AUTOMATIC GENERATION OF ROAD NETWORK DATA FROM TAXI GPS TRAJECTORIES

By

Alborz Soltankhah-Bidkhti

B.Eng, Tehran University, Tehran, Iran, 2012

A MRP

Presented to Ryerson University

In partial fulfillment of the requirements

for the degree of Master of Engineering

in the program of Civil Engineering

Toronto, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this MRP. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

AUTOMATIC GENERATION OF ROAD NETWORK DATA FROM TAXI GPS TRAJECTORIES

Alborz Soltankhah-Bidkhti

Master of Engineering 2017

Civil Engineering (Geomatics Engineering)

Ryerson University

Abstract

Keeping road network databases up-to-date is crucial to Geographical Information System (GIS) applications such as road networking. The vector road centerlines extracted from field surveys and satellite images are expensive and labor intensive with long updating processes. The GPS data crowd-sourced by public transportation users, provides an expanding source for enhancing road maps because of its rich spatial-temporal coverage and reasonable level of accuracy. The overall objective of this project is to implement an optimized methodology, which generates road centerline from GPS data obtained from taxis in Beijing without using any reference plans.

Since the dataset used in this project has longer time intervals between trajectories compared to previous studies, the extracted road network on straight road segments are more accurate than the extracted road network on highway ramps in this project.

Acknowledgements

This project would not have been successfully completed without the efforts and valuable input from many people to whom I owe my deepest gratitude. I have been very fortunate with my supervisor, Dr. Songnian Li. I will forever be thankful to Dr. Li for his continuous support and patience over the last two years, as well as valuable guidance and effort in keeping my study in the right direction.

I also owe many thanks to my best friends who helped me through the project. I would like to acknowledge special thanks to Dr. Wai-Yeung Yan. Wai-Yeung has given me constructive and meticulous comments, which were an enormous help to my graduate research since 2014. Advice and comments given by Wai-Yeung were helpful in learning and developing practical programming skills of Python and ARCGIS.

I have greatly benefited from education at Ryerson University Geomatics Engineering program, which forms the basis of this project. I would like to show my greatest appreciation to all former and current professors in our faculty, especially Dr. Michael Chapman, Dr. Ahmed Shaker, and Dr. Ahmed El-Rabbany. I also want to thank Rachel Peluso and Desmond Rogan for their generous support and warm concern.

Finally, my heartfelt gratitude goes to my brother Alvand and my parents because of their love, understanding, encouragement and sacrifice in all forms.

Table of Contents

Table of Figures

Chapter 1. Introduction

1.1 Motivation

Keeping road network databases up-to-date is crucial to Geographical Information System (GIS) applications such as road networking. The more traditional methods for collecting vector road centerlines from field surveys and satellite images are expensive and labor intensive with long updating processes. On the other hand, advances in global positioning systems and their measuring techniques, makes GPS data crowd-sourcing an accurate and cost effective method of data collection. Many municipalities all around the globe have installed GPS receivers in their taxis for tracking, managing and improving their services. This sort of data can be interpreted and processed to identify traffic patterns, road geometry and network connectivity.

Many studies have investigated road network extraction from GPS points, one of which is automatic road network extraction introduced by Niu (2013). In this project, the motive is to apply Niu's (2013) method on Beijing taxi GPS data and examine if his method can be adapted to this datasets and if any modifications need to be made. Therefore, road centerlines are automatically generated by implementing tools and scripts developed by Niu (2013) in ArcGIS. The main challenge in this project is to modify the Python scripts written by Niu (2013) to work with the Beijing taxi data.

1.2 Objective

Automatic road network generation method developed by Niu (2013) is originally designed to create road network from GPS data collected by smart phone users in southern Ontario, Canada. In comparison with Niu's (2013) dataset, attributes in crowed-sourced data obtained from China also include longitude, latitude, speed, azimuth, timestamp and status of vehicle whether it is in service or not. However, it is missing accuracy of measurements. Even though both datasets have the same attributes, the range of their values are very different. For instance, speed values in this project are in km/h, where in the original research conducted by Niu (2013) speed values are in m/s. Also, coordinates in this project are in decimal degrees and they need to be changed to meters. Therefore, to automatically extract the road network, the objective is to modify the dataset as well as Python scripts written by Niu (2013).

1.3 Limitations

The real world GPS dataset used in this project was originally obtained from www.datatang.com, which was recorded during the month of November in 2012. Due to long period of data collection, the data size was over 60 GB that was in .txt format. Working with such huge dataset requires relatively large random access memory for processing. After all, it is impossible to use the whole dataset due to memory limitations when running 32-bit PythonWin on the 64-bit Microsoft Windows 7 operating system. In order to resolve this issue a sample study area was selected for processing.

Since time intervals of the dataset used in this project is longer compared to the study done by Niu (2013), connectivity of points in highway ramps are not as accurate as straight segments.

1.4 Project Organization

This report is consisted of six chapters. The first chapter covers introduction to this project and motivation and why this topic was selected, objective and what the focus is as well as limitations controlling the extent of this project. Second chapter looks into previous work done and previous studies conducted in road network extraction. In the third chapter, methodology is explained as well as modifications made to the dataset acquired from Datatang[1] website, who is a global data provider. The fourth chapter looks deeper into the raw data, its attributes and descriptions along with a statistical analysis. This chapter also describes the selected study area. The results of the implemented methodology, overlaid with open street map (OSM) data, are described in the fifth chapter. Finally, Chapter 6 concludes this project and shares some insights for future work.

[1] www.datatang.com

Chapter 2. Literature Review

Recently many efforts have been made to reconstruct road networks from trajectory data. In comparison to traditional methods such as, field survey and satellite image processing, extraction of road centerlines from crowdsourced GPS trajectories has many advantages regarding labor cost, real-time and data completeness. However many challenges are faced for constructing road network while using crowdsourced GPS trajectory data due to sparse sampling and extensive data volume and large number of outliers (Wei et al., 2016). This chapter summarizes some of these methods and compares them to the method introduced by Niu (2013), which is used in this project.

Ai et al. (2016) presented a method based on Delaunay triangulation to detect change in transport land-use by using crowdsourced taxi trajectory in Beijing, China. Their method consists of three steps. First step includes preprocessing the trajectory data to eliminate the noise and also create a track line from GPS points obtained from taxis. For this step they used an interpolation method to add more points to improve triangulation by setting a threshold value for adjacent trace points on trajectory line. Then, by using interpolated track line Delaunay triangulation was created. Second step is determining road area by comparing the triangles constructed in the previous step. Triangles with smaller area and shorter edges are on road area while triangles with larger area and longer edges are distributed in non-road area. Third step extracts the road network by organizing created polygons and removing the short edges of the triangles. By comparing their result to existing road data such as, OpenStreet map, their method is proven to be fast and accurate.

Wu et al. (2016) proposed an algorithm based on hidden Markov model (HMM) to identify problem road segments (PRS). This approach helps reduce the extensive computational processes. To extract the road segments in problem neighborhoods. They introduced a clustering method to group all points in sample area to find skeleton points, which would represent underlying road segments. In this technique, it is assumed that all the sampling points are on the same path therefore it is required to implement a preprocessing step to select points based on speed and direction of movement. Later on, they used partitioning around medoids (PAM) clustering method, which was introduced by Newson et al. (2009) because this method is robust to noise and outliers and is very effective on smaller datasets. Li et al. (2012) generated

trajectories from raw GPS points and then uses the trajectory as the basic processing unit, which is the first step of road generation process. Their methodology consists of two other steps, which are trajectory processing and post-processing. In the second step, all the points in each trajectory created in the first step are processed to extract a candidate road network by merging new trajectory points into them. After processing all trajectories, final road network is extracted by confidence filtering, similar roads merging, road smoothing and road linking. The purpose of applying confidence filtering is to eliminate extracted roads with low accuracy or essentially lower probability of being a real road. In order to eliminate recurrent roads, similar roads are merged together. Road smoothing is implemented to smooth sawteeth effect in generated road networks. Since the density of points is not uniform in all areas the extracted road networks might be broken in some parts of the study area. In order to solve this problem, they performed road linking by setting a threshold value for difference in angle between broken road parts and distances between them.

Cao and Sun (2014), presented a method to extract road centerlines with the incorporation of satellite images and GPS data. Their method selects road GPS data as the initial road position for generation of road network. This method has a faster computing speed and is capable of dealing with complex geometries and occluded regions. The only limitation of this method is the GPS overlapping rate, which will be improved by advances in GPS receivers and further use of volunteered geographical datasets.

Niu (2013) introduced a method to extract the road network by using GPS trajectories from smart phone users in southern Ontario, Canada. The basis of his methodology was to restructure the GPS trajectories on each road in a manner that ideally each lane contains at least one new GPS trajectory. To extract the centerline of the road he grouped the reconstructed GPS trajectories on the same road segment into one cluster and merged them into one polyline segment, which would represent the road centerline. Finally, the geometric relationship between extracted end points of road centerlines were considered for linking the centerlines together to create an integrated road network.

Wang et al. (2011) introduced a weighted clustering algorithm based on the physical attraction model adopted by Cao and Krumm (2009) in which vehicles are assigned different weights with reference to their speed and change in direction. According to Wang et al. (2011), vehicles with

higher speed have less derivation from the road they are moving on. Therefore, a higher weight is assigned to its trace in the physical attraction model. This assignment causes the clusters to be shifted closer to the roads. To keep the adjusted points on each trace consistent, an angle threshold filter was set to be able to distinguish between the points on the same road.

