The use of crowd-sensed Global Positioning System trajectory data to improve Transport Modelling practice

Adham Badran

Department of Civil Engineering

McGill University, Montreal

December 2023

A thesis submitted to McGill University

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy (Ph.D.) in Civil Engineering

© Adham Badran, 2023

Abstract

Transport models are decision making tools used to evaluate current and potential transport system conditions and to optimize its development. They can evaluate the impact of policies, sociodemographic changes, and infrastructure projects on the transport system. Large scale transport models, also known as macroscopic transport models, are composed of three components. The first component is the supply, which is a digital representation of the transport network for all modeled transport modes. The second component is the transport demand, representing all the trips that need to be made. The third component is the performance, which represents the network conditions when the demand is assigned to the transport network, therefore having a significant influence on route choice and traffic assignment. Developing these models is limited by the important resources required.

Recently, global positioning systems (GPS) trajectory data have been collected by GPS-enabled smartphones, creating large databases of GPS trajectories. This emerging data source has the potential to provide high coverage information to different applications such as transport modelling. In practice, the ability to extract transport system related variables for large networks will reduce model development resources and increase the model update frequency and quality. This thesis aims to explore this potential by developing methods to extract transport system features from GPS trajectory data.

First the thesis presents a comprehensive review of work examining the use of GPS data to extract road network features. Although some studies have developed methods to extract road networks from GPS data, it was found that the level of detail was insufficient from the

transport modelling perspective. Studies were mostly focusing on extracting and updating road centerlines.

Second, a reproducible method is proposed to extract road network topology and connectivity for modelling purposes. The resulting network model accuracy was greater than 95% for road segments compared to the ground truth dataset (Google Maps and Streetview). Moreover, extraction accuracy was very high for intersection movements except for intersections with very low GPS trajectory sample size.

Third, a combined GIS-machine learning method is proposed to extract the number of traffic lanes, essential in determining road capacity. The number of lanes information was inferred with an accuracy of 92% based on the lateral distribution of GPS points with respect to each road segment in addition to the sample size variable.

Fourth, road intersection control type is extracted. The proposed method used nearest neighbors' classifiers technique with intersection level speed and count statistical variables to predict intersection control type with an accuracy of 96%.

Finally, a method is proposed to improve traffic assignment route choice by standardizing the extraction of intersection movement delays to improve turn delay modelling in large scale transport models. Adjustment factors are proposed based on turning movement and road type to be integrated directly into volume delay functions used in transport models.

In sum, crowd sensed GPS trajectory data is a great source to standardize the extraction of road network features and improve transport model development. A combination of GIS and machine learning techniques were necessary to process the raw data and extract the

necessary information. However, the main limitation of the GPS data used in this research was the limited sample size which reduced data extraction locally where the data sample size was small or null.

Résumé

Les modèles de transport sont des outils d'aide à la décision utilisés pour évaluer les conditions actuelles et potentielles du système de transport pour optimiser son développement. Ils peuvent évaluer l'impact des politiques, des changements sociodémographiques et des projets d'infrastructure. Les modèles de transport à grande échelle, macroscopiques, sont composés de trois composantes: l'offre, une représentation numérique du réseau de transport pour tous les modes de transport modélisés, la demande de transport, représentant l'ensemble des déplacements à effectuer et performance représentant les conditions du réseau lorsque la demande est affectée au réseau de transport. Cette dernière influence significativement le choix de l'itinéraire et l'affectation du trafic. Le développement de ces modèles est limité par les importantes ressources nécessaires.

Récemment, des sources de données de parcours de type système de positionnement par satellite (GPS) ont été collectées par des téléphones intelligents équipés par GPS, créant de grandes bases de données de parcours GPS. Cette source de données émergente a le potentiel de fournir des informations à couverture élevée à différentes applications telles que la modélisation des transports. Cette thèse vise à explorer ce potentiel en développant des méthodes pour extraire les caractéristiques du système de transport à partir des données de parcours GPS.

Tout d'abord, la thèse présente une revue complète des travaux examinant l'utilisation des données GPS pour extraire les éléments du réseau routier. Bien que certaines études aient développé des méthodes pour extraire les réseaux routiers à partir des données GPS, il a

été constaté que le niveau de détail était insuffisant du point de vue de la modélisation des transports. Les études se concentraient principalement sur l'extraction et la mise à jour du réseau routier filamentaire.

Deuxièmement, une méthode reproductible est proposée pour extraire la topologie et la connectivité du réseau routier à des fins de modélisation. La précision du réseau routier modélisé était supérieure à 95 % pour les segments de route en comparant les résultats aux jeux de donnée de validation (Google Maps et Streetview). De plus, la précision d'extraction était très élevée pour les mouvements d'intersection, sauf pour les intersections avec un échantillon de parcours GPS très limité.

Troisièmement, une méthode combinée SIG-apprentissage machine est proposée pour extraire le nombre de voies de circulation, essentiel pour déterminer la capacité routière. L'information sur le nombre de voies a été prédite avec une précision de 92 % grâce à une approche d'apprentissage d'ensemble basée sur la distribution latérale des points GPS dans chaque segment de route en plus de la variable de taille d'échantillon.

Quatrièmement, le type de contrôle d'intersection des routes est extrait. La méthode proposée utilise la technique des classificateurs des voisins les plus proches en utilisant des variables de vitesse et de fréquence d’observation pour prédire le type de contrôle de l'intersection à une précision de 96 %.

Enfin, une méthode est proposée pour améliorer le choix de l'itinéraire d'affectation du trafic en standardisant l'extraction des retards de mouvement aux intersections. Des facteurs d'ajustement sont proposés en fonction du mouvement de virage et du type de

route à intégrer directement dans les fonctions volume-délai utilisées dans les modèles de transport.

En conclusion, les données de parcours GPS détectées par la foule sont une excellente source pour normaliser l'extraction des éléments du réseau routier et améliorer le développement de modèles de transport. Cependant, la principale limite des données GPS utilisées dans cette recherche était la taille limitée de l'échantillon qui réduisait la qualité de l'extraction des données localement lorsque la taille de l'échantillon de données était petite ou nulle.

Acknowledgements

I would like to acknowledge the generous support of McGill University's Faculty of Engineering through the McGill Engineering Doctoral Award and other travel awards and the Vadasz Scholars Program that allowed me to pursue my studies.

My deepest gratitude and appreciation go to my advisor Dr. Luis Miranda-Moreno for all his support. He has always provided me with constructive feedback and guidance and supported my participation in conferences and workshops to learn and share our research.

My deepest gratitude and appreciation go to my co-supervisor Dr. Ahmed El-Geneidy for all his help. He has helped me and provided constructive feedback throughout my studies.

I would also like extend my acknowledgments to my academic and professional colleagues and members of the McGill staff who supported me in different ways.

My heartfelt gratitude goes to my parents who are always at the foundation of all my achievements. This work is dedicated to them and to my children Reem and Fares.

Table of Contents

Index of Figures

Index of Tables

Contribution to Original Knowledge

This thesis employs emerging crowd sensed GPS trajectory data to improve current transport model development practice. It builds on previous research by proposing methods to build a topologically sound transport network model and extract different road network features such as number of lanes, road intersection control type, intersection turning permissions, and intersection movement delay per road, turn type, and intersection control type. The proposed methods are adapted to transport model development for use in large scale transport modelling and planning.

Publication Details

This thesis is presented in a manuscript-based format and is composed of the following five manuscripts authored by the candidate. These manuscripts are co-authored by the candidate's supervisor and co-supervisor Dr. Luis Miranda-Moreno and Dr. Ahmed El-Geneidy, respectively. The table below presents the manuscripts' titles, journal and/or conference of publication or presentation, the year, and the status.

Title	Journal / *Conference*	Year	Status
A Review of Techniques to Extract Road Network Features from Global Positioning System Data for Transport Modelling	Transport Reviews Journal / *WCTR**	2023	Published / *Presented*
Creating a Road Network Model from Global Positioning System Trajectory Data for Macroscopic Simulation	WCTR - Transportation Research Procedia / WCTR and *TRBAM***	2023	Accepted / *Presented*

Title	Journal / *Conference*	Year	Status
Global Positioning System Data to Model Network-Wide Road Segment Level Number of Lanes Using Machine Learning	TRBAM**	2023	To be presented
Inferring Road Intersection Control Type from GPS Data	Transport Findings	2022	Published
Intersection Movements Delay Modelling Based on Crowd-sensed Global Positioning System Trajectory Data	Canadian Journal of Civil Engineering	2023	Under Review

* World Conference on Transport Research

** Transportation Research Board Annual Meeting

Contribution of Authors

The contributions to this thesis can be divided into two main parts: the contributions to the introduction, discussion and conclusion of the full thesis and the contributions to each of the manuscripts. For the first part, it was written by the candidate and revised by the supervisor and the co-supervisor. For the manuscripts, the candidate performed the literature review, the data analysis, the interpretation of results, and the manuscript write-up. The co-authors, Dr. Luis Miranda-Moreno and Dr. Ahmed El-Geneidy, contributed intellectually to the research and provided critical comments and edits to the manuscripts.

Chapter 1 - Introduction

1.1 Context

In 2018, 55% of the world's population was living in urban areas. It was 30% in 1950 and is forecasted to reach 68% by 2050. In addition to the phenomenon of urbanization, global population growth, population aging, and international migration are the mega trends dictating the demography of today and tomorrow (UN, 2019). To adapt to demographic changes and support the economy, urban area transport agencies are continuously maintaining and developing their infrastructure. These investments represent important costs to communities, for example, the planned transport investments for the province of Quebec, Canada for 2022-2032 are estimated at 47,7 Billion dollars (*Plan québécois des infrastructures 2022-2032*, 2022). In addition, they are usually justified by economic development and growth. However, to ensure this positive outcome, effective use of transport infrastructure in addition to the infrastructure quality and type are essential determinants that need to be accounted for (Deng, 2013). For this reason, large-scale transport models, known as macroscopic transport models, are essential tools to help in the decision-making process.

A transport model is a digital representation of the transport system used at the planning stage of transport projects to simulate the impact of policies, sociodemographic changes, and infrastructure projects on the transport system (Wegener et al., 1991). The outputs of these models are used in project evaluation procedures such as cost-benefit analysis and multicriteria analysis (Sinha & Labi, 2011) or meta-analysis (Hrabec et al., 2022; Liu et al., 2023) .

Transport models can be divided into three main categories: macroscopic, mesoscopic, and microscopic models (Figure 1-1). Macroscopic models are used for regional planning and cover the road network of an entire metropolitan area. They require less detailed information about the road network. Mesoscopic models are used for sub-regional planning, for example, only the island of Montreal area.

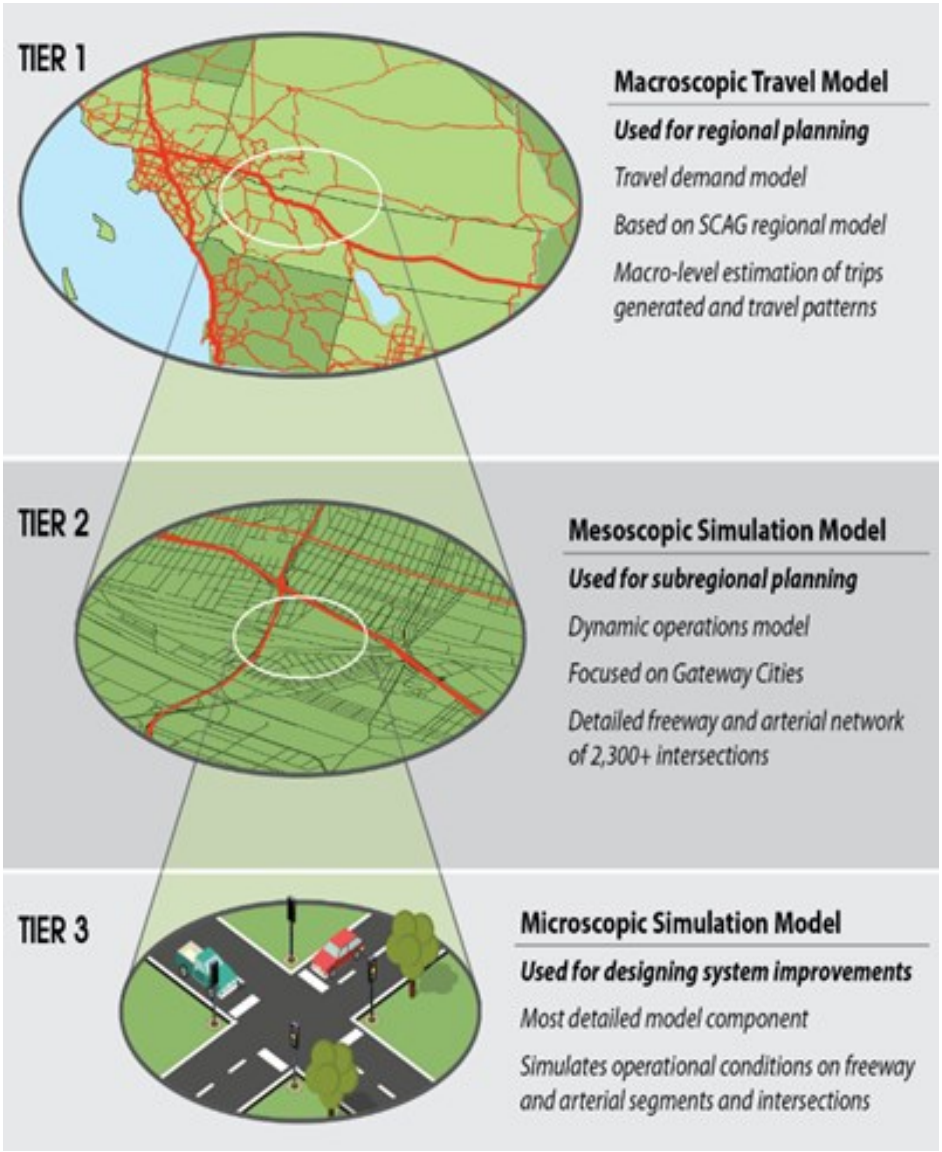


Figure 1-1. Transport Model Categories (Source: California Department of Transportation)

This type of model includes more detailed information about the road network, for example signal timing plans at each intersection and can be used to optimize operations at a subregional level. Finally, microscopic models can cover from one intersection to a small network or corridor of intersections, they are used in traffic operation optimization and in traffic design. This type of model requires much more detailed information about the road network, for example, it needs a detailed road geometry compared to the two other model categories. This dissertation focuses on macroscopic models.

Macroscopic transport models, are composed of three main components: transport demand, transport supply, and performance functions (Ortúzar & Willumsen, 2011). Transport demand represents the users of the transport system in terms of spatial and temporal distribution of the trips that need to be made. This information is generally obtained through traditional origin destination surveys (*ARTM Faits saillants EOD 2018*, 2018) or derived from data collected using cellular towers (Alexander et al., 2015; Iqbal et al., 2014; Ma et al., 2013), GPS devices (Demissie & Kattan, 2022), transit smartcard readers (Munizaga & Palma, 2012), or Bluetooth and Wi-Fi devices (Carpenter et al., 2012; Lesani & Miranda-Moreno, 2019). In addition, researchers have proposed to infer origin destination travel demand based on outdated origin destination surveys and recent traffic counts (Freytes, 2022).

Transport supply is the digital representation of the transport network and service. It is represented in terms of directional links and nodes. These elements also contain additional attributes used to describe road segments and intersection' properties. For example, each segment has a specific number of lanes, a road type, and a link performance function.

Intersection properties are also required to indicate permitted movements, turn penalty functions, and control type. The digital road network representation is usually obtained through manual extraction or inference using other data sources such as satellite imagery, lidar, vehicle imagery, and GPS data (Banqiao et al., 2020). For public transit, transit routes, schedules, and vehicle capacities are also modeled as part of the transport supply. Recently, studies have used General Transit Feed Service to extract the public transit supply (Fortin et al., 2016).

The last component of macroscopic transport models is the demand-supply interaction, known as link performance functions. They are specified for each directional link and represent the relationship between traffic flow and travel time (Kucharski & Drabicki, 2017). Calibration of these function relies on observed traffic counts and travel time data is generally done by collecting travel time data using GPS equipped floating vehicles driving along the main corridors of the modelled region (TRANS, 2014). This signifies that turning movement speeds are not observed and used to calibrate modelled intersection movements.

