Processing Improvement of Map-Matching for Travel Time Prediction Model

by

Ai Teng

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in
Transportation Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

For trajectory-based travel time prediction model, map matching shows its excellence in terms of GPS data processing by providing an efficient technique to generate the vehicle trajectory on the digital map. The transit vehicle trajectory contains essential information about arrival time at bus stops and delay at major intersections. An understanding of reliable map-matching method is necessary for the development of the real-time prediction result accuracy. This thesis provides an enhanced map-matching method, which has better performance in terms of accuracy of path inference and link identification, compared with Spatial-temporal matching method, a well-recognized map-matching method used in previous literature. Compared with the existing map-matching method, a reference point file is added to original digital map, converting the point-to-curve match to point-to-point match. The map is also divided into equal digital grids by latitude and longitude to narrow down the matching scale. The feasibility and the accuracy of the method are evaluated in different traffic environment using real field geometric information and GPS data. The last part of the thesis is the comparison analysis between single transit trajectory prediction results derived from from both map-matching methods. The field test is conducted on 23rd Avenue corridor from Legar transit center to Century Park transit center in Edmonton.

Acknowledgement

I sincerely appreciate all those who support my graduate studies generously.

First, I would like to send my respect to all committee members for their advising and comments. I would like to express my deep appreciation to my supervisor, Professor Tony Qiu. This thesis could not have been completed without his guidance and encouragement.

I would also thank Yuwei Bie, Jiangchen Li, Chenhao Wang, Dr. Peng and all the team members in Center of Smart Transportation. Thank you for providing me with assistance and suggestions about the research during my master program.

Second, I deeply appreciate the help from Xiaoyi Lan, Lian Gu, Yan Ma, Yiyi Qian, Difei He, Will P. Roschefild, Matthew S. Woo and all my friends who support, comfort me all the time and encourage me to go forward. Their understanding and trust are the best motivation for me to finish my graduate study.

Finally, I wish to thank my families Mr. Hongzhi Xu, Mrs. Qijie Wu and Xiaohui Zhao, Yixuan Xiong, for their confidence and continuing love during all these years.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

This chapter presents the background of map-matching method and its application in trajectory prediction model. In this part, the author also describe the research motivation, objectives and structure of the thesis.

1.1 Background

The increasing ownership of private vehicles over the last 30 years simulates 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. With the development of the GPS collection techniques, transit agencies equip the transit vehicles and taxis with GPS transmitters and receivers to collect the real-time location for better management and services. However, the GPS trajectories cannot be directly matched to the digital road network due to the different link width and misreported misreported data. The accuracy of the map-matching results directly affects the prediction results.

The philosophy of map-matching is to identify the location and the trajectory that matches the digital map. The decision of link identification can be made based on the distance, speed, direction of GPS data, or other effective methods that can infer the precise location of the GPS on the digital map. Most of existing map-matching methods are developed with the GPS data with insufficient quality and the sampling intervals of collection is higher than 60 seconds. The methods can generate the results that which road the GPS points belong to, but there are still limitations to the precise locations. The limitations also include the implementation problems. Either the method is cost-inefficient or the method is not modifiable according to different traffic environment.

1.2 Problem Statement and Research Motivation

This thesis focuses on a developed reference point-based map-matching method. The key point is the link identification and distance calculation to determine the precise locations of the GPS points on the digital network. Fig. 1.1 shows the main idea of the map-matching problem. Fig. 1.1(a) shows the real road network with the GPS record. However, the GPS point does not belong to any links in the citywide because the width of the real links cannot be reflected on the digital map, therefore, no available information of the traffic status can be acquired through the GPS data.

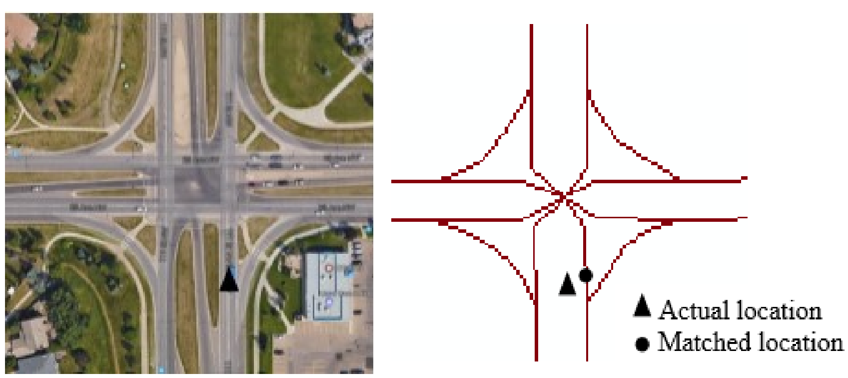

Figure 1.1 (a) Set of Actual Road Condition; (b) Set of Estimated Arcs and Matched Results

Although, there are plenty of map-matching methods nowadays, the application is still limited by the data format and traffic conditions. Some software can only recognize certain format of the input data. Each method has its preferable road network to generate the accurate results.

A travel time prediction model is proposed to test the application of the proposed map-matching method. The model is established based on transit data (GTFS data) within different time intervals to the current state. The development is that the proposed model combines the varying impacts of historical data on the current data to obtain the predicted results more close to the real traffic status.

1.3 Research Objectives

This thesis looks into developing map-matching method with three angles. There are three specific objectives of this thesis:

a. Establish the dataset including reference point file in the digital map and integrate the GPS data input;
b. Develop the map-matching algorithm and compare its accuracy with spatial-temporal matching method. The case study is conducted in different road conditions
c. Analysis of the performance of the transit travel time prediction model using the matching results from both methods.

1.4 Structure of thesis

The structure of this thesis is as follows:

Chapter 1 presents the introduction of the map-matching methods and travel time prediction background and the main problem discussed in this thesis.

Chapter 2 shows the literature review about the existing map-matching methods and travel time prediction models. The limitations of existing methods are also discussed.

Chapter 3 describes the data format used in this thesis, which is the input of the map-matching algorithm, including digital map information and GTFS data.

Chapter 4 presents a RP-based map matching with a reference point file in the digital map. The accuracy of the map-matching results is compared with ST-matching method. The comparison shows that RP-based map-matching method outperforms the ST-matching method in terms of link identification accuracy and path inference.

Chapter 5 introduces the processing improvement for single trajectory prediction model based on transit GPS data of different sampling intervals.

Chapter 6 is the conclusion and contribution summary reached in this thesis and the proposed future work.

CHAPTER 2 LITERATURE REVIEW

This chapter reviews the previous literature about existing map-matching methods and travel time prediction model that are widely used.

2.1 Review on Map Matching Method

Map-matching process identifies the proper link sequence based on the collected positioning data and roadway centerlines in the digital map (Mohammed A. Quddus, 2007). Most existing researches about map matching have focused on both the user's location and the map that is known for a high degree of accuracy (Christopher E. White, 2000) and can be categorized into four groups listed below.

The map matching is considered as the search problem and simply integrates the geometric information and features of the digital map. Certain amount of researches match the GPS locations to the "nearest" given point, also referred as *range query* (Maurer, 1980). The method is easily implemented and efficient to operate. However, the connection between the digital links is not considered and the matching results highly rely on the layout of the digital nodes. Normally, intersections and major turning points are treated as the nodes. Links with more digital nodes are more likely to be matched to (Christopher E. White, 2000).

The map-matching problem can be considered as 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 matched results. Honey et al. first introduced this model to match the positions for a position sensor and a map (Honey, 1989). Ochieng et al. (2004) develops an optimal estimation algorithm to determine the matched locations of users on a link and evaluates the impacting factors of GPS data on the accuracy of matching process (Ochieng, 2004).

Connectivity and contiguity of the links are helpful to find the link sequence, therefore, the topological analysis is developed. Greenfeld et. al. proposes a weighted topological method based on analysis of the digital map and position information of users (Greenfeld, 2002). First is to find the possible path and use the probability analysis to find the most likely path.

Other map matching methods include those which use sophisticated concepts. Syed et. al. develops a map-matching method based on the fuzzy logic theory. The results show that fuzzy logic can be effectively used for map matching in urban canyons because of its ability to generate precise output from noisy (error prone) navigation input obtained from GPS (Syed, 2004).

Obradovic et. al proposes a two-step map-matching approach to integrate the sensor-collected data, GPS data and digital map information. The first step is to update the user-movement model using installed odometer and GPS signal based on the Kalman filter.Second step is to compare the candidate trajectories with improved user-movement (from *step 1*) and find the best match (Obradovic, 2006).

The limitations of existing map-matching methods include following. The usefulness of map-matching method highly depends on the accuracy of GPS point and digital map like the following situations in Fig. 2.1.

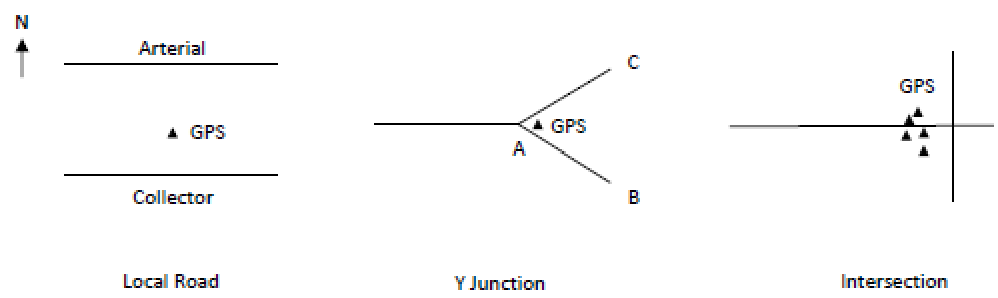

Figure 2.1 Hypothetical road network

In scenario 1, there are three lanes in eastbound of the arterial and one lane for each direction of the collector road. Obviously, the GPS point probably has shorter distance to the collector road in the digital map. In this case, GPS point will be matched to the collector road. In this case, speed information of the GPS point can distinguish the interference. In scenario 2, it is hard to determine if the probe vehicle has passed point A or not. Direction of the GPS can help determine the right match. When the vehicle is stopping before the signal line, the GPS will have slight off the current road even if it is not actually moving like shown in scenario 3. In this case, it is hard to match all the points to the same location. The reference point file can help collect these points to the same point. Besides the scenario mentioned in Fig. 2.1, other situations like overpass, underpass, turning restriction should be taken into consideration to increase the accuracy of map-matching method.

The low accuracy of the GPS data is another problem with path inference. For the road network like Fig. 2.2(a), the GPS receiver is operating on two parallel roads with distance of 23.4 meters in between. The accuracy of the GPS points is within the rough range of 12 meters, which results in the confusion about what the route of the vehicle is like.

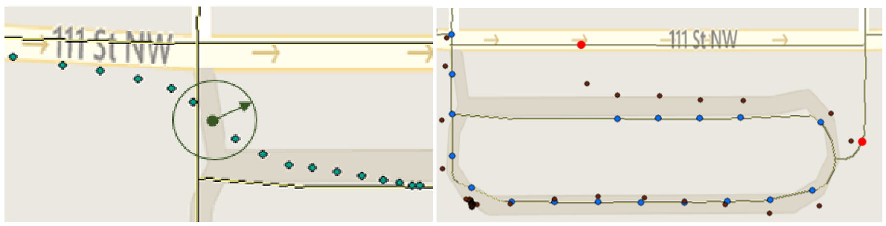

Figure 2.2 (a) False Identification Example; (b) False Identification Results.

The map-matching is usually conducted for the further research on the trip information on the entire path or corridor. The matching results are the basic evidence to show the operation condition of the vehicles on the researched links, paths or corridors. Existing map-matching methods are mostly rely on the matching consistency of the direction and the shortest distance between the GPS points and the links in the digital map. Normally, the connectivity between adjacent links and adjacent GPS points is not taken into consideration, therefore, each matching result is independent from the others. The results for the single trip may not be able to maintain the consistency. Fig. 2.2(b) shows the transit is arriving at the transit center. Dark purples are the GPS records, while the blues are correct matched results. There are two mismatched results marked as the red spots. If considering the consistency of the links and GPS points, this kind of mistake can be avoided.

Traditional map-matching method focuses on finding the links with shortest distance and smallest direction difference from the GPS points. However, the digital network contains a large amount of the links, it would take long processing time to compare the GPS point with every links in the network. Therefore, the digital map is simplified to save the processing time in the traditional methods, which may result in the inaccurate matched results.

2.2 Review on Travel Time Prediction Model

Advanced Public Transportation Systems (APTS) related technologies are widely expanding, for instance global positioning systems (GPS), automatic vehicle location (AVL) systems (AVL), and automatic passenger counters systems (APC), advanced traveler information system (ATIS). Accurate transit arrival and departure information should be provided to passengers for arranging their trip plan and to transit operators to properly arrange the transfers and schedule plan (Kalaputapu & Demetsky, 1995; Khan & Abdelfattah, 1998; Chien, Ding, & Wei, 2002). Therefore, it is vital to conduct the accurate prediction of transit travel time. A variety of prediction models mentioned in previous researches are reviewed.

2.2.1 Historical data model

Historical data models assume that travel time of predicted trip is related to the previous trips and the traffic condition of the links remain consistent. It is assumed by Chen et. al. that the traffic condition within the citywide changed cyclically and the ratio of current and previous travel time remained stable (Furth, Brendon, Theo, & Strathman, 2003). The prediction model can be calibrated based on the real-time transit data. However, the model is based on massive historical data and exclusive to certain area. The hybrid model is presented to combine the historical data and recent data by Gong et. al. (Gong, Liu, & Zhang, 2013). The higher weight is given to the more recent is data. This model is closer to the real traffic condition, but the distribution of the weight has linear relationship with the time series instead of the traffic condition.

2.2.2 Regression model

The regression model takes multiple variables (i.e. passenger number, stop number, link length, delay, etc.) related with the traffic condition build the regression function. The passenger number and delay at the stops are obtained based on the APC (Automatic Passenger Counters) and built the regression model with variables of distance, control delay, stops number and trip starting time (Patnaik, Chien, & Bladikas, 2004). Fuzzy regression is used to build the travel time prediction model and evaluated the model using the transit data from Shenzhen, China (Yang, Bao, & Zhu, 2004).

Regression model can minimize the effects of varying traffic condition on the predicted results. However, the variables in the model are supposed to be independent from the others (or the relevance is lower than the preset threshold), which cannot be guaranteed in the real-time situation.

2.2.3 Time series model

Time series model is an extension of historical data model based on the pattern of changing traffic flow. The cyclical characteristics of the traffic flow are captured to establish the nonlinear regression model to predict the travel time (Sherif & Al-Deek, 2002). The delay at the bus stops and control delay at the intersections can be considered into travel time prediction as well (Zhu, Ma, Ma, & Li, 2011).

The accuracy of time series model is highly related to the similarity between current traffic situation and historical situation. If the traffic condition experiences significant changes (e.x. traffic flow, signal control plan, priority plan), the model will more likely create errors in prediction.

2.2.4 Kalman-filter model

Kalman filter model is to solve linear filter problem of discrete data based on recursion method. Shalaby and Farhan collected the AVL and APC data from transit vehicles in Toronto, and applied Kalman filter to build the prediction model. Data from the first four days was used as the model training, and data from the following day was used as the test (Shalaby & Farhan, 2003). Wang et al improved the adaptability of the Kalman filter model by adding a "forgotten factor" to the procedure to restrain the influence of the old data on the model (Wang, Wang, Yang, & Gao, 2012). Kalman filter model requires less historical data than other methods and more reliable in short-term prediction. However, the model needs meeting higher standard of the equipment of data collecting and more calculation. The results are not reliable in long-term prediction.

2.2.5 Artificial neural network model (ANN Model)

ANN model emulates the learning process of human brain, which is good at pattern recognition, prediction, classification, etc. ANN models are calibrated using two steps, including training and testing . Gurmu et al. input the GPS data as the only data resource to the dynamic travel time prediction ANN model and then to predict the arrival time. Predicted results indicated that the prediction accuracy and robustness of this model outperformed the historical data-based models in terms of predicting the travel time between current location and certain downstream bus stop (Gurmu & Fan, 2014).

2.2.6 Support vector machine model

Traditionally, many studies focus on the application of SVM to document classification and pattern recognition (Jeong & Rilett, 1999). Recently, with the application of SVM to time-series forecasting, called support vector regression (SVR) shown many breakthroughs and plausible performance, Chun-Hsin Wu et al. used SVR to predict travel time for highway users, which demonstrated that SVR was applicable to travel-time prediction and outperformed many previous methods (Wu, Ho, & Lee, 2004).

2.3 Summary of literature review

Map-matching methods can be categorized in to four groups based on the research angle. Geography-based methods are easy and efficient to be implemented. When the road network becomes complicated and diversified, the method is more likely to lose the accuracy. The accuracy of probability-based model is highly depended on the decision of confidential region, which is hard to determine, especially when navigation sensor is influenced by external factors. Topological method considers the connection within the road network. And some other novel technology for map-matching process. The problems with the existing map-matching methods are discussed in this chapter, and the solutions to these problems are the focus of the proposed method in this thesis.

Travel time prediction models are also reviewed in this chapter. Historical data model, regression model, and Kalman filter model are empirical methods the shortcomings of which is highly influenced by data quality and external factors. For instance, the accuracy of the regression model requires the independence of each variables, however, most variables are related to others in real traffic environment. The historical data model requires the stable traffic status, which is difficult to maintain in the real traffic environment. With the development of the data-collecting systems and ITS systems, numerous real-time travel time data is accessible, especially for the transit vehicles. Edmonton transit system (ETS) equips the transit vehicles with GPS transmitters and receivers, which makes the real-time locations of transit are available to both fleet managers and passengers.