Zhang et al. (2010) presented a method to integrate traces to create relatively accurate and detailed road network. In order to extract the road centerline, they started off by averaging all the traces of one road. The actual geometry is later on derived by use of open street map (OSM) data as initial information to differentiate between several roads that are nearby. To extract the true centerlines, they created profiles perpendicular to the initial road and wherever samples intersect with the GPS traces sampling points for the road centerlines are selected to generate the road centerline. This method incrementally improves the existing road network.

However, the rapid development of GPS and wireless communication technologies provides an alternative data source for extracting road geometric data for road network database updating and road maps refinement. This new approach entails a fast, inexpensive way of updating existing road maps and refining road maps with real-time changes.

Among all of the studies summarized above, Niu (2013) was the only one to perform quantitative evaluation for assessing the accuracy of his results. He evaluated his results by comparing them to the actual geometric road alignment data. As main challenge in existing methods is extracting road centerline accurately, Niu's (2013) method is still a desirable practical approach with accordance to measured horizontal accuracy as a root mean square of 1.252 m for straight road segments and 1.424 m for curved road segments and is better than that of some existing datasets. In this project, Niu's (2013) method is used for road extraction on a different dataset to prove his method can be used for taxi GPS trajectory datasets in other parts of the world as well as GPS trajectory datasets provided by smartphone users.

Chapter 3. Methodology

The methodology of automatic extraction of road network is presented by Niu (2013), which identifies GPS trajectories that are restricted to the road network without using map-matching technology. In this report, I have used the modified methodology presented by Niu (2013). The following steps have been conducted to extract the road centerline in selected regions of Beijing, China.

First, GPS points is collected regardless of where the taxi is (e.g. driving on roads, idling (in parking lots or traffic jams), waiting at traffic lights), so it is difficult to completely filter out point clouds that are not moving on the road. Second, the distribution of GPS points of moving taxis could be normal, scattered or multimodal within a certain width along the road centerline. However, it is not reasonable to utilize a fixed value of road width to clustering GPS points that belong to the same road segment, because there might be at a road split or merge.

The objective of this chapter is to accomplish the following tasks:

1. Reduce the data size without affecting the extraction of road centerline.
2. Merging the GPS points towards the middle of each lane.
3. Extracting representative points also known as cluster centers.
4. Connecting the cluster centers and creating polyline segments.
5. Merging polyline segments and creating the final road centerline on each direction.

General procedures adapted in this project to extract road network centerline from GPS trajectories of moving vehicles are explained in Section 3.1. Section 3.2 describes preprocessing techniques implemented in this project to reduce the data size. Smoothing algorithm adapted in this project is described in Section 3.3. Section 3.4 covers clustering method utilized in this project for further reduction of data size. Furthermore, Section 3.5 explains trajectory reconstruction and to complete this chapter Section 3.6 describes road centerline extraction from polyline segments created in the previous step.

3.1 Overall Workflow

As it is stated in the previous chapters, algorithms as well as the methodology presented by Niu (2013) have been used to complete this project. Fig 3.1 shows the overall workflow introduced by Niu (2013) and in the following sections, changes and modifications implemented to his methodology in order to extract the road centerlines from a different dataset will be discussed.

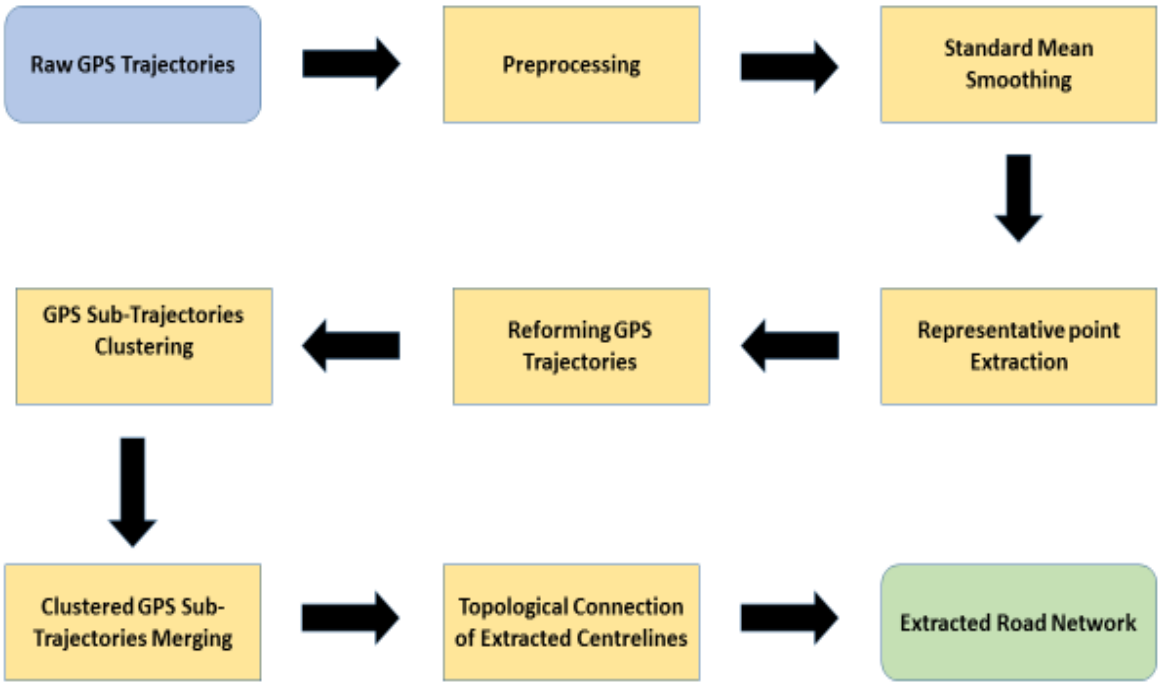

Fig 3.1: Overall Workflow of Automatic Extraction of Road Network.

The original methodology consists of 7 steps that include:

1. Preprocessing.
2. Standard mean smoothing.
3. Representative point extraction.
4. Reforming GPS trajectories.
5. GPS sub-trajectories clustering.
6. Clustered GPS Sub-trajectories merging.
7. Topological connection of extracted centerlines.

The original models and scripts written by Niu (2013) in ArcGIS software where designed to extracted road network from GPS dataset acquired from cellphone users in southern Ontario. In this project, the dataset was obtained from GPS receivers installed in taxis in Beijing. In order to implement the same tools and models created by Niu (2013) the models and Python scripts had to be modified to extract the road centerline in selected study area. More details regarding changes to models and scripts are available in Chapter 4.

3.2 Preprocessing

Due to limited random access memory (RAM), it is not possible to use all of the GPS points in the acquired dataset from taxis, since the original data size is 66 GB. Therefore, the preprocessing step intends to reduce the data size without impacting the underlying road network. In this step idling taxis were filtered out by eliminating points with speed value of 0 Km/Hr. Fig 3.2, illustrates the effect of existing idling points in the dataset.

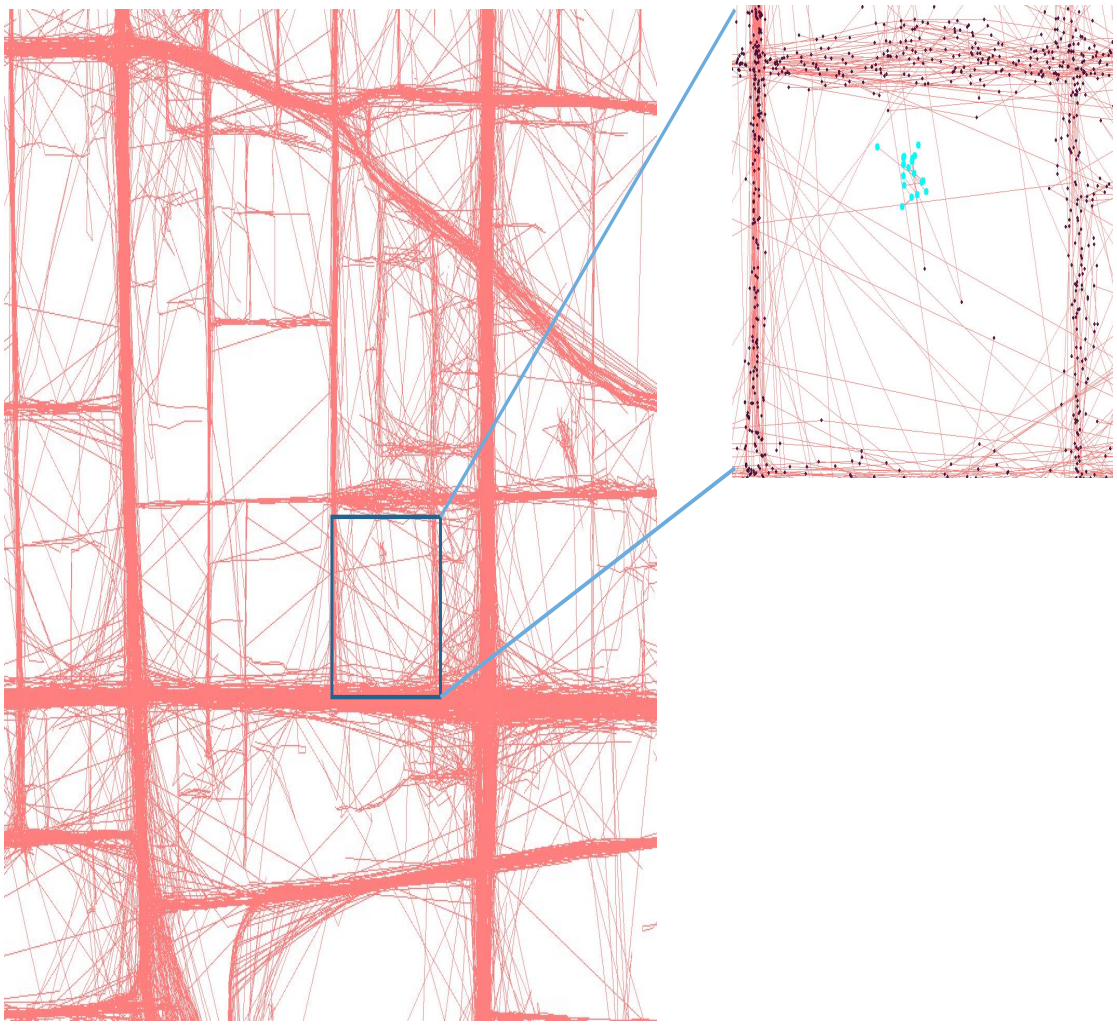

Fig 3.2: Individual Traces (red lines), GPS Points (black dots) and Idling Taxis (Highlighted Blue Dots).

