outliers Identification

An outlier is defined as an observation that"appears" to be inconsistent with other observations in the data set[6]. Identification of potential outliers is important for the following reasons.

i. An outlier may lead to inaccurate map matching. For example, the data may have been coded incorrectly due to various GPS error sources. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be reconciled to have the accurate map matching.

ii. And, of course, in some cases, outliers identification may indicate something scientifically interesting, so that it is possible to analyze various GPS error sources, if the data contains significant outliers.

There are various approaches to outliers identification depending on the application and number of observations in the data set. Statistical methods were developed to accommodate outliers and to reduce their impact on the analysis. Z score is one of the simplest statistical approach to identify outliers. This approach is adopted for this GPS Trajectory data set by using the following formula.

$$z = (x - \mu)/\sigma$$

where, x is the sample from the GPS Trajectory data set μ is the mean of the GPS Trajectory data set; σ is the standard deviation of the GPS Trajectory data set. A data value with z score less than -3 or greater than +3 might be considered an outlier. It presents a reasonable criterion for identification of the outlier when data follow the normal distribution.

B. Outliers Reconciliation

By using z score, it is possible to identify meaningful outliers in large GPS Trajectory data sets. It is also necessary to reconcile them, because all the map matching methods are centered around improving the accuracy.

Lagrangian Interpolation is one of the best predictors, when the data points in the data set are known and also, it yields the perfect accuracy at the data points. The Lagrangian interpolation formula is

Let x_0, x_1, ..., x_n be (n + 1) pairwise distinct points and let there be given (n + 1) arbitrary numbers y_0, y_1, ..., y_n. Further define the fundamental Lagrange polynomials by

$$l_i^n(x) = \frac{(x-x_0)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_0)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}.$$

By using Lagrangian Interpolation method, the outliers are reconciled to normal.

IV. IMPLEMENTATION

The GPS Trajectory data set received from Garmin etrex GPS receiver. The modules were developed in Java with Netbeans IDE tool. The system consists of two modules.

i. Outliers Identification: All the outliers in the GPS Trajectory data set were identified by implementing Z score method.

ii. Outliers Reconciliation: The identified outliers were reconciled with the data centric to outliers by implementing Lagrangian Interpolation method.

The output GPS Trajectory data set was tracked in the digital map, which shown the improvement of accuracy in map matching.

V. EXPERIMENTS AND RESULTS

The original fix positions were collected in the city Chennai, Tamilnadu, India, Asia. Data collection was done with the distance of 10 meters. The GPS Trajectory data set in GPS receiver in the format of XML files were transferred to the computer system and they were processed with the above mentioned modules. The following table of the data set and graphical representation exhibits the outliers in the sample GPS Trajectory data set.

ISBN 978-93-82338-22-2 | © 2012 Bonfring

Table II: Outliers Identification in GPS Trajectory Data Set

Data Id(i)	Latitude(x_i)	Z score	Normal/Outlier
520	13.0077155	1.2932	Normal
521	13.0076915	1.1492	Normal
522	13.0076686	1.0122	Normal
523	13.0076489	0.8942	Normal
524	13.0076304	0.7833	Normal
525	13.0076131	0.6799	Normal
526	13.0075967	0.5821	Normal
527	13.0075836	0.5033	Normal
528	13.0075730	0.4400	Normal
529	13.0075653	0.3939	Normal
530	13.0075617	0.3723	Normal
531	13.0075592	0.3572	Normal
532	13.0075597	0.3602	Normal
533	13.0075644	0.3883	Normal
534	13.0080580	3.3438	Outlier
535	13.0075753	0.4536	Normal
536	13.0075766	0.4616	Normal
537	13.0075783	0.4716	Normal
538	13.0075774	0.4661	Normal
539	13.0075716	0.4315	Normal
540	13.0075601	0.3627	Normal
541	13.0075482	0.2915	Normal
542	13.0075372	0.2252	Normal
543	13.0075265	0.1615	Normal
544	13.0075087	0.0546	Normal
545	13.0074805	-0.1140	Normal
546	13.0074567	-0.2566	Normal
547	13.0074431	-0.3379	Normal
548	13.0074310	-0.4106	Normal
549	13.0074186	-0.4844	Normal
550	13.0074005	-0.5928	Normal
551	13.0069113	-3.5219	Outlier
552	13.0069319	-3.3985	Outlier
553	13.0073315	-1.0058	Normal
554	13.0073357	-0.9810	Normal
555	13.0073403	-0.9531	Normal
556	13.0074523	-0.2826	Normal
557	13.0074631	-0.2179	Normal
558	13.0074744	-0.1506	Normal
559	13.0074839	-0.0934	Normal
560	13.0074934	-0.0367	Normal
561	13.0074078	-0.5491	Normal
562	13.0074227	-0.4598	Normal
563	13.0074325	-0.4011	Normal
564	13.0074422	-0.3434	Normal
565	13.0074486	-0.3052	Normal
566	13.0074537	-0.2746	Normal
567	13.0074564	-0.2586	Normal
568	13.0074570	-0.2546	Normal
569	13.0074580	-0.2485	Normal

Outliers identified in the given sample GPS trajectory data set and they were reconciled to the following values by using Lagrangian Interpolation method.