Given the large scale of macroscopic models, their development requires important resources which limit their update frequency and their ability to represent the observed conditions. Therefore, it is important to seek new data sources and techniques that can reduce the required resources required to develop the model and increase its accuracy and update frequency.

Recently, GPS trajectory data has been collected using GPS-enabled smartphone devices. It should be noted that given the location of the research in North America, this dissertation refers colloquially to the Global Navigation Satellite System (GNSS) as GPS which is understood to be just the American part of the GNSS. This data contains the location longitude, latitude, and timestamp information recorded by each device at a specific rate. This information is used in numerous location-based services such as navigation applications, for example Google Maps. Activity tracking apps such as Strava also track and record GPS trajectories for active transport and makes the data available to help in active transport infrastructure planning (Lee & Sener, 2021; Sun & Mobasheri, 2017). In another study, crowd-sensed GPS trajectory data collected using a usage based insurance program was used for traffic safety studies (Stipancic et al., 2021).

One of the main challenges in using crowd-sensed GPS trajectory data is the need to infer transport mode since it is not explicitly provided. However, extensive research has already been carried out and multiple studies have developed algorithms to infer transport mode based on trajectory characteristics (Dabiri & Heaslip, 2018; J. Li et al., 2021).

Transport data providers have also developed techniques to integrate GPS data from multiple sources to be used in transport planning activities as it is the case with StreetLight Data (Turner et al., 2020; Yang et al., 2020).

Research effort in the computer science and geography fields have researched the use of GPS trajectory data to infer road network map and topology. The works by Ahmed et al. (2015a, 2015b) provide a good overview of the different categories of map inference

algorithms. Although these research efforts are performed from a different perspective, they have a similar objective to transport network modelling, which is to develop a road network representation. However, in transport network modelling the modelled network requires the extraction of more detailed information which has not been sufficiently explored in past research using GPS data.

1.2 Research Objectives

The general goal of this research is to propose crowd-sensed GPS data-driven methods to help build on current large scale model development practice by increasing network modelling accuracy while facilitating its' updateability and reducing the required resources. More specifically, the objectives of this research are:

a. To provide a comprehensive literature review on the use of GPS trajectory data to extract road network features and maps.
b. To propose a method to extract road network topology and connectivity features including road segments' number of lanes information and road intersection control type based on GPS trajectory data.
c. To propose a simple method to calibrate turn delay functions per road type, per control type, and turn type based on GPS data.

The relationship between the different objectives of the research, the main input data, and the main modelling tool is presented in Figure 1-2.

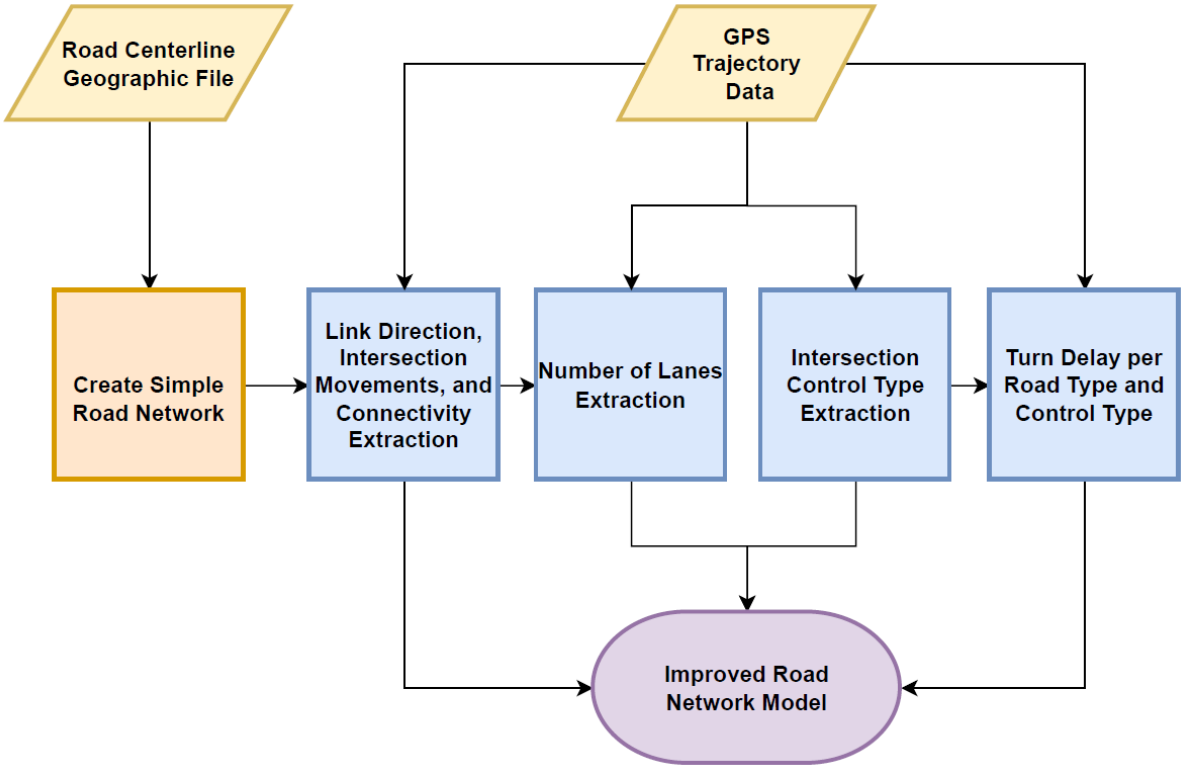


Figure 1-2. Overview of Research Structure and Contributions

1.3 Data Sources and Computing Tools

This research is based on the use of multiple data sources that are described in this section.

a. GPS data was collected during the spring of 2014 in Quebec City, Canada. It was collected during 21 days by 2000 voluntary users through the Mon Trajet smartphone app, made available by the Municipality. Each point is described by the following attributes: raw and map matched X and Y coordinates (reported with six decimal places), trip ID, speed, and timestamp (Year-Month-Day-Hour-Minute-Second). Figure 1-3 presents a map of the study area showing raw GPS points (226,000 points) inside the study zone, which consists of 81 intersections.

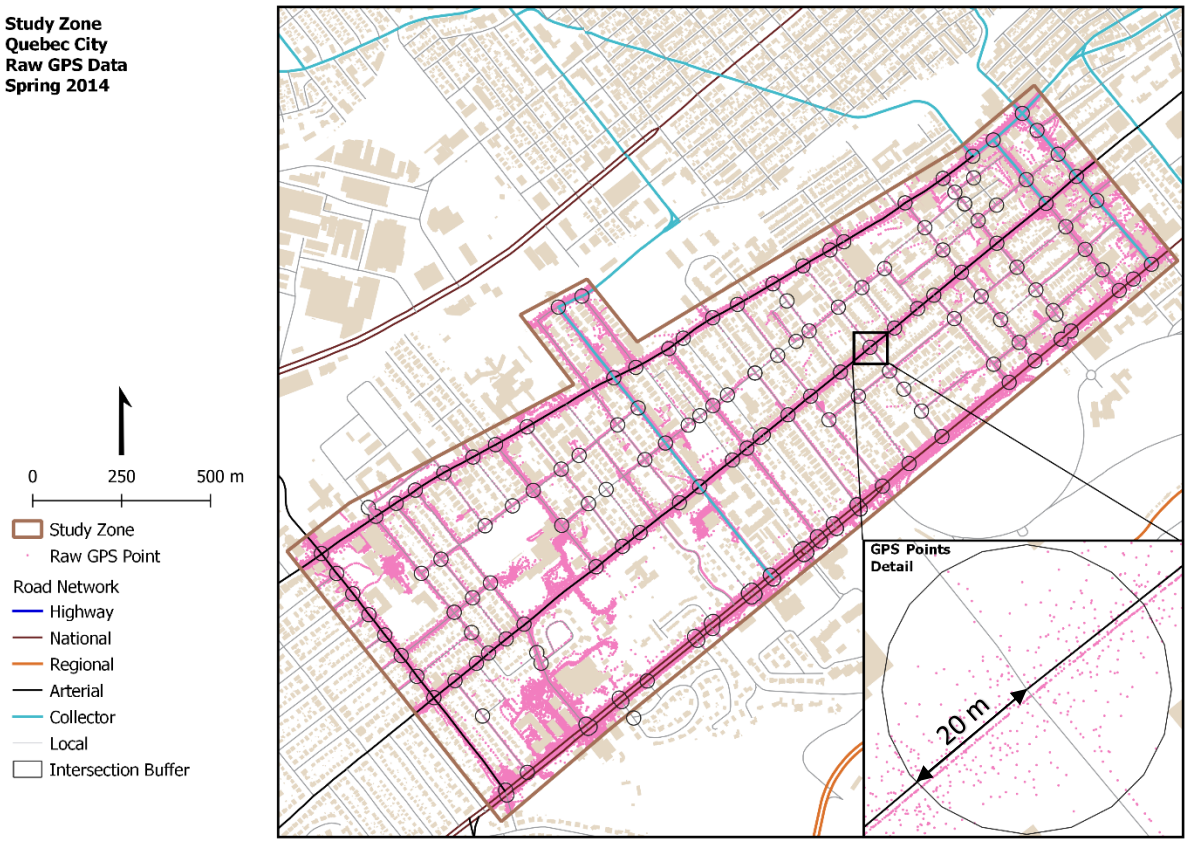


Figure 1-3. Study Area - GPS points

b. A shapefile file of the up to date study area road network, as seen in Figure 1-4, was available online ("Adresses Québec - AQréseau," 2022). The location of all intersections was obtained from the municipality ("Données ouvertes ", 2022). Finally, Google maps and street view were used to manually extract ground truth information.

Figure 1-4. Simple Road Network Geographic File

c. Traffic counts were collected between the years 2013 and 2019 and provided by the Municipality of Quebec City. Traffic counts were available for a one-day period per intersection for 15-min time intervals from 7:00 to 10:00 and from 15:00 to 18:00. These periods were selected by the municipality to be able to cover peak traffic periods.

Completing this study was possible using the following software:

- Quantum Geographic information systems (QGIS)[1]: was used to visualize geographic data, perform visual validations, and create maps.

[1] https://qgis.org/en/site/

- ArcGIS by ESRI[2]: was used to construct travel paths from trajectory points using the network analyst extension.
- Feature Manipulation Engine (FME)[3]: was used to manipulate data and perform geographic operations.
- Equilibre Multimodal Multimodal Equilibrium (EMME)[4]: macrosimulation software used to build the road network model.

1.4 Thesis Structure

This dissertation is organized according to McGill University's guidelines for manuscript-based thesis. It is composed of one literature review manuscript and four manuscripts that address the objectives of this research. Each manuscript includes a more specific literature review that emphasizes the context of the study with respect to past research. The five manuscripts are followed by a comprehensive scholarly discussion of all the findings including the limitations and future research directions.

Chapter 2 is a general and comprehensive literature review regarding research on the extraction of road network information using GPS trajectory data. This helped in identifying research gaps and guided this research work.

Chapter 3 proposes a method to extract network-wide road direction and turning movement rules. In addition, it serves as a proof of concept by building a road network model under a widely used macroscopic transport modelling software, EMME. Sensitivity

[2] https://www.arcgis.com/index.html
[3] https://www.safe.com/
[4] https://www.inrosoftware.com/en/products/emme/

analysis was carried out to determine the output quality and to recommend future improvements.

Chapter 4 presents a method to extract the number of lanes from GPS trajectory data by defining the problem as a supervised machine learning decision trees classification. The proposed method is divided into two main steps, a spatial analysis step and a machine learning modelling step. It presents the model accuracy in addition to the best predictors for intersection control type.

Chapter 5 presents a method to develop a model inferring road intersection control type (traffic light, stops on all approaches, or stops on the secondary approach). Data was used to train and validate supervised machine learning classification models. It presents the model accuracy in addition to the best predictors for intersection control type.

Chapter 6 presents a method to improve intersection movement delay modelling using crowd-sensed Global Positioning System (GPS) data. This is done through spatial analysis by providing a general definition of turning movements and extracting travel times from GPS trajectory points. A method was also provided to integrate the observed delays per movement type (right turn, through movement, and left turn) into volume delay functions commonly used in large scale transport models.

Chapters 7 presents a scholarly discussion of this research by presenting findings, limitations, and future work directions.

Chapter 8 presents the conclusion of this research and a summary of how the objectives were met and discusses the implication of findings.

Chapter 2 - A Review of Techniques to Extract Road Network Features from Global Positioning System Data for Transport Modelling

A Review of Techniques to Extract Road Network Features from Global Positioning System Data for Transport Modelling

Adham Badran [a]*, Ahmed El-Geneidy [b], and Luis Miranda-Moreno [a]

[a] Civil Engineering Department, McGill University, Montreal, Canada; [b] School of Urban Planning, McGill University, Montreal, Canada

Contact: adham.badran@mail.mcgill.ca, Department of Civil Engineering, McGill University, 817 Sherbrooke Street West, Montreal H3A 0C3, Canada.

2.1 Abstract

With the spread of smartphones and mobile internet, Global Positioning System (GPS) data from vehicles has become widely available. This data represents a unique opportunity to automatically extract road network features and generate detailed maps that can be used in the creation of transport network models, while minimizing the quantity of resources usually invested in that task. Accurate transport network models can be used in a variety of applications either in transport simulation models or autonomous vehicles navigation. Although two relevant literature reviews were performed during the last decade, they were not systematic and did not explore the road network inference methods from a transport network modelling point of view. The objective of this research is to perform a systematic and reproducible literature review on the use GPS data in transport network modelling and provide limitations and future work to extract a road network representation for transport models and autonomous vehicles navigation. This was done by systematically examining the studies' different approaches with respect to relevant criteria. Most studies produced a simple representation of the road network, not detailed enough for transport models. Other limitations were the bias introduced by the GPS sample and the reproducibility of the different methods.

Keywords: map inference; GPS data; transport model; road network; intersection movements.

2.2 Introduction

Data and knowledge of detailed transport network features are important for multiple fields such as traditional and autonomous vehicle navigation, traffic safety, urban planning, and transport modelling. Although a basic road centreline network representation is sufficient for certain applications, other applications can require additional and more detailed information, which is the case for transport models. In fact, transport models are tools developed by transport engineers and planners to help in the decision-making process of transport infrastructure planning. This type of model can be divided into three main components: supply, demand, and performance where the supply component is mainly represented by a detailed digital road network. It represents road segments as directional links and intersections as nodes. It also contains additional attributes used to describe road segments and intersection' properties. For example, each link has a specific number of lanes, a road type, and a link performance function. Intersection properties are also required to indicate permitted movements, turn penalty functions, and traffic control type. Additionally, the road network is dynamic in nature, since traffic rules can prohibit a subset of road users from using a specific road lane or making a specific movement at an intersection, depending on a temporal criterion. Therefore, the modelled road network should also represent this characteristic. The digital road network representation is usually obtained through manual extraction or inference using other data sources such as satellite imagery, lidar, and vehicle imagery (Banqiao et al., 2020). The high cost and labour associated to these methods is the main limiting factor to data quality and update frequency.

To improve the transport network modelling process, transport modelling software providers have provided tools to automatically construct transport networks based on digital maps. While improving some aspects of the network modelling process, achieving a satisfactory network model quality still relies on manual intervention and additional data sources to validate and input some of the essential attributes. For example, traffic control information at intersections, permitted intersection movements, and number of lanes are usually unavailable in digital maps. In addition, digital maps require continuous maintenance and update, which also requires important resources.

Thanks to location-based services, global positioning systems (GPS) data has become widely available in terms of spatial coverage and sample size, providing an immense potential for transport network modelling. This potential lies in the possibility to automatically extract road network features from GPS trajectory information. GPS trajectory data is defined as a set of chronological location points data where each point is described using longitude and latitude coordinates, a timestamp, and a trip ID. Depending on the parameters of the GPS device recording the points, the sampling rate or frequency can be set in terms of time or distance. For example, the sampling rate can be set to record the location point every 1 sec which is equivalent to a frequency of 1 Hz, or to record a location point every 10 meters.