Also, different GPS trajectories were split into individual traces by assigning individual Route ID to each taxi in any given day. However, by connecting consecutive GPS points together, unreasonable trajectories can be generated amongst different trips made by taxis in any given day. Fig 3.3 demonstrates these unreasonable trajectories, which are off the actual road network and are considered being outliers.

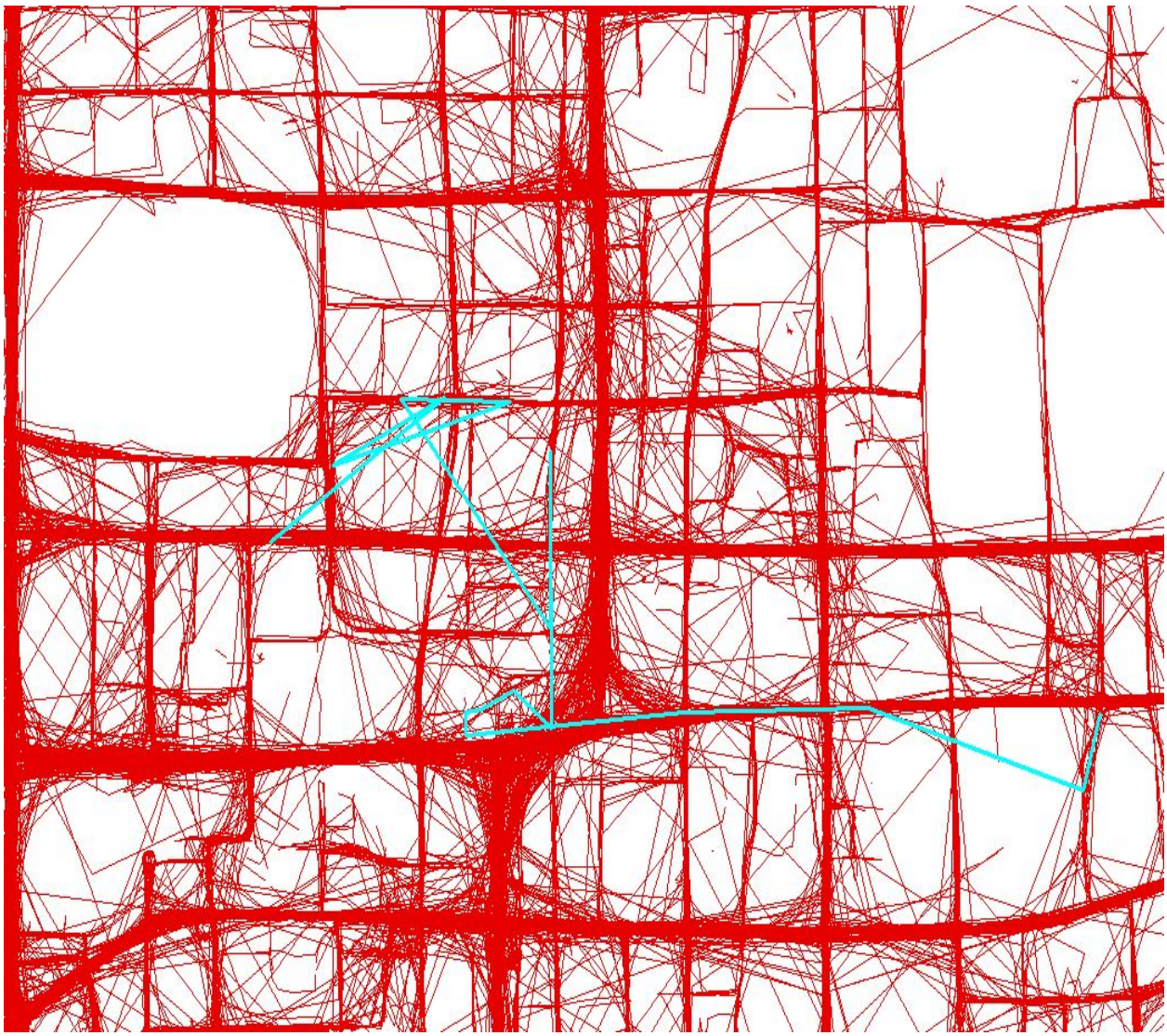

Fig 3.3: Unreasonable Connections Within Raw GPS Trajectories.

In order to resolve this issue, every trace is subjected to two checks.

1. Change in direction
2. Distance between every two consecutive positioning points.

Each trip traveled by every taxi must be split into discrete GPS trajectories whenever the gap between any two time-stamped positioning points is larger than the distance threshold or change in their direction of driving is greater than the allowable threshold. In this project, since the time intervals between measurements are 1 minute, distance threshold is set to 2000 meters and directional change threshold is set to 180 degrees. Depending on how fast taxis are moving on

highway, in 1 minute they can be more than 2000 m away from previous recorded position. Further analysis regarding distances and threshold values can be found in Chapter 4.

Duplicated positioning points were also eliminated to further reduce the data size. These points are considered to be outliers since they are representing the same values multiple times. In case such positioning points are found in the dataset, only the first one is considered for processing purposes and the rest are deleted from the data set. Implementing this step reduced the data size significantly.

Since this project utilizes Python scripts and ArcGIS tools created by Niu (2013) for data processing it was necessary to make sure that all the dictionaries created in Python scripts include their keys in the dataset. Since the origin of data in this project was different from Niu (2013), field titles in the raw dataset had to be given the appropriate name, which Niu (2013) used in the original script. This way there is no need to change the keys for every step.

3.3 Standard Mean Smoothing

The main purpose of this step is to bring GPS points representing moving vehicles towards the middle of the road. Therefore a four-meter buffer was set on every positioning point to cover the lane width ranges between 3.5 and 3.7 m (TAC, 1999) and the new position of the points are changed to the mean of all the positioning points in similar direction within the four-meter buffer. In the original method developed by Niu (2013), the weighted mean of the positioning points was calculated for finding the new position of the points. In the original research, due to nature of the dataset, which was collected by smart phones, one of the attributes in the dataset was the accuracy of GPS measurements. Niu (2013) used the accuracy as a weight for calculating the mean of positioning points. In this project since the data does not include accuracy of measurement, mean was calculated to shift the positioning points towards the middle of each lane.

3.4 Representative Point Extraction

A modified density-based point clustering method is adopted to reduce the size of data by extracting a smaller number of positioning points as representative of smoothed points, without affecting the underlying road geometry. This step is similar to the data reduction method adopted by Guo et al. (2010), but it is mainly based on the initial bearing of individual positioning points.

The smoothed positioning points within a 4-meter (definition of 4-m refers to Section 3.3) cluster can be represented by the center of that cluster, but there must be at least one smoothed positioning point within that cluster. Smoothed positioning points always belong to the cluster that has the shortest distance from its center to that smoothed positioning point. It should be noted that all the smoothed positioning points within a cluster must have the similar direction. Niu (2013) provided more details regarding extracting representative points also known as cluster centers.

3.5 Refining GPS Trajectories

Extracting representative points reduces the data size without affecting the underlying road network. In this section, the objective is to reconstruct the GPS trajectories to ideally create at least one new GPS trajectory in each lane. This step helps outline the road width by evenly distributing GPS trajectories along the road. To remain faithful to the underlying road network, representative points on the same lane must be connected to each other based on their topological relationships such as their distance and direction.

According to Niu (2013), the modified version of the overall logic flowchart of reforming GPS trajectories to better comply with this project is represented in Fig 3.4. Starting with any single unconnected representative positioning points, its surrounding representative positioning points are checked based on three conditions. First, the surrounding representative points must be within 500 meters of the current representative point, which is determined as the best empirical value based on series of experiments. Second, there must be at least one preprocessed trajectory passing within 4 meters of the current representative point and its surrounding representative points. Third, the difference between the directions of the current point and its surrounding representative points must be smaller than 11 degrees, based on definition of degree of curve described by Ghilani and Wolf (2002). As per third condition in this step, 11 degrees is the value originally calculated by Niu (2013) and this value is used in this project. This condition is different from the allowable directional change discussed in the preprocessing step. In the preprocessing step, traces were created for each vehicle where in this step trajectories are being created based on topological relationship between different GPS points obtained from different vehicles.

Furthermore, in case a near representative point meets all of these prerequisites, it is marked as a candidate point for construction of a new connection from/to the current representative positioning point. If a representative positioning point is selected as the current representative positioning point but has no candidate point, the next representative positioning point becomes the available current point. The GPS trajectory reconstruction algorithm determines the optimal succeeding and preceding points from candidates to construct new connections along their bearings. New polyline segments are constructed. The algorithm is iterated until all representative positioning points are connected.

