Improved Map-Matching Method for Transit Trajectory Reconstruction based on Transit GPS Data: Case Study in Edmonton, Canada

Ai Teng
Shenzhen Transportation Design & Research Institute Co. LTD
1006 Dongmen N. Rd. Lohu Dist., Shenzhen, China
526394085@qq.com

ABSTRACT
Map matching shows its excellence in terms of generating the vehicle trajectory on the digital map, which is the foundation of trajectory-based bus travel time prediction model. An understanding of reliable map-matching method is necessary for developing the accuracy of real-time prediction results. This thesis provides an improved map-matching method, which has better performance in finding single bus trajectory than Spatial-temporal matching method, a well-recognized map-matching method used in previous research. Link identification and path inference are two key components for this method. The field test is conducted on a main arterial in Edmonton, Canada from Legar transit center to Century Park transit center.

KEYWORDS: Map-matching, Link identification, Path inference, Single trajectory, Spatial-temporal matching

1 INTRODUCTION
The increasing ownership of private vehicles over the last 3 decades stimulates the transformation from public transit to private trips, causing severe traffic problems. To increase the attraction of the public transit, travel time is increasingly critical for Advanced Traveler Information Systems (ATISs). Such trajectory-based travel time prediction models regularly require the map-matching results as the fundamental data input. However, the GPS trajectories cannot be directly matched to the digital road network due to the different link width and misreported data. The accuracy of the map-matching results directly affects the prediction results.

Map-matching is commonly considered as a ‘search’ problem and simply integrates the geometric information and features of the digital map by matching the GPS locations to the ‘nearest’ given point, also referred as range query (Maurer, 1980). Map-matching can be otherwise considered as a statistical model. The analysis requires the definition of an elliptical or rectangular confidence region around a fixed position obtained from a navigation sensor (Mohammed A. Quddus, 2007). Each link within this range is given a probability of matching. The path with the highest probability will be chosen as the matching results. Traditional map-matching method focuses on finding the links with shortest distance and smallest direction difference from the GPS points. The method in this paper minimize the error of angle comparison and converts the ‘point-link’ match to ‘point-point’ match.

Some of the existing bus travel time prediction models are based on scattered historical space-time points, i.e. time series model (Zhu, Ma, Ma, & Li, 2011), artificial neural network model (Gurmu & Fan, 2014). The core is also to reconstruct the previous trajectory and train the model predict the future one. Other models are based on the historical trip data, i.e. support vector machine model (Wu, Ho, & Lee, 2004), Kalman filter model (Wang, Wang, Yang, & Gao, 2012). The accurate map matching results could avoid the error brought by the discrete points and generate the trajectory closest to truth.

The improved map-matching method in this paper is a reference-point based searching process. The links on digital map are divided into small sections at reference points of intersections, and
specific distance from start and ending points of links. Each GPS point is given an array of distances between the GPS point and each reference points within certain range. The reference points with 3 shortest distances are considered as potential match, then conduct the angle comparison. The link to which the final matching reference point belong is where the GPS point passed.

2 DATA DESCRIPTION

The data used in this paper is described in this chapter.

2.1 GTFS Data

GTFS is short for General Transit Feed Specification, which is the definition of a common format for public transportation schedules and associated geographic information (Google, Static Transit, 2015). GTFS data is is used as GPS input in this paper.

Trip update and vehicle position are two parts of the dataset. Trip update contains information of each trip made by transit vehicles, including fields of TripID (unique identifier), RouteID, Vehicle_Label, Start_time and Stop_Seq, recording every stop the bus makes. Vehicle position contains the information of physical location of 30 seconds frequency, including fields of TripID, Vehicle_Label, Timestamp, Longitude and Latitude.

2.2 Digital Map Data

Map-matching requires a base map to show the matching results. The digital map used in this paper is in standard format provided by City of Edmonton. The roads are divided into virtual links at inflection nodes. Digital map is made of links including fields of Link_ID (unique identifier), starting_point ID, ending_point ID, Length, Road_name, Direction. The digital map also contains the information of link connectivity, including fields of Link_ID, Connected Link_ID and Sequence to show how many and which links are connected to one.