This systematic literature review explores research that used large-sample GPS data to automate the network construction process, by extracting road shape, topology, number of lanes, and permitted intersection movements. A special focus is placed on transport network features extraction usable for large scale transport model development.

In the geography and computer science fields, extracting a road map from GPS data, also known as map inference, has been explored since the 1990s. Within the last decade, two literature reviews were published on map inference techniques using GPS data by Ahmed et al. (2015a) and Chao et al. (2022). Map inference can be defined as the process of constructing the digital road map (roads location, intersections, topology, connectivity, etc.) based on specific data sources such as aerial images or GPS trajectories. The produced map can be as simple as a line representing the roads' centerline. In contrast, transport network modelling requires the construction of digital road network model that describes the road network in detail to enable its use in transport modelling and simulation. An inferred map where the road network is created in a standardize directional link (road segment) and node (intersection) format containing the required attributes, such as the number of lanes, turning permissions, road type, intersection control type, can be defined as a road network model. can be defined as a road network model.

The work by Ahmed et al. (2015a) benchmarks map inference algorithms by performing a comparison and evaluation using multiple GPS datasets and various quality measures. These algorithms have a common objective; to use GPS data points or trajectories as an input to create directional links and nodes representing the road network. The output is usually compared to a ground truth map. The algorithms were classified under three distinct categories based on the technique used: 1. Point Clustering, 2. Incremental track insertion, and 3. Intersection linking. In addition, algorithm performances were evaluated using four quality measures: 1. Directed Hausdorff distance, 2. Path based distance, 3. Shortest path-based distance and graph-based sampling distance. This work is complemented by the book

authored by Ahmed et al. (2015b). Although the review is insightful and comprehensive in terms of map inference techniques, it is not systematic and does not approach the question from the transport modelling point of view, which requires specific road network features to be included in the network model. In fact, the review does not assess if the examined papers are extracting network features usable for transport network modelling, such as turning movement permissions, intersection controls, or the number of lanes available for traffic. In addition, it does not discuss the reproducibility of the different works reviewed. Furthermore, the review does not present the necessary future work to improve on the techniques and extract more detailed information from GPS data. Finally, Given the time elapsed since 2015 and the increasing availability of GPS data in recent years, an updated review of the work is beneficial to explore new work.

More Recently, the literature review by Chao et al. (2022) explored more recent studies in the map inference context. Their focus was placed on the proposition of a new categorization of algorithms while assessing the existing map inference quality measures and the effect of GPS errors on the inference results. They proposed to classify map inference algorithms as: 1. Road abstraction, 2. Intersection linking and 3. incremental branching. Despite a minor change in the category names, these categories are not significantly different from the ones proposed by Ahmed et al. (2015a) and do not change the classification of the different algorithms. In addition, the work identifies the best algorithms in terms of scalability, accuracy, and suitability to update. This review is not reproducible and does not discuss map inference from the transport modelling point of view. Thus, it cannot assist in determining which technique is preferred to extract network

features for transport modelling. In fact, the review focuses on the performance of the available algorithms and does not present future works required to be able to extract more detailed network features from GPS data.

Although past literature reviews are a good place to explore the work done in map inference, it was usually performed from the optic of the geography and computer science fields. Overall, there was no discussion about the ability of current algorithms to extract more detailed network features or the necessary research towards this objective. The literature review method performed in both works was not systematic, thus not reproducible. Finally, past literature reviews provide limited guidance for transport modelers in selecting the best techniques to model road networks or in determining future work since this was not the objective of their research. Therefore, the objective and contribution of this work is to build on previous research by developing a systematic and reproducible literature review that surveys the work done in the map inference field and the additional work required to be able to extract detailed road network features to support in transport network modelling.

2.3 Methods

To systematically review all relevant research while ensuring a high level of reproducibility of this research effort, this work was inspired by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, 2015). This technique requires the presentation of the study identification process, clearly indicating the sources and the screening steps and justifications. The research scope design, including objective, input and output data, and inclusion and exclusion criteria are presented below.

2.3.1 Search criteria

The objective of the studies had to be the development of inference techniques of road network features based on regular / commodity GPS data. This excludes the use of high precision GPS or differential GPS, which is not feasible for large scale applications. Studies in the fields of geography, computer science, and transport planning and engineering using GPS points or trajectories as the main input regardless of the data collection device (in-vehicle, smartphone, etc.) were included. All studies aiming to construct (infer) a road network were included. The final output had to be a map of the road network. Only English and French publications were selected given the authors' language abilities. If an author produced multiple publications, only the most recent was selected. In addition, only publications from the last 10 years were included (2012-2022). Publications without full texts were discarded.

2.3.2 Search strategy

The search strategy was developed by the authors in consultation with the librarian associated to the Civil Engineering department. Multiple trial searches were conducted to determine all synonyms. These trials were critical to the keyword selection as this research effort included different fields of research that do not use the same terminology. For example, the main research objective could be network modelling, map inference, map generation, map construction, or map extraction depending on the research field (computer science, geography, or transportation engineering and planning). The chosen keywords were then selected and searched in the following bibliographic databases: Scopus, Web of Science, Compendex, and Transport Research International Documentation

(TRID). The searches were performed on February 24th, 2022. The exact keyword specification is presented below:

> ("GPS") AND ("network inference" OR "inference of network" OR "network extraction" OR “extraction of network” OR "network mining" OR “mining of network” OR "network generation" OR “Generation of network” OR "Road extraction" OR “Extraction of Road” OR "Road inference" OR “Inference of road” OR "Road Mining" OR “mining of road” OR "map extraction" OR “extraction of map” OR "map inference" OR “Inference of map” OR "Map mining" OR “mining of map” OR "lane reconstruction" OR “reconstruction of lane” OR "intersection reconstruction" OR “Reconstruction of intersection” OR "lane mining" OR “mining of lane” OR "intersection mining" OR “Mining of intersection” OR "lane inference" OR “inference of lane” or "intersection inference" OR “Inference of intersection” OR "intersection detection" OR “detection of intersection”)

2.3.3 Selection of studies

Following the removal of duplicates, the titles and abstracts were screened systematically by the author using the Rayyan web platform (Ouzzani et al., 2016). The full texts of the remaining publications were retrieved for an in-depth selection assessment. Finally, all studies respecting the inclusion criteria stated above were selected for data extraction and further analysis.

2.3.4 Data Extraction

A global extraction form was developed and used to systematically extract all relevant information from the publications. The form was then used to analyze all studies on the same standardized basis. This form was completed by the author and contained, when available, the following information: author, year, title, journal / conference, study setting (country, city), field of study, research question, sample description, comparative methods, techniques used, detailed output, coverage, validation, comprehensibility, reproducibility, and limitations.

2.4 Results

Following the keywords' selection, the database search identified 500 publications. Duplicate articles and publications before 2012 were removed. The title and abstract of the remaining 158 articles were screened, resulting in the exclusion of 110 articles. The final screening step was the full report retrieval and examination of the 48 publications. Following the screening process, 17 articles were included in this literature review. Reports were excluded when the research paper was a literature review, a book, not building a road network, requiring additional resources such aerial images, newer work was published by the same author, or the GPS sampling frequency was greater than one minute. Figure 2-1 presents a breakdown of the search and screening process.

A summary of the selected papers is presented in Table 2-1. It can be noted in the Journal/Conference column that most of the work done is in the field of geography and computer science. As for the experimental data that was tested, it was mainly collected in the United States and China. The main research question for all the studies was the

construction of a road network using GPS points or trajectories as input, by developing different algorithms and methodologies that can outperform previous research efforts.

Out of the 17 studies, the most popular approach is clustering (n = 11). The intersection linking approach is the most recent to be explored by researchers (n = 4). Finally, the least popular approach is track alignment (n = 2).

This work presents the different publications by approach as in Ahmed et al. (2015a). The selected studies are summarized in the following section under each of these approaches. The summarized information relates to the following elements: a) road network definition (network components, directionality, number of lanes, and turning movement permissions), b) output quality (if and how the output quality was evaluated), c) experimental data characteristics (sample size, sampling rate, collection method, and coverage), d) method clarity and reproducibility (if the article is sufficient to understand the method and be able to reproduce it.).

The discussion goes further by analyzing the results from a transport network model point of view and presenting the opportunities for further research to extract road network features.

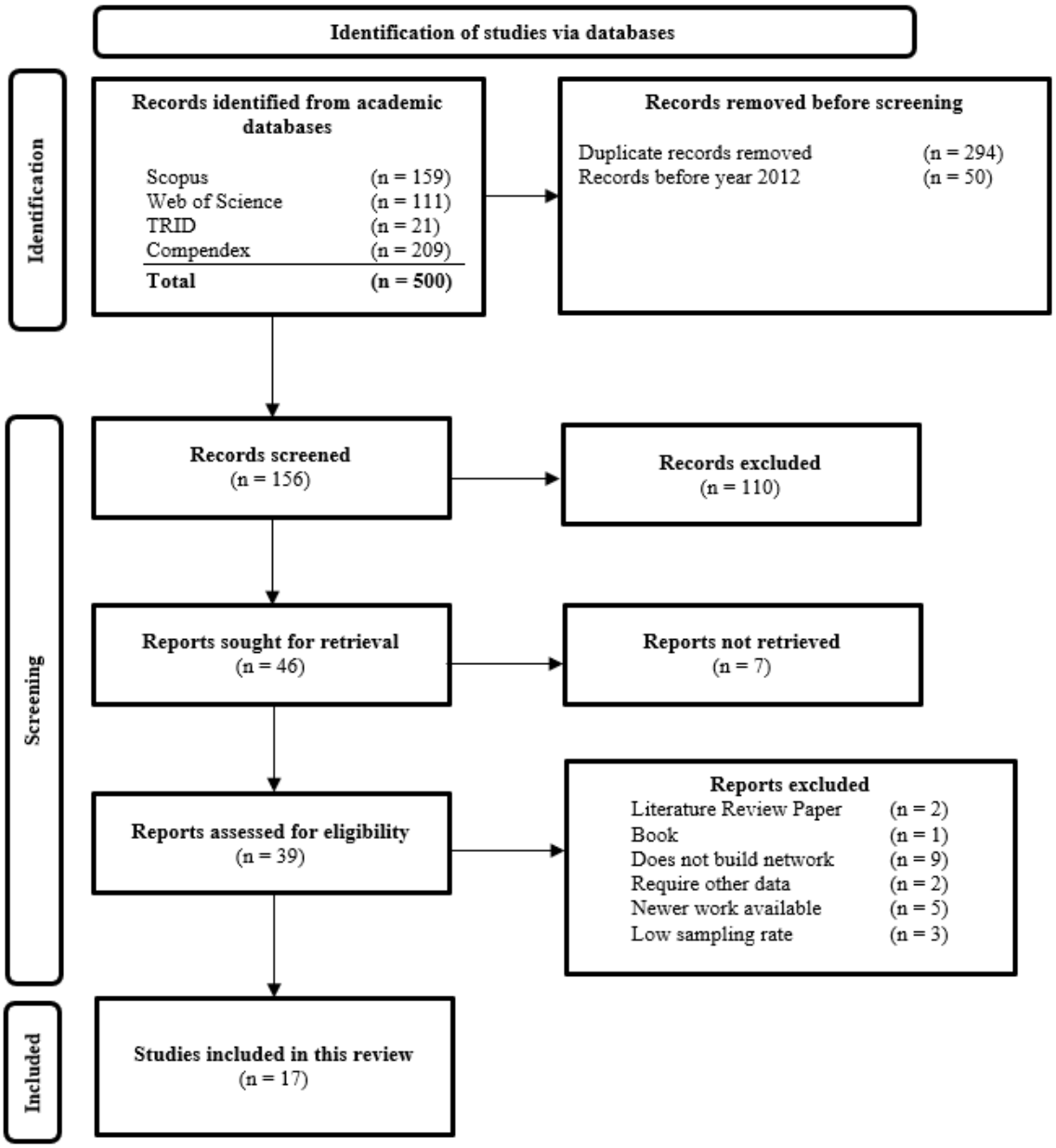


Figure 2-1. PRISMA diagram - Study identification process

Paper	Journal / Conference	Data Location	Research question(s)	Approach
Guo (2021)	Geo-spatial Information Science	Wuhan, China	Develop a novel method of extracting road maps from floating car data.	Clustering
Chen (2021)	ISPRS International Journal of Geo-Information	Shenzhen, China	Automatically generate road maps.	Clustering
Zhang (2020)	ISPRS International Journal of Geo-Information	Shenzhen, China	Incrementally extract urban road networks from spatio-temporal trajectory data.	Clustering
Arman (2020)	Procedia Computer Science	Antwerp, Belgium	Identify lanes on highway segments based on Mobile Phone GPS.	Map inference: Intersection Linking Lane detection: Gaussian Mixture Model
Zhang (2019)	ISPRS International Journal of Geo-Information	Chicago, USA and Wuhan, China	Intersection-first approach for road network generation based on low-frequency taxi trajectories.	Intersection Linking
Leichter (2019)	Applied Sciences-Basel	Joensuu, Chicago, Berlin, Athens	Fast and straightforward method for the extraction of road segment shapes from trajectories of vehicles.	Track alignment
Hashemi (2019)	IEEE Transactions on Intelligent Transportation Systems	Cary, USA, and Beijing China	Automatic inference of road and pedestrian networks from spatial-temporal trajectories.	Clustering
Daigang (2019)	ISPRS International Journal of Geo-Information	Chicago, USA and Dongguan, China	Two-stage approach for inferring road networks from trajectory points and capturing road geometry with better accuracy.	Clustering
Zhongyi (2018)	ISPRS International Journal of Geo-Information	Nanning, China	A road network generation method based on the incremental learning of vehicle trajectories.	Track alignment
Stanojevic (2018)	SIAM International Conference on Data Mining	Doha, Qatar and Chicago, USA	Inferring the road network of a city from crowd-sourced GPS traces.	Clustering
Ezzat (2018)	Journal of Computational Science	Cairo, Egypt	A clustering-based technique to extract the road map from GPS tracks.	Clustering
Dorum (2017)	ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems	San Francisco and Knoxville, USA	A comprehensive end-to-end unsupervised method based on principal curves for creating bi-directional road geometry from sparse probe data yielding a complete double-digitized road network from raw probe sources without prior map information.	Clustering
Li (2016)	ACM International on Conference on Information and Knowledge Management	Chicago, USA and Porto, Portugal	A Spatial-Linear Clustering (SLC) technique to infer road segments from GPS traces.	Clustering
Jia (2016)	ISPRS International Journal of Geo-Information	Chicago, USA and Wuhan, China	A new segmentation and grouping framework for road map inference from GPS traces.	Clustering
Xingzhe (2016)	ISPRS International Journal of Geo-Information	Chicago, USA	A method to infer the topology of the road network through intersection identification, and to extract the geometric representation of each road segment by track alignment.	Intersection Linking
Elleuch (2015)	INNS Conference on Big Data	Tunisia	Infer the geometry of road maps in Tunisia and the connectivity between them.	Clustering
Karagiorgou (2012)	International Conference on Advances in Geographic Information Systems	Athens, Greece,	Automatic road network generation algorithm that takes vehicle tracking data in the form of trajectories as input and produces a road network graph.	Intersection Linking

Table 2-1. Summary of findings

2.4.1 Clustering approach

This method uses GPS points or segments to fit the road centreline according to the data density distribution. Two main methods are used to cluster GPS data. The first covers the entire region with a grid and computes the GPS data density for each grid cell. Based on that information, it is possible to infer road segment or intersection locations. An example of density-based clustering is the Kernel Density Estimation (KDE) method used by B. Q. Chen et al. (2021).

The second method clusters the GPS data by averaging it based on proximity and direction criteria to determine road segments and intersections. Examples of this method are the k-means algorithm used by Stanojevic et al. (2018) and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) used by Ezzat et al. (2018). Eleven publications are classified under this approach of map inference. A summary of the experimental data description and validation results of these papers is presented in Table 2-2. The data collection method provides information regarding how the GPS trajectory data was collected, for example, it could be collected using GPS-enabled smartphones, commercially available GPS devices, or in-vehicle GPS trackers. Moreover, the GPS trajectory data sample size, which is the number of collected GPS points, is also presented in the table to give an idea about the scale of the sample. Finally, it is important to mention that sampling rate, or the frequency at which GPS points are collected during a trip, has a direct influence on the resolution of the GPS trajectory data and it is also reported in Table 2-2. The samples and collection methods are not the same for all the studies, therefore, the results cannot be compared directly. However, some studies have tested their algorithm on a dataset previously used in another study to enable comparability. Additionally, the work by Ahmed et al. (2015a) has tested different map inference algorithms using common

GPS trajectory datasets and compared the output using multiple indicators to provide a comparison between algorithms.