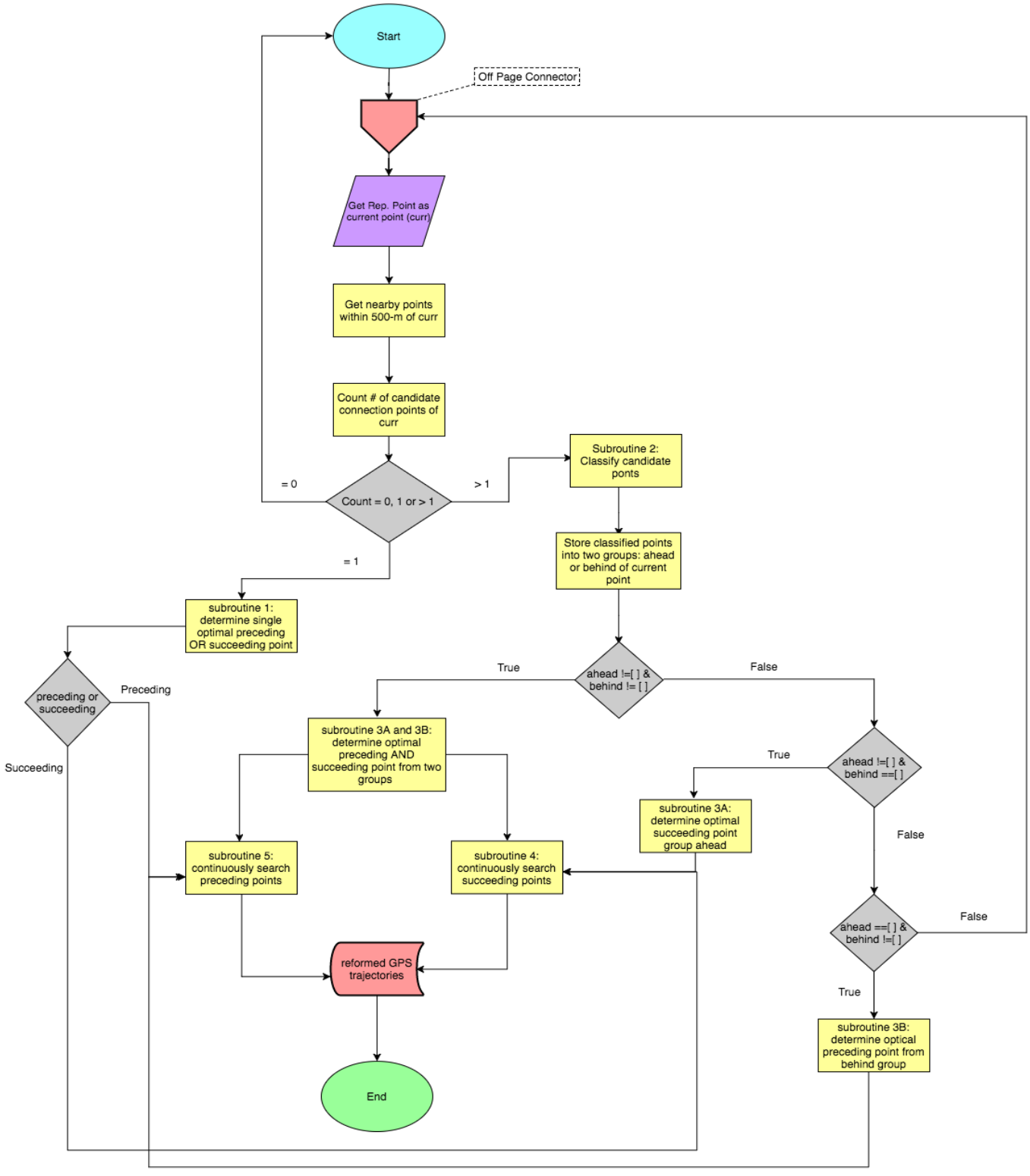

Fig 3.4: Main Logic Flowchart of Reconstructing GPS Trajectories.

In this step, every point goes through five different processes to finally construct the GPS trajectory. The basis of these processes is to find succeeding and preceding point for each representative point extracted in the previous step based on the direction of each point. Niu (2013) referred to these processes as subroutines. In this process, the first step is to find candidate points to process them in the subroutines and determine if they fit within the criteria of being a succeeding or preceding point. In this project, as oppose to Niu's method I decided to choose 500 meters as my distance threshold for creating the nearby list, since in a 50-meter buffer there were not enough points to reconstruct the GPS trajectory. Nearby list includes all the points within a 500-meter buffer of each point. Next is to check the number of candidates in the nearby list. If there are no points in the list, another representative point will be selected to as current point to go through the process. In case there is only one candidate in the nearby list it goes through the Subroutine 1 which defines if the candidate is a succeeding or preceding point. After this depending on whether the point is succeeding or preceding, the next succeeding or preceding points will be defined by continuously searching for them in either Subroutine 4 or Subroutine 5. In case there is more than one candidate in the nearby points, they are subject to two checks so they could be classified in two groups of points ahead or behind (Subroutine 2). Subroutine 3 finds the optimal preceding and succeeding points. Subroutine 4 and 5, as mentioned, continuously search for succeeding or preceding points for connection until all the representative points are processed.

3.6 GPS Sub-trajectories Clustering

GPS trajectories that were extracted in the previous step are consisted of multiple polyline segments. In this step, I need to separate the created trajectories of the same road from the other trajectories on nearby roads and the ones on the same road but on the opposite direction. The original method introduce by Niu (2013) consists of two different algorithms, sweep-line algorithm and recursive polyline searching algorithm.

The sweep-line algorithm works around a reference polyline segment. The reference polyline segment has the most number of segments and every other trace or segment is compared to the reference. Then, the nearby segments are grouped as a cluster. The distance threshold from the reference is the longest distance between the traces on the other side of the road. After using this algorithm to find nearby polyline segments, a recursive polyline searching algorithms is utilized.

The aforementioned algorithm selects and groups nearby segments close to the reference segment based on directional change as the threshold value.

3.7 Clustered GPS Sub-trajectories Merging

Section 3.6 creates the clustered polyline segments. This section aims at extracting the road centerline by merging all divisional polylines of the same cluster. The algorithm implemented in this section runs in three steps. The first step is to extract road centerline segment by merging all polyline segments in each cluster to one polyline. The second step in this section is to estimate the road width. Niu (2013) discussed that by outlining a polygon based on four vertexes, it is possible to estimate the road width on the extracted centerlines. The third step is to update the road centerline segment by performing topological analysis.

3.8 Topological Connection of Extracted Centerlines

As mentioned in the previous section, the generated GPS trajectory with the highest number of segments is selected as the reference line. According to Niu (2013), wherever the direction changes the extracted centerlines are to be separated, which usually occurs at splits or merges. These two separations are technically referred to as Y-splits and intersections. This step is designed to create integrated road network by topologically connecting extracted road centerlines. Intersections of extracted road centerlines are linked together based on their geometric relationship to create the final centerline, only if the GPS trajectories generated in previous steps go through intersecting centerlines.

Chapter 4. Experimental Data and Study Area

4.1 GPS Data Collection

The real world GPS dataset collected by taxis in Beijing, China was obtained from www.datatang.com. The actual dataset included different attributes such as longitude, latitude, speed, azimuth (direction of flow), timestamp and taxi condition (i.e., whether it was in service or not). Data structure is as follows.

OBJECTID	Taxi Id	Time	Longitude	Latitude	GPSSpeed (km/hr)	Azimuth
1	486328	20121108001228	116.3038635	39.9029655	0	176
2	63490	20121108001233	116.302002	39.9116364	36	102
3	426466	20121108001235	116.3045197	39.9082947	0	0
4	162418	20121108001230	116.3038864	39.9002419	0	282
5	68174	20121108001236	116.3041077	39.9046173	75	178
6	154728	20121108001235	116.3044815	39.9027939	62	0
7	426336	20121108001237	116.3042145	39.9078102	25	292
8	164532	20121108001240	116.304451	39.9080925	32	356
9	204858	20121108001238	116.3004837	39.9118462	0	188
10	194189	20121108001243	116.3039017	39.9152565	54	178

The dataset spans from November 1st to November 30th in 2012 and covers Beijing, China. Fig. 4.1 demonstrates the spatial distribution of 35,200,000 GPS points from taxis in Beijing. Since the data was collected by the taxi services, it is assumed that all the points are on the roads. Nevertheless, collected information can be incorrect due to nature of GPS collection methods.