CHAPTER 3 TRANSIT PROBE DATA DESCRIPTION

The data input used in following chapters contains the digital road network information and the transit GPS information. The GIS information from the digital map are processed in ArcGIS, including distance, direction, coordination, alignment, etc. Transit data is collected from GTFS public data resource.

Digital map information for map-matching mainly contains the information of the distance and direction of the road segments. Under special circumstances, like overpass, underpass and local minor roads, the recognition can be mismatched. In this thesis, the concept of reference points is introduced and generated using software to be considered as the main measurement of the GPS data to diminish the error that may happen during the matching process. The example data used in the following chapters and the relevant explanations can be found in the Fig.3.1-3.9.

The data used in this thesis was collected from the test site of Edmonton, capital city of Canadian province Alberta. Fig. 3.1 shows the digital road network within the whole city scale. GPS data contains all the routes of transit vehicles equipped with GPS transmitters and receivers, covering the arterials in the urban municipality scale.

3.1 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). The sharing information contained in GTFS is in format of a series of text files with different fields separated by comma (Google, General Transit Feed Specification Reference, 2012). GTFS data obtained from Edmonton open data is used as GPS input in this chapter. The available data contains two datasets in terms of trip update and vehicle position.

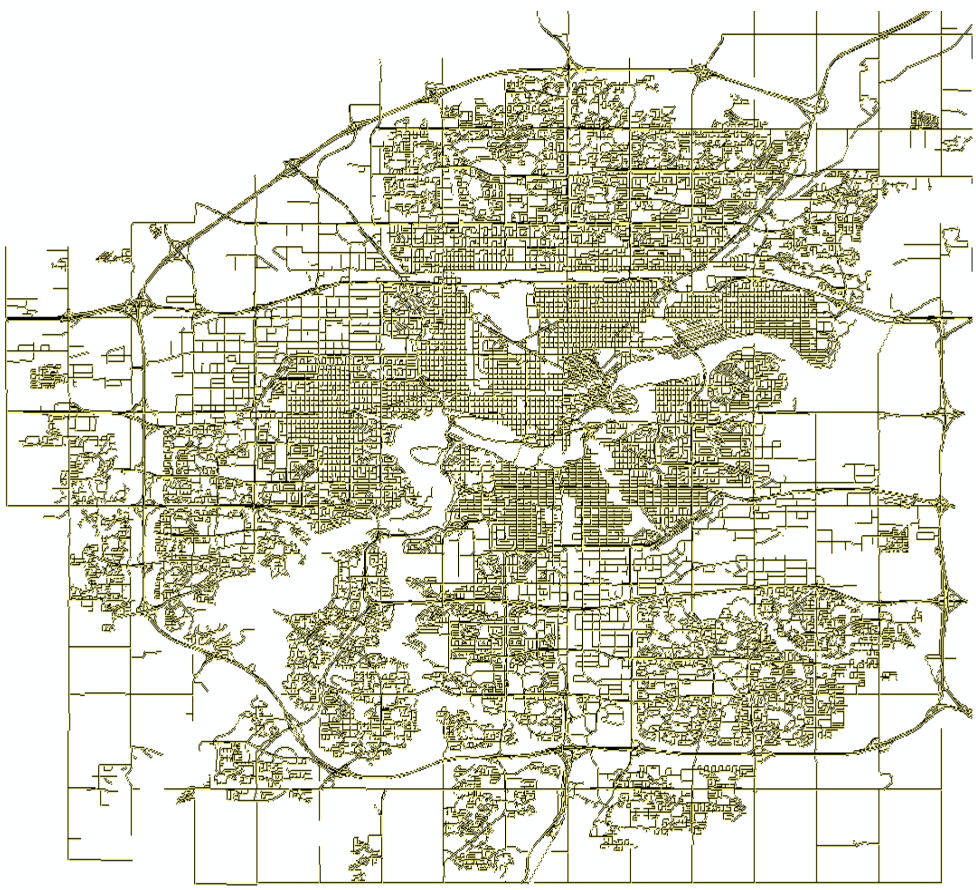

Figure 3.1 Digital Road Network_City of Edmonton (2016)

Each trip is made of a set of GPS points, which is the basic unit for map-matching. The trip update information contains the following fields like shown in Fig. 3.2. Following terminology and notification of fields are explained for further use.

TripID	RouteID	VehicleID	VehicleLabel	StartDate	StartTime	StopSequence	StopID	DepartureTime	Delay
11234897	5	2098	4160	20160509	12:54:00	28	1035	1462821240	-120
11234897	5	2098	4160	20160509	12:54:00	29	1271	1462821360	-120
11234897	5	2098	4160	20160509	12:54:00	30	1322	1462821480	-120
11234897	5	2098	4160	20160509	12:54:00	31	1336	1462821720	0
11234897	5	2098	4160	20160509	12:54:00	32	1429	1462821720	0
11234897	5	2098	4160	20160509	12:54:00	33	1256	1462821720	0
11234897	5	2098	4160	20160509	12:54:00	34	1196	1462821780	0
11234897	5	2098	4160	20160509	12:54:00	35	1393	1462821840	0
11234897	5	2098	4160	20160509	12:54:00	36	1188	1462821900	0
11234901	5	2098	4160	20160509	13:32:00	1	1328	1462822320	0

Figure 3.2 Trip Update Information City of Edmonton, 2016

- Field 1: Trip ID: A trip identification, which is a sequence of two of more stops occurring during the specific time period.
- Field 2: Route ID: The number of the bus route. Note: the route ID may not exist in current bus operation schedule.
- Field 3: Vehicle ID: A user-visible and unique identification of the vehicle, which corresponds to system-internal vehicle ID.
- Field 4: Vehicle Label: A unique identifier for transit vehicles in internal system.
- Field 5: Start Date: The scheduled start date of the trip instance. This field must be provided to disambiguate trips that are so late as to collide with a scheduled trip on a next day.
- Field 6: Start Time: The scheduled start time of the trip instance.
- Field 7: Stop Sequence: The sequence number of the stop for the trip.

- Field 8: Stop ID: A unique internal system of identification for the stop.
- Field 9: Departure Time: The time when the bus leaves the stop, formatted in POSIX time (i.e. number of seconds since January 1st 1970 00:00:00 UTC). This departure time can be either a predicted or historical one.
- Field 10: Delay: Departure time, measure by minute (60 seconds).

The vehicle position information contains the following fields like shown in Fig. 3.3. Each position stands for one GPS point containing information mainly including longitude, latitude and timestamp. Following terminology and notification of fields are explained for further use.

TripID	VehicleLabel	Timestamp	Longitude	Latitude
11234897	4160	1462821524	-113.49813	53.540924
11236250	4193	1462821534	-113.413124	53.604355
11236156	4198	1462821530	-113.49219	53.59003
11235667	4204	1462821527	-113.63037	53.51766
11236939	4207	1462821524	-113.4539	53.57024
11231212	4210	1462821529	-113.62628	53.52913
11233074	4212	1462821535	-113.56993	53.542244
11235669	4252	1462821536	-113.57814	53.5386

Figure 3.3 Vehicle Position Information City of Edmonton, 2016

- Field 1: Trip ID: In GTFS, a trip is a sequence of two of more stops occurring at a specific time. The trip ID is the unique identity of a trip.

- Field 2: Vehicle Label: A user-visible identification of the vehicle, which corresponds to system-internal vehicle ID. It is unique to the vehicle.
- Field 3: Time Stamp: Moment at which the vehicle's position was measured, formatted in POSIX time (i.e. number of seconds since January 1st 1970 00:00:00 UTC).
- Field 4: Longitude: Degrees East, in the WGS-84 coordinate system.
- Field 5: Latitude: Degrees East, in the WGS-84 coordinate system.

Transit agencies schedule the departure frequency for their transit vehicles to achieve the most usage efficiency of operation. The distinct trip counts of each weekday from May 8 to May 12 are collected to show the trend of the transit departure pattern, which indicates the demand of the passengers for the transit system. Fig. 3.4 shows the general trend of the trip counts of weekdays during one week. There are most trip information collected during 7:00 AM-8:00 AM and 3:00 PM-4:00 PM, which are rush hours in morning and evening respectively. The scheduled departure frequency is higher than usual according to the higher demand. Fig. 3.5 shows the total trip counts on 8-12 in May.

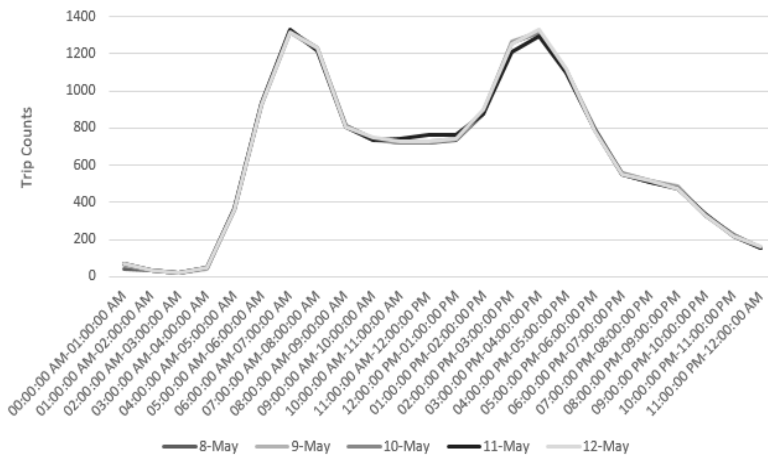

Figure 3.4 Hourly Distinct Trip_ID Counts

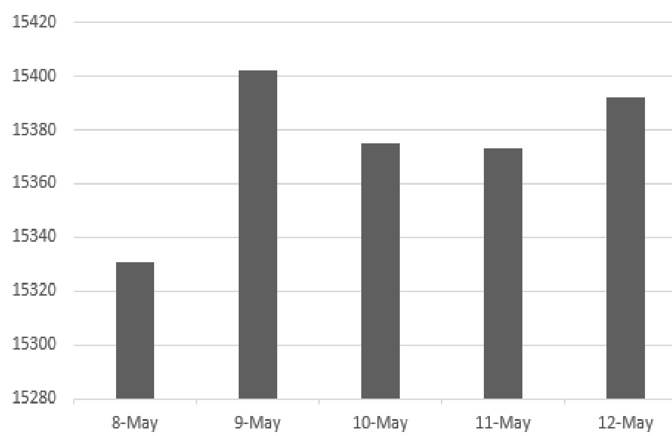

Figure 3.5 Weekday Daily Total Trip_ID Counts

3.2 Digital Road Network Data Description

The map-matching algorithm proposed in this thesis requires certain standard format for the data. GPS data used is based on General Transit Feed Specification (GTFS). This chapter will provide the description of the input data and briefly introduce the meanings and functions of the data. The network data used in this paper is based on the digital map provided by City of Edmonton (Fig. 3.1), including the information of links, reference points and grid arrangement.

3.2.1 Link Information

In the digital map, the roads are divided into virtual links at certain inflection point and traffic nodes, like intersections and roundabouts. Every link has a set of information acquired from the pre-processing of the digital map using ArcGIS (shown in Fig. 3.6). Following terminology and notification of fields are gathered to complete the information of links.

ID	FROM_N	TO_N	LENG	TrafficS	R_NAME	DIR	R_TYPE	ROAD_ID
4038	2924	2925	473.15537	0	ST ALBERT TRAIL NW	NB	Service Road	15927
4039	2925	2924	473.15537	0	ST ALBERT TRAIL NW	SB	Service Road	15927
4042	2928	2929	250.42011	0	21 STREET NE	NB	Roadway (Standard)	15930
4043	2929	2928	250.42011	0	21 STREET NE	SB	Roadway (Standard)	15930
4000	2896	2897	457.30143	0	33 STREET	NB	Roadway (Standard)	15951
4001	2897	2896	457.30143	0	33 STREET	SB	Roadway (Standard)	15951
4002	2897	1234	949.88265	0	33 STREET	NB	Roadway (Standard)	15952
4003	1234	2897	949.88265	0	33 STREET	SB	Roadway (Standard)	15952
3990	2889	2890	217.77735	0	74 AVENUE NW	WB	Roadway (Standard)	15946
3991	2890	2889	217.77735	0	74 AVENUE NW	EB	Roadway (Standard)	15946
3992	2891	2889	201.06875	0	48 STREET NW	NB	Roadway (Standard)	15947
3993	2889	2891	201.06875	0	48 STREET NW	SB	Roadway (Standard)	15947
4101	2973	2976	184.97439	0	99 STREET NW	SB	Roadway (Standard)	20893
4106	2981	2982	1287.3775	0	101 STREET SW	SB	Roadway (Standard)	20897
4107	2982	2981	1287.3775	0	101 STREET SW	NB	Roadway (Standard)	20897
4096	2973	2974	258.94671	0	101 STREET NW	NB	Roadway (Standard)	20889
4097	2974	2973	258.94671	0	101 STREET NW	EB	Roadway (Standard)	20889
4098	2973	2975	219.77752	0	99 STREET NW	NB	Roadway (Standard)	20890
4099	2975	2973	219.77752	0	99 STREET NW	SB	Roadway (Standard)	20890
4100	2976	2973	184.97439	0	99 STREET NW	NB	Roadway (Standard)	20893

Figure 3.6 Link Information (Partially)

- Field 1: ID: A unique identifier for every link.
- Field 2: FROM_N: The unique identifier for the starting node of the link, the information of which can be found in reference point information.
- Field 3: TO_N: The unique identifier for the ending node of the link.
- Field 4: LENG: The length of the link in NAD_1983 coordinate system.
- Field 5: TrafficS: Numbers of traffic signal along the link.
- Field 6: R_NAME: The road name to which the link belongs.

- Field 7: DIR: The direction of the link calculated with assumption of each link is considered as a straight line, and the direction is 0-360 degree.
- Field 8: R_TYPE: The type of the road to which the link belongs.
- Field 9: ROAD_ID: A unique identifier of the road to which the link belongs.

The link length is one of the critical parameters for the following map-matching process and travel time prediction model. The proper distance will lower the error for the matching and prediction results, since links with too long or too short length may contain the curves or other special alignment situations affecting the distribution of the reference points and the travel speed consistency in the sample data, which is the basic assumption for the prediction model.

Fig. 3.7 shows the distribution of the link length. 92.5% of the recorded links fall in the category of $[60,130]$ meters. When generating the reference points in the digital network, the link will be further divided into smaller segments.

Besides the basic information, there is also connectivity information between links like shown in Fig. 3.8. The connectivity between the adjacent links can indicate the most possible routes of the vehicles, which is path inference. Following terminology and notification of fields are gathered to complete the connectivity. For example, for the link #4038, there are three connected links, and the IDs of these three links are: #4039, #31672, #31671, with the sequence number 0, 1, and 2 respectively.

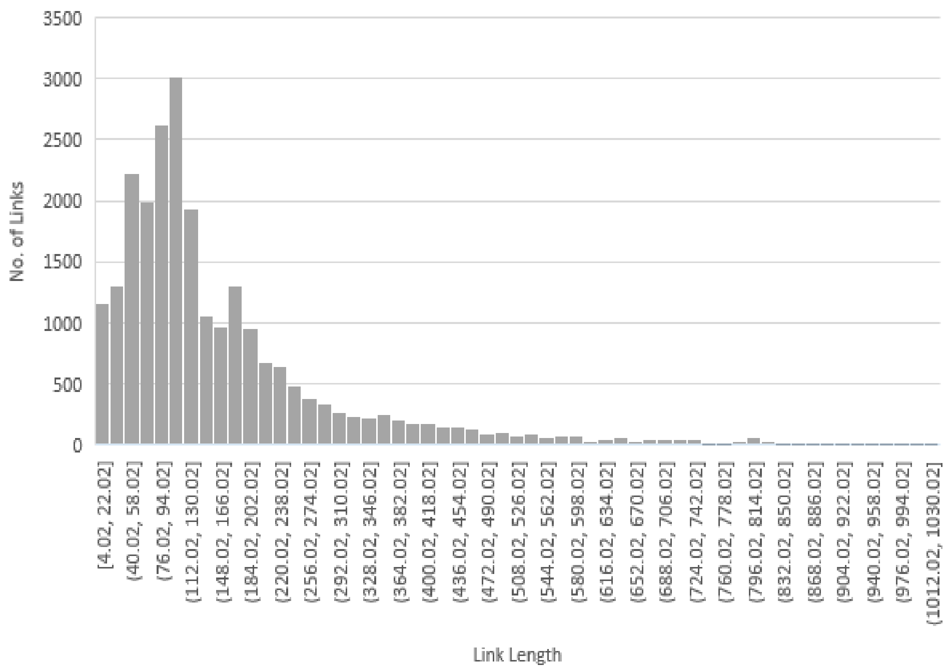

Figure 3.7 Probability Distribution of the Virtual Link Length

LINK_ID	SEQ_Link	Connected Link_ID
4038	0	4039
4038	1	31672
4038	2	31671
4039	0	4038
4039	1	25490
4039	2	25900
4042	0	4043
4042	1	30079
4042	2	30312
4043	0	4042
4043	1	28541
4000	0	4001
4000	1	4002
4001	0	4000
4001	1	40779
4001	2	31524

Figure 3.8 Information of Connected Links (Partially)

- Field 1: Link_ID: A unique identifier of every link connecting to others.
- Field 2: SEQ_Link: The sequence number of the connected link.
- Field 3: Connected Link_ID: The unique identifier of the link connected to the link in *Field* 1.

3.2.2 Reference Point Information

There are two types of reference points. One type is the starting and ending nodes of the links and the other type is added manually to the map. The added reference points separate the oversize links to the small segments with similar distance. In this paper, 20-meter is chosen as the distance between two adjacent reference points. However, due to the actual length and road alignment, the length of the segments may have slight differences.