In network modelling, a detailed network model is essential to ensure the correct connectivity, topology, and capacity of roads and intersections. Therefore, road direction, turning movement permissions at intersections, and number of lanes are essential features to know. Research effort by Elleuch et al. (2015) has simply created an undirected road network without formally creating road segment and intersection representations. The produced shape of the road network is insufficient for use in road network modelling since it is missing most of the basic essential details, such as connectivity and topology. Meanwhile, several research efforts go further by generating directional road segments and intersection location (B. Q. Chen et al., 2021; Ezzat et al., 2018; Y. Guo et al., 2021; Y. F. Zhang et al., 2020). However, none of the studies implementing a clustering approach extract an explicit representation of intersection movements nor have they developed a lane-level road network, essential in determining the network's vehicular capacity.

Paper	**Sample Description (Location, Collection Method, Sample Size, Sampling Rate)**	**Validation Results**
Guo (2021)	Wuhan, China, GPS device by researchers, 1.4 million points, 20 to 60 seconds	Intersection Detection: Precision: 0.914 - 0.929 Recall: 0.787 - 0.975 F-score: 0.846 - 0.951 Road centerline extraction: Precision: 0.754 - 0.802 Recall: 0.805 - 0.812
Chen (2021)	Shenzhen, China, Taxi GPS, 75 million points, 26 seconds	Road centerline extraction: Precision: 0.966 Recall: 0.943 F-score: 0.850
Zhang (2020)	Shenzhen, China, Taxi GPS, 1.2 million points, 60 to 100 seconds	96% of extracted road length fell within 15m buffer w.r.t. ground truth
Hashemi (2019)	Cary, USA, and Beijing China, N/A, Multiple datasets, 9 to 40 seconds,	Completeness, Precision, and Topology Correctness Variable results reported for 33 datasets
Daigang (2019)	Chicago, USA and Dongguan, China, University Campus Shuttles and taxis, respectively, 118364 and 280253 points, respectively, 3.61 and 50.13 respectively	Length of extracted road: 83.6% - 87.4% Precision: 0.78 Recall: 0.6 F-score: 0.68

Paper	Sample Description (Location, Collection Method, Sample Size, Sampling Rate)	Validation Results
Stanojevic (2018)	Doha, Qatar and Chicago, USA, Fleet of vehicles with GPS-enabled devices. 5.5 million and 200 000 points, respectively, N/A	Geometry: F-score: 0.53 - 0.60 Topology: F-score: 0.80 to 0.85
Ezzat (2018)	Cairo, Egypt, Two user contributed datasets, 302 000 and 12.7 million points, 11 to 15 seconds and 1 to 3 seconds	Precision: 0.92 Recall: 0.68 F-score: 0.79
Dorum (2017)	San Francisco and Knoxville, USA, Commercial fleets and consumer devices, 43 million and 850 million points, respectively, N/A	Link Count % (reported per road type) 65% - 98.6% Link Length % (reported per road type) 71.9% - 99.4%
Li (2016)	Chicago, USA and Porto, Portugal, University Shuttles and Taxis, respectively, 118 000 and 296 573 points respectively, 3.6 seconds and more than 15 seconds, respectively	Precision: 0.68 - 0.98 Recall: 0.45 - 0.65 F-Score: 0.56 - 0.78
Jia (2016)	Chicago, USA and Wuhan, China, University Shuttles and Taxis, respectively, 118 000 and 350 000 points respectively, 3.6 seconds and more than 37.4 seconds, respectively	Precision: 0.902 - 0.975 Recall: 0.679 - 0.734 F-Score: 0.775 - 0.838
Elleuch (2015)	Tunisia, GPS receivers in 10 000 vehicles, > 100 Gb, N/A	N/A

Table 2-2. Clustering approach - sample description and validation results

Although researchers are continuously improving map inference techniques to obtain higher quality results, input data characteristics remain a main determinant of output quality. The variety of data sources used in the 11 studies makes it difficult to compare them and determine the best map inference method. This is caused by the differences in GPS data collection devices (in-vehicle, GPS enabled smartphone, GPS tracker, etc.), differences in sampling rates, differences in the number of points or trajectories available, and differences in collection environments (various levels of GPS signal interference and availability). For Example, B. Q. Chen et al. (2021) uses a dataset of 75 million points collected by taxi GPS devices in Shenzhen, China with an average sampling rate of 26 seconds, while one of the two datasets used by Daigang et al. (2019) is composed of 118 000 points collected by university shuttles in Chicago, United States at an average sampling rate of 3.6 seconds. The same algorithm applied to both datasets can result in different output quality levels. GPS data used in most of the studies was obtained using

GPS-equipped taxis or shuttles, which introduces bias by not representing an average motorist's behavior. In the case of shuttles, this bias can be in terms on spatial coverage since they have fixed routes and might also be permitted to drive on private roads such as campuses. Thus, the inferred map based on this data might not reflect the whole network available to all motorists. Additionally, shuttles usually have a fixed schedule and cannot provide a good temporal coverage for all periods of the day. On the other hand, GPS-equipped taxis can have adequate temporal coverage, however, some road networks have dedicated lanes and turning permissions for taxis to encourage their use. Therefore, this introduces some spatial bias if the extracted network is to be used by a private motorist.

In the study by Elleuch et al. (2015), insufficient information was provided regarding the experimental data. In parallel, some researchers have used well known benchmark datasets to enable the comparability of their algorithm's performance. For example, some researchers have evaluated the execution of their algorithms on the Chicago dataset (Daigang et al., 2019; Jia & Ruisheng, 2016; H. F. Li et al., 2016; Stanojevic et al., 2018). However, this dataset is obtained from university shuttles and has spatial and temporal limitations.

The most common evaluation method, initially introduced by Biagioni and Eriksson (2012), was the harmonic mean of precision and recall, also known as F-score or F-value. It is calculated as follows:

$$Precision = \frac{Correctly\ Extracted}{Correctly\ Extracted + Incorrectly\ Extracted}$$

$$Recall = \frac{Correctly\ Extracted}{Correctly\ Extracted + Not\ Extracted}$$

$$F - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

where *Correctly Extracted + Incorrectly Extracted = Extracted* or inferred network elements and *Correctly Extracted + Not Extracted = Ground Truth*. A higher F-score (closer to one) indicates a better inference and match to the ground truth. Typically, the ground truth was selected to be an open-source map from Open Street Maps, a road map that relies on the public for update. Using it as the ground truth assumes that does not contain errors, which is not always true. Therefore, this introduces a bias in the output quality measurement.

Distance and direction angle thresholds are used to determine if two elements (road segments or intersections) match. In addition, the sampling can be in terms of points or entire segments. For example, H. F. Li et al. (2016) samples every segment (or link) while Biagioni and Eriksson (2012) sample points throughout the inferred and ground truth networks. Eight papers out of eleven use this indicator to quantify the output network quality, while Dorum (2017) and Y. F. Zhang et al. (2020) only report recall values. Recall values are unable to quantify the number of network elements that were incorrectly extracted. The study by Elleuch et al. (2015) does not report any quantitative measures, which does not allow the author to assess the output quality.

The output quality assessment was also reported for different threshold values with lower thresholds making the ground truth matching stricter. This explains the different values presented for precision, recall, and F-score for a given method.

As presented in Table 2-2, the method proposed by B. Q. Chen et al. (2021) for centerline extraction achieved the highest F-score (0.850), followed closely by Jia and Ruisheng (2016)

(0.838). Meanwhile, the method proposed by Daigang et al. (2019) resulted with the lowest F-score (0.68).

Overall, F-score is found to be the best indicator method output quality since it takes into account the number of correctly extracted, incorrectly extracted, and not extracted network features.

Most studies are easy to read and understand and graphics, tables, and GIS components are relatively well presented (Dorum, 2017; Ezzat et al., 2018; Y. Guo et al., 2021; Jia & Ruisheng, 2016; Y. F. Zhang et al., 2020). However, only the works by Hashemi (2019) and Ezzat et al. (2018) are presented in a reproducible fashion.

2.4.2 Intersection linking approach

This approach divides the network inference process into two main steps: 1) detecting intersections using the GPS data, for example, based on turning angles, 2) using GPS trajectories to link the intersections together and form a network.

This technique can be seen in (Karagiorgou & Pfoser, 2012; Xingzhe et al., 2016; C. Zhang et al., 2019). A variation is presented by Arman and Tampere (2020) where intersections are determined by finding merge and diverge locations. In fact, this paper also uses the Gaussian Mixture Method to estimate the number of lanes based on the distribution of GPS points within a road segment.

Four publications are classified under this approach of map inference. A summary of the experimental data description and validation results of these papers is presented in Table 2-3.

Paper	Sample Description (Location, Collection Method, Sample Size, Sampling Rate)	Validation Results

Arman (2020)	Antwerp, Belgium, Mobilis smartphone app, 21 100 trajectories, 1 second	On average within 4% in term of speed and 14% in term of lane share w.r.t ground truth
Zhang (2019)	Chicago, USA and Wuhan, China, University Shuttles and Taxis, respectively, 118 364 and 800 000 points respectively, 3.6 seconds and more than 40 seconds, respectively	Intersection Detection: more than 90% Road centerline extraction: Precision: 0.932-0.980 Recall: 0.704 - 0.886 F-score: 0.820 - 0.908
Xingzhe (2015)	Chicago, USA, University Shuttles, 118 000 points, 3.6 seconds	Intersection Accuracy: F-Score: 0.02 - 0.91 Connectivity Accuracy: F-Score: 0.19-0.95
Karagiorgou (2012)	Athens, Greece, GPS devices, N/A, 30 seconds	Shortest paths comparison

Table 2-3. Intersection linking approach - sample description and validation results

The intersection linking approach has the advantage of explicitly defining intersections by default, since it is the first step of the method. The four papers produce a directional road network. While three of the methods infer road centerlines, the work by Arman and Tampere (2020) is the only one to propose a method that determines the number of lanes. Intersection movements are only determined using the methods proposed by Karagiorgou and Pfoser (2012) and Xingzhe et al. (2016).

Different GPS data sources were used to propose intersection linking map inference methods. The Sampling rate varies between one second and thirty second in the works by Arman and Tampere (2020) and Karagiorgou and Pfoser (2012), respectively. Meanwhile, Xingzhe et al. (2016) and C. Zhang et al. (2019) use the same benchmark dataset, which enables their comparability. It is important to note that the work by Arman and Tampere (2020) limits the experiment to a small section of a highway corridor. This is insufficient to determine if the proposed method will perform well in more complex environments.

Network inference quality was evaluated using three different methods. Arman and Tampere (2020) compared the results with speed and count data while Karagiorgou and Pfoser (2012) used a shortest path based distance. In fact, this measure computes the shortest path distance for a set of OD pairs for both inferred and ground truth maps. The similarity between these distances indicates a similarity between the two maps in terms of geometry and connectivity. This method is not deterministic and can lead to false similarity conclusions. The final two papers by Xingzhe et al. (2016) and C. Zhang et al. (2019) use the harmonic mean of precision and recall, to assess the output quality. Both methods produce a very good F-score (>0.90), however, the method proposed by Xingzhe et al. (2016) has a high variability in the output quality. In terms of clarity, methods proposed by Karagiorgou and Pfoser (2012) and C. Zhang et al. (2019) are well explained. However, only the work by Karagiorgou and Pfoser (2012) contains sufficient details to be deemed reproducible.

2.4.3 Track alignment approach

Map inference using track alignment incrementally adds GPS tracks to an initially empty map. This approach can also be seen as an incremental averaging of the GPS tracks. Two publications are classified under this approach of map inference. A summary of the experimental data description and validation results of these papers is presented in Table 2-4.

The proposed methods focus on extracting a directional road network, represented by the centerline of the road. Therefore, intersections are not formally defined, and the number of lanes information is not determined.

In Zhongyi et al. (2018), experimental GPS data is obtained from a logistics company trucks. The use of truck GPS data can introduce a bias in terms of road coverage, as trucks are usually limited to drive on a subset of the entire road network due to their size, nuisance, and material they transport. The work by Leichter and Werner (2019) does not specify the experimental data details. In fact, this paper was written as part of competition oriented towards map inference algorithms efficiency and speed.

The inferred map quality was not evaluated by Zhongyi et al. (2018) since no ground truth was available. Meanwhile, Leichter and Werner (2019) evaluated the quality of inferred map using the HC-SIM, which measures the overlap of two lines (inferred and ground truth). An HC-SIM measure of 0.612 was obtained which ranked this method among the best in the competition. The explained methods lack some details to be fully understandable. The work by Leichter and Werner (2019) does not present the algorithm, while Zhongyi et al. (2018) does not present sufficient description, figures, and diagrams. Therefore, none of the two works is reproducible.

Paper	Sample Description (Location, Collection Method, Sample Size, Sampling Rate)	Validation Results
Leichter (2019)	Joensuu, Chicago, Berlin, Athens, N/A, Multiple datasets, N/A	HC-SIM of around 0.66
Zhongyi (2018)	Nanning, China, Logistics company trucks, 451 537 points, 10 seconds	N/A (no ground truth)

Table 2-4. Track alignment approach - sample description and validation results

2.5 Discussion

A detailed road network representation is essential for multiple tasks such as traditional navigation, autonomous vehicle navigation, and transport modelling. A transport model relies on the road network model as one of its main components. In more detail, the road network

representation needs to accurately depict the road's geographic location, direction, type, number of lanes, connectivity, and intersection control type, and permitted turning movements. Additionally, the actual road network is dynamic in nature, since traffic rules can prohibit a subset of road users from using a specific road lane or segment or making a specific movement at an intersection, depending on the temporal criteria. Therefore, the modelled road network should also consider this characteristic.

The reviewed studies demonstrate that research has been carried out on the topic of road network feature extraction. This review found that two main approaches are the most popular: clustering and intersection linking, as can be seen in Table 2-3 and Table 2-4. They can reconstruct a road network model from GPS data with high accuracy (Y. Guo et al., 2021). However, it is not possible to conclude if one approach is better than the other since within one approach, different methods achieve different accuracies. Moreover, different methods have used GPS data from different sources and different validation methods which makes them not directly comparable. The work done by Ahmed et al. (2015a) tested the main algorithms using different GPS trajectory datasets and compared the output quality using different indicators. They found that, in general, algorithms that produce maps with higher accuracy have a lower coverage and the opposite is also true. However, the algorithm by Karagiorgou and Pfoser (2012) produced maps that have good accuracy and coverage.

The reviewed research used multiple measures to evaluate the accuracy of the constructed networks in comparison to ground truth maps. The most relevant and common measure was the F-score introduced by Biagioni and Eriksson (2012). It evaluates the similarity between the extracted network and the ground truth by relating the number of correctly extracted features,

with the number of incorrectly extracted features and the number of unextracted features. Although these findings are a good basis for road network features extraction from GPS, the following limitations were noted and need to be addressed in future research to be able to extract road network models usable in transport modelling and autonomous vehicle navigation:

- The constructed network is only a representation of directed road centrelines, and intersection locations. This level of detail is insufficient for road network model requirements as described above.
- Given the multitude of GPS data sources used in past research to extract network features, it is impossible to select the best method simply based on the F-score. In fact, GPS data used in the studies was obtained via shuttles, taxis, trucks, fleets, researcher initiative, or crowdsourcing. This results in variable spatiotemporal sampling characteristics rendering a direct comparison of the results impossible. Ideally, all methods should be evaluated using the same GPS sample and compared to the same ground truth.
- Not all GPS data sources provide the same level of road network representativity. For example, using GPS data collected by a specific fleet such as trucks, transit vehicles, or shuttles introduces bias with respect to the type of roads or routes that are permitted for them. Multiple studies used university shuttles to extract road network features, the most recent being the effort by Daigang et al. (2019). This limits the coverage of the extracted network features to fixed routes or road types.
- Several studies were found to be irreproducible since the method is not clearly detailed or due to data unavailability.