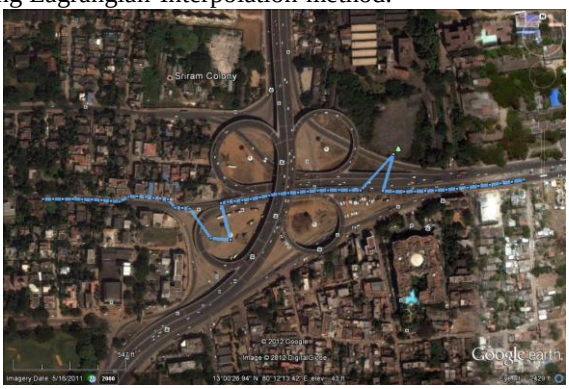

Figure 1: Outliers Identification in GPS Trajectory Data Set

Table III: Outliers Reconciliation in GPS Trajectory Data Set

Data Id(i)	Latitude(x_i)	Z score	Normal/Outlier
534	13.0075704	0.6755	Normal
551	13.0073766	-1.4037	Normal
552	13.0073504	-1.6848	Normal

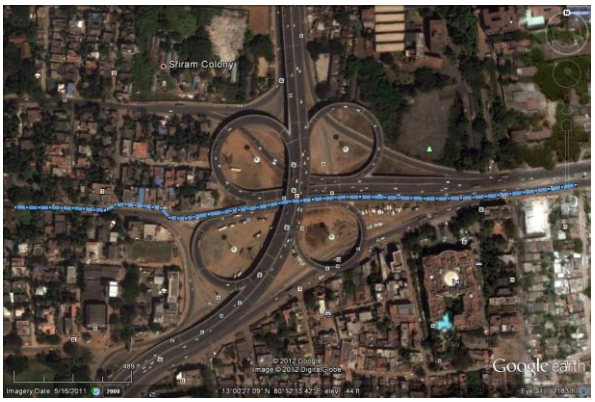

Figure 2: Outliers Reconciliation in GPS Trajectory Data Set

VI. CONCLUSION

In this research, GPS trajectory data set was received from a etrex, Garmin basic GPS receiver. This data set was collected for a known busy place in the city with the well defined road structure. The modules were evaluated to identify outliers in the data set and reconcile them. The accuracy in map matching achieved. Thus, the position measurement errors in the form of outliers were considerably removed.

REFERENCES

[1] V. Barnett and T. Lewis, Outliers in statistical data, Wiley, 1994.
[2] I. Ben-Gal, Data Mining and Knowledge Discovery Handbook: A Com- plete Guide for Practitioners and Researchers, Kluwer Academic Pub- lishers, 2005.
[3] R. Brandt, Comparing classical and resistant outlier rules, Journal of theAmerican Statistical Association, 1990.
[4] David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, Prentice Hall of India Private Limited, 2001.
[5] Greenfeld, J.S., Fuzzy logic-based map matching algorithm for vehicle navigation system in urban canyons, Matching GPS observations to locations on a digital map. In proceedings of the 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.
[6] B. Iglewicz, D.C. Hoaglin How to Detect and Handle Outliers, American Society for Quality Control, WI, 1993.
[7] M. R. Mosavi, K. Mohammadi, M. H. Refan, Fuzzy Processing on GPS Data to Improve Positioning Accuracy, before and after S/A Is Turned off, The Asian GPS Conference, India, 2002.
[8] M. R. Mosavi, Fuzzy Point Averaging of the GPS Position Components, Map Asia Conference, China, 2004.
[9] P. J. Rousseeuw, A. M. Leroy, Robust Regression and Outlier Detection, John Wiley, New York, 1987.
[10] S. Syed, M. E. Cannon, Fuzzy logic-based map matching algorithm for vehicle navigation system in urban canyons, In proceedings of the Institute of Navigation (ION) national technical meeting, USA, 2004.

ISBN 978-93-82338-22-2 | © 2012 Bonfring