Fig. 3.9 shows the general idea of the reference points in the digital map. The entire road network can be divided into segments with the similar length. The reference points only exist on the links covered by the transit routes while there is not reference points on the links (like collectors, local driveways) without transit route covered.

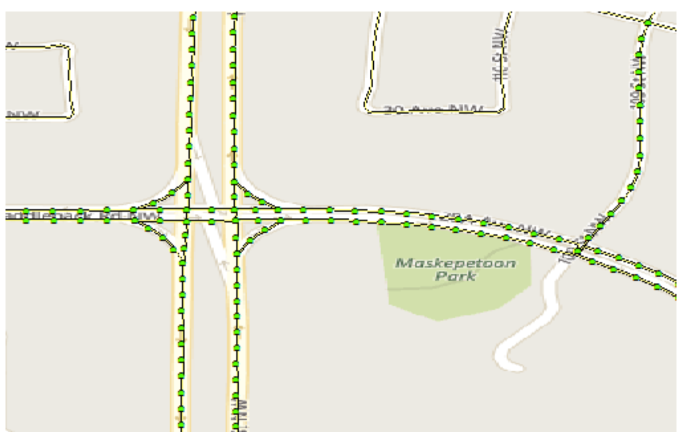

Figure 3.9 Illustration of Reference Points in Digital Map

The reference point information contains the following fields like shown in Fig. 3.10. Following terminology and notification of fields are gathered to complete the information.

Rep_ID	Link_ID	SEQ_REP	FROM_Dist	TO_Dist	COR_X	COR_Y	DIR	Long	Lat
37936	23654	1	18.41	128.93	326241	5937034	75	-113.6206066	53.53115699
37937	23654	2	36.83	110.51	326258	5937040	69	-113.6203517	53.53122326
37938	23654	3	55.25	92.09	326275	5937047	71	-113.6200958	53.53128571
37939	23654	4	73.67	73.67	326293	5937052	74	-113.6198387	53.53134317
37940	23654	5	92.09	55.25	326310	5937057	76	-113.6195793	53.53138885
37941	23654	6	110.51	36.83	326327	5937061	77	-113.6193197	53.53143339
37942	23654	7	128.93	18.41	326345	5937065	67	-113.6190596	53.53147487
37943	23654	8	147.35	0	326361	5937076	57	-113.618825	53.5315739
37945	23655	1	18.37	183.76	326068	5936932	67	-113.6231503	53.53019198
37946	23655	2	36.75	165.39	326084	5936943	56	-113.6229228	53.53029192
37947	23655	3	55.13	147.01	326099	5936954	56	-113.6226954	53.53039186
37948	23655	4	73.5	128.63	326115	5936964	57	-113.622468	53.5304918
37949	23655	5	91.88	110.26	326130	5936974	57	-113.6222384	53.5305866
37950	23655	6	110.26	91.88	326146	5936984	58	-113.6220086	53.53068106
37951	23655	7	128.63	73.5	326162	5936993	60	-113.6217769	53.53077046
37952	23655	8	147.01	55.13	326178	5937002	62	-113.6215432	53.53085459
37953	23655	9	165.39	36.75	326193	5937011	62	-113.6213082	53.53093506
37954	23655	10	183.76	18.37	326209	5937019	56	-113.621073	53.53101511
37955	23655	11	202.14	0	326224	5937032	49	-113.6208613	53.53113225

Figure 3.10 Reference Point Information (Partially)

- Field 1: REP_ID: A unique identifier for every reference point.
- Field 2: Link_ID: The identifier of the link to which the reference point belongs.
- Field 3: SEQ_REP: The sequence number of the reference point on the link.
- Field 4: FROM_Dist: The distance between the starting nodes of the links to which the reference point belongs, and the reference point.
- Field 5: TO_Dist: The distance from the reference point to the ending point of the link.
- Field 6, 7: COR_ X, Y: The coordinates of the reference point on X and Y-axis in plain coordinate system respectively.
- Field 8: DIR: The direction of the reference point range from 0-360 degree, calculated based on the tangent of the short straight line between two adjacent reference points.

- Field 9, 10: Long, Lat: The coordinates of the reference point in WGS_1984 coordinate system (degrees East/North, in the WGS-84 coordinate system).

3.2.3 Grid Arrangement

The digital map contains a large amount of information. When the GPS logs in, there will be huge amount of calculation to do to search for the matching. But map-matching process is expected to provide the matching results within short time to do the further research. To improve the efficiency of the method, the large-scale digital map is divided into small grids to downsize the database that the GPS point is programmed to match.

The map is divided into 197 × 159 grids like shown in Fig. 3.11. Following terminology and notification of fields are gathered to complete the information.

X_0	Y_0	X_1	Y_1	COL	ROW
318472	5958854	350272	5919454	197	159

Figure 3.11 Grid Arrangement Information

- Field 1: X0: The x coordination of the first range point in plain coordination system.
- Field 2: Y0: The y coordination of the first range point in plain coordination system.
- Field 3: X1: The x coordination of the second range point in plain coordination system.
- Field 4: Y1: The y coordination of the second range point in plain coordination system.

- Field 5: COL: The number of the columns.
- Field 6: ROW: The number of rows.

The grid arrangement of the Edmonton urban area is shown in Fig. 3.12. The digital map is divided into grid-based units and labeled in the geometry order. Fig. 3.13 shows the general grid arrangement in the digital map. The coordination of first and second range points is used to determine the map range.

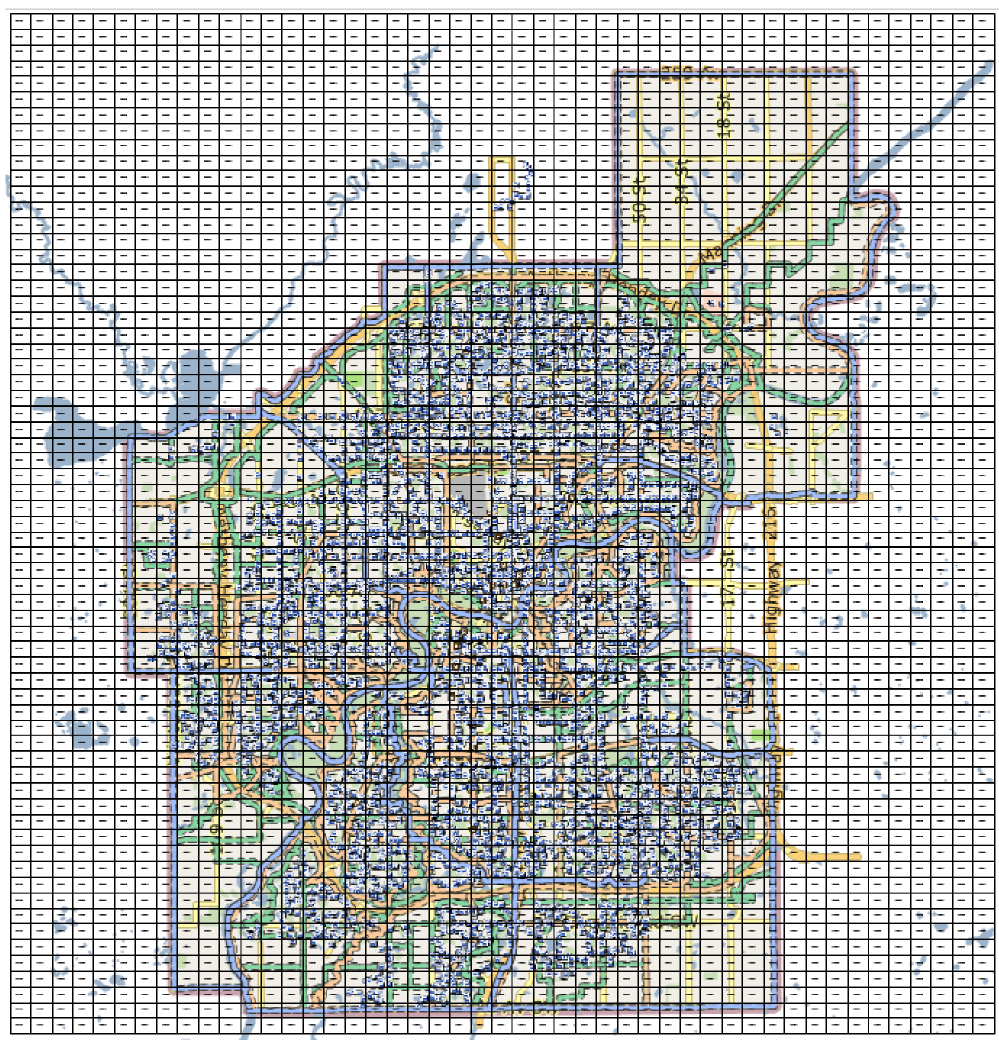

Figure 3.12 Grid Arrangement of Edmonton Urban Area

The grids are generated to narrow down the matching scale to improve the processing efficiency. The reference points that are included in each grid are formatted as the grid dataset. Fig. 3.14 shows the reference point ID contained in each grid.

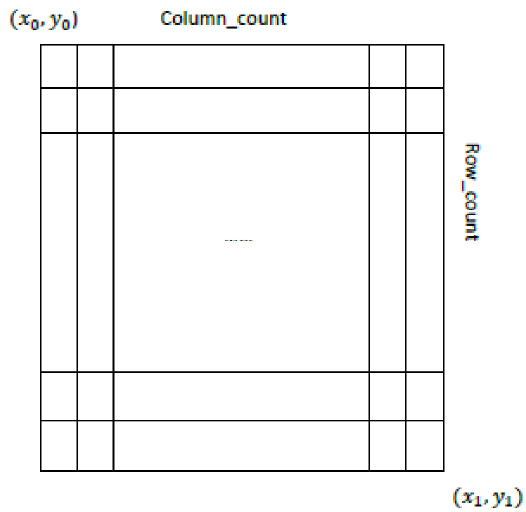

Figure 3.13 A Schema of Grid Arrangement

GRID_ID	REP_COUNT	REP_ID									
1378	20	23336	23337	23338	23339	23340	23341	23342	23343	23344	...
1379	28	23342	23343	23344	23345	23346	23347	23348	23349	23350	...
1380	28	23352	23353	23354	23355	23356	23357	23358	23359	23360	...
1381	30	23362	23363	23364	23365	23366	23367	23368	23369	23370	...
1382	46	23372	23373	23374	23375	23376	23377	23378	23379	23536	...
1383	28	24217	24218	24219	24220	24221	24222	24223	24224	24225	...
1384	28	24207	24208	24209	24210	24211	24212	24213	24214	24215	...
1385	28	24197	24198	24199	24200	24201	24202	24203	24204	24205	...
1386	28	24187	24188	24189	24190	24191	24192	24193	24194	24195	...
1387	28	24177	24178	24179	24180	24181	24182	24183	24184	24185	...
1388	28	24167	24168	24169	24170	24171	24172	24173	24174	24175	...
1389	28	24157	24158	24159	24160	24161	24162	24163	24164	24165	...
1390	40	24152	24153	24154	24155	24156	24157	24158	24159	24160	...
1391	28	168646	168647	168648	168649	168650	168651	168652	168653	168654	...
1392	28	168636	168637	168638	168639	168640	168641	168642	168643	168644	...

Figure 3.14 Reference Point Information in Grids (Partially)

- Field 1: GRID_ID: A unique identifier for each grid. Only grids containing reference points are included in the table.

- Field 2: REP_COUNT: The numbers of reference points which belong to the grid. Those 20 meters outside the grid edges are considered as “belong” as well.
- Field 3: REP_ID: The identifiers of reference points which belong to the grid with identifier shown in Field 1.

After the preparation of the map, the information of reference points within each grid is essential for the map matching. The database partially shown in Fig. 3.14 categorizes all the references by the grids. 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. 3.15, GPS point belongs to grid 1414, and is expected to match to the reference point #62239, which is not within the grid 1414. Therefore, the database shown in the Fig. 3.14 contains not only the reference points within the grid, but also those around 20 meters outside the grid. The relationship between points, grids can be found in Fig. 3.16. For example, the grid #1405 contains 6 reference points, including No. 26618, 26619, 26620, 27641, 27642, and 27643.

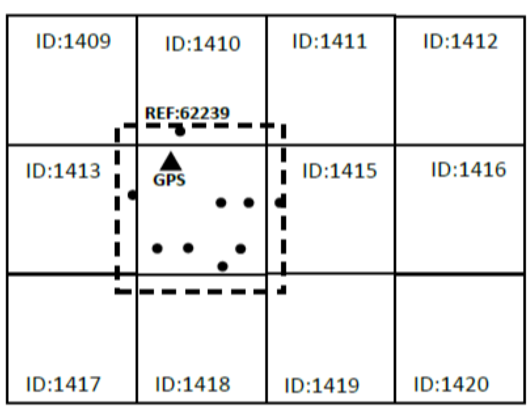

Figure 3.15 A Schema of Grid_RP

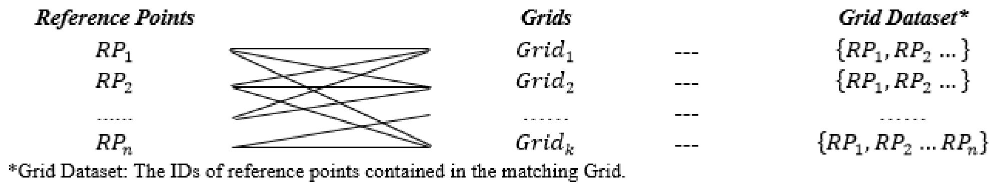

Figure 3.16 Relationship between Grids and Reference Points (RP)

The precise matching scale for each GPS point is shown like dotting line square in Fig. 3.15. The number of the reference points in every grid follow the probability distribution in the Fig. 3.17. 89.6% grids contain the even number reference points. 1470 grids contains 28 reference points in each of them and there is higher probability to have integer multiple of hundred reference points in the grid.

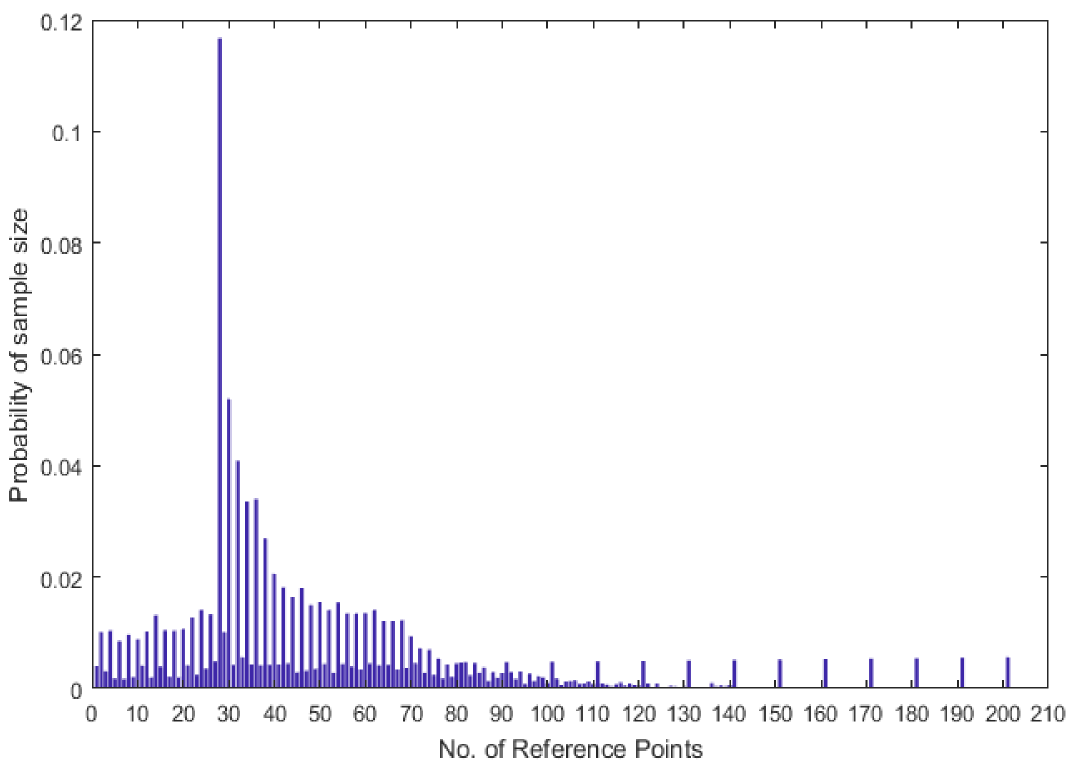

Figure 3.17 Histogram of Number of RP in Grids

3.3 Summary of Data Description

This section introduces the data that will be used in this thesis, including the road network data and GTFS data. The digital road network is a simplification of the actual streets, only keeping the main characteristics of the road condition, including centerline, direction, starting and ending point locations. Reference point file and grid file are added to the original map provided by City of Edmonton.

GTFS data is used as the GPS input in map-matching method in this thesis. Given the fact that GTFS data has low collecting sampling interval, which is around 30 seconds. The direction of the GPS points is not included in the GTFS dataset and the calculation of the angle may not be precise especially when the distance between two adjacent GPS point covers more than one link. The reference points are used as the standard to do the matching, converting point-to-curve comparison to point-to-point comparison. In this case, the calculation of the direction is no longer needed. The performance of the map-matching will be more reliable. The reliability will be evaluated in next chapters.

CHAPTER 4 REFERENCE POINT-BASED MAP MATCHING METHOD USING LOW-FREQUENCY TRANSIT DATA

Conducting the map-matching method is the first step in the prediction of the transit travel time in this thesis. 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 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 field tests based on the large amount dataset collected from the real traffic situation, providing the reliable evaluation of the method. The case study shows the matching results of the method and the comparison results with the Spatial-temporal matching method presented in previous literature. The proposed method is evaluated in terms of accuracy of path inference and the link identification. The conclusion comes to that the proposed method outperforms the ST-matching algorithm.