These limitations need to be addressed to extract road network features with sufficient detail for use in transport simulation models. The following steps can help achieving this goal and contribute to the current research:

- The use of large GPS datasets collected by light private vehicles to reduce the road network coverage bias.
- The development of methods to extract road segment related features from GPS data such as road type, posted speed, and number of lanes.
- The development of methods to extract intersection related features from GPS data such as turning movement permissions and control type.
- The consideration of the dynamic nature of the road network which affects road segment or intersection related variables.
- Making detailed and reproducible methodology available for future researchers to build on.

2.6 Conclusion

This paper extends past literature reviews by viewing the map inference problem from the transport network modelling point of view. The search strategy was shared to render the search reproducible. It has been found that two main approaches are popular to extract network features from GPS data. However, the extracted output is limited to the road centreline, including directionality, and intersection locations. It was also found that the main accuracy indicator used to assess the similarity between the extracted network and the ground truth is the F-score. Additionally, some of the reviewed methods achieve high, but improvable accuracy.

GPS data, depending on its sampling coverage and frequency is rich and can be further explored to extract more detailed road network features. For example, future research can explore the extraction of road segment type, posted speed and number of lanes in addition to intersection control type and turning movement permissions. Being able to extract all road network features required for large scale transport modelling from GPS data will be of immense value as it will improve model quality and update frequency while reducing the required resources. Such data will be valuable for accurate navigation systems of automated vehicles.

Acknowledgements

The author would like to acknowledge the generous support of McGill University's Faculty of Engineering and the Vadasz Scholars Program.

Disclosure Statement

The authors declare that there are no financial or non-financial competing interests to report.

References

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. 2015a. A comparison and evaluation of map construction algorithms using vehicle tracking data. Geoinformatica, 19(3). 31

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. 2015b. Map Construction Algorithms. Springer Publishing Company, Incorporated.

Arman, M. A., & Tampere, C. M. J. 2020. Road centreline and lane reconstruction from pervasive GPS tracking on motorways. Procedia Computer Science, 170. 8

Banqiao, C., Ding, C., Wenjuan, R., & Guangluan, X. 2020. Extended Classification Course Improves Road Intersection Detection from Low-Frequency GPS Trajectory Data. ISPRS International Journal of Geo-Information, 9(3). 181 (120 pp.)

Biagioni, J., & Eriksson, J. 2012. Inferring Road Maps from Global Positioning System Traces Survey and Comparative Evaluation. Transportation Research Record(2291). 61-71

Chao, P., Hua, W., Mao, R., Xu, J., & Zhou, X. 2022. A survey and quantitative study on map inference algorithms from GPS trajectories. IEEE Transactions on Knowledge and Data Engineering, 34(1). 14

Chen, B. Q., Ding, C. B., Ren, W. J., & Xu, G. L. 2021. Automatically tracking road centerlines from low-frequency gps trajectory data. ISPRS International Journal of Geo-Information, 10(3). 26

Daigang, L., Junhan, L., & Juntao, L. 2019. Road network extraction from low-frequency trajectories based on a road structure-aware filter. ISPRS International Journal of Geo-Information, 8(9). 17

Dorum, O. H. 2017. Deriving double-digitized road network geometry from probe data. 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

Elleuch, W., Wali, A., & Alimi, A. M. 2015. An investigation of parallel road map inference from big GPS traces data. 2015 INNS Conference on Big Data, San Francisco, CA, United states.

Ezzat, M., Sakr, M., Elgohary, R., & Khalifa, M. E. 2018. Building road segments and detecting turns from GPS tracks. Journal of Computational Science, 29. 13

Guo, Y., Li, B., Lu, Z., & Zhou, J. 2021. A novel method for road network mining from floating car data. Geo-spatial Information Science. 16

Hashemi, M. 2019. Automatic inference of road and pedestrian networks from spatial-temporal trajectories. IEEE Transactions on Intelligent Transportation Systems, 20(12). 17

Jia, Q., & Ruisheng, W. 2016. Road map inference: A segmentation and grouping framework. ISPRS International Journal of Geo-Information, 5(8). 20

Karagiorgou, S., & Pfoser, D. 2012. On vehicle tracking data-based road network generation. 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, United states.

Leichter, A., & Werner, M. 2019. Estimating road segments using natural point correspondences of GPS trajectories. Applied Sciences-Basel, 9(20). 11

Li, H. F., Kulik, L., Ramamohanarao, K., & Acm. 2016. Automatic generation and validation of road maps from GPS trajectory data sets. 25th ACM International on Conference on Information and Knowledge Management.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. 2016. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 5(1). 210

PRISMA. 2015. Transparent reporting of systematic reviews and meta-analyses. PRISMA Ottawa, ON, Canada http://www.prisma-statement.org/

Stanojevic, R., Abbar, S., Thirumuruganathan, S., Chawla, S., Filali, F., & Aleimat, A. 2018. Robust road map inference through network alignment of trajectories. 2018 SIAM International Conference on Data Mining (SDM), San Diego, CA, United states.

Xingzhe, X., Wenzhi, L., Aghajan, H., Veelaert, P., & Philips, W. 2016. A novel approach for detecting intersections from GPS traces. Piscataway, NJ, USA.

Zhang, C., Xiang, L., Li, S., & Wang, D. 2019. An intersection-first approach for road network generation from crowd-sourced vehicle trajectories. ISPRS International Journal of Geo-Information, 8(11). 26

Zhang, Y. F., Zhang, Z. X., Huang, J. C., She, T. T., Deng, M., Fan, H. C., Xu, P., & Deng, X. S. 2020. A hybrid method to incrementally extract road networks using spatio-temporal trajectory data. ISPRS International Journal of Geo-Information, 9(4). 15

Zhongyi, N., Lijun, X., Tian, X., Binhua, S., & Yao, Z. 2018. Incremental road network generation based on vehicle trajectories. ISPRS International Journal of Geo-Information, 7(10). 19

Chapter 3 - Creating a Road Network Model from Global Positioning System Trajectory Data for Macroscopic Simulation

Creating a Road Network Model from Global Positioning System Trajectory Data for Macroscopic Simulation

Adham Badran [a], Ahmed El-Geneidy [b], Luis Miranda-Moreno [a]

[a] Department of Civil Engineering, McGill University, 817 Sherbrooke Street West, Montreal H3A 0C3, Canada

[b] School of Urban Planning, McGill University, 815 Sherbrooke Street West, Montreal H3A 0C2, Canada

3.1 Abstract

Emergence of road users' global positioning system (GPS) trajectory data is increasing research interest in knowledge discovery to improve transport planning related methods and tools. In fact, the widespread use of GPS enabled smartphones and mobile internet has increased the availability and size of such data. With the increase in GPS data coverage and availability, some research has expanded its use to estimate state-wide vehicle-miles travelled, to classify driving maneuvers for road safety assessment, or to estimate environmental performance indicators, such as vehicular fuel consumption and pollution emissions. In computer science, research has used GPS data to infer road network maps. Although the inferred maps provide a correct topology and connectivity, they lack essential details to be used for transport modelling. Therefore, this work proposes a method to extract network-wide road direction and turning movement rules. In addition, it serves as a proof of concept by building a road network model under a widely used macroscopic transport modelling software, EMME. Sensitivity analysis was carried out to determine the output quality and recommend future improvements. Road

segment geometry and directionality were extracted accurately, however, turning movement rules can be extracted more accurately using a larger GPS trajectory sample.

Keywords: GPS, Transport Model, Road Network, Intersection Control, Map Inference, Road Direction, Turning Movement, EMME.

3.2 Introduction

Transport network modelling requires large quantities of data, depending on the project size and level of detail. For example, building a micro-simulation network model for a neighborhood, requires detailed road geometry, road type, transport demand matrices, intersection control type and traffic light phasing, to name a few. The model results also require validation, usually done by comparing the model output to traffic counts and observed travel times and delays. This data is collected by different means and for different sample sizes depending mainly on modelling needs and available resources.

Traditionally, in the transport field, global positioning system (GPS) data was obtained from floating vehicles and probe vehicles to estimate travel time, queue length, and traffic volume as in the works by Zito and Taylor (1994) and Zhao et al. (2019). This technique estimates trip characteristics for a specific fleet or for predefined corridors which can introduce bias when the sample is limited spatially (only a few corridors) or in terms of fleet (only buses, taxis, or commercial vehicles). In one study, Tantiyanugulchai and Bertini (2003) used GPS-equipped transit vehicles to determine if transit vehicle speeds and travel times are a good proxy for general traffic conditions to be used in real-time advanced traffic management and traveler information systems. A second study by El-Geneidy and Bertini (2004) used probe transit vehicle

GPS data to determine the optimal time resolution and speed measure to report traffic conditions obtained from loop detector data. Although these methods are useful to answer specific questions, Mennis and Guo (2009) found that an increase in the sample size and the coverage of GPS data enables researchers to perform data mining and increase geographic knowledge discovery.

Recently, the widespread use of GPS enabled smartphones and mobile internet made collecting and saving GPS data simple and relatively inexpensive. As this data is becoming more widely available, it is attracting a lot of research interest in the transport field. High-sample GPS databases are being built and knowledge discovery research from GPS data has already started. For example, Fan et al. (2019) examined the use of GPS data to estimate vehicle miles travelled within the state of Maryland in the United States. In another study, Phondeenana et al. (2013) used GPS data to classify driving maneuvers to improve road safety. In the environmental field, studies proposed methods to estimate congestion, vehicle fuel consumption, and pollution emissions using GPS data (Gately et al., 2017; Kan et al., 2018; Lin et al., 2019).

In parallel, computer science and geography researchers have been exploring the use of GPS data to infer road network's geometry, topology, and connectivity. Some studies have compared the different algorithms to infer road network from GPS data. These algorithms were divided into three categories: point clustering, intersection linking, and incremental track insertion. Through a clustering approach, few studies in China have developed techniques to automatically extract the road network from GPS points or segments (B. Q. Chen et al., 2021; Y. Guo et al., 2021; Y. F. Zhang et al., 2020). The main idea was to fit the road centerline according to the GPS data density distribution. Other researchers explored intersection linking techniques to generate road

segments from vehicle trajectory data (Karagiorgou & Pfoser, 2012; Xingzhe et al., 2016; C. Zhang et al., 2019). This approach divides the network inference process into two main steps: 1) detecting intersections using the GPS data, for example, based on turning angles, 2) using GPS trajectories to link the intersections together and form a directed road network. Finally, the track alignment approach was studied by some researchers(Leichter & Werner, 2019; Xingzhe et al., 2015; Zhongyi et al., 2018). This technique incrementally adds GPS tracks to an initially empty map and can also be seen as an incremental averaging of the GPS tracks. These techniques can serve as the base for future research that aims to build a detailed road network for transport modelling or autonomous driving. In fact, road network building is a very active research area in autonomous driving. Providing detailed network features is essential for autonomous vehicles' operation since they require precise knowledge of network topology and geometry. For example, Bender et al. (2014) aimed to develop the first map model usable by autonomous vehicles by representing road lanes and intersections not only in terms of directional lines but also in terms of drivable surfaces by introducing right and left bounds. The map model also needed to integrate driving rules.

Although, the past developed work is useful for generating simple road networks, based on GPS data, with a correct topology and connectivity, there is a need to develop methodologies that help extract detailed network features for transport network modelling. For example, map inference methods are lacking the ability to extract detailed information of network-wide features such as turning movement permissions at intersections or road segment number of lanes, essential input data for transport network models. The development of macroscopic

network models is labor- and data-intensive. Therefore, the development of automated methods can help reduce significantly the resources required in the transport modeling tasks.

The objective of this work is to develop a method to extract road network features from GPS Data to be used in transport network modelling. In addition, it aims to provide proof of concept by building a road network model under a widely used macroscopic transport modelling software, EMME. More precisely, GPS data is used to extract network-wide road direction and turning movement information. This information is essential in transport modeling and land use studies when the study area is large, and network features cannot be collected as efficiently using other methods.

3.3 Methods

Four main input datasets are used: 1) GPS trajectory points, 2) Geographic representation of the road network, 3) Geographic location of all intersections, and 4) Google maps and Street View. GPS data was collected during the spring of 2014 in Quebec City, Canada. It was collected during 21 days by 2000 voluntary users through the Mon Trajet smartphone app, made available by the Municipality. Each point is described by the following attributes: map matched X and Y coordinates, trip ID, speed, and timestamp (Year-Month-Day-Hour-Minute-Second). Figure 3-1 is a map of the raw GPS points (226,000 points) inside the study zone, which consists of 81 intersections.

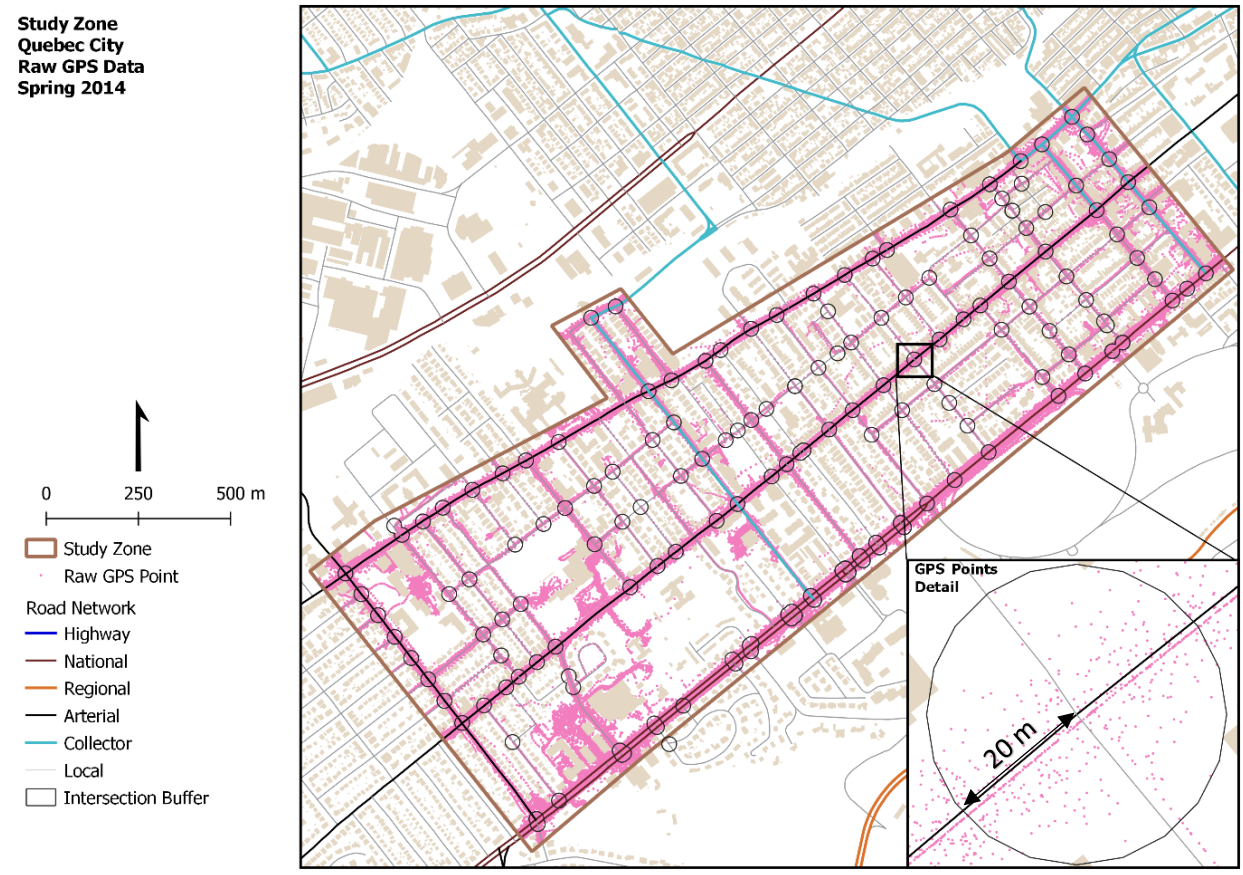


Figure 3-1. Study area – GPS data points

A shapefile file of the up to date study area road network was available online ("Adresses Québec - AQréseau," 2022). The location of all intersections was obtained from the municipality ("Données ouvertes ", 2022). Finally, Google maps and street view were used to validate the results by serving as the ground truth for road segment direction and intersection movement permissions.