2.3 Reference Point

This paper proposes a map-matching method based on introduction of reference points. Reference points are generated on the digital map with 20 meters in them between. Information includes fields of ID (unique identifier), Link_ID, Distance from link begin, Distance to link end, Direction, Longitude and Latitude.

2.4 Grid

The digital map is cut into 197x159 small grids to narrow the matching scale. The information contains Grid_ID (unique identifier), the amount of reference point contained in this grid and all these reference point ids. The GPS points sometimes are located very close to the edge of the grid, and the matching process will generate the error results if the correct reference point is not included in the grid. In the case shown in Fig. 1, GPS point belongs to grid 1414, and is expected to match the RP #62239, which is however not within the grid 1414. The GPS point is the closest to the reference point outside the grid. Therefore, each grid contains not only the RPs within the grid, but also those around 20 meters outside the grid. The belonging relationship between RPs and grid is shown like below.
3 IMPROVED MAP-MATCHING METHOD

In this section, the map-matching method contains two part: (1) link identification procedure, which is the projection process to match the GPS point to the digital network; (2) path inference, which is to confirm path link sequence that the vehicle uses to complete the trip between two adjacent GPS points.

The major contribution of this method is: (1) create the reference point file in the digital map. It makes it more efficient and direct for the matching process of the low-frequency data; (2) perform the experiments field tests based on the large amount dataset collected from the real traffic situation, providing the reliable evaluation of the method.

3.1 Link Identification

The GPS points are normally off the segments in the digital network. The link identification procedure is conducted to match the GPS points to the digital segments to obtain the trajectory information. The method proposed in this paper thesis can be divided into three parts, including database design, data extraction and projection analysis like shown in Fig. 2.

![Data Flow for Link Identification Procedure](image)

**Definition 1:** Gridded set: The database containing all the reference points categorized based on which grid they belong to.

**Definition 2:** Candidate set: Output of the data extraction that shows all the reference points belonging to the grid in which the GPS point is.

**Definition 3:** Target reference point: The reference point closest to the GPS point with the smallest direction differential.

**Definition 4:** Target link: The link to which the target reference point belongs.

**Database Design**

The database is designed based on the digital road network of City of Edmonton. The map is projected in GCS_WGS_1984 geographic coordinate system. The goal of the database is to downsize
the matching scale for the GPS point to one grid instead of the whole network. The database contains two parts of information, including the basic grid arrangement and reference point identifier belonging to each grid.

The range of the network is in plain coordinate system, and offline divided into 193 columns and 159 rows based on the standard format of map-matching inputs. Each grid has a unique identifier (Grid_ID) and the reference points existing belonging to the grid are picked out.

**Data Extraction**

When GPS logs in, the database will automatically search for the grid ID, to which the GPS point belongs, and pick out extract all the reference points within and closely around the grid to create the candidate link set, which is the potential match scale. The process is shown in Fig. 3.

**Projection Analysis**

Projection analysis is the final step of the map-matching method proposed in this paper. The purpose of projection analysis is to match the GPS point to the digital link within the selected grid by the former steps based on the comparison of distances and directions, like shown in Fig. 4.

![Figure 3 Data Flow of the Data Extraction Process](image)

![Figure 4 Map-Matching Projection Analysis](image)

### 3.2 Path Inference

Path inference is defined as the determination of the most likely trajectory of the trip given the sequence of matched GPS points (Rahmani, 2013). After matching the GPS points to the digital road network, the trajectory that is connected by all the projections in certain sequence needs to be inferred.

Transit trajectory is the key information for passengers and the bus agency for efficiency and security purposes. The methodology used in this paper has three steps:

**Build the possible paths**
This step is based on the projection of the GPS points and the connectivity information of the digital network to find all the possible paths. The possible paths include all the accessible link combinations that connect every two adjacent projected points.

**Using shortest path algorithm to calculate the costs of possible paths**
In the previous literature review, there are several shortest path algorithms (SPA) merging in the past ten years. The SPA used in this paper is classic Dijkstra algorithm (Dijkstra, 1959). The inferred path is mainly decided by the cost of each link.