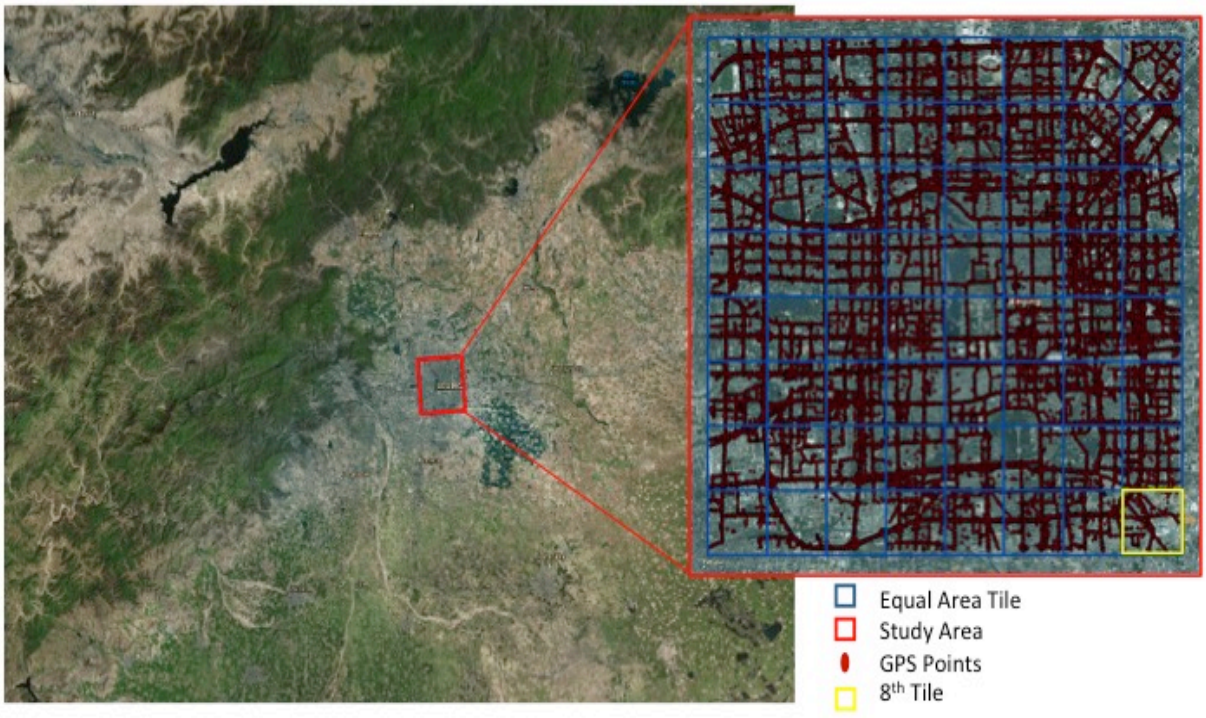

Fig 4.1: Spatial Distribution of Collected GPS Data

4.2 Case Study Area

The original GPS data retrieved from the source were stored in comma delimited text format (.txt) files. Since the Environmental Systems Research Institute (ESRI) shape file is not capable of handling such huge dataset (66GB), the experimental GPS data was split into 64 equal-area tiles covering downtown Beijing as shown in Fig 4.1. In order to better test algorithm, the 8th tile was selected as the actual study area since there are straight, curved and merged highway segments with large number of GPS points found there, Fig. 4.2 represents 1,204,328 GPS points in selected study area. Road network was constructed on three different regions as illustrated in Fig. 4.2. Region 1 is selected for testing the adaptability of methodology on a typical straight segment. In Region 2, there are merges and splits in some areas and the whole region is on a curve. Region 3 is a more complex segment of the highway to further check the adaptability of the methodology.

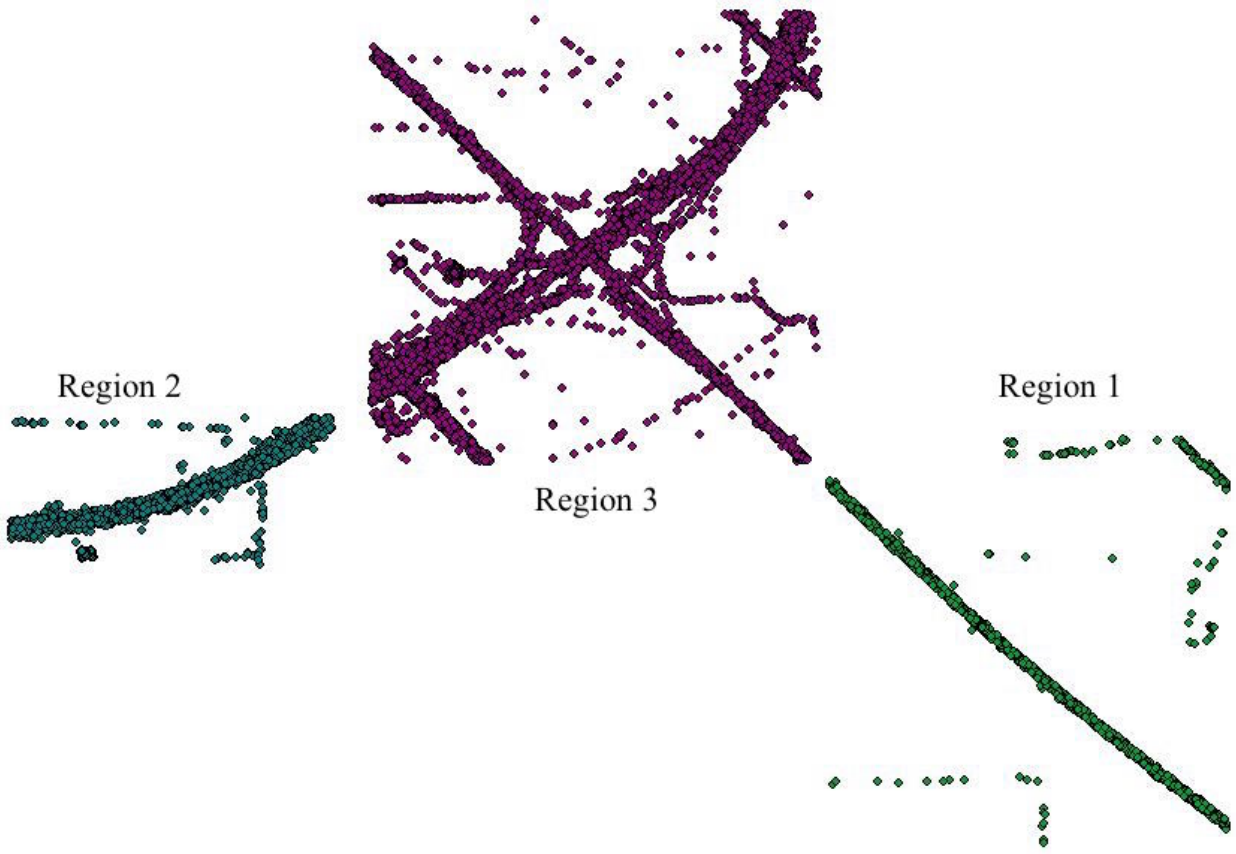

Fig 4.2: Study Area and Spatial Distribution of GPS Data in The 8th Tile

4.3 Raw GPS Data Analysis

The accuracy of GPS data can be affected by many factors such as: the GPS receiver unit quality, the position of the satellites at the time the recording was made and the characteristic of the surrounding landscape. Unfortunately, there is no information on any of the factors mentioned above, anyhow since receivers installed in taxis collected the data and despite recording errors it is assumed all the points in the dataset are on the road network. Therefore, all of the points share the same accuracy value.

The extracted road centerlines are off the actual road geometry due to the nature of the raw GPS data. Therefore, it was necessary to improve the quality of the data. In this section, the optimum parameters to preprocess the data were obtained based on statistics of original raw data collected by the taxis.

According to Niu, Z (2013), and inspired by Liu et al. (2012) and Ghiliani and Wolf (2002), the distance and directional change thresholds must be applied in order to improve the quality of the results. In this project, data timestamps are one minute apart where in Niu's study the timestamps are one second apart. So, depending on the road speed limit any given consecutive taxi points can be as far as two kilometers from one another, as shown in Fig. 4.3.

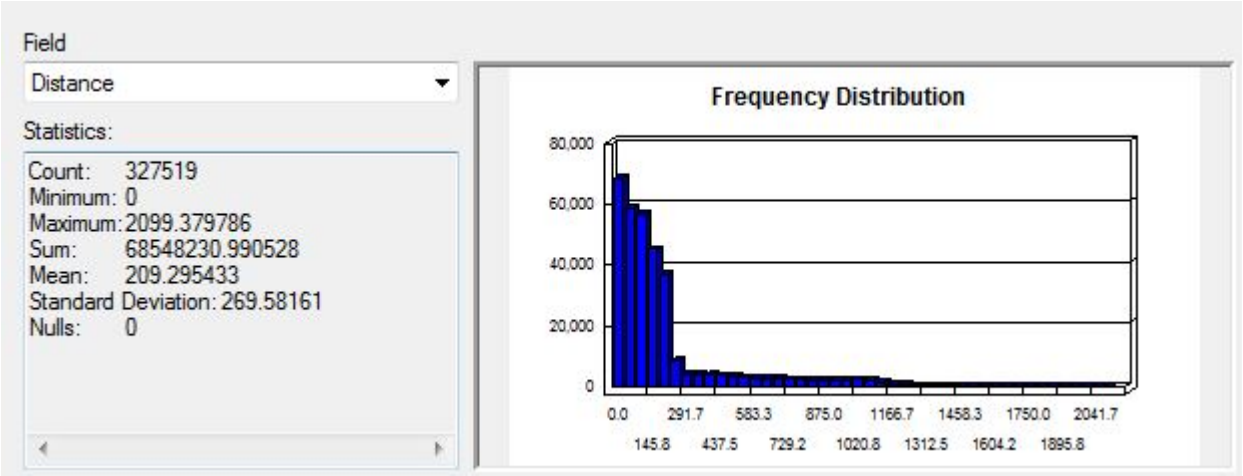

Fig 4.3: Statistics of Distances between Every Measurement for Each Taxi

As illustrated above, the maximum distance that any taxi traveled in 1 minute is 2099.38 m, with an average of 209.29 meters and standard deviation of 269.58 meters. In this project, segments of different highways are selected as the study area and the speed limit is assumed as 120 km/hr. Therefore, each taxi can travel 2000 meters within 1 minute that is the time interval between GPS measurements. A series of studies (Li et al., 2012; Karagiorgou & Pfoser, 2012; Wang et al., 2011; Zhang et al., 2010; Cao and Krumm, 2009) suggested that the best empirical value of direction threshold could be determined by series of experiments. Therefore, in this project threshold value of 2000 meters for distance and directional change of 180 degrees give the best

result for this dataset, as shown in Fig 4.4.