4.1 Link Identification Procedure

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 thesis can be divided into three parts, including database design, data extraction and projection analysis like shown in Fig. 4.1. Given the complicated road network in the real traffic situation, there would be a huge amount of calculation if we match the GPS point to every single segment. Therefore, it is important to minimize the matching scale before the projection. The algorithm is required to extract the grid information matched to the given GPS data.

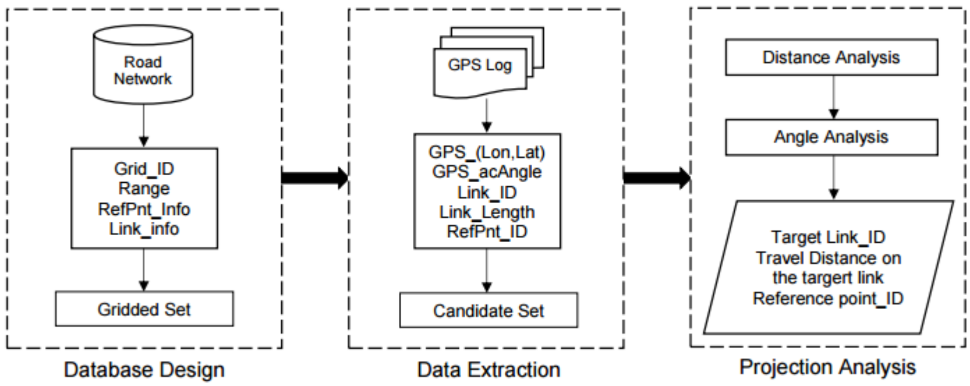

Figure 4.1 Data Flow for Link Identification Procedure

The method proposed in the thesis combines the consideration for both distance and angle matching results like shown in *projection analysis* in Fig. 4.1. Angle analysis helps calibrate the direction possibility and define the most likely candidate for the best match (Eq. 4.1).

$$\mathcal{L}\left(d_i, \vec{\theta}\right) = \alpha(\min\{d_i\}) + \beta\left(\min\left|\overrightarrow{\theta_i} - \overrightarrow{\theta_{GPS}}\right|\right) \qquad (Eq. 4.1)$$

The weights of distance and direction are expressed as α and β respectively. The goal of the theoretical function is to find the minimum distance between the reference points and the GPS point and the minimum angle differential between reference points and the GPS point. Fig. 4.1 shows the data flow in link identification. The following terminology and notification are used.

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. The set S of reference points $n_{(lon,lat)}$within the pre-defined grid N to which the GPS point $t_{(Lon,lat)}$ belongs.

$$\mathrm{S} = \{\mathrm{n}: \begin{cases} n_{lon} - Lon_N \leq \epsilon_{lon} \\ n_{lat} - Lat_N \leq \epsilon_{lat} \end{cases} \} \qquad \text{(Eq. 4.2)}$$

Lon_N, Lat_N refer to the longitude and latitude of the left and down edge of the grid to determine the locations of reference points. ϵ refers to the threshold that indicates if the reference point is within the range of the grid.

Definition 3: Target reference point: The reference point closest to the GPS point with the smallest direction differential as well. The most likely reference point, n^* is selected from the set S as the one with the smallest distance. d_j refers to the distance between the GPS point and the target reference point.

$$n^* = \{n: d_j = min\, d_n\, , n \in S\} \qquad \text{(Eq. 4.3)}$$

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

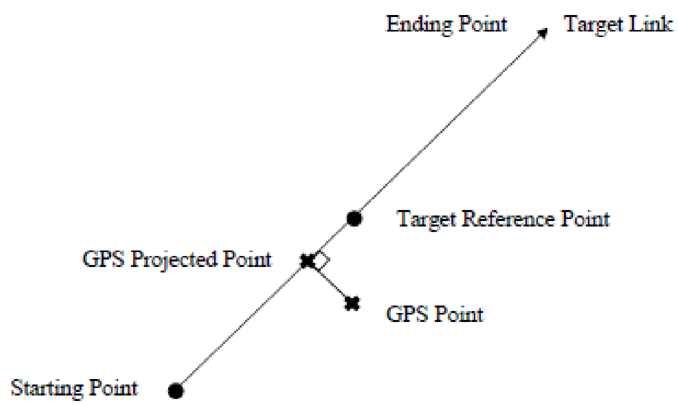

Figure 4.2 Map-Matching Result Schematic

When the GPS signal logs in, the coordination and direction of the GPS point are entered in the database to decide to which grid the GPS point belongs. The information of reference points in each grid is picked out to create the candidate set, which is called potential matching scale. Projection analysis firstly compares the point-to-point distance between the GPS point and every reference point in the grid, and then picks out three reference points closest to the GPS point.

Projection analysis secondly compares the direction of GPS point and three chosen reference points. The connection line of adjacent two high-frequency GPS data is assumed as the straight line. The direction can be calculate based on the latitude and longitude. The reference point with the smallest angle differential from the GPS point is considered as the target reference point. The target link, to which the target reference point belongs can be identified based on the candidate link set. Projection analysis finally calculates the distance between the projected point and the start point of the link like shown in Fig. 4.2.

4.1.1 Database Design

The database contains two parts of information, including the basic grid arrangement and reference point identifier within 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 within and closely around the grid are picked out and organized in the database.

4.1.2 Data Extraction

When GPS logs in, the database will automatically search for the grid ID to which the GPS point belongs, and extract all the reference points within the grid to create the candidate link set, which is the potential match scale. The process is shown in Fig. 4.3.

4.1.3 Projection Analysis

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.

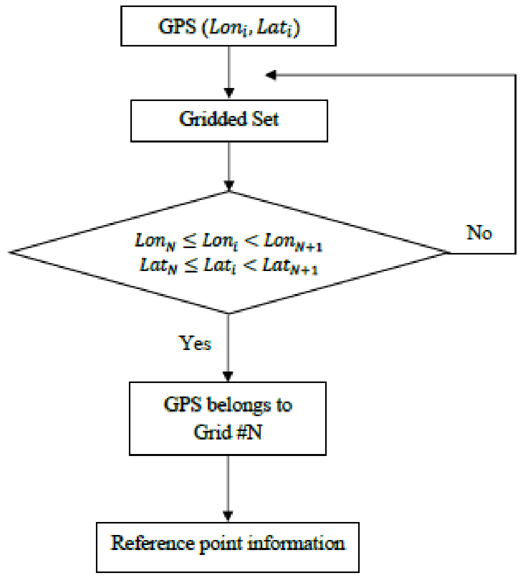

Figure 4.3 Dataflow of the Data Extraction Process

Fig. 4.4 shows the data flow of the projection analysis. Given the GPS information and the candidate link set, the first step of

the projection is comparing the distance between the GPS point and the potential reference points (PRPs), which are the reference points within the potential matching scale. The procedure will pick out three reference points with the shortest distances to the GPS point, which are defined as potential matches (PMs). Next step is to compare the angle of the GPS point and the PMs. The PM with the smallest angle differential will be the target reference point. The target reference point can be retrieved from the database to obtain the information of the target link to which the target reference point belongs, including link ID, from node ID & coordination and length. Then the procedure will calculate the distance from the starting point of the target link to the GPS projected point.

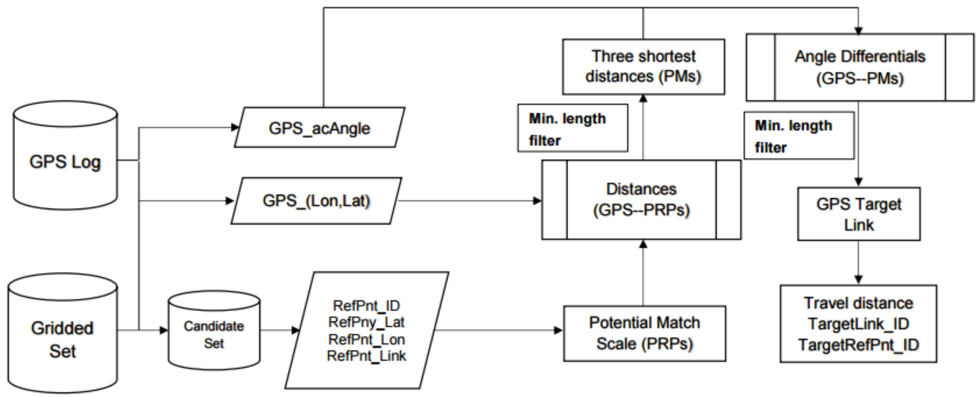

Figure 4.4 Map-Matching Projection Analysis

4.2 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. Fig. 4.5 shows the description of the path inference problem.

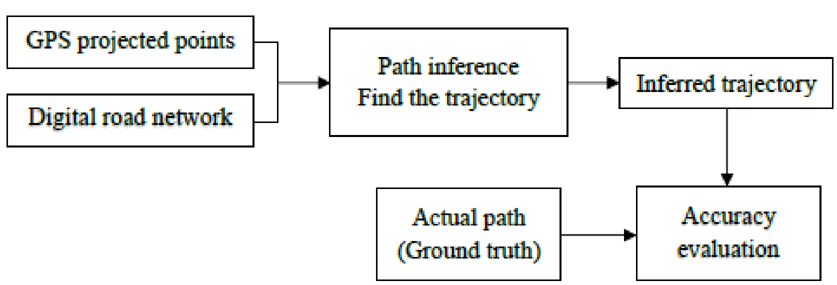

Figure 4.5 A schema of Path Inference Problem

4.2.1 Methodology

The main idea is to find the most likely trajectory the trip was made by the vehicle, which is the sequence of the links that the GPS points have been matched to, and then compare the trajectory with the ground truth recorded by other methods, like handheld GPS recorder. The method for path inference used in this paper mainly includes three steps.

1. Build the possible path.

This step is based on the projection of the GPS points and the topological information of the digital network, which is already known from link identification. The link to which the projected points (n_x) belong. The projected point $n_x(t)$ of GPS point $t_{(Lon,Lat)}$ on link x and the starting and ending points of link x (x_o, x_D respectively) are both known. The possible paths include all the accessible link combinations that connect every two adjacent projected points, like shown in Fig. 4.6(a). The paths with the consistent direction are considered as the possible paths. The paths which require the behaviors of detour and restricted accessibility are not considered. Let T be the possible path set, and t_p be the travel time of path alternative p. t_{min} is denoted as the minimal travel time, which is the ratio of Euclidean distance of the OD pair on the minimal free flow speed of the links belonging to the whole path. ε denotes the parameter related to the acceptance level of trip-makers to the travel cost. If the cost is higher, ε will be higher to contain more qualified possible paths to make sure all the choices will be considered. For example, if the congestion is severe on path #1 and #2, trip-makers may take detour path to avoid the hot spots. In this case, ε should be higher to include the detour path in the possible path set.

$$\mathrm{T} = \left\{p: t_p - t_{min} \leq \varepsilon\right\} \qquad \text{(Eq. 4.4)}$$

The restricted path (Fig. 4.6 a) contains the restricted left turn at the intersection north, therefore, the path cannot be accessed. The detour path containing the inconsistent direction should be included in the possible path set if the travel time of the path satisfy the Eq. 4.4.

Connectivity file is the main evidence to filter the possible paths. Each link can only be connected with the links within the connectivity vectors. Assume link A_i^n is connected with i links B_x^1 to B_x^i ($\boldsymbol{A_i^n} = \left\{B_x^1, \dots B_j^n, \dots B_x^i | x \in Q\right\}$) and link B_j^n is connected with j links C_x^1 to C_x^i ($\boldsymbol{B_i^n} = \left\{C_x^1, \dots C_k^n, \dots C_x^j | x \in Q\right\}$) etc. If GPS $t_0(Lon_i, Lat_i)$ is matched to target link A_i^n and $t_1(Lon_{i+1}, Lat_{i+1})$ is matched to C_k^n, the path inference should generate the target sequence of $\Theta = \left\{\boldsymbol{A_i^n}, \boldsymbol{B_j^n}, \boldsymbol{C_k^n}, \dots\right\}$.

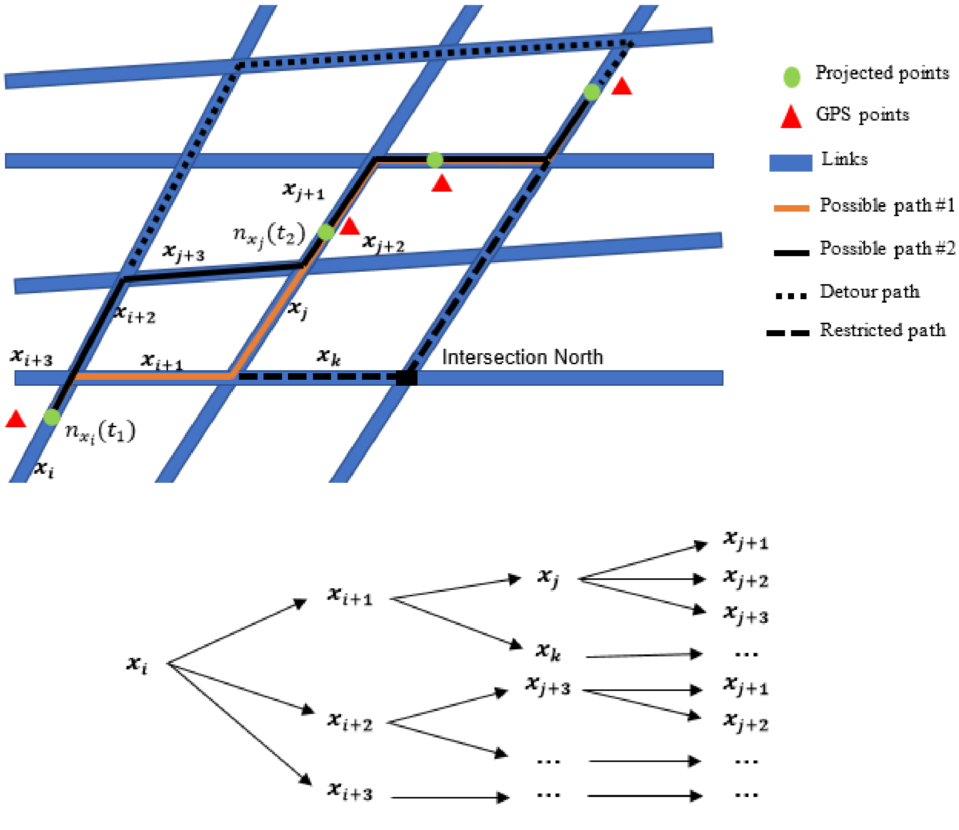

Figure 4.6 Possible trajectories of the Projected Points Sequence

Two matched results show the GPS points have been matched to link x_i and x_{j+1} respectively. Given the connectivity of the road network, all possible paths between the origin and the destination of the trip are built: $n_{x_1}(t_1) \rightarrow x_{i+1} \rightarrow x_j \rightarrow n_{x_{j+1}}(t_2)$, or $n_{x_1}(t_1) \rightarrow x_{i+2} \rightarrow x_{j+3} \rightarrow n_{x_{j+1}}(t_2)$ (i.e. possible path #1 and #2).

2. Using shortest path algorithm to calculate the costs of possible paths.

In the preliminary literature review, there are several shortest path algorithms (SPA) merging in the past ten years. The SPA used in this paper is Dijkstra' algorithm (Dijkstra, 1959). The inferred path is mainly decided by the cost of each link. The shortest path is the "cheapest" path. The cost of the shortest path from origin projected point to destination projected point is denoted as $C_{(o,d)}$.

$$C_{(O,D)} = c_{(O,t_1)} + c_{(t_1,t_2)} + \cdots + c_{(t_X,D)} \qquad \text{(Eq. 4.5)}$$

t_x is denoted as the projected point at sequence of x. There are totally X projected points in the trip t. $c_{(t_1,t_2)}$ is denoted as the cost of the first link from projected point t_1 to t_2.

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). The cost on the link x (from t_x to t_{x+1}) can be expressed as: $c_{(t_x,t_{x+1})} = \frac{\sum tt_x}{K} + k_1 \cdot \sum_{i=1}^{x} Signal_x + k_2 \cdot \sum_{i=1}^{x} turns_x$, which is the function of the travel time of the trips from t_x to t_{x+1}. K is the number of observed trips. The first item is the average travel time on the link x. The second item is the delay at the signalized intersections. k_1 depicts the delay penalties for each signalized intersection on the link x. $Signal_x$ denotes the average control delay of the ith signalized intersection. Similarly to the third item, k_2 is the delay penalties for each conflicting turn. $turns_x$ depicts the average delay for ith conflicting turn. The delay penalties can be calibrated based on the history data or traffic field survey.

3. Find the most likely trajectory.

To identify the most likely trajectory among the possible paths, first step is to settle the criteria based on the travel time on each link, travel cost (e.g. congestion time, fuel price, extra charging) and distance. The model used in this paper mainly uses the criteria of travel time and travel distance respectively based on the assumption that the trip-makers normally follow the path with the least travel time and travel distance, which is called "generalized cost" in the previous literature (Rahmani, 2013). The most likely trajectory is selected from the possible path set based on the shortest path model. Let p^* be the most likely path. d_p is the distance of the path p, which belongs to the set T.

$$p^* = \{j: d_j = \min(d_p), p \in T\} \quad \text{(Eq. 4.6)}$$

4.2.2 Special Considerations

The existing GPS capturing technologies intend to record low frequency GPS data given the limited storage space, therefore, the distance between two adjacent GPS points normally covers several links, which makes it difficult to determine which path the vehicle was on. There are three special considerations discussed in this thesis about the possible errors that may happen during the map-matching and path inference.