The cost of each link can be decomposed to three parts based on the road condition and traffic situation, including average travel time, delay related to the signalized intersections and delay due to turns (conflicting time with opposing traffic or merging traffic).

**Find the most likely trajectory**
The most likely trajectory is selected from the possible path set based on the shortest path model to be the path with the lowest cost.

**Summary**
The existing GPS signal technologies intend to record low frequency (15 seconds or more) GPS data given the limited storage, therefore, the distance between two adjacent GPS points normally covers several links, which makes it difficult to determine which path the vehicle was actually on. Under circumstances like equal cost paths, overpass (i.e. bridges, skyline) or underpass (i.e. tunnel, underground highway) facilities, high density road network.

However, in this paper, the focus is placed on the map-matching for transit vehicle. The route is fixed and the punctuality and reliable prediction are the key.

4 **ACCURACY COMPARISON WITH SPATIAL-TEMPORAL MATCHING METHOD**
The benchmark for comparison is ST-Matching method, which was first proposed by Lou et al. in 2009 (Lou Y., et al., 2009). The algorithm in the method targets the map-matching for the large sampling interval GPS data. The matching procedure consists of: (1) spatial analysis uses both geometric and topological information to pick out the most likely candidate points; (2) temporal analysis employs the average speed between two consecutive GPS points to exclude the interference candidate options. ST-matching requires the average speed, which is calculated as the ratio of the distance between two consecutive GPS points over the time interval. It should be noted that ST-matching method is based on the probability model to find the most likely path that matches the GPS sequence. The candidate points are picked out from pre-determined area with radius of 30 meters. It assumes the probability of each candidate point of one GPS point follows normal distribution, and the distance and speed are used to calibrate the probability to find the most likely candidate sequence.

In order to quantify the comparison, two indices of accuracy are defined as below:

The identification of the correct links to which the GPS points belong can be evaluated by comparing correctly identified link with the manually recorded trip route. The correctly identified percentage (CI %) can be computed by Equation 1.

\[ CI(\%) = \frac{Number of Correctly Identified Points}{Sample size} \times 100\% \]  

(1)

Given the method can accurately identify which link the GPS belongs to, there still is error for the precise location, therefore, the average distance error (ADE) is defined, like shown in Equation 2.

\[ \text{Average Distance Error} = \frac{\sum |d_{mp} - d_{upp}|}{\text{sample size}} \]

(2)
In this paper, vertical projection point is used as the matching location of the GPS point on the digital road network. $d_{mp}$ denotes the distance between the starting point of the target link and the target reference point. Since the GPS point has already been matched to the target link, the location of vertical projection point is available for each GPS point. $d_{vpp}$ denotes the distance between the starting point of the target link and the vertical projection point.

A field study is conducted on 23 Ave. which is an essential arterial located in the south of Edmonton, Canada. The corridor from Terwillegar Drive (in the west) to the Calgary Trail (in the east). The vehicle position and trip update of 169 trips are available for the thesis. There are 85 trips are in eastbound and 84 trips are in westbound. Since there are data missing and device error in the GPS transmission and storage, this thesis only uses 68 validate trips in eastbound.

For the accurate link identification, the field test using mobile device was conducted to collect GPS acting with sampling interval of 1 second as the ground truth. To test the relationship between the method accuracy and the data collection frequency, three more datasets were generated from the high-sampling interval data, which includes 15-second data and 60-second data. Partial trips from Legar Transit Center to Century Park Transit Center are used for map-matching process evaluation shown in Table 1 and the matching results used reference-point-based method can be found in Table 2.