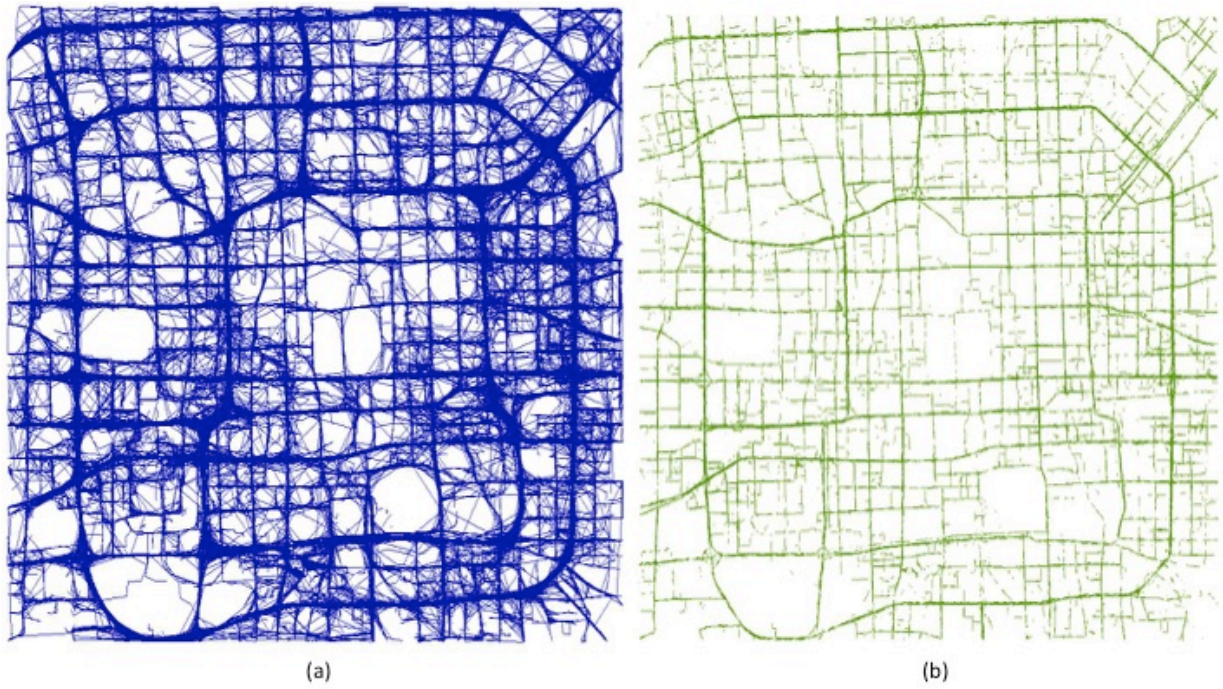

Fig 4.4: Before (a) and After (b) Applying the Threshold Values.

Other than the timestamp, since the data was collected in a different part of the world the appropriate coordinate system to choose in ArcMap is WGS 1984 UTM zone 50N. Latitude and longitude were also recorded in decimal degrees, which needed to be calculated in meters so the data would match the models and scripts created by Niu (2013). As mentioned earlier and according to previous studies, taxis with speed of 0 Km/Hr were called outliers and were eliminated from the dataset before any processing.

As mentioned in the previous chapter for refining GPS trajectories, threshold value of 500 meter was selected. This value gives the best result since other threshold values such as 50, 100, and 250 have been tested but extracted polyline segments are too short to be used for extracting final road centerline and 500 m results in longer polyline segments. Higher values can also be selected for creating more polyline segments but the bigger threshold value gets, processing time increases.

Chapter 5. Results and Analysis

This chapter includes the results of implementing the automatic road network extraction algorithm and also a visual comparison with an open street map data is provided.

5.1 Experimental Results

The data size was reduced by 82.5% after applying the standard circular window-smoothing algorithm, due to the 2-Gigabyte memory limitation of PythonWin, three typical regions were selected as the final case study in this project.

Fig. 5.1 demonstrates the overview of the collected GPS data and the final extraction of road networks in three different regions. Region 1 is selected to test effectiveness of the algorithms and methodology on straight roads, Region 2 to the results on the curved roads and the third region is selected to check the validity of the methodology at highway merges. It is apparent that centerlines of roads can be extracted via the implemented methodology. However, road centerlines of some minor roads and parts of highway ramps are not extracted because of the lower density of GPS points.

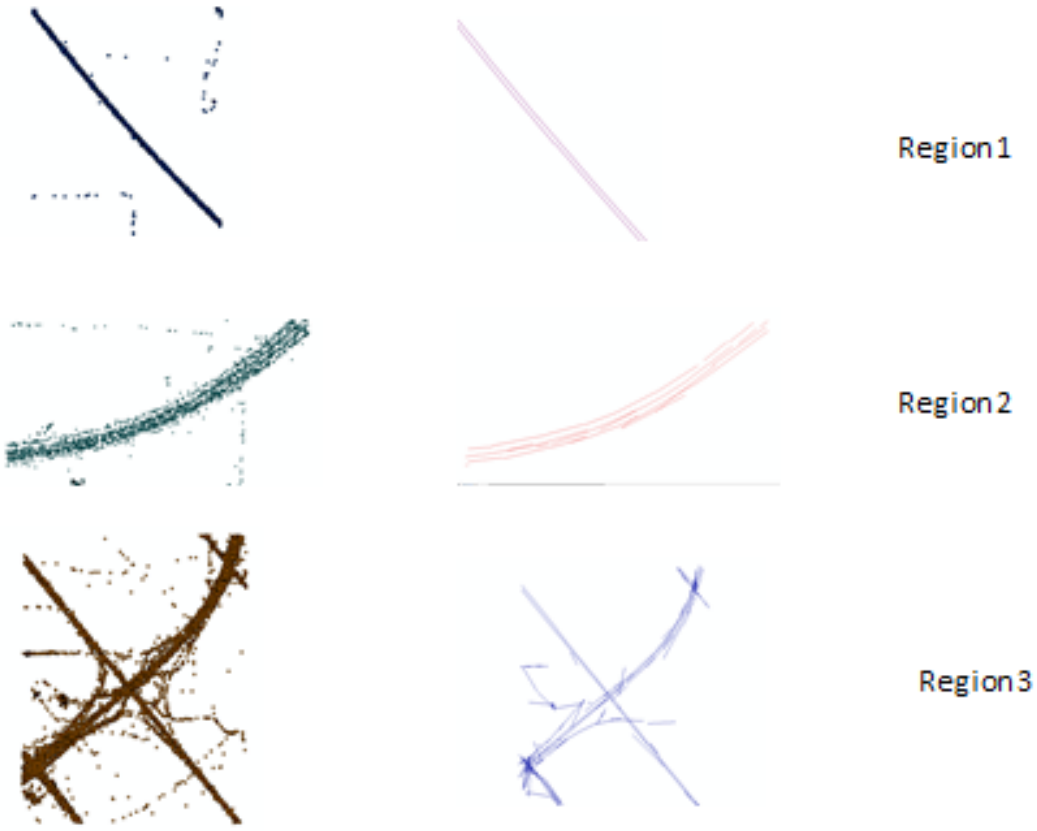

Fig 5.1: Overview of Collected GPS Data (left) and Extracted Road Network (right) of Three Typical Regions.

Furthermore, Fig. 5.2 to 5.4 illustrates the process of road centerline construction from raw GPS data more clearly. After preprocessing and smoothing, extraneous points were removed from the collection and also the remaining points were shifted to the middle of the road. Consequently, the distance between roads seems longer which helps to verify GPS points on different roads. The representative points were extracted in order to minimize the number of remaining outliers in dataset, while preserving the geometric shape major roads. It is clear that the constructed road network captured the direction and connectivity of the road network on each section.