1. **Equal distance choice**

Based on the path inference model mentioned previously, there would be a chance that the model provides double or triple results, which shows that there are several results with the same travel distance. In this case, the trip-makers are reasonably assumed to make choices based on their own preferences. Driver's route choice model is used to conduct the path inference.

Miwa et. al. developed a driver's route choice concept (Tomio Miwa, 2012) using the developed logit model. The utility of one path is considered related with four factors including the number of signalized intersections, the number of stop signs and pedestrian walkway, road surface condition and the connectivity. The number of the traffic facilities on the path mostly affect the travel time, which is proven to be within the range (Eq. 4.3). It is assumed that the utility function shows the attraction level of each path to the trip-makers. The utility function of path p is $U_p = \alpha_0 \overline{TT} + \alpha_1 Signal_p + \alpha_2 Stop_x + \alpha_{RC} + \alpha_3 \sum_{i=i-3}^{p-1} \beta_i x_i + \epsilon_p$. The parameters α can be calibrated by regression experiments based on historical data. α_{RC} denotes the road condition based on the classification of the road, arterial (4), minor (3), collector (2) and local road (1). The connectivity is indicated by the effects of 3 previous links on the path choice like shown as the 5th term in the utility functions.

The probability of the trip-makers taking one of the model-generated path p is shown as $P_p = \exp(\theta U_p) \Big/ \sum \exp(\theta U')$. U' is denoted as the utility function of every path resulting from model Eq. 4.5. θ is the scale parameter in the Gumbel distribution which the random error term ϵ follows.

2. Overpass / Underpass

Overpass and underpass are common transportation facilities in urban environment. Such facilities are unlikely shown very clear on the digital map, therefore, the map-matching may not be able to identify the proper candidate links and target link without three-dimensional map. The GPS signal may be blocked when going through the tunnel or other underpass conditions.

3. High Density Local Road Network

The high density local road network is common in the residential neighborhoods. Most of the links in such road network are narrow, single lane in each direction or shared direction and sometimes one-way, therefore, the connectivity and restriction rules become very useful when it comes to the map-matching and path inference.

4. Transit Path Inference

This thesis is focused on the transit travel time prediction. The transit data remains the main source for the research. Given the fixed route of the transit vehicles, the path inference of the transit is easier to conduct. However, errors may occur in the map-matching process. If the target link is not identified properly, the path inference may show the route is different from the schedule, which can be used to evaluate the accuracy of the transit map-matching method.

4.3 Evaluation of Factors Affecting Performance of the Map-Matching Method

There are several factors affecting the performance of the map-matching method, including the map-matching process and path inference. The main source of the error comes from the model calibration, data source and accuracy evaluation process, and other error like mechanical failure, weather impact.

4.3.1 Indices of accuracy

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 Eq. 4.7.

$$\mathrm{CI}(\%) = \frac{Number\ of\ Correctly\ identified\ points}{Sample\ size} \times 100\% \quad \text{(Eq. 4.7)}$$

Given that 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 Eq. 4.8.

$$\text{Average distance error} = \frac{\left|d_{mp} - d_{vpp}\right|}{Sample\ size} \quad \text{(Eq. 4.8)}$$

In this thesis, vertical projection point is used as the precise location of the GPS point on the digital road network compared with the matched results to evaluate the performance of the method. 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.

The accuracy of the path inference means the ratio that the plots correctly matched to the digital road network. ARR index (accuracy ratio of matched) is the most commonly used (Tomio Miwa, 2012). The evaluation of the path inference is based on the following criteria.

$$\mathrm{ARR} = \frac{length\ of\ correctly\ matched\ path}{length\ of\ actual\ path} \qquad (\mathrm{Eq.\,4.9})$$

However, ARR cannot express all the error it may have in the method. Like shown in Fig. 4.7, when the length of AB become small enough, the ARR will become closer to 1, which indicates the method becomes more and more acceptable. ARR is adjusted to PFI (proportion of false identification) and IAR (inaccurate length of matched) is defined to ensure the evaluation is valid.

$$\mathrm{PFI} = \frac{False\ identified\ link\ number}{Total\ identified\ link\ number} \times 100\% \qquad (\mathrm{Eq.\,4.10})$$

$$\mathrm{IAR} = \frac{length\ of\ wrong\ matched\ path}{length\ of\ matched\ path} \qquad (\mathrm{Eq.\,4.11})$$

The accuracy of map-matching process is measured by correct link identification, which indicates how many GPS points have been correctly matched to the target links, compared with the high-frequency data as the ground truth. For the transit data, the route of transit vehicles has been predetermined, therefore, the route can be considered as the ground truth to conduct the accuracy evaluation.

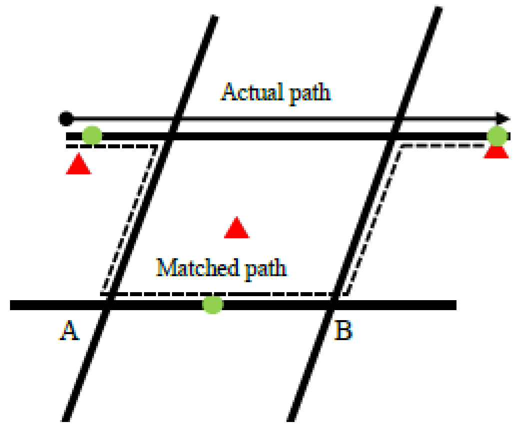

Figure 4.7 Schema of ARR Error

4.3.2 GPS signal and movement criterion

The accuracy of the prediction results highly depends on the accuracy of GPS signal. Fig. 4.8 shows that if two adjacent GPS points have distance of 25 meters, there will be probability of 95% that the transit vehicle is on movement. The results are based on the existing GPS data of transit vehicles and field test.

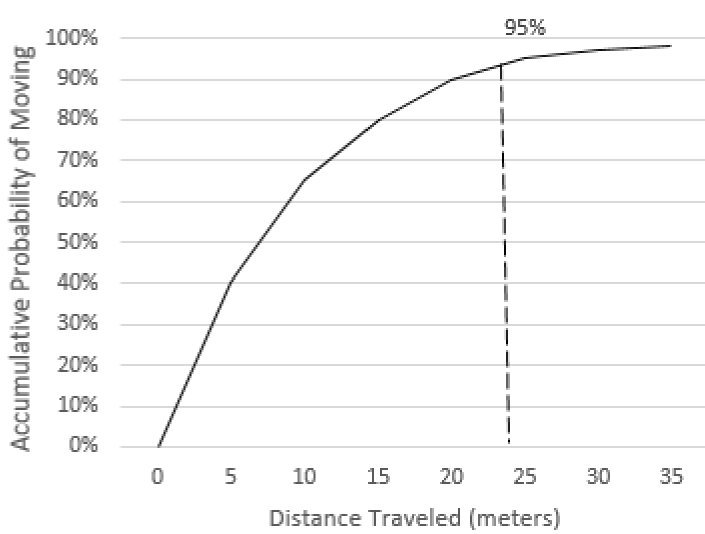

Figure 4.8 GPS Error Distribution

Normally, the GPS signal is accurate from 60 to 300 feet (Bajaj, Ranaweera, & Agrawal, 2002). One of the major factors that affecting the accuracy is the propagation delay caused by the radio signal speed changing in different atmospheric particles. When the signal bounces among the objects on the ground, there will be fading of the signal strength. Other factors can also affect the accuracy of the GPS signal, like receiver noise, unreliable distance measurements between satellites the GPS receiver is connected with.

4.3.3 External factors

Other external factors may affect the accuracy of the map-matching method. Weather condition will have impact on road condition and may lead to the change of the drivers' behaviors. Trip-makers are more likely to take a longer distance detour to avoid the congested path resulting from the weather. The choice model of the trip-makers may vary by changing the utility functions. Moreover, extreme weather may also affect the GPS signal, the error distribution of the GPS data may change and result in the errors.

Road condition also has the impact on the choice model by changing the utility functions. The temporary road construction may increase the access restriction. Sometimes, not all the information in the digital map will be updated in time when conducting the map-matching method.

Many trips are completed by multi transportation modes, like park & ride, cycle & transit, walk & transit, etc. When the transportation modes are changed, the GPS data will lose the consistency at the transfer point. Error may happen due to this reason.

4.4 Case Study

To validate the map-matching method and path inference model, a case study based on the GPS data of real trips and digital network data is conducted.

4.4.1 Test Site Description

23 Ave. is an essential arterial located in the south of Edmonton. 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 trips in eastbound.

4.4.2 Results and Discussion

The data was collected from transit vehicles operating on 23rd Avenue equipped with GPS receivers by City of Edmonton to evaluate the performance of the map-matching method proposed in this thesis. 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. The transit route map was retrieved from the city website as the ground truth for path inference. 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 4.1.

Table 4.1 Field Trips Used in Case Study

Trip ID	From Time	From Station	Vehicle ID	Route ID	Sample Size
11414576	13:43:28 Jul.28	CPTC	4673	36	666
11414614	14:05:49 Jul.28	LTC	4362	36	558
11414153	14:31:39 Jul.28	CPTC	4840	23	628
11414616	15:01:10 Jul.28	LTC	4346	36	702
11414580	15:18:04 Jul.28	CPTC	4346	36	686
11414172	15:46:13 Jul.28	LTC	4626	23	720
11414156	16:06:48 Jul.28	CPTC	6003	23	660

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

The snapshot of the matching results can be found in Fig. 4.9. The case shows the matching result in Century Park transit center in South Edmonton. The trip 11414156 starts from Century Park Station at 16:06:48 pm on July 28, 2016. The route is known from the trip update data. The road network in such transit center includes minor roads and one-way roads of high-density. GPS points are more likely to be matched incorrectly under such circumstances.

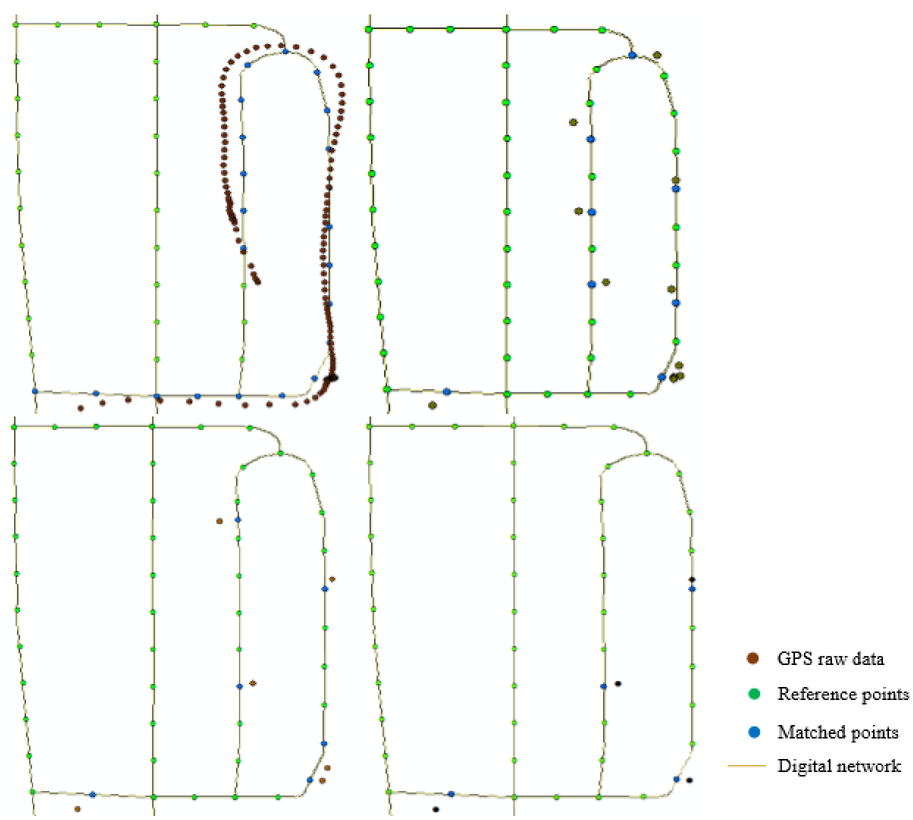

Figure 4.9 Match Results w/ Different Sampling Intervals

The accuracy results are shown in Table 4.2. There are totally 4620 GPS points collected from the field tests. Based on the map-matching method proposed in this thesis, the correctly matched points are 4609 out of 4620, 99.8% correct identification. Reducing the data sampling interval to 15 seconds and 30 seconds, the matching results still provides over 99.1% correct identification. Since the reference points have the fixed location, the high sampling interval GPS points within the small range will be clustered to one reference point. In this case, there is distance error between the matched reference point position and the vertical projected position like shown in Fig. 4.10.

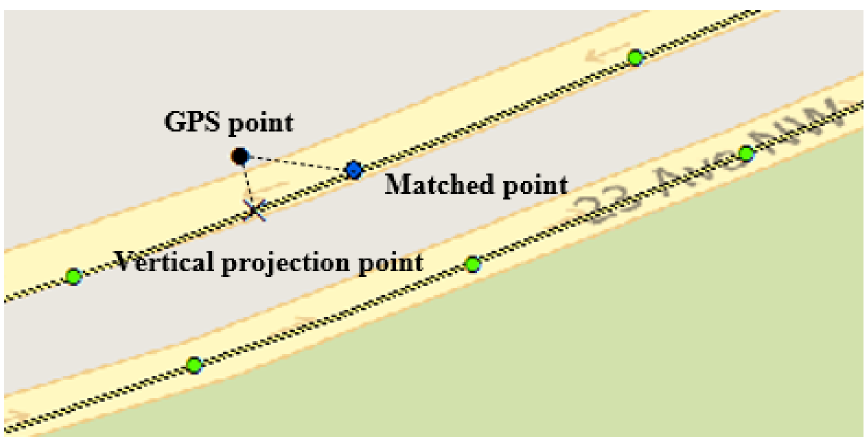

Figure 4.10 Average Distance Error

Table 4.2 shows that higher-sampling interval dataset provides more average distance error. This is because the reference points have the fixed locations and all the GPS points are projected to these reference points no matter if it is the vertical projection point. If one trip contains more GPS points, it is more likely to create larger distance error, like the occasion shown in Fig. 4.10.

Table 4.2 Performance of Map-Matching Method with Different Frequency GPS Data

Variables	Frequency of GPS data			
	1s	15s	30s (GTFS data)	60s
Sample size	4620	315	161	84
Correct link identification	4609	312	160	84
Average distance error	4.67 m	4.02 m	3.97 m	3.32 m
Processing time	< 25 s	< 20 s	< 15 s	< 15 s

Fig. 4.11 shows the snapshot of partial path inference results. The sequence of the segments to which the GPS point set belongs (trip 11414156 with sampling interval of approximately 30 seconds) is obtained from the map-matching process. Fig. 4.11 (b) shows that there are two matching errors with the possible path compared with the ground truth (Fig. 4.11 (a)). The connectivity relationship between adjacent two target links is interfered at two marked sites. Due to the mismatched results, the path inference will provide the different match results from the ground truth (Fig. 4.11(c)).

If the adjacent target links are connected, the false identified links will be excluded from the path inference (Fig. 4.12). The right out way is included in the map-matching results. However, link 22181 and 70414 are the continuous route, therefore, the link 22188 is excluded from the path inference results. The datasets of 14 trips (July 28, 2016), including trips in Table 4.1 with different sampling intervals are conducted for the performance evaluation and comparison of the path inference process. The accuracy of the results is evaluated in terms of IARR and PFI value based on the datasets of different sampling intervals (Fig. 4.13).

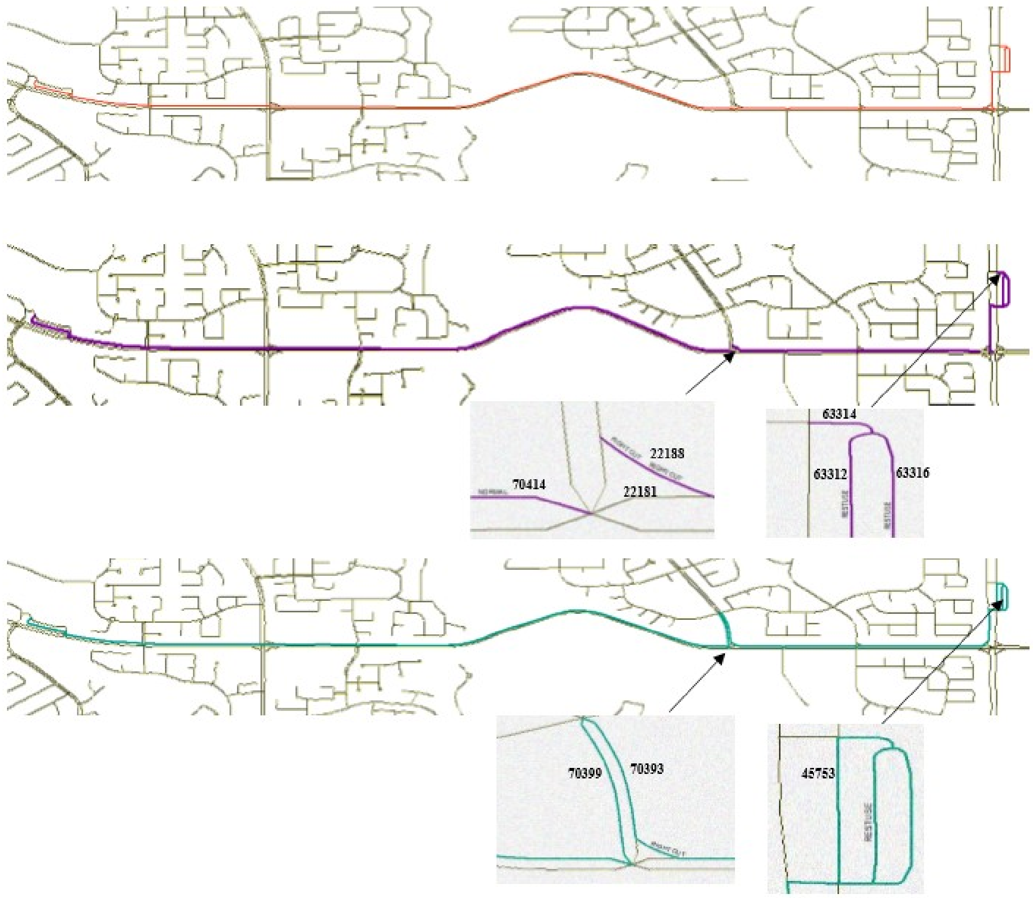

Figure 4.11 Path Inference (a) Ground truth; (b) Map-matching results (Identified link sequence); (c) Path inference results

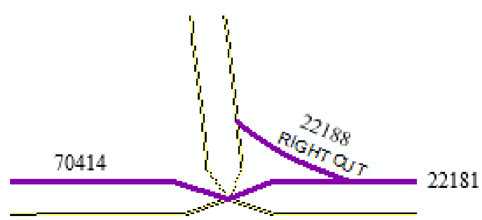

Figure 4.12 A Scheme of Path Inference Results

Fig. 4.13 shows the average IARR and false identification percentage of the datasets of different sampling intervals. The comparison shows the ascending trend of both measurements, indicating that if the sampling interval is smaller, the accuracy of the path inference is higher. Since the GTFS data consists 0.3% missing and invalid data, the false identification is slightly higher than 60-second dataset by 0.4%. Besides the inaccuracies, all the trips found in the dataset can identify the correct route by 83.5%.