Completing this study was possible using QGIS, ArcGIS, FME, and EMME. QGIS was used to visualize geographic data, perform visual validations, and create maps. While travel paths were constructed using the network analyst extension in ArcGIS. FME was used to manipulate data and perform geographic operations. Finally, EMME is the macrosimulation software used to build the road network model.

Extracting road network features from GPS data can be divided into three main parts: 1) initialization, 2) link direction extraction, and 3) turning movement permission extraction.

3.3.1 Initialization

The first part of the process is the initialization. It consists of creating an initial base network using the EMME software and the simple road network shapefile. This creates a digital network representation composed of links and nodes (see Figure 3-2). Each node is uniquely identified and located exactly at the intersections depicted in the simple road network shapefile. On the other hand, links are created assuming that all roads are two-way streets, and each link is represented using its origin and destination nodes. It should also be noted that the initial road network model created by EMME allows all movements at intersections except for U-turns.

In parallel, the GPS points are filtered to remove outliers, defined as consecutive points separated by more than 30 meters. This threshold was determined following the visual inspection of GPS trajectory points. The outlier removal created small gaps within the GPS trajectories which were connected using a shortest path algorithm and the simple road network shapefile using the network analyst extension in ArcGIS. This produced full trip trajectories that were geographically snapped to the simple road network, which enabled the following geographic processing steps.

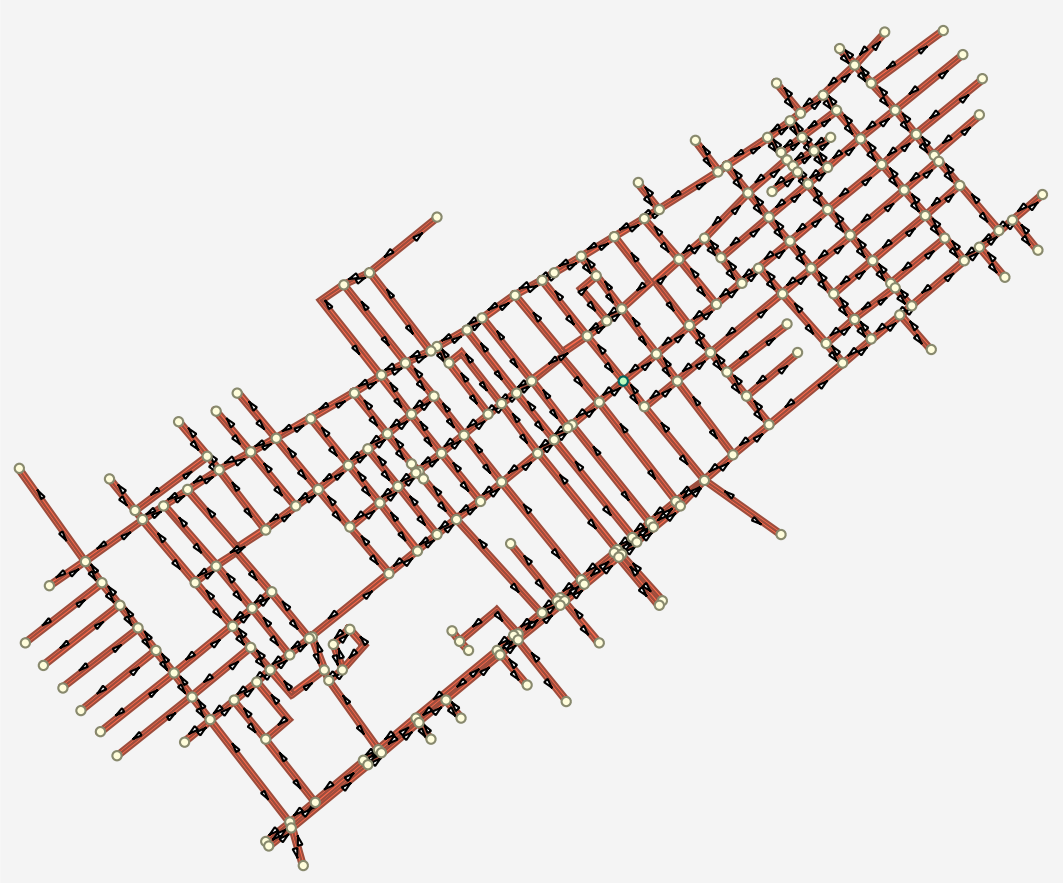


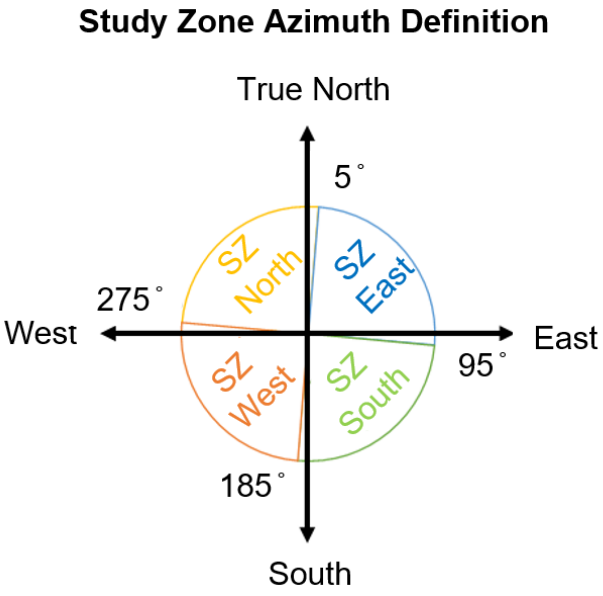
Figure 3-2. EMME initial road network model – links and nodes

3.3.2 Link direction extraction

The road network model created during the initialization step assumed 2-way links for all road segments. However, this is not always true as some roads are only one-way. The link direction extraction process aims to extract the directionality information from the observed GPS data to remove modelled links that do not exist.

Following the initialization phase, trip trajectories are divided into straight segments for which segment azimuth is calculated. The azimuth corresponds to the angle between the segment orientation and the North measured clockwise. After examining the study area, a direction dictionary was created to associate different azimuth ranges to cardinal directions (Figure 3-3). Each segment was then associated to a cardinal direction depending on its azimuth. The same

procedure was applied to the initial EMME link table to determine link directions. Once the directions are determined for the GPS trajectory segments and EMME links, a geographic operation was carried out to determine the nearest link for each GPS trajectory segment. The segment direction was used as a criterion to only select the nearest link with the same direction.



Azimuth angle range (degrees)			Direction
275	-	4	North
5	-	94	East
95	-	184	South
185	-	274	West

SZ: Study Zone

Figure 3-3. Correspondence between azimuth angle and direction

For each link, the number of observed GPS segments associated to it was calculated and put in relation with the number of segments associated to the reverse link by computing their ratio. For example, a ratio value of 0.05 (or 5%), for a given link, signifies that the number of GPS segment observations for that link is equal to 5% of the observed number of segments on the reverse link. This indicates a high likelihood of that link (or direction of travel) to not exist, since it is expected to have a similar count magnitude in both directions for a given road.

To determine the optimal ratio value indicating the presence/absence of a link, a sensitivity analysis was carried out by testing different ratio value limits between 1% and 10% and

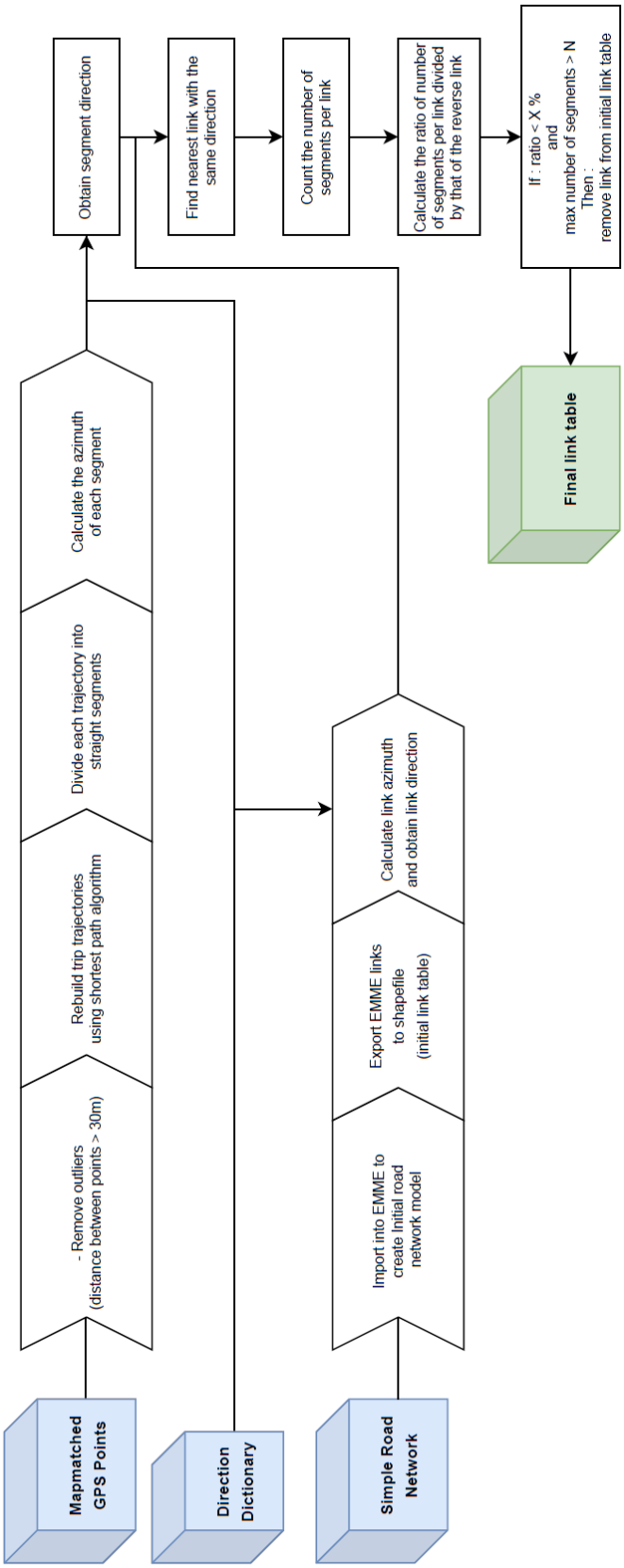


Figure 3-4. Link direction extraction process

comparing the results with the ground truth obtained from Google Maps. For a given ratio limit value limit, a link having a ratio value smaller than the ratio limit value is considered to not exist, while a ratio value greater or equal to the ratio limit value is considered to exist. Once the prediction is made for each link, the accuracy was calculated as the number of correct link direction predictions divided by the total number of links. The ratio limit value producing the highest accuracy was selected as the optimal ratio limit value. Once the optimal value was determined, the absolute number of observed segments for each road was analyzed to determine the impact of sample size on link direction prediction accuracy. A second sensitivity analysis was performed by introducing different segment count cutoff thresholds. In other words, prediction accuracy was computed on a subset of links that have at least a minimum number of observed segments in one of the two link directions. The tested cutoff threshold values were: 0 (or all links),10, 20, 30, 40, 50, 100, and 200. At each of the cutoff thresholds, link directions were predicted and compared with the ground truth to calculate prediction accuracy. A summary of the steps, including initialization, is presented in Figure 3-4.

3.3.3 Turning movement permission extraction

Having created a correct link and node representation of the road network, the following step was to determine the permitted turning movements at intersections. In the initial road network model created using EMME, all intersection turning movements are allowed except for U-turns, however, the objective of this step is to extract and allow only the turning movements that were observed within the GPS trajectories. Figure 3-5 presents a summary of the process to extract turning permissions from observed GPS trajectories.

Since not all nodes are intersections, the intersection locations obtained from the Municipality were used to create 20-meter radius buffers, selected through inspection of the study area, and select the nodes within these buffers as intersection nodes. The selected nodes were then used to create new 3-meter radius buffers. These intersection node buffers were used to clip only the parts of GPS trajectories located within each buffer. Since the modelled road network is geographically based on the simple road network and the GPS trajectories were snapped on the same simple road network during the shortest path operation, a good superposition of both geographic features was ensured. The clipping operation removed GPS trajectory segments that were considered to not be intersection movements. The remaining trajectory segments within the node buffers were then divided into two segments, inbound (towards the node) and outbound (outwards from the node). The following step was to determine the azimuth for the inbound and outbound segments per node per trajectory segment. The azimuth was then used to determine the direction of every segment using the correspondence between the azimuth angle and the direction established in the previous step (see Figure 3-3). The next geographic operation was to find the nearest link to each inbound and outbound segment while insuring a matching direction between them. At this point each trajectory with an inbound and outbound segment, within a node buffer, is associated to two links (inbound and outbound) and can be expressed in terms of intersection node, origin node (from the inbound link) and destination node (from the outbound link). A compilation of all observed movements at the different nodes provides the number of times that each movement has been made. A turning movement was predicted to be permitted if there was at least one observation from the GPS trajectories for that specific movement.

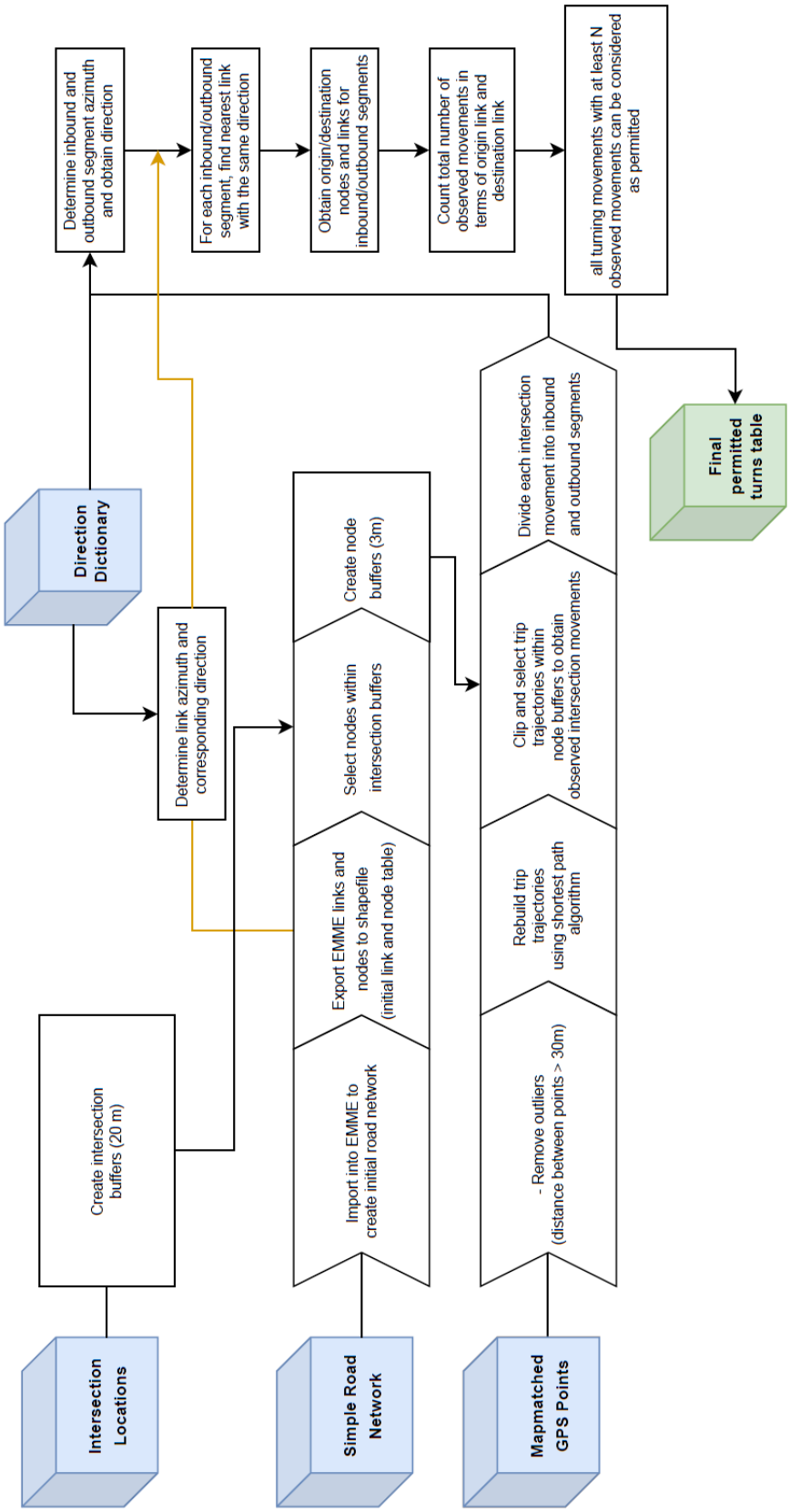


Figure 3-5. Turning permission extraction process

To determine intersection movement prediction accuracy, the extracted turning movements for a subset of nine intersections (90 turning movements) were compared to the ground truth obtained from Google Maps and Street View. A sensitivity analysis was performed to assess the effect of sample size on prediction accuracy by evaluating prediction results for turning movements that have at least one and two extracted observations.