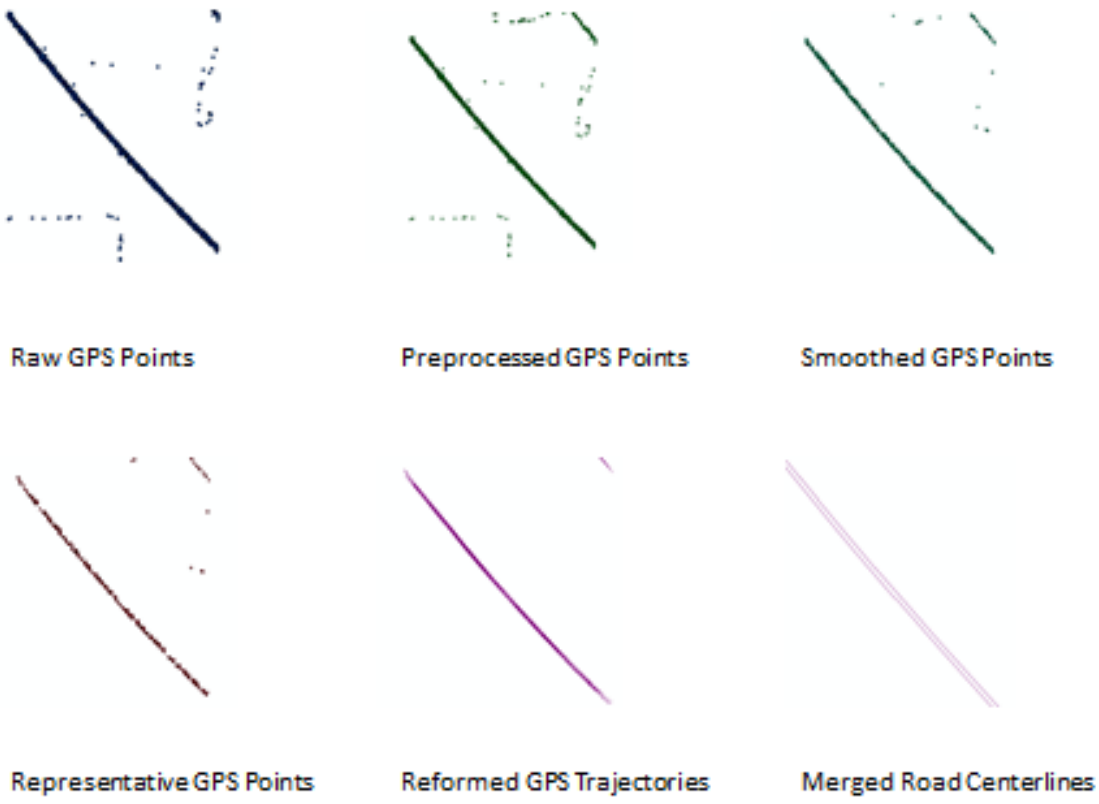

Fig 5.2: Results of Each Step of the Overall Workflow in Region 1

Region 1 is a typical straight segment of the highway in the 8th tile and is mainly selected to determine the adaptability of the method introduced by Niu (2013) with taxi GPS in Beijing, on a straight segment. As shown in the figure, points are filtered and merged to the center of the road and final centerline are extracted.

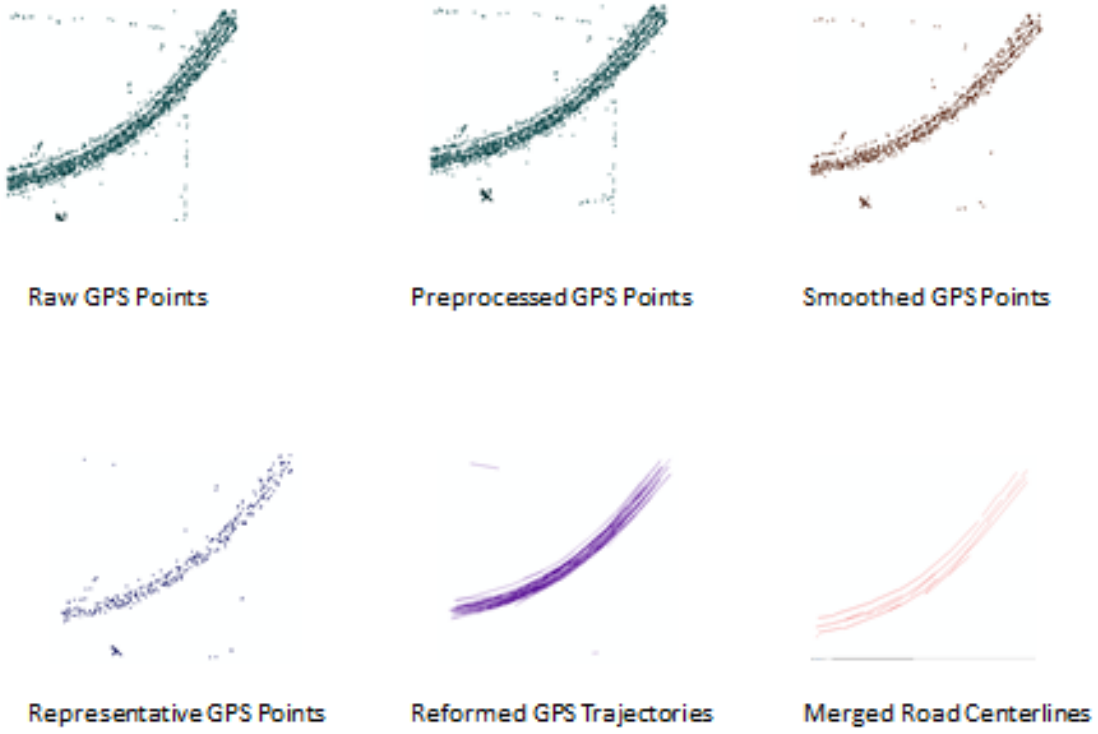

Fig 5.3: Results of Each Step of The Overall Workflow in Region 2

Region 2 is a typical curved segment of the highway in the 8th tile and is mainly selected to determine the adaptability of the method introduced by Niu (2013) on taxi GPS trajectories especially at splits and merges in between lanes. As shown in the figure, points are filtered and merged to the center of the road and final centerline are extracted.

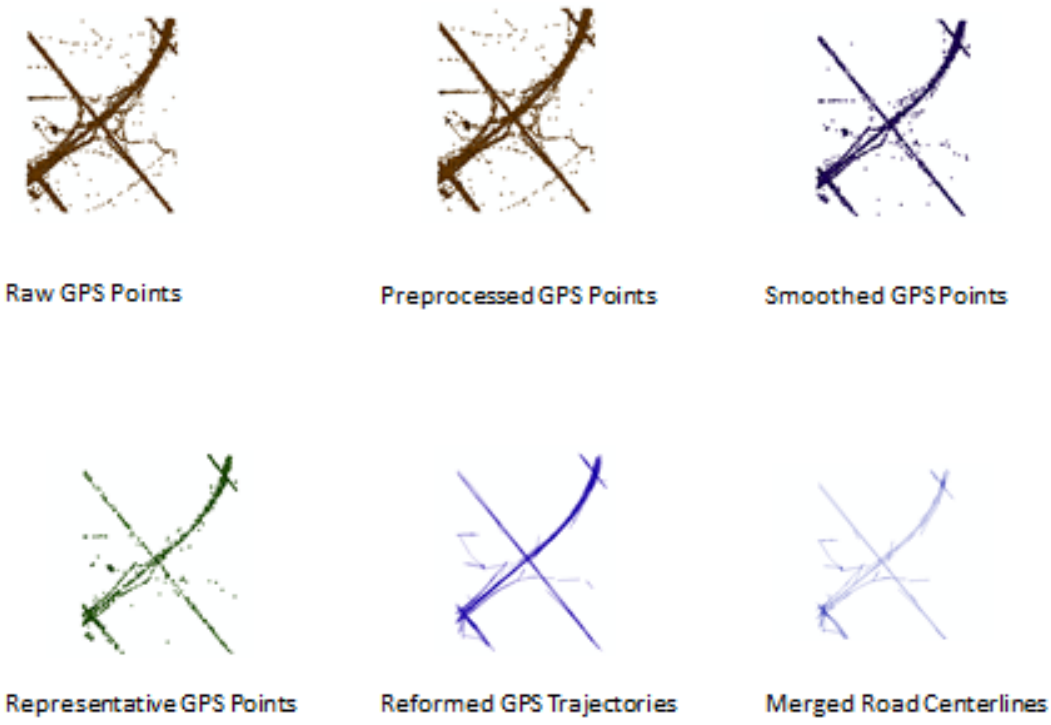

Fig 5.4: Results of Each Step of The Overall Workflow in Region 3

Region 3 is more complex compared to the other two Regions. Without the preprocessing step, the noise in the dataset can cause a number of GPS trajectories offsetting the road and overlap with other trajectories and also trajectories of opposite direction, as illustrated in Fig. 4.4. Therefore, the generation of road centerlines was utilized by connecting the representative points in similar movement routine.

5.2 Visual Inspection

According to the data acquired from OpenStreetMap (OSM) database, the extracted road network matches well with road features. OSM data (OSM) is supported by OpenStreetMap Foundation is a crowd-sourced dataset that is obtained from manual surveys, GPS data, aerial photographs and other free sources. Fig. 5.5, 5.6 and 5.7 show the overview of correspondence between the extracted road network and features in three typical regions. The majority of the

road centerlines match the geometry of road centerlines. In the following figures, I have used the base images provided by ESRI available for ArcGIS users, which was last updated in January 2017. World Imagery is the name of the service providing satellite imagery of world with one-meter resolution in some parts available from GeoEye IKONOS, Getmapping, and AeroGRID.

The data format obtained from OSM (.osm) is specific to OpenStreetMap so it needs to be processed with ArcGIS OSM editor in order to import them in ArcMap and manipulate the data. After editing and using specific tools provided by ArcGIS to handle the .osm file, the result is a set of polylines or possibly polygons.

Fig 5.5: Visual Inspection of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by ESRI, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 1.

Fig 5.6: Visual Inspection of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by ESRI, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 2.