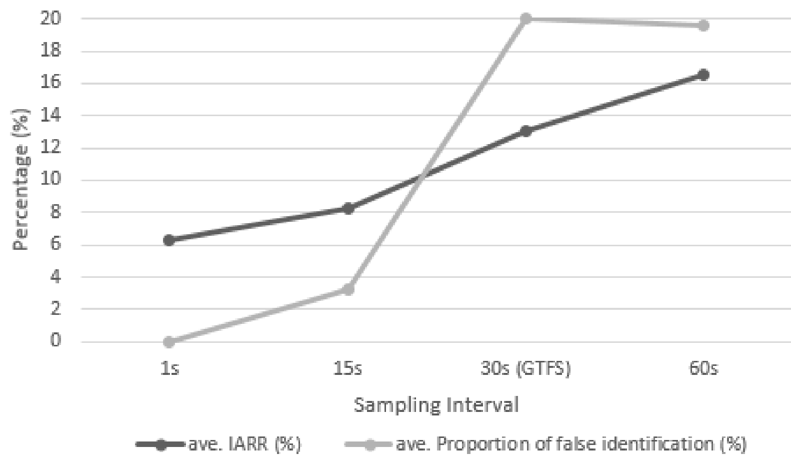

Figure 4.13 Performance of Path Inference Method

4.4.3 Comparison with ST-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 parameter and speed parameter are used to calibrate the probability to find the most likely candidate sequence.

Fig. 4.14 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.

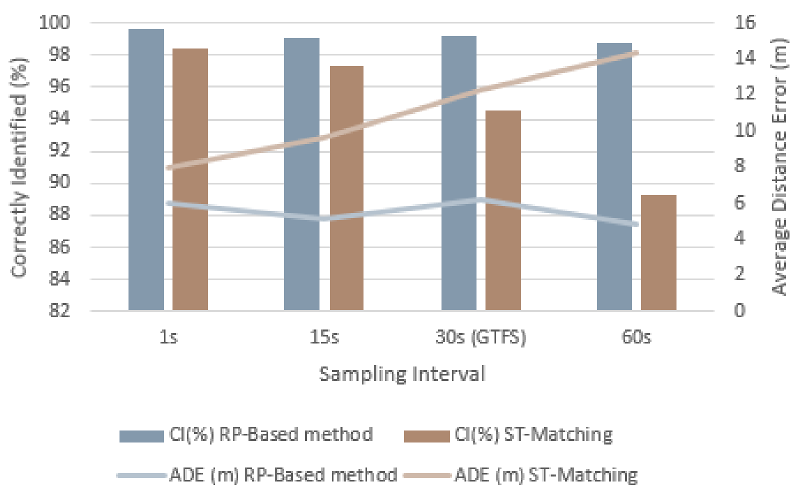

Figure 4.14 Performance Comparison w.r.t. sampling interval

To evaluate the performance of the proposed method in different traffic conditions, five locations are chosen, including transit centers, CBDs in downtown area, residential communities, highway and urban arterials. Fig. 4.15(a) shows the types of test sites and valid GPS data at each location. The matching accuracy is shown in Fig. 4.15(b).

	Location	Description	Valid GPS counts
1.	**Century park transit center**	Transit center	4620
2.	**City center**	Downtown CBD	4237
3.	**Calgary trail & 41 Ave.**	Highway ramp	3906
4.	**Calgary trail & Whitemud**	Urban arterial	4492
5.	**Blue Quill community**	Residential community	4023

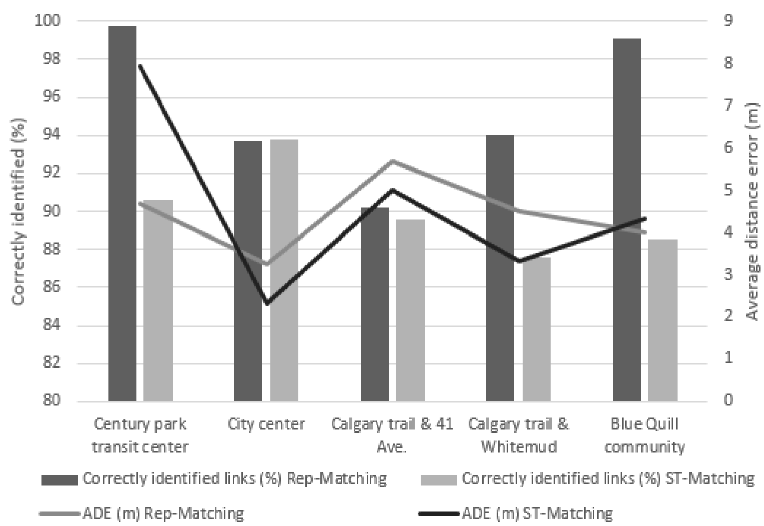

Figure 4.15 (a) Description of Test Sites; (b) Performance of Map-Matching Method with Different Link Sets

Fig. 4.15 shows that the performance of the proposed method works better for the complicated high-density road network like local communities and transit centers. According to the field tests, there are entries to the underground parking within the city center area and overpass/underpass on the urban arterial and highway ramp, which impact the accuracy of the method. However, Rep-based method still outperforms St-matching method in most cases by 5.4% on average. ST-matching method has better performance in terms of average distance error in

City center, highway ramp, and urban arterial. Because the matching process starts with the identification of the vertical projected candidate points on the links which are within its match scale. Rep-based method works better in residential communities and transit centers by 3.83 meters.

The evaluation shows the rep-based method is more suitable for the areas with high-density road network like residential communities and transit centers, because the high-density reference point files can help exclude the interference of the near links. ST-matching method is more suitable for the complicated road network like overpass, underpass and roundabouts. The temporal analysis can help find the most suitable link in the partially overlapped digital map.

The path inference is compared with ST-matching method as well in terms of IARR and PFI. The comparison results are shown in Fig. 4.16.

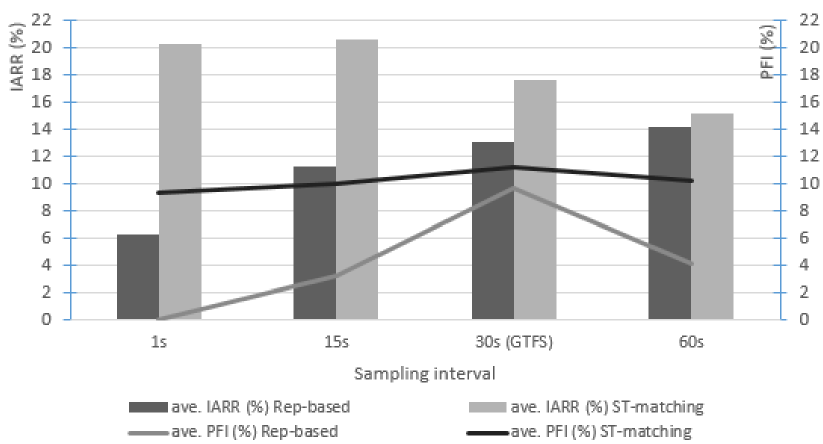

Figure 4.16 Accuracy of Path Inference Comparison w.r.t. Sampling Interval

IARR of ST-matching has the descending trend with the higher interval of the sample, while the rep-based method is increasing. The trends of IARR intends to be stable for both methods when the sampling interval becomes higher than 60 seconds, which indicates both methods perform well with high sampling interval data. The PFI shows the stable trend of ST-matching method. When the sampling interval turns to 30-second, PFI value of rep-based method is higher than other sampling intervals, which is because GTFS dataset has more missing or low-qualified data than the dataset recorded by handheld-devices. In general, the rep-based method generates more accurate results than ST-matching method for the same sampling interval. According to the previous literatures, the performance of ST-matching is better with frequency of 120 second or lower, however, with the development of the collection and storage technique, data with sampling interval of 60 seconds or lower is more popular for the future research.

4.5 Summary

In this chapter, a developed reference point-based map-matching method is introduced, including map-matching process and path inference. The development is that the reference point files is created and included in the digital network artificially. The reference points make it quite effective and accurate to match the GPS points to the correct target links, especially when the sampling interval of GPS records is high and the quality of the data source is limited. The case study uses the GTFS dataset provided by City of Edmonton. The matching results show the method provides over 99.1% of correct link identification and the average distance error is lower than 5 meters, which indicates the accuracy is acceptable compared with traditional ST-matching method, which is well-recognized in the previous literature (Rahmani, 2013). It is possible to reduce the distance error if increasing the density of the reference points in the digital network in the future research. For the path inference, the rep-based method is more suitable for the data with sampling interval of 30-second or lower.

CHAPTER 5 INDIVIDUAL BUS TRAJECTORY PREDICTION MODEL

Map- matching results can construct the GPS sequence on the digital network, therefore, the vehicle trajectory can be matched to the road network in ArcGIS, which is the foundation for most trajectory-based transportation applications. An application of the map-matching results in single trajectory prediction model is introduced in this chapter.

Transit trajectory prediction is the useful information for both transit users to plan the trip and for the transit agencies to make the most efficient schedule. With the provision of the accurate travel time prediction, the transit users can take advantage of the transit system in the most efficient way and reduce their waiting time at the stops or stations. Transit trajectory prediction model used in this thesis is based on known trajectory of transit vehicles that is the results of map-matching method. The algorithm based on the timestamps and locations of GPS points can generate the arrival time of the transit vehicles at stops and major intersections. Data of 1-second sampling interval is collected from field tests used as the ground truth to evaluate the accuracy of the prediction.

5.1 Transit Vehicle Trajectory Reconstruction

The GPS points illustrate the route of the transit operation. The timestamp and distance information are available. However, the key point to the travel time prediction is the delay at the specific traffic nodes, which mainly include intersections and stops. Signalized intersections and uncontrolled intersection (i.e. stop-sign control, yield control, non-control, etc.) will create the control delay and stop delay. Delay at the stops will be related to the time for the passengers alighting and boarding. This section focuses on modeling the arrival time at specific traffic nodes along the transit trajectories. The time-space diagram is shown like Fig. 5.1.

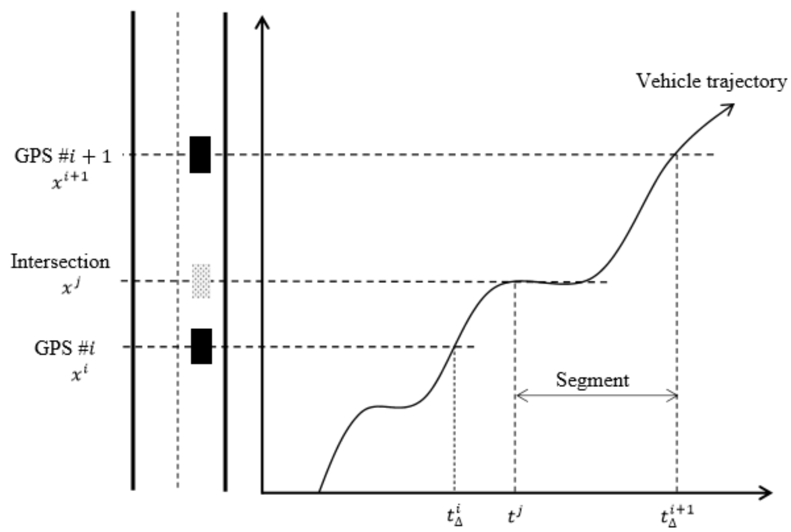

Figure 5.1 Vehicle Trajectory Reconstruction

5.2 Sample Data Integration

Sample data is defined as the dataset connecting the road network with the traffic condition in this thesis. Given the geometric network structure, it is assumed that the traffic parameters (travel speed, density) on single link are constant. The sample data in this thesis is used to describe the traffic condition on each link, including the arrival time at stops and intersections, travel time on each segment and estimated travel speed. The sample data is the key component of travel time prediction, which is also the focus of the trajectory reconstruction.

The speeds at the intersections and stops are estimated based on the linear trajectory algorithm. The reconstructed trajectory is truncated to pieces at intersections, stops and GPS points.

Given that the transit vehicles are more likely to remain the fixed route and speed when there is no disturb during the operation, it is reasonable to assume that the speed between adjacent GPS point and traffic node is constant.

The timestamps of two adjacent, ith and $(i+1)$th, GPS points are expressed as t_Δ^i and t_Δ^{i+1}, Δ is denoted as the time interval between two adjacent GPS points. The distances between the GPS points and the starting point of the trajectory are denoted as x^i and x^{i+1} respectively. $x^j(t)$ denotes as the distance between the traffic node j and the starting point of the corridor at time t. If there is a traffic node in between these two GPS points like shown in Fig. 5.2, the arrival time at intersection #1 $t^j(x)$ can be described as Eq. 5.1.

$$t^j(x) = t_\Delta^i + \frac{x^j(t) - x^i}{x^{i+1} - x^i} \cdot \left(t_\Delta^{i+1} - t_\Delta^i\right) \qquad (Eq. 5.1)$$

$x^j(t)$ denotes as the distance between the traffic node j and the starting point of the trajectory. The portion of the vehicle traveled is the same as the travel time portion on the piecewise trajectory. Therefore, the travel time on segments i_j and $j\_(i+1)$ can be expressed like Eq. 5.2-5.3 respectively.

$$TT_{i\_j} = t^j(x) - t_\Delta^i \qquad (Eq. 5.2)$$

$$TT_{j\_(i+1)} = t_{\Delta}^{i+1} - t^{j}(x) \qquad (Eq. 5.3)$$

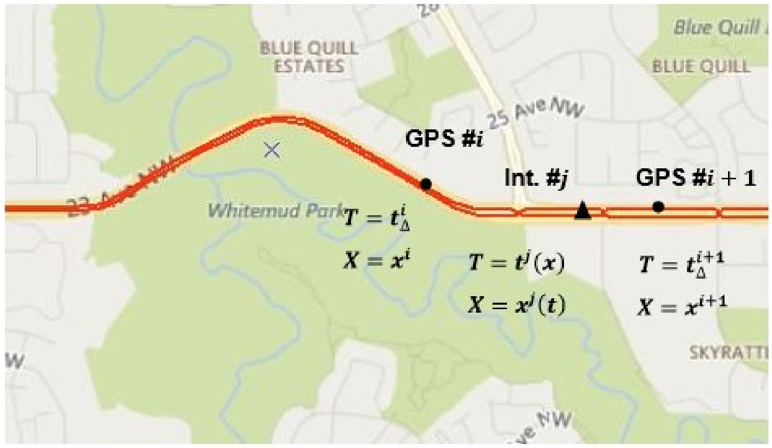

Figure 5.2 Piecewise Trajectory Reconstruction (Map-Matching Results)

If there are more than one traffic nodes between two adjacent GPS points, the first traffic node will be treated as the first estimator for the second traffic node, like shown in Fig. 5.3.