<table>
<thead>
<tr>
<th>Trip ID</th>
<th>From Time</th>
<th>From Station</th>
<th>Vehicle ID</th>
<th>Route ID</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>11414576</td>
<td>13:43:28 Jul 28</td>
<td>CPTC</td>
<td>4673</td>
<td>26</td>
<td>666</td>
</tr>
<tr>
<td>11414614</td>
<td>14:05:49 Jul 28</td>
<td>LTC</td>
<td>4362</td>
<td>26</td>
<td>558</td>
</tr>
<tr>
<td>11414653</td>
<td>14:31:39 Jul 28</td>
<td>CPTC</td>
<td>4840</td>
<td>23</td>
<td>628</td>
</tr>
<tr>
<td>11414616</td>
<td>15:01:10 Jul 28</td>
<td>LTC</td>
<td>4346</td>
<td>36</td>
<td>702</td>
</tr>
<tr>
<td>11414680</td>
<td>15:13:04 Jul 28</td>
<td>CPTC</td>
<td>4346</td>
<td>36</td>
<td>686</td>
</tr>
<tr>
<td>11414712</td>
<td>15:48:13 Jul 28</td>
<td>LTC</td>
<td>4826</td>
<td>23</td>
<td>720</td>
</tr>
<tr>
<td>11414156</td>
<td>16:06:46 Jul 28</td>
<td>CPTC</td>
<td>6003</td>
<td>23</td>
<td>660</td>
</tr>
</tbody>
</table>

CPTC: Century Park transit center; LTC: Legar transit center

Table 2 Performance of RP-based Method with Different Frequency GPS Data

<table>
<thead>
<tr>
<th>Variables</th>
<th>Frequency of GPS data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1s</td>
</tr>
<tr>
<td>Sample size</td>
<td>4620</td>
</tr>
<tr>
<td>Correct link identification</td>
<td>4609</td>
</tr>
<tr>
<td>Average distance error</td>
<td>4.97 m</td>
</tr>
<tr>
<td>Processing time</td>
<td>&lt; 25 s</td>
</tr>
</tbody>
</table>

Fig. 5 shows the performance of the datasets with different sampling intervals. Both methods can maintain the high accuracy (over 98% correctly identified link percentage) with small sampling interval GPS data. With the sampling interval becomes higher, the accuracy of ST-matching method declines. According to previous literature, the performance of ST-matching method will become stable when the sampling interval increases to 120 seconds or higher (Lou Y. et al., 2009). For the reference points-based method, the accuracy is not affected significantly by the sampling interval.
5 CONCLUSION

Map-matching methods are receiving increasing attention because they are the foundation data input for the trajectory-based transportation applications; i.e. the reliability of the matching results directly affect the accuracy of the applied models, like prediction models, estimation models. This paper focuses on a reference point-based map-matching method. Compared with traditional methods, the improved method coverts the point-to-curve match to the point-to-point match. The conversion excludes the interference of the curved road alignment on the match results. The distance calculation and projection is more accurate and easier for point-to-point than point-to-curve since the direction information can not directly obtained from most sources of raw GPS data. The major contributions of this paper include the follows:

A method is created to generate the reference point file in the original digital map using ArcGIS. The file includes the geometric information of each reference point and the matching relationship between the reference point file and original link file.

This paper uses a new way to narrow down the matching scale by dividing the digital map into square grids. Traditional methods search the candidate links in an ellipse or circle with predefined axis length or radius. It cannot be sure if the target link is in this range. The grids defined in this thesis include all the reference points that may be matched to the GPS points. The matching process is conducted after the identifying to which grid the GPS belongs.

An algorithm is developed to realize the improved map-matching method including locating the GPS points in the grids, distance and direction comparison between reference points and GPS points.

The paper proposes a new map-matching method to generate better map-matching results for further transit prediction and estimation of travel time and arrival time.

6 ACKNOWLEDGEMENTS

Deep appreciation and respect for my mentor, Dr. Ge, my boss, Bin, and my supervisor of master program, Dr. Tony Qiu.

REFERENCES

Zhu, T., Ma, F., Ma, T., & Li, C. (2011). The prediction of bus arrival time using global positioning system data and dynamic traffic information. Wireless and Mobile Networking Conference (WMNC) 4th Joint IFIP (pp. 1-5). IEEE.


About author

Teng, Ai, (M.Sc.) is now working as a junior transportation engineer in TDRI (Shenzhen Transportation Design & Research Institute Co. LTD) with emphasis on public transit planning, transit-related policy-making and transit operation evaluation. She could be reached at Tel.: 0755-25193461 or Email: ateng1@ualberta.ca.