3.4 Results

3.4.1 Link direction extraction accuracy

The highest link direction extraction accuracy was 95% obtained using a ratio limit value of 5%. In other words, a link that has a GPS segment count less than 5% of that of the reverse link can be considered to not exist with a 95% accuracy.

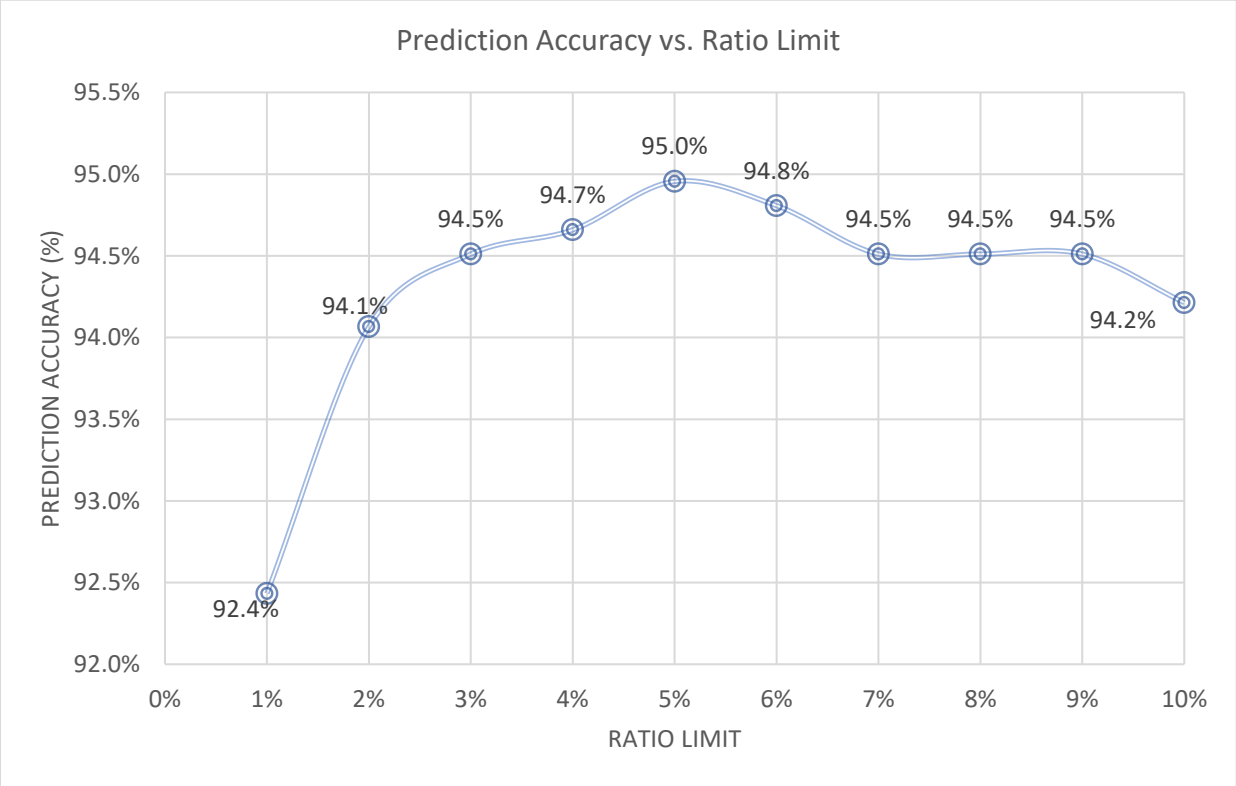


Figure 3-6. Link direction prediction accuracy - sensitivity analysis

The sensitivity analysis results for the different ratio limit values are presented in Figure 3-6.

Following the selection of the optimal ratio limit value (5%) an attribute was added to the initial modelled road network to indicate whether the directional road segment, represented by a link, exists or not. The resulting link representation of the road network is presented in Figure 3-7. Links presented in blue were determined to be non-existent since they do not have enough GPS segment observations compared to the reverse link (ratio < 5%). A special case is also presented in caption A of Figure 3-7 where the initial road network model was created as four parallel links (compared to two links in regular situations). This is explained by the way the simple road network, used as an input, represented that road. Since it has a large median, it was represented

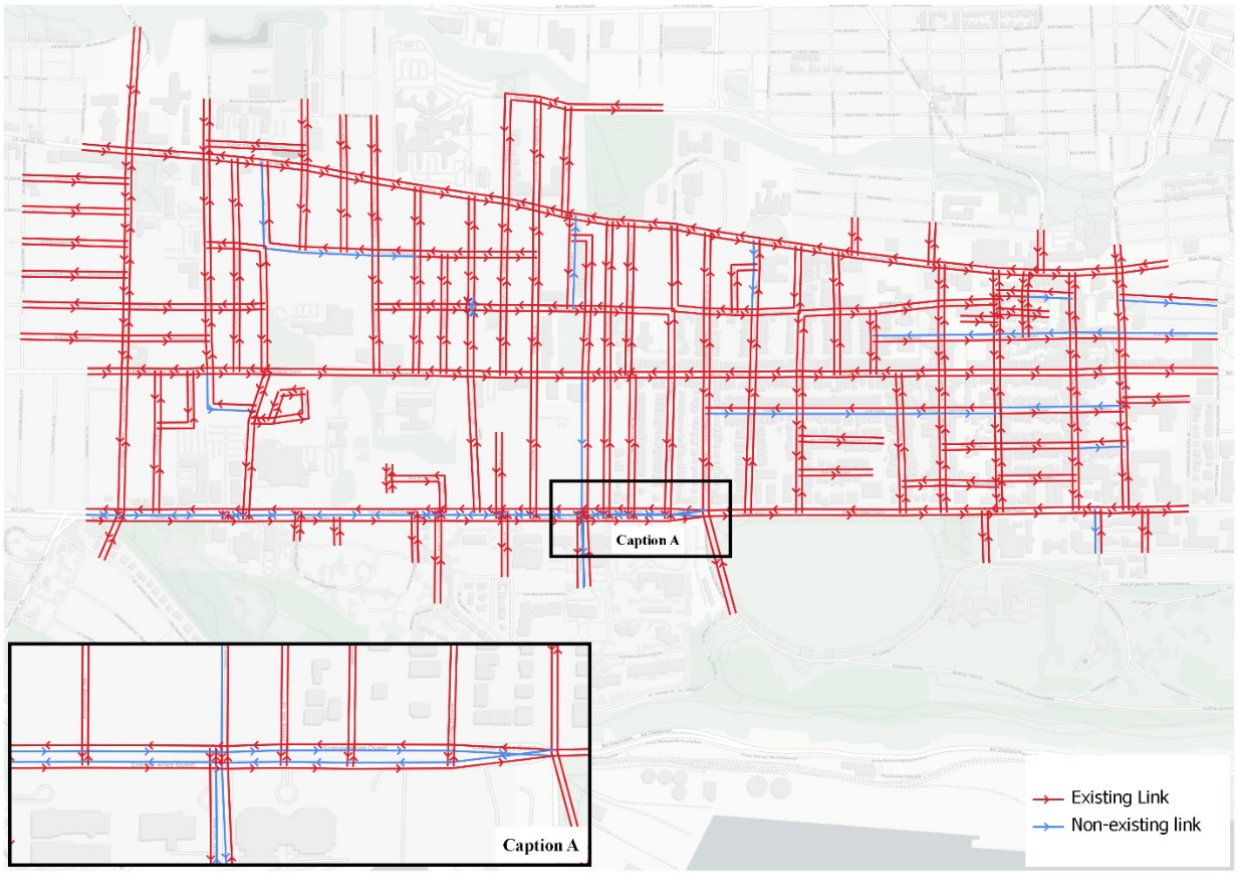


Figure 3-7. Final network model – extracted link result

as two lines in the simple road network and therefore understood as two different roads by EMME. However, the developed method was able to determine which link corresponds to an existing road segment and filter the non-existing links.

Considering sample size in the prediction accuracy assessment was found to have an impact. The introduction of a threshold on the segment count ensured that only roads with a minimum number of GPS segment count were considered. The results in Figure 3-8 show that increasing the minimum threshold is correlated with an increase in link direction prediction accuracy. For example, using a minimum threshold of 200 segments for a given road segment results in a prediction accuracy of 98.7% as opposed to not having a minimum threshold which results in 95% accuracy. However, for this study area and GPS dataset, setting the highest threshold implies that prediction can only be made for 304 links instead of all the 674 links as seen in Table 3-1.

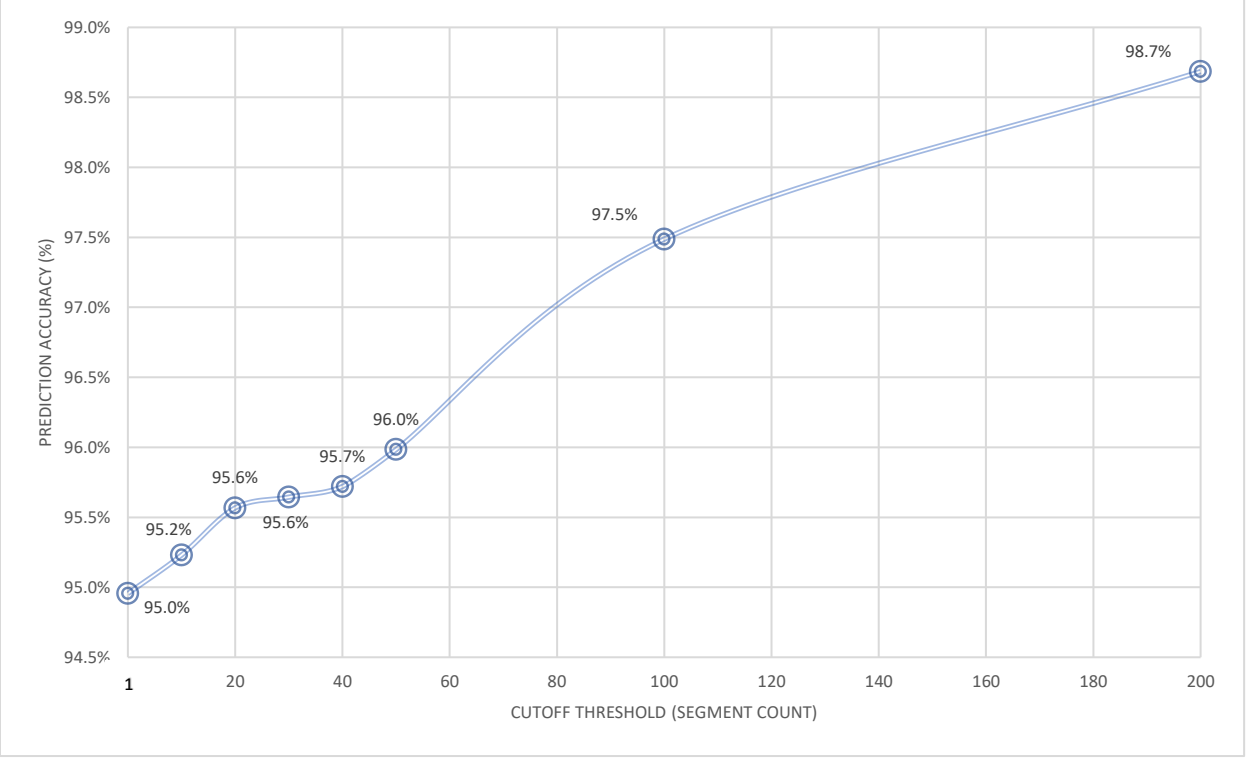


Figure 3-8. Prediction accuracy vs. cutoff threshold

Threshold	Number of Links
0	674
10	608
20	564
30	528
40	514
50	498
100	398
200	304

Table 3-1. Impact of the cut-off threshold on the number of links

3.4.2 Turning movement extraction accuracy

After comparing the extracted turning movements (n = 90) to the ground truth obtained from Google maps and Street View, an accuracy of 68% was found. Meanwhile, 97% of the wrong predictions correspond to turning movements that are permitted within the ground truth dataset but for which no observation was extracted from the GPS dataset. Furthermore, the prediction accuracy was 98% when only turning movements with at least one observation were examined. However, this restriction reduced the number of turning movements for which a turning movement is predicted by 37 and reduces the probability of detecting prohibited turning movements. Lastly, prediction accuracy for turning movements with at least 2 observed movements was 100% but prediction could only be performed for 51% of the total number of turning movements. Figure 3-9 presents an example of the result for one intersection that was extracted with a 100% accuracy. Permitted turning movements are presented in red while the prohibited movements are presented in green. It should also be noted that no U-turns were extracted from the GPS data sample, therefore, it was not possible to determine the turning permissions for that turn type.

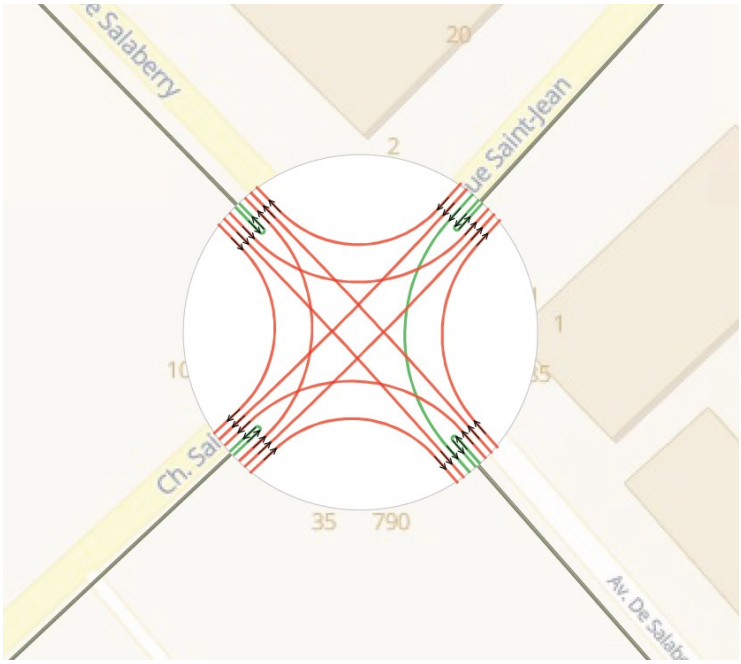


Figure 3-9. Extracted intersection movement permissions

3.5 Conclusion

In this work, we develop a method that can extract link directions and turning movement rules from GPS trajectory data with a high accuracy. Considering a link, corresponding to a directional road segment, with an observed GPS segment count of less than 5% than that of the reverse link is a good indicator of the absence of that road segment. This resulted in a minimal prediction accuracy of 95%. The performance of a sensitivity analysis on the sample size (GPS segment count per road segment) proved that an increase in sample size will only improve prediction results (up to 99 %). This level of accuracy is adequate for macroscopic models that require this type of information for large regions. The contribution of this method is the automatic extraction of directional road segments for very large regions, assuming a good coverage of GPS data observations.

At intersections, turning movement permissions prediction achieved a lower accuracy (68%) than link direction extraction. This is due to the lower number of observations for each intersection turning movement. In fact, turning movements with at least one observation were predicted at 98% accuracy. However, 37% of the permitted turning movements did not have any observation extracted from the GPS data. Therefore, an increase in sample size will allow better coverage of intersection turning movements.