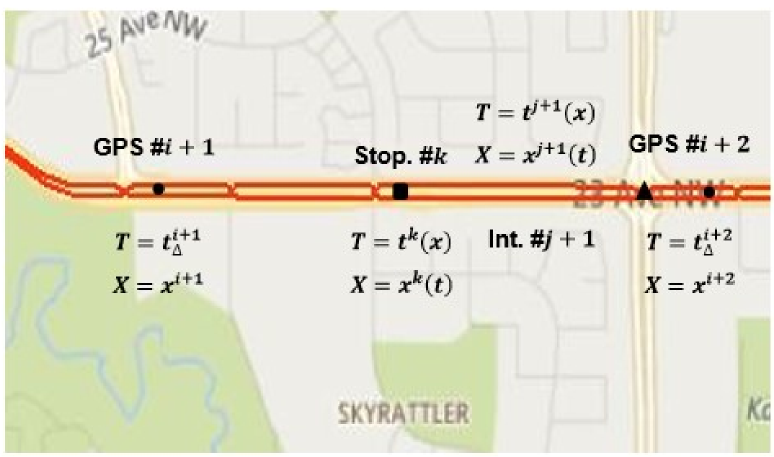

Figure 5.3 Piecewise Trajectory Reconstruction - Multi-nodes with Low Frequency Data

In Fig. 5.3, there is one stop k and one intersection $(j+1)$ between GPS $(i+1)$ and $(i+2)$. The arrival time at the stop k can be described using Eq. 5.4.

$$t^{k}(x) = t_{\Delta}^{i+1} + \frac{x^{k}(t) - x^{i+1}}{x^{i+2} - x^{i+1}} \cdot \left(t_{\Delta}^{i+2} - t_{\Delta}^{i+1}\right) \qquad \text{(Eq. 5.4)}$$

For intersection No. $(j+1)$, the estimation of the arrival time will be based on the arrival time at stop k and GPS No. $(i+2)$.

$$t^{j+1}(x) = t^k(x) + \frac{x^{j+1}(t) - x^k(t)}{x^{i+2} - x^k(t)} \cdot \left(t_\Delta^{i+2} - t^k(x)\right) \quad \text{(Eq. 5.5)}$$

The travel time on these three segments can be expressed as equation set 5.6.

$$\begin{cases} TT_{(i+1)_k} = t^k(x) - t_\Delta^{i+1} \\ TT_{k_{j+1}} = t^{j+1}(x) - t^k(x) \\ TT_{(j+1)_{i+2}} = t_\Delta^{i+2} - t^{j+1}(x) \end{cases} \quad \text{(Eq. 5.6)}$$

5.3 Individual Trajectory Prediction Model

Multi-interval prediction concept was first proposed by Chang et. al. in 2010. (Chang, Park, Lee, Lee, & Baek, 2010). The predicted results contain series of travel time of the future trips. The multi-interval travel time prediction model is applied for single trajectory prediction in this thesis. The trips happen in five weekdays are used to conduct the multi-interval prediction model. The departure frequency of the transit vehicles can be retrieved from the schedule on city website. The travel time of the about-to-departure bus at present on the first link can be predicted based on the historical data of previous buses scheduled at the same departure time in previous days. According to the basic principle of the model, the weight of the more recent day is supposed to have more positive impact on the accurate prediction. Since the traffic condition is similar during the short term, the travel time of the previous trips in the same day is considered as the current data. And the historical data is used to calibrate the current data. The historical travel time can provide the general pattern of the travel time during multi time interval. It is reasonable to assume that the road condition and traffic situation on the same link are similar during the continuously time period. The main data resource is historical transit data during weekdays and the high-frequency field data collected by the handheld GPS device. Fig. 5.4 shows the concept of the prediction model. The current trip is about to leave at present, and completed trips happen in the previous time with different time interval from the current trip. The future trip is to be predicted based on both current trips and completed trips.

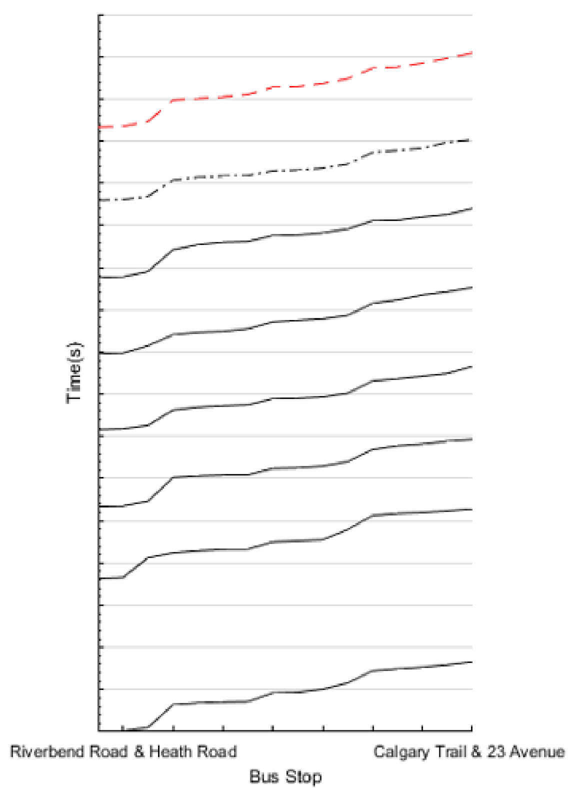

Figure 5.4 Temporal-Spatial Diagram of 23 Ave. on May 12.

There are three elements in the model, including current data, historical data, and weight measurement. The current data is the travel time information of the previous trips happen at the same day. Historical data is the completed trips of the vehicles departure at the same time in the previous days, in which the traffic situation is considered similar. The historical data is used to calibrate the current data. Weight measurement is used to identify the impact of historical data within multi time interval on the current data.

The departure time of the transit vehicles is fixed during the time period T with interval of f. The number of vehicles departure during T in one day is $T/_f$. To predict the travel time of the future trip on link p happen in day n, the historical data in day $[1,2,\dots,(n-1)]$ on link p is shown in Table. 5.1.

Table 5.1 Historical Data Integration

Day (i)	*Time period*	*Departure time*	*Trip ID*	*Travel time*
1	T_1^p	$\begin{cases} f_1^1 \\ f_1^2 \\ \dots \\ f_1^{T/f} \end{cases}$	V_1^1, V_1^2, ..., $V_1^{T/f}$	t_1^1, t_1^2, ..., $t_1^{T/f}$
2	T_2^p	$\begin{cases} f_2^1 \\ f_2^2 \\ \dots \\ f_2^{T/f} \end{cases}$	V_2^1, V_2^2, ..., $V_2^{T/f}$	t_2^1, t_2^2, ..., $t_2^{T/f}$
...	...	...	...	...
$n-1$	T_{n-1}^p	$\begin{cases} f_{n-1}^1 \\ f_{n-1}^2 \\ \dots \\ f_{n-1}^{T/f} \end{cases}$	V_{n-1}^1, V_{n-1}^2, ..., $V_{n-1}^{T/f}$	t_{n-1}^1, t_{n-1}^2, ..., $t_{n-1}^{T/f}$

The travel time data of current trips on link p on day n has the similar format to the historical data like the vector shown in Eq. 5.7.

$$\left[t_n^1, t_n^2, \dots, t_n^{T/f}\right] \qquad (Eq. 5.7)$$

The matched departure time can be shown like in Eq. 5.8.

$$\left[f_n^1, f_n^2, \dots, f_n^{T/f}\right] \qquad (Eq. 5.8)$$

The historical and current travel time data can be obtained from *sample data integration* process mentioned in last section.

The weight measurement is based on the Euclidean distance, which is a virtual distance between the current status and the historical status of different days respectively, representing the relevance between two statuses. It is reasonable to assume that if the historical status is closer to the current status, more weight should be given to this historical status. The distance is computed as Eq. 5.9.

$$D_i^p = \sqrt{(t_n^1 - t_i^1)^2 + \cdots + \left(t_n^{T/f} - t_i^{T/f}\right)^2} \qquad (Eq. 5.9)$$

Average value of the historical data is used as the prediction base, which is calibrated by three elements mentioned above. A parameter ω_t stands for the time series is used to calibrate the impacts of historical data on current status.

$$\overline{t_i^p} = \frac{\sum_{j=1}^{T/f}(\omega_t \cdot t_i^j)}{T/f} \qquad (Eq. 5.10)$$

The inverse of the distance input is used in the forecast function as the weight measurements, which has been evident to outperform the direct average of the dependent variables (Smith, Williams, & Oswald, 2002). The predicted travel time of vehicle $(T/f+1)$ in day *n* on link *p* can be expressed by Eq. 5.11.

$$t_n^{T/f+1} = \frac{\sum_{i=1}^{n-1}\left(\overline{t_i^p}\Big/D_i^p\right)}{\sum_{i=1}^{n-1}\left(1\Big/D_i^p\right)} \qquad (Eq. 5.11)$$

5.4 Case Study

A case study is conducted to evaluate the validation and accuracy of the travel time prediction method. The tested site is 23rd Avenue in South Edmonton. The corridor from Terwillegar Drive (in the west) to the Gateway Boulevard (in the east) is an essential arterial covered by several major transit route.

5.4.1 Tested Corridor Description

According to the GIS data, there are 15 signalized intersections and 16 stops along the eastbound 23rd Avenue. Among the stops, there are two transit centers including Century Park transit center which is located off the 23 avenue, and Legar transit center. The specific locations of stops and intersections in WGS-84 coordinate system and the distances in between are shown in Table 5.2-5.5.

In TABLE 5.2, the field of INT_ID is the sequence number of the intersection along the corridor. Intersection_name shows the intersecting roads. IF_SIG shows if the intersection is signalized. Flag 1 means the intersection is signalized, and flag 0 means the intersection is not signalized.

Table 5.2 Intersection Location

Intersection Location				
INT_Seq	Intersection_name	IF_SIG	Int_lat	Int_lon
0	Starting point	0	53.460794	-113.593996
1	23Ave.&Terwillegar Dr.	1	53.457709	-113.58999
2	23Ave.&Terwillegar Dr.	1	53.457228	-113.588368
3	23Ave.& legar Rd.	1	53.455474	-113.582431
Legar transit center				
4	23Ave.& legar gate	1	53.454009	-113.573713
5	23Ave.& Rabit Hill Rd.	1	53.453965	-113.565747
6	23Ave.& Hodgson Way	1	53.453967	-113.559294
7	23Ave.& 119St.	1	53.453834	-113.53425
8	23Ave.& 118St.	1	53.453831	-113.530553
9	23Ave.& Saddleback Rd.	1	53.453814	-113.525647
Century Park transit center				
10	23Ave.& 111St.	1	53.453821	-113.516607
11	23Ave.& 110St.	1	53.453826	-113.51322
12	23Ave.& 109St.	1	53.45382	-113.509183
13	23Ave.& 105St.	1	53.453838	-113.501593
14	23Ave.& Calgary Trail	1	53.453556	-113.493545

Table 5.3 Distance between Adjacent Intersections

Link_Seq	Length (M)	From_INT	To_INT
0	583.74	0	1
1	120.34	1	2
2	440.01	2	3
3	605.77	3	4
4	529.31	4	5
5	428.63	5	6
6	1791.11	6	7
7	245.59	7	8
8	325.87	8	9
9	600.31	9	10
10	1307.00	10	11
11	268.21	11	12
12	504.20	12	13
13	536.47	13	14
SUM	8286.56	--	--

Table 5.3 shows the distance between two adjacent intersections. The field of Link_ID is the sequence number of the connection between two adjacent intersections. From_INT and To_INT are the sequence number of the starting and ending intersections respectively. The summary amount of the length is the total length of the tested corridor.

Figure 5.5 Illustration of Intersection Locations

Table 5.4 Stop Location

Stop Location				
Stop_Seq	stop_name	IF_TC	stop_lat	stop_lon
0	Starting Point	0	53.460794	-113.593996
1	Riverbend Road & Heath Road S	0	53.46048304	-113.5938232
2	Haddow Drive & Riverbend Road	0	53.45867657	-113.5919195
3	Legar Transit Centre	1	53.4549475	-113.5804914
4	Tegler Gate & 23 Avenue	0	53.45398468	-113.5730102
5	Rabbit Hill Road & 23 Avenue	0	53.45396844	-113.5649155
6	Magrath Road & 23 Avenue	0	53.45396699	-113.5600423
7	Magrath Road & 23 Avenue	0	53.45396715	-113.5587047
8	119 Street & 23 Avenue	0	53.45383275	-113.5333987
9	118 Street & 23 Avenue	0	53.45382938	-113.5298698
10	Saddleback Road & 23 Avenue	0	53.45382982	-113.5248035
11	Century Park Transit center	1	53.45779333	-113.5157402
12	111 Street & 23 Avenue	0	53.45382054	-113.5151387
13	110 Street & 23 Avenue	0	53.45381877	-113.5124478
14	109 Street & 23 Avenue	0	53.45382751	-113.5085557
15	105 Street & 23 Avenue	0	53.45382648	-113.5008778
16	Calgary Trail & 23 Avenue	0	53.45355775	-113.4936281
0	Ending point (Int. #14)	0	53.453556	-113.493545

In the Table 5.4, Stop_ID is the sequence number of the stop along the tested corridor. IF_TC shows if the stop is the transit center. Flag 1 shows the stop is the transit center, while flag 0 shows it is not the transit center.

In TABLE 5.5, Link_Seq is the sequence number of the connection between two adjacent stops. From_Stop and To_Stop are the sequence number of the starting and ending stops respectively.

Figure 5.6 Illustration of Bus-Stop Locations

Table 5.5 Distance between Adjacent Stops

Link_Seq	Length (M)	From_Stop	To_Stop
0	176.20	0	1
1	239.37	1	2
2	870.11	2	3
3	510.92	3	4
4	537.77	4	5
5	323.70	5	6
6	88.85	6	7
7	1808.52	7	8
8	234.40	8	9
9	336.64	9	10
10	1022.09	10	11
11	701.43	11	12
12	178.77	12	13
13	258.55	13	14
14	510.05	14	15
15	483.41	15	16
16	5.78	16	0
SUM	8286.56	--	--

5.4.2 Test Results and Evaluation

Data used in this case was retrieved from May 9 (Mon.) to May 13 (Fri.), 2016 in GTFS dataset. No. 23 is the route from West Edmonton Mall to Mill Wood Transit Center, via 23^{rd} Avenue (corridor: Legar Transit center - Century park Transit center). The eastbound trips are selected and the corridor contains 8 links. Trips on Monday to Thursday are considered as the historical data, while the trips on Friday are considered as the current data.

Three trips scheduled to departure from the original stop between 10.10 am and 11.10 every day from Monday to Thursday are used as the historical data, while the three trips happened between 9.10 am and 10.10 am on Friday are considered as the current data. To predict the travel time between each two adjacent bus-stops, first is to estimate the travel time of the historical trips on the path using the method in section 5.2. The estimated results of travel time on each link are stored in the travel time vector $\{tt_1, tt_2, \ldots, tt_8\}$.

Table 5.6 Introduction to Travel Time Vector

Travel time	From_Stop	To_Stop	Distance (m)
tt_1	Legar Transit Centre	Tegler Gate & 23 Avenue	510.92
tt_2	Tegler Gate & 23 Avenue	Rabbit Hill Road & 23 Avenue	537.77
tt_3	Rabbit Hill Road & 23 Avenue	Magrath Road & 23 Avenue, West	323.70
tt_4	Magrath Road & 23 Avenue, West	Magrath Road & 23 Avenue, East	88.85
tt_5	Magrath Road & 23 Avenue, East	119 Street & 23 Avenue	1808.52
tt_6	119 Street & 23 Avenue	118 Street & 23 Avenue	234.40
tt_7	118 Street & 23 Avenue	Saddleback Road & 23 Avenue	336.64
tt_8	Saddleback Road & 23 Avenue	Century Park Transit center	1022.09

Table 5.7 Travel Time Information of GTFS data

Day (*i*)	Time period	Sch. departure time	Trip ID	Travel time vector (Sec.)
Mon. May 9, 2016	10:10-11.10 am	10:10 10:40 11:10	11229247 11229248 11229249	{139,99,40,22,160,23,38,89} {167,68,33,12,117,18,46,96} {101,52,32,62,159,23,72,123}
Tue. May 10, 2016	10:10-11.10 am	10:10 10:40 11:10	11230345 11230346 11230347	{136,77,42,20,151,28,30,86} {169,72,29,15,120,19,37,84} {94,47,39,56,159,26,54,107}
Wed. May 11, 2016	10:10-11.10 am	10:10 10:40 11:10	11231433 11231434 11231435	{116,110,52,19,140,16,40,95} {162,70,35,17,134,22,48,103} {99,57,42,41,164,28,49,92}
Thur. May 12, 2016	10:10-11.10 am	10:10 10:40 11:10	11232529 11232530 11232531	{120,97,45,20,157,26,40,87} {173,68,33,9,114,18,46,96} {98,42,35,58,167,15,75,116}
(current) Fri. May 13, 2016	9.10-10.10 am	09:10 09:40 10:10	11233247 11233248 11233249	{156,69,33,21,142,12,30,91} {104,56,24,21,165,37,35,81} {100,122,56,17,138,15,50,97}

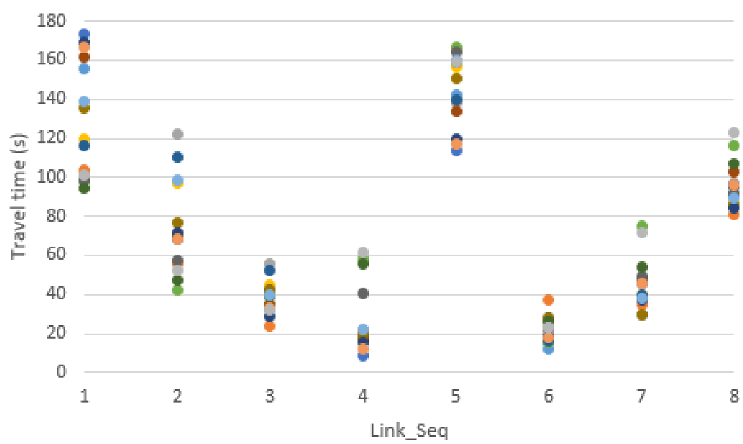

Figure 5.7 Estimation Results of Travel Time on Links

The travel time vectors shown in Table 5.7 is the estimated travel time on each link of the entire corridor. The travel time on each link with average standard variance of 16.17 seconds, which is resulted from the signal control strategy, variable traffic situation and other traffic accidents. The model input are concluded in the Table 5.8, including Euclidean distance and historical weighted travel time.