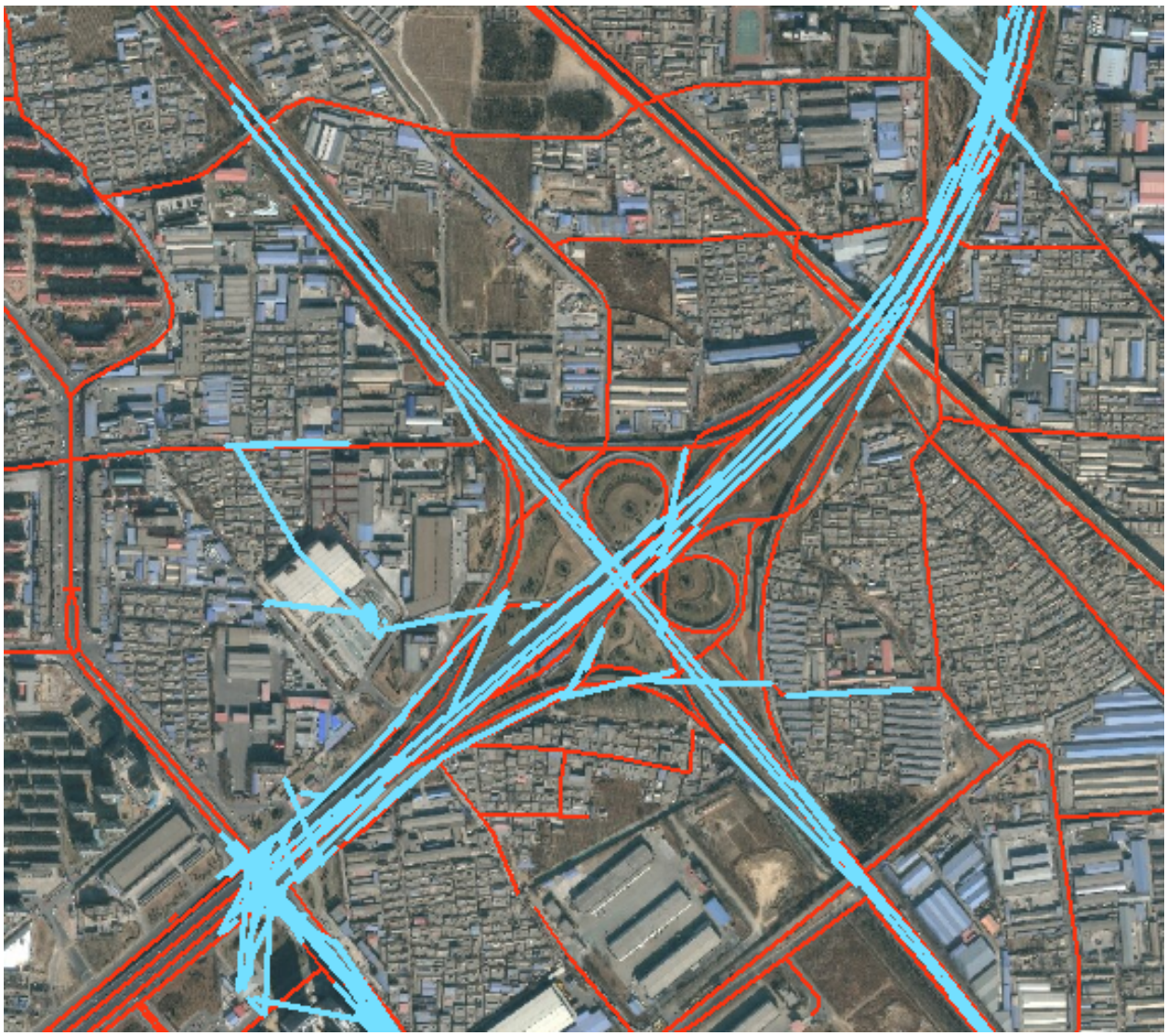

Fig 5.7: Visual Inspection of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by ESRI, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 3.

Figures 5.8, 5.9 and 5.10 show close-up views of extracted road centerlines at three typical regions. Roads centerlines can be constructed from massive GPS points for close to reality presentation of the road network. While, location of road centerlines is slightly oscillated in the road boundary most of them are close to their actual positions along the straight road segments.

Fig 5.8: Close-Up Views of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by Esri, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 1.

Fig 5.9: Close-Up Views of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by ESRI, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 2.

Fig 5.10: Close-Up Views of Extracted Road Network (Highlighted Blue Line) Overlaid with Aerial Photograph Provide by ESRI, and Road Centerline from www.openstreetmap.org (Red Lines) in Region 2.

After comparing the vector data obtained from OSM and extracted road centerline, the maximum perpendicular distance between the two centerlines on any straight segment was calculated to be 4.23 meters. This value was obtained by using the measure tool in ArcMap and measuring the longest distance between the extracted road centerline and the acquired OSM data.

Chapter 6. Conclusions and Future Work

The GPS data crowd-sourced by public transportation users provides an expanding source for enhancing road maps because of its rich spatial-temporal coverage and reasonable level of accuracy. The overall objective of this project is to implement an optimized methodology, which generates road centerline from GPS data obtained from taxis in Beijing without using any reference plans.

The overall workflow presented in the previous chapters along with results and objectives of this project are summarized in this chapter. Future work, which can improve the quality of the extracted road network, is also discussed.

6.1 Conclusions

As mentioned in Chapter 1, GPS trajectories are being used to extract road network for improving and updating the databases in road maps refinements. The method implemented in this project is a fast and inexpensive way of updating maps along with real-time changes. However, the presence of outliers and uncertain accuracy of the recorded GPS trajectory and also relatively long timestamp for GPS measurements in taxis cause major obstacles for extracting road network without the use of reference maps.

As summarized in Chapter 2, the main concern regarding the generation of road centerlines is how to effectively collect accurate geometry and also connectivity of the actual road network. This project was focused on the integration and modification of a point-based approach to extract road network without use of any reference map. The method implemented in this project was consisted of 5 main stages.

1) Preprocessing and standard mean smoothing algorithms were used to remove extraneous points.
2) Representative points were extracted by implementation of modified density based point clustering method.
3) Connecting representative points on the same lane.
4) Deriving road centerlines by density based clustering method to merge the reformed trajectories.
5) Connecting road centerlines topologically to create a completed road network.

The contribution of this project is to approve Niu, Z (2013)'s methodology and algorithm and also to overcome the main challenges found in similar studies as discussed in Chapter 2.

1) Overall, an 82.5% of GPS points were determined to be noise and were removed.
2) Due to a long timestamp of 1 minute between every GPS measurement the search radius and threshold value for distance between two consecutive points were set to 500m and also the threshold value for directional change set to 180^{o} which would better represent change in direction of flow.

6.2 Future Work

This project approves automatic self-learning GIS application for updating existing road maps and refining road maps with real-time changes developed by Niu, Z (2013).

The proposed thesis by Niu, Z (2013) incorporated threshold values speed and change in moving direction where time gap between any two consecutive points were 1 second unlike this project where the time gaps were 1 minute. It seems a value in between 1 second and 60 seconds can provide a better result for road network extraction. Since 1 second time gap is short and can cause point clouds at traffic lights or congested roads at a certain time of the day. Also 1 minute seems to be unreasonably too long since on a high way with speed limit of 120km/h two consecutive points can be 2000 meters apart which causes inaccurate extraction of road network on highway ramps as demonstrated in Fig.5.10.

References

Ai, T., & Yang, W. (2016). The Detection of Transport Land-Use Data Using Crowdsourcing Taxi Trajectory. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 785-788.

Cao, C., & Sun, Y. (2014). Automatic road centerline extraction from imagery using road GPS data. *Remote Sensing*, *6*(9), 9014-9033.

Cao, L., & Krumm, J. (2009). From GPS traces to a routable road map. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 3-12.

Dal Poz, A. P., Zanin, R. B., & Do Vale, G. M. (2006). Automated extraction of road network from medium-and high-resolution images. *Pattern Recognition and Image Analysis*, 16(2), 239-248.

Datatang, Beijing Datatang Technology Co. Ltd. www.datatang.com

ESRI 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Guo, D., Liu, S., & Jin, H. (2010). A graph-based approach to vehicle trajectory analysis. *Journal of Location Based Services*, 4(3-4), 183-199.

Karagiorgou, S., & Pfoser, D. (2012). On vehicle tracking data-based road network generation. *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, 89-98.

Li, J., Qin, Q., Xie, C., & Zhao, Y. (2012). Integrated use of spatial and semantic relationships for extracting road networks from floating car data. *International Journal of Applied Earth Observation and Geoinformation*, 19, 238-247.

Liu, X., Zhu, Y., Wang, Y., Forman, G., Ni, L. M., Fang, Y., & Li, M. (2012). Road recognition using coarse-grained vehicular traces. *HP Labs, HP Labs2012*.

Newson, P., & Krumm, J. (2009, November). Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 336-343). ACM.

Niu, Z. (2013). Automatic Generation of Road Network Data from Smartphone GPS Trajectories (Unpublished Master of Applied Science Dissertation). Ryerson University, Toronto, Ontario, Canada.

OpenStreetMap Foundation (OSMF), www.openstreetmap.org

Transportation Association of Canada (TAC). (1999). Geometric Design Guide for Canadian Roads, *TAC*, Ottawa, Ontario.

Wang, J., Rui, X., Song, X., Wang, C., Tang, L., Li, C., & Raghvan, V. (2011). A weighted clustering algorithm for clarifying vehicle GPS traces. *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, 2949-2952.

Wei, Y. A. N. G., & Ting-hua, A. I. (2016). Road Centerline Extraction from Crowdsourcing Trajectory Data. *Geography and Geo-Information Science*, *3*, 001.

Wolf, Paul R., and Charles D. Ghilani. (2002). *Elementary Surveying: An Introduction to Geomatics*. Upper Saddle River, NJ: Prentice Hall.

Zhang, L., Thiemann, F., & Sester, M. (2010). Integration of GPS traces with road map. *Proceedings of the Second International Workshop on Computational Transportation Science*, 17-22.