Using this GPS dataset, it was not possible to extract network features (link directions and turning movements) for different times of the day simply due to the sample size. A larger dataset will allow for better knowledge discovery by providing a larger temporal coverage of the different times of day. This is of importance for road networks with varying number of lanes available for traffic at different times of the day. For example, some road networks have restricted lanes for transit use during peak periods, or for parking during off-peak periods. Similarly, some intersections have varying turning movement permissions by time of day for traffic optimization and safety purposes.

Overall, the developed method demonstrates the feasibility of automatic road network feature extraction for modelling and macrosimulation purposes. This work presents the required input data and the proposed methodology to achieve this objective. A proof of concept was also made by building a road network model using the EMME software for the study area in Quebec City, Canada, using GPS trajectory data collected by motorists. Data tables in the EMME format were created, indicating which links and turns had to be removed from the base network to better represent real road network features within the study zone.

In sum large datasets of GPS points/trajectories can be used to extract road network features to build road network models. Extraction accuracy was found to depend mainly on the sample size. Therefore, the main limitation of this work is the GPS trajectory sample size. The increased use of GPS enabled devices and availability of larger GPS datasets will only increase prediction accuracy by providing greater spatial and temporal coverage. Spatial and temporal coverages dictate the area for which network features can be extracted and the possibility to extract features for different periods of the day.

In addition to the use of larger datasets, future works include the use of machine learning techniques, such as classification learners, to determine intersection movement permissions. Additionally, future research can explore the possibility to extract more network features required for macroscopic modelling from GPS data, such as the number of lanes, road types, and link performance relationships. Autonomous driving can also benefit from the extracted network features by adding them to maps used in autonomous vehicles.

Acknowledgements

The author would like to acknowledge the generous support of McGill University's Faculty of Engineering and the Vadasz Scholars Program.

References

Adresses Québec - AQréseau. 2022. https://adressesquebec.gouv.qc.ca/aqreseau.asp

Bender, P., Ziegler, J., & Stiller, C. Year. Lanelets: Efficient map representation for autonomous driving. 2014 IEEE Intelligent Vehicles Symposium Proceedings.

Chen, B. Q., Ding, C. B., Ren, W. J., & Xu, G. L. 2021. Automatically tracking road centerlines from low-frequency gps trajectory data. ISPRS International Journal of Geo-Information, 10(3). 26

Données ouvertes 2022. https://www.ville.quebec.qc.ca/services/donnees-services-ouverts/index.aspx

El-Geneidy, A. M., & Bertini, R. L. Year. Toward validation of freeway loop detector speed measurements using transit probe data. Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749).

Fan, J., Fu, C., Stewart, K., & Zhang, L. 2019. Using big GPS trajectory data analytics for vehicle miles traveled estimation. Transportation research part C: emerging technologies, 103. 298-307

Gately, C. K., Hutyra, L. R., Peterson, S., & Wing, I. S. 2017. Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone GPS data. Environmental pollution, 229. 496-504

Guo, Y., Li, B., Lu, Z., & Zhou, J. 2021. A novel method for road network mining from floating car data. Geo-spatial Information Science. 16

Kan, Z., Tang, L., Kwan, M.-P., & Zhang, X. 2018. Estimating Vehicle Fuel Consumption and Emissions Using GPS Big Data. International Journal of Environmental Research and Public Health, 15(4). (

Karagiorgou, S., & Pfoser, D. 2012. On vehicle tracking data-based road network generation. 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, United states.

Leichter, A., & Werner, M. 2019. Estimating road segments using natural point correspondences of GPS trajectories. Applied Sciences-Basel, 9(20). 11

Lin, C., Zhou, X., Wu, D., & Gong, B. 2019. Estimation of Emissions at Signalized Intersections Using an Improved MOVES Model with GPS Data. International Journal of Environmental Research and Public Health, 16(19). 3647

Mennis, J., & Guo, D. 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. Computers, Environment and Urban Systems, 33(6). 403 - 408

Phondeenana, P., Noomwongs, N., Chantranuwathana, S., & Thitipatanapong, R. 2013. Driving Maneuver Detection System based on GPS Data. 1-6

Tantiyanugulchai, S., & Bertini, R. L. Year. Arterial performance measurement using transit buses as probe vehicles. Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems.

Xingzhe, X., Wenzhi, L., Aghajan, H., Veelaert, P., & Philips, W. 2016. A novel approach for detecting intersections from GPS traces. Piscataway, NJ, USA.

Xingzhe, X., Wong, K. B. Y., Aghajan, H., Veelaert, P., & Philips, W. 2015. Inferring directed road networks from GPS traces by track alignment. ISPRS International Journal of Geo-Information, 4(4). 26

Zhang, C., Xiang, L., Li, S., & Wang, D. 2019. An intersection-first approach for road network generation from crowd-sourced vehicle trajectories. ISPRS International Journal of Geo-Information, 8(11). 26

Zhang, Y. F., Zhang, Z. X., Huang, J. C., She, T. T., Deng, M., Fan, H. C., Xu, P., & Deng, X. S. 2020. A hybrid method to incrementally extract road networks using spatio-temporal trajectory data. ISPRS International Journal of Geo-Information, 9(4). 15

Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., & Liu, H. 2019. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transportation Research Part C: Emerging Technologies, 107. 70 - 91

Zhongyi, N., Lijun, X., Tian, X., Binhua, S., & Yao, Z. 2018. Incremental road network generation based on vehicle trajectories. ISPRS International Journal of Geo-Information, 7(10). 19

Zito, R., & Taylor, M. A. P. 1994. The use of GPS in travel-time surveys. Traffic Engineering and Control, 35. 685-685

Link Between Chapters

In the optic of building on current tools, techniques, and data availability Chapter 3 presented a method to refine a simple road network model by extracting from GPS trajectory data the road directionality and turning movement permissions at intersections. This is the first contribution of this research as it enhances the output of current transport modelling tools, such as EMME. The proposed method is directly applicable to road networks that are designed primarily in a grid-type layout. Applying this method to road networks that are not in a grid-type format requires some adjustments to account for the added complexity in the road network layout. At the link and intersection levels, the same logic remains applicable, however, the azimuth-direction relationship becomes more complex and needs to be accounted for in the 3-67evelop3-67ry.

At this point, the road network model is still incomplete for transport modelling purposes, it requires additional attributes to adequately describe the road. The following chapter complements the developed network model by proposing a method to determine the number of lanes based on the lateral distribution of observed GPS trajectory points with respect to each road segment. This additional attribute enables the estimation of road capacity and is used as an essential input of volume delay functions that estimate segment travel time during traffic simulation procedures.

Chapter 4 - Global Positioning System Data to Model Network-Wide Road Segment Level Number of Lanes Using Spatial Analysis and Machine Learning

Global Positioning System Data to Model Network-Wide Road Segment Level Number of Lanes Using Spatial Analysis and Machine Learning

Adham Badran [a*], Ahmed El-Geneidy [b], and Luis Miranda-Moreno [c]

[a] Civil Engineering Department, McGill University, Montreal, Canada (adham.badran@mail.mcgill.ca)

[b] School of Urban Planning, McGill University, Montreal, Canada, (ahmed.elgeneidy@mcgill.ca)

[c] Civil Engineering Department, McGill University, Montreal, Canada (luis.miranda-moreno@mcgill.ca)

[*] Corresponding Author

4.1 Abstract

One of the main features required in transport network modelling is the number of lanes used to estimate the road capacity and predict vehicular travel times based on traffic flows. Traditionally, the number of lanes information is collected manually or more recently extracted using computer vision techniques, which are two resource intensive methods. This research proposes the use of emerging crowd-sensed Global Positioning System (GPS) data to predict the number of lanes per road segment for large scale transport models through geographic operations and machine learning. The developed method consists of i) a spatial analysis to analyze the GPS trajectory data and estimate predictors and ii) a supervised machine learning model development to predict the number of lanes per road segment.

It was found that the method predicts the number of lanes at an accuracy of 91% using two predictors: number of GPS points per road segment and a lateral distance variable containing

60% of the GPS data points, centered around the lateral distance distribution median. The best prediction model was obtained using decision trees classifier. It was also found that most of the local roads did not have sufficient data points to obtain a stable lateral distance distribution, therefore, the model was limited to a subset of road segments with sufficient observations. Given the availability of high spatiotemporal coverage GPS data, the method can be adapted and applied to large scale road network models and predict the number of lanes accurately and cost-effectively.

Keywords: Global Positioning System, Transport Model, Road Network, Number of Lanes, Road Capacity, EMME.

4.2 Introduction

Knowledge of the number of lanes on road segments within the transport network is essential for the planning and operation of the transport system. For example, conventional and autonomous vehicle navigation, transport modelling and simulation, road safety applications all require the number of lanes information as an input. In fact, lane-level digital maps are critical for advanced driver assistance systems and continuous research is being performed to improve their4-70eveloppment (C. Guo et al., 2016). Moreover, autonomous vehicle navigation requires prior knowledge of the road network in addition to real-time detection of the road lanes to select the trajectory appropriately (Bounini et al., 2015). In transport modelling, the number of lanes information is essential for all modelling scales. Macroscopic models include the number of lanes information into volume-delay functions to determine the road's vehicular capacity and evaluate road segment level travel time. Meanwhile microscopic transport models consider the number

of lanes through lane changing models (Treiber & Kesting, 2013). Another example is the use of the number of lanes when analyzing pedestrian-vehicle interaction at crossings and the relationship with road-user safety (Kadali & Vedagiri, 2020).

The challenge in obtaining the number of lanes information is for large-scale road networks and maps. In fact, local transport departments are generally uncapable to develop and maintain a large scale road network. The main reasons, among others, can be attributed to the decentralization of road network data and road network jurisdictions, the inconsistent format and level of the data depending on the source, and the lack of connection between transport planning and modelling teams with infrastructure construction teams to enable a smooth and timely update of the digital road network. Ideally, an entity needs to be constantly aware of the road network characteristics for the whole metropolitan area to be able to maintain a road network model, which is not the case. With the advances in technology, new data sources and techniques are emerging and present a potential to extract transport network-related information. Global Positioning System (GPS) trajectory data is being collected by different organizations through GPS-enabled smartphones and stored on servers using cellular internet. For example, the city of Montreal has provided its residents a smartphone application that records their trajectories for a limited period to analyze the trajectory data and improve transport planning and reduce traffic delays (Montréal).

Extracting the number of lanes has been tackled in the past using different data sources. The most frequent method to extract the number of lanes for large-scale networks is based on aerial imagery and computer vision techniques. Multiple studies have been looking at extracting road network features automatically using different data sources. First, high-resolution imagery, in

combination with computer vision methods have enabled the large-scale detection and extraction of road network-related attributes. One of the research groups has done extensive work using road segmentation to detect different visible features such as the road, sidewalks, vegetation, buildings, and cars to augment OpenStreetMap by adding more features (Mattyus et al., 2015). The main challenges were found to be the presence of trees, shadows, cars, as they increase heterogeneity in the images in addition to misalignment issues with respect to the road centreline file used as a priori of road segments' location. The same research group further expanded the analysis by collecting and incorporating street-level imagery in the number of lanes recognition algorithm which increased its prediction accuracy (Máttyus et al., 2016). Recognizing that collecting street-level imagery presents high collection and processing costs, they proposed a more resource-friendly version that only employs satellite imagery but takes advantage of new methodological advances in deep learning to improve the model accuracy (Máttyus et al., 2017). Another study has also extracted the number of lanes information from satellite imagery using an SVM classifier for lane identification based on brightness levels. Although they predicted the number of lanes at an accuracy of 100%, the experiment was only presented for six road segments (Tang et al., 2014). Although satellite imagery has been used to detect the number of lanes and improved by collecting street level high-resolution imagery, it is not without limitations. Data availability is limited due to the collection costs, moreover, occlusions, illumination variability and unmarked road lines reduce the capacity of such techniques (Kasmi et al., 2018). The best number of lanes prediction accuracy obtained was 83 %.

Recent research efforts have been studying the extraction of road networks from GPS data using different spatial analysis algorithms. Three main approaches were used to extract road networks:

Clustering, intersection linking, and track alignment. For example, the work by Y. Guo et al. (2021) proposes a clustering method to extract road network centreline and intersections with the accuracy of 92%. Clustering is in fact the most popular method to extract road networks from GPS trajectory data. Another study by C. Zhang et al. (2019) employs the intersection linking method to detect the road network and intersections at an accuracy greater than 90%. Although not very popular, studies by Leichter and Werner (2019) and Zhongyi et al. (2018) have also used the track alignment method to generate road networks. However, accuracy was either low or not compared to the ground truth. Although these road network inference methods are able in some cases to extract the road network centreline and intersections with high accuracy, they do are not designed to extract more detailed road network features such as the number of lanes.

Very few studies have examined the use of sole GPS trajectory data to extract the number of lanes. A study by Arman and Tampere (2020) proposes a method that extracts lane locations on a highway corridor. However, the number of lanes extracted is not explicitly validated by comparing to the ground truth. Therefore, no accuracy was provided. One attempt by L. Zhang et al. (2010) used GPS traces and a road centreline map from OpenStreetMap to improve the map quality and estimate the number of lanes. The main limitation was the assumption of normal distribution of GPS traces with respect to the road centre, which is not the case and resulted in number of lanes prediction accuracy of less than 60%.

Moreover a study by Y. Chen and Krumm (2010) fits Gaussian mixture models to GPS trajectory data to determine the number of lanes. Although the study attempts to preserve the continuous nature of road segments, it is limited by the sample size and the fact that this method requires

prior knowledge of the number of Gaussian distributions to fit. Thus, the study resulted in relatively low accuracy predictions.

In sum, the main limitations of past studies extracting the number of lanes are the high cost of imagery data collection and the output accuracy. In fact, the high cost reduces the frequency of map updates which can result in maps not representing the continuously evolving nature of the road network. In addition, the output accuracy of past studies can potentially be improved by using large-scale GPS trajectory data.

Considering the general availability of road centreline data or algorithms to infer them from different data sources, the objective of this study is to propose a method that uses GPS trajectory data to extract the number of lanes with a relatively high accuracy. This is done through spatial analysis of GPS trajectory points to extract variables that feed into a machine learning classification algorithm that predicts the number of lanes for road segments for use in large-scale transport models.

4.3 Methodology

GPS data treatment can be divided into two main parts based on the analysis type. The first part of the analysis was the spatial analysis using Geographic Information System (GIS) software carried out in the FME software. This software was selected since it is a very efficient data integration platform capable of managing, combining, and transforming big data with advanced spatial data analysis capabilities. The second step was the number of lanes prediction model development and visualization carried out in MATLAB. The general assumption of this study is that although GPS accuracy is between 7 and 13 meters (Merry & Bettinger, 2019), GPS trajectory

points will be distributed around the middle of traffic lanes when the sample size is large. Therefore, this method proposes to determine the distance distribution of GPS points for each directional link with respect to a reference line and infer the number of lanes based on the distribution properties through a machine learning classification method.

4.3.1 Spatial Analysis

The first step requires raw GPS trajectory data, a road network model (links and nodes), and an azimuth-direction dictionary table as input. A summary of the spatial analysis steps can be seen in Figure 4-1. The yellow boxes present the input data sources required to carry out the spatial analysis steps. In this study, each GPS trajectory point had two sets of longitude and latitude points; raw, and map matched coordinates, which were both used at different stages of the analysis.

The process can be divided into four main steps: 1. Determine the direction of each GPS trajectory point, 2. Remove GPS trajectory points located at intersections, 3. Associate each GPS trajectory point to a directional road segment, and 4. Calculate the lateral distance between each GPS point and the reference line. The number corresponding to each step is also presented in Figure 4-1.

Firstly, the azimuth of each GPS trajectory point was calculated based on its location and the location of the consecutive point within the same trip. The azimuth is defined as the orientation, in degrees, between two points as the number of degrees clockwise from the north reference. The azimuth was selected as the measure to define trip segment directions and an azimuth-direction dictionary was created for that purpose as seen in Figure 4-2. Map matched coordinates were used to calculate the azimuth to ensure consistent direction results and remove the

fluctuations found in raw GPS point data. Following the azimuth calculation for each point, the direction was calculated using the azimuth-direction dictionary.