Table 5.8 Prediction Model Input Calculation

Link_Seq		1	2	3	4	5	6	7	8
Euclidean Distance (D_i^p)	Mon. May 9	65.26	77.10	26.57	45.90	55.40	23.37	25.87	30.08
	Tue. May 10	68.27	77.10	19.87	39.47	50.47	26.48	4.47	11.58
	Wed. May 11	70.46	78.12	26.04	24.41	40.51	20.25	16.43	22.91
	Thur. May 12	77.85	85.60	25.81	42.73	60.56	23.60	29.09	24.54
Weighted Historical Travel Time ($\bar{t}_i^p$)	Mon. May 9	135.67	73.00	35.00	32.00	145.33	21.33	52.00	102.67
	Tue. May 10	133.00	65.33	36.67	30.33	143.33	24.33	40.33	92.33
	Wed. May 11	125.67	79.00	43.00	25.67	146.00	22.00	45.67	96.67
	Thur. May 12	130.33	69.00	37.67	29.00	146.00	19.67	53.67	99.67

The predicted results of the trips happen in Fri. May 12 starting at 10.40 am is shown in Fig. 5.8, along with the model inputs of all trips on each link. The predicted results remain the consistency with the trend of weighted historical travel time. The weighted travel times of all trips happen on link 1 are lower than 140 seconds, but the distance and weighted travel time have been proven to be closer and higher with the time proceeding respectively, which means the travel time should be increasing during this week. The predicted result on link 1 is 145 seconds. The trend is proven to be true according to the ground truth, which is 167 seconds, higher than all the historical records.

The prediction results shown in Fig. 5.8 is based on fourth-interval model ($n = 5$). To evaluate different interval prediction model, other n values are introduced to test the similarity to the ground truth. The predicted results based on RP map-matching method are compared with the results generated by ST-matching method in Fig. 5.9, and the results show that the RP method is closer to the ground truth, especially for the long-distance links like link 1 and link 5.

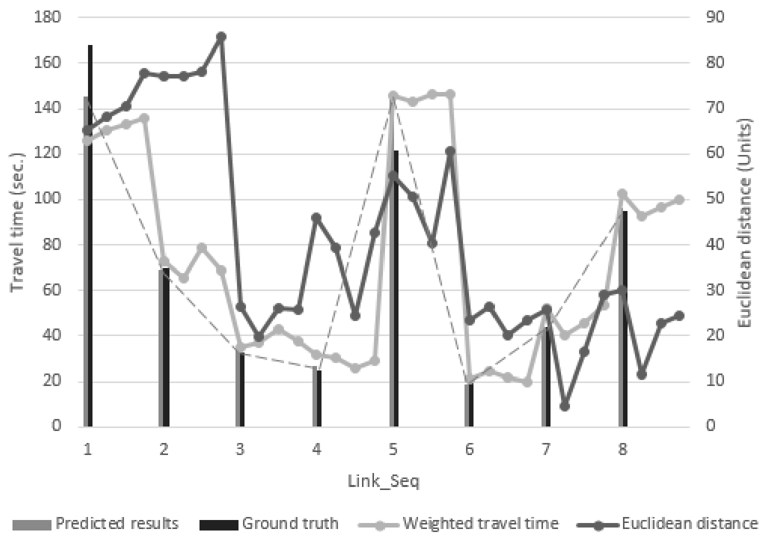

Figure 5.8 Prediction Results of the Trip (at 10.40 AM. May 12, 2016)

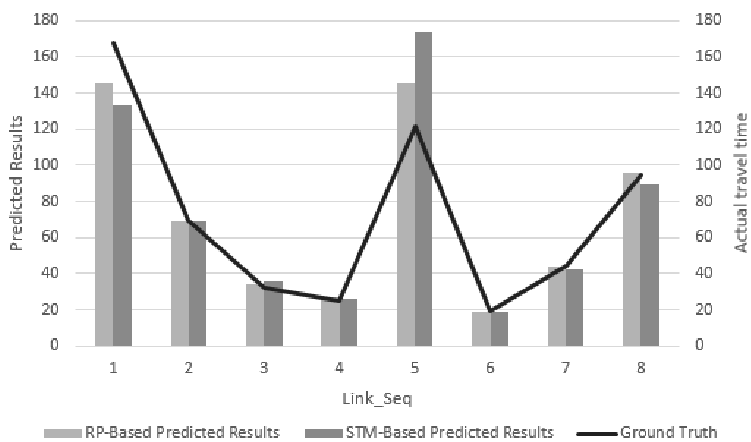

Figure 5.9 Predicted Result Comparison between Both Map-Matching Methods

Fig. 5.10 shows the 4-interval prediction result has the closest distance to the zero line, which means the prediction is closest to the ground truth. Because 4-interval model takes more historical data into consideration, avoiding the extreme situations. 3-interval prediction results are next closest to the zero line. 2-interval and single-interval prediction results are highly impacted by the extreme data and cannot combine the historical results in different days. The error is due to the unpredictable traffic accidents. (Chang, Park, Lee, Lee, & Baek, 2010).

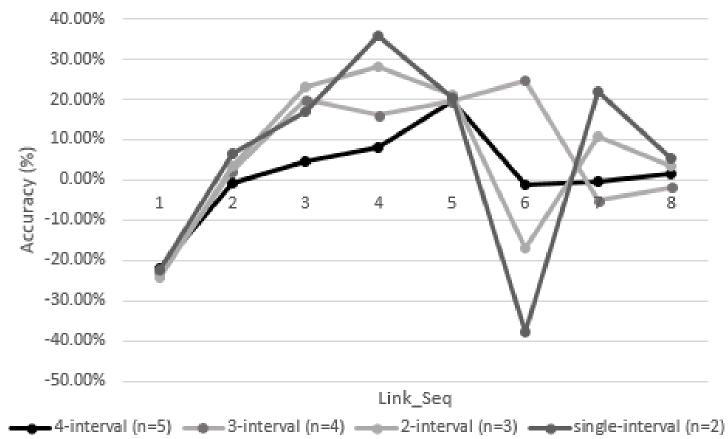

Figure 5.10 Link Prediction Accuracy (%) w.r.t. Multi-intervals

Based on travel time predicted results on each link, the trajectory of the predicted trip can be sketched in temporal-spatial diagram (Fig. 5.11). T-test is conducted to evaluate the relevance between the ground truth and predicted results. The returned value indicates that t-test does not reject the null hypothesis at the 5% significance level (Fig. 5.12).

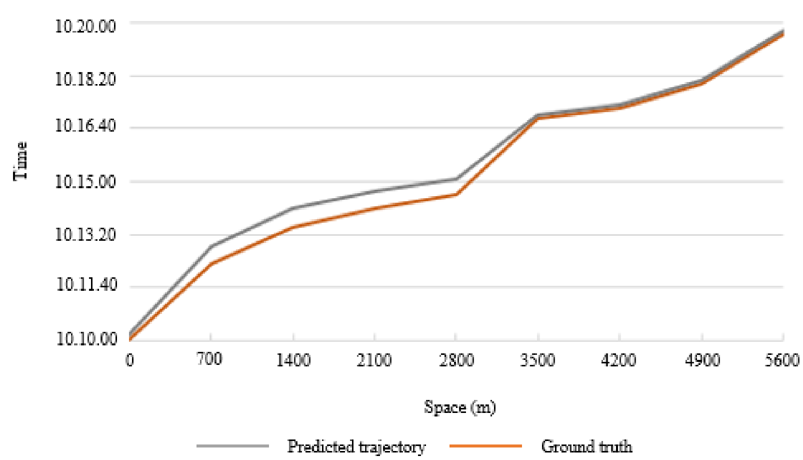

Figure 5.11 Trajectory Predicted Results (May 13, scheduled at 10.10 AM)

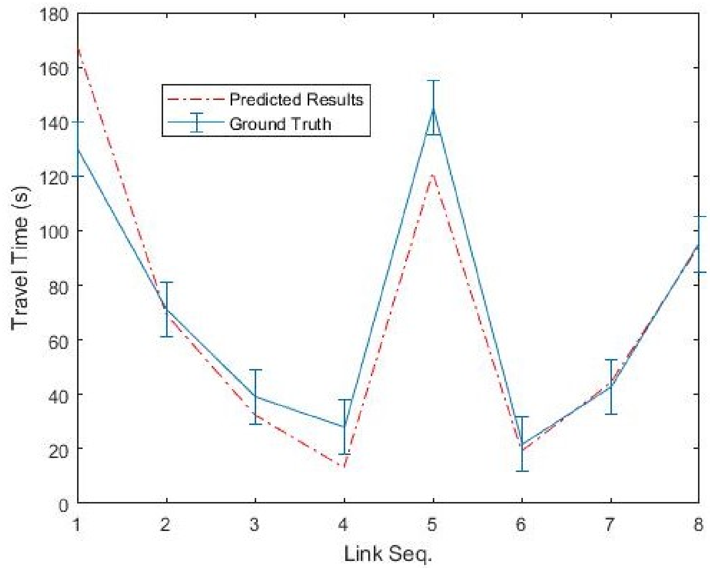

Figure 5.12 T-Test w/ Error Bar

5.5 Summary

This chapter presents a multi-interval travel time prediction model to test the impact of map-matching results on accuracy of the model performance. The model inputs integrate the different impacts of multi interval historical data. The raw data is processed using the map-matching method proposed and evaluated in chapter 4. The prediction is based on the estimation of the historical travel time using the RP-based map-matching method proposed in Chapter 4. The predicted results indicate the better performance than using ST-matching method. The prediction results are compared with the ground truth which is the high frequency and high-quality GPS data collected by handheld GPS device. The T-test shows that the results of the model are acceptable for the research. The evaluation results also contain the comparison in terms of different time-intervals. Multi-interval model generates more accuracy predicted results than single interval model.

CHAPTER 6 CONCLUSION AND RECOMMENDATION FOR FUTURE WORK

6.1 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 thesis focuses on an improved 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 raw GPS data. The major contributions of this thesis 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 thesis 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 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 algorithm is executed using real field GPS data collected from five traffic environment. The matching results show the algorithm is compatible with different types of road networks. The proposed method is proven to be outperforming a traditional spatial-temporal matching method by a case study conducted on 23rd Avenue corridor, Edmonton. A travel time prediction model is used to prove the prediction improvement.

6.2 Recommendations for Future Work

The proposed map-matching method depends highly on the digital map information, including the original road network and the newly added files. There are still some opportunities for future research including the follows.

- This thesis generates the reference points using 20 meters as the distance in between to limit the amount of reference points in each grid. In the future work, it is possible to evaluate the impact of the distance between reference points on the matching results.

- This thesis uses the GPS data from two sources, including GTFS dataset and handheld GPS collection devices. The quality of GPS data is not testified. The future work can be related to the impact of the GPS quality on the matching results.
- Spatial-temporal matching method is used as the benchmark in this thesis. In the future work, more traditional map-matching method can be compared with the proposed method.
- One travel time prediction model is the only example to evaluate the improvement of the proposed map-matching method. In the future work, more applications can be conducted based on the matching results from proposed method to see if the accuracy is better.

REFERENCES

Bajaj, R., Ranaweera, S. L., & Agrawal, D. P. (2002). GPS: location-tracking technology. *Computer*, 92-94.

Chang, H., Park, D., Lee, S., Lee, H., & Baek, S. (2010). Dynamic multi-interval bus travel time prediction using bus transit data. *Transportmetrica, 6*(1), pp. 19-38.

Chien, S. I.-J., Ding, Y., & Wei, C. (2002, September). Dynamic Bus Arrival Time Prediction with Artificial Neural Networks. *Journal of Transportation Engineering, 128*(5), 429-438.

Chien, S.-J., & Kuchipudi, C. M. (2003, November). Dynamic Travel Time Prediction with Real-Time and Historic Data. *Journal of Transportation Engineering, 129*(6), pp. 608-616.

Christopher E. White, D. B. (2000). Some map matching algorithms for personal navigation assistants. *Transportation Research Part C*, pp. 91-108.

Cui, Y. G. (2003). Autonomous vehicle positioning with GPS in urban canyon environments. *IEEE Transactions on Robotics and*, pp. 15-25.

Dieter, P., & Jensen, C. (1999). Capturing the uncertainty of moving-object representations. *Advances in Spatial Databases*, pp. 111-131.

Dijkstra, E. (1959). A note on two problems in connecion with graphs. *Numerische Mathematik*, 1 (1), 269–271.

Furth, P. G., Brendon, H. J., Theo, M. H., & Strathman, J. G. (2003). *Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Review and*

Potential. Retrieved from TCRP Web Document 23 (Project H-28): Contractor's Final Report.

Gong, J., Liu, M., & Zhang, S. (2013). Hybrid dynamic prediction model of bus arrival time based on weighted of historical and real-time GPS data. *Control and Decision Conference (CCDC)* (pp. 972-976). IEEE.

Google. (2012, October 16). *General Transit Feed Specification Reference*. Retrieved November 11, 2012, from https://developers.google.com/transit/gtfs/reference

Google. (2015). *Static Transit*. Retrieved from https://developers.google.com/transit/gtfs/?hl=en.

Greenfeld, J. (2002, January). Matching GPS observations to locations on a digital map. *Transportation Research Board*.

Gurmu, Z. K., & Fan, W. (2014). Artificial Neural Network Travel Time Prediction Model for Buses Using Only GPS Data. *Journal of Public Transportation, 17*(2), 45-65. doi:http://dx.doi.org/10.5038/2375-0901.17.2.3

Hargrove, S. (2015). Self-Learning License Plate Matching Algorithm: Some Enhancements and Its Role in Travel Time Ground Truth Measurements.

Hellinga, B., Izadpanah, P., Takada, H., & Fu, L. (2008, December). Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C: Emerging Technologies, 16*(6), 768-782.

Honey, S. Z. (1989). Vehicle navigational system and method. *United States Patent No., 4796191*.

Jeong, R., & Rilett, L. R. (1999, September). Bus Arrival Time Prediction Using Artificial Neural Network Model. *IEEE Transactions on Neural Networks, 10*(5), 1048-1054. doi:10.1109/72.788645

Kalaputapu, R., & Demetsky, M. J. (1995). Modeling Schedule Deviations of Buses Using Automatic Vehicle-Location Data and Artificial Neural Networks. *Transportation Research Record*, 44-52.

Khan, A. M., & Abdelfattah, A. M. (1998). Models for Predicting Bus Delays. *Transportation Research Record*, 8-15.

Kwong, K., Kavaler, R., Rajagopal, R., & Varaiya, P. (2009, December). Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C: Emerging Technologies, 17*(6), 586-606.

Lappin, J. (2000). *What do ATIS Customers?* (U. S. Transportation, Producer) Retrieved December 1, 2007, from Advanced Traveler Information Service (ATIS): http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_TE/12284.pdf

Liu, H. X., & Ma, W. (2009). A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. *Transportation Research Part C*, pp. 11-26.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., & Huang, Y. (2009). Map-Matching for Low-Sampling-Rate GPS Trajectories . *17th ACM SIGSPATIAL international conference on advances in geographic information systems ACM*, (pp. 352-361).

Lou, Y., Zhang, Y., Zheng, Y., Xie, X., Wang, W., & Huang, Y. (2009). Map-matching for low-sampling-rate GPS trajectories. *In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, (pp. 352-361).

Maurer, J. L. (1980, February). Efficient worst-case data structures for range searching. *Acta Informatica, 13*(2), pp. 155–168.

Meng, Y. (2006). Improved positioning of land vehicle in ITS using digital map and other accessory information. *Diss. The Hong Kong Polytechnic University,* .

Mohammed A. Quddus, W. Y. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C*, pp. 312–328.

Obradovic, D. L. (2006). Fusion of map and sensor data in a modern car navigation system. *Journal of VSLI Signal*, 112–122.

Ochieng, W. Q. (2004). Map-matching in complex urban road networks. *Brazilian Journal of Cartography (Revista Brasileira de Cartografia)*, pp. 1–18.

Oh, J.-S., & Chung, Y. (2006). Calculation of Travel Time Variability from Loop Detector Data. *Transportation Research Record: Journal of the Transportation Research Board, 1945*, 12-23. doi:10.3141/1945-02

Patnaik, J., Chien, S. I.-J., & Bladikas, A. (2004). Estimation of Bus Arrival Times Using APC Data. *Journal of Public Transportation, 7*(1). doi:http://dx.doi.org/10.5038/2375-0901.7.1.1

Rahmani, M. H. (2013). Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, 41-54.

Shalaby, A., & Farhan, A. (2003). Bus Travel Time Prediction Model for Dynamic Operations Control and Passenger Information System. *Transportation Research Board*.

Sherif, I., & Al-Deek, H. (2002). Performance evaluation of short-term time-series traffic prediction model. *Journal of Transportation Engineering, 128*(6), 490-198.

Smith, B., Williams, B., & Oswald, R. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C, 10*, 303-321.

Sun, L., Yang, J., & Mahmassani, H. (2008, January). Travel time estimation based on piecewise truncated quadratic speed trajectory. *Transportation Research Part A: Policy and Practice, 42*(1), 173-186.

Syed, S. C. (2004). Fuzzy logic-based map-matching algorithm for vehicle navigation system in urban canyons. *Institute of Navigation (ION) national technical meeting.* California.

Tomio Miwa, D. K. (2012, June). Development of map matching algorithm for low frequency probe data. *Transportation Research Part C: Emerging Technologies, 22*, 132–145.

Wang, B., Wang, W., Yang, M., & Gao, L. (2012). An Approach to Bus Travel Time Prediction Based on the Adaptive Fading Kalman Filter Algorithm. *CICTP 2012: Multimodal Transportation Systems—Convenient, Safe, Cost-Effective, Efficient*, (pp. 1652-1661).

Wu, C.-H., Ho, J.-M., & Lee, D. (2004, December). Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems, 5*(4), 276-281. doi:10.1109/TITS.2004.837813

Yang, Z.-s., Bao, L., & Zhu, G. (2004). An Urban Express Travel Time Prediction Model Based on Fuzzy Regression. *Journal of Highway and Transportation Research and Development, 21*(3), 50-75.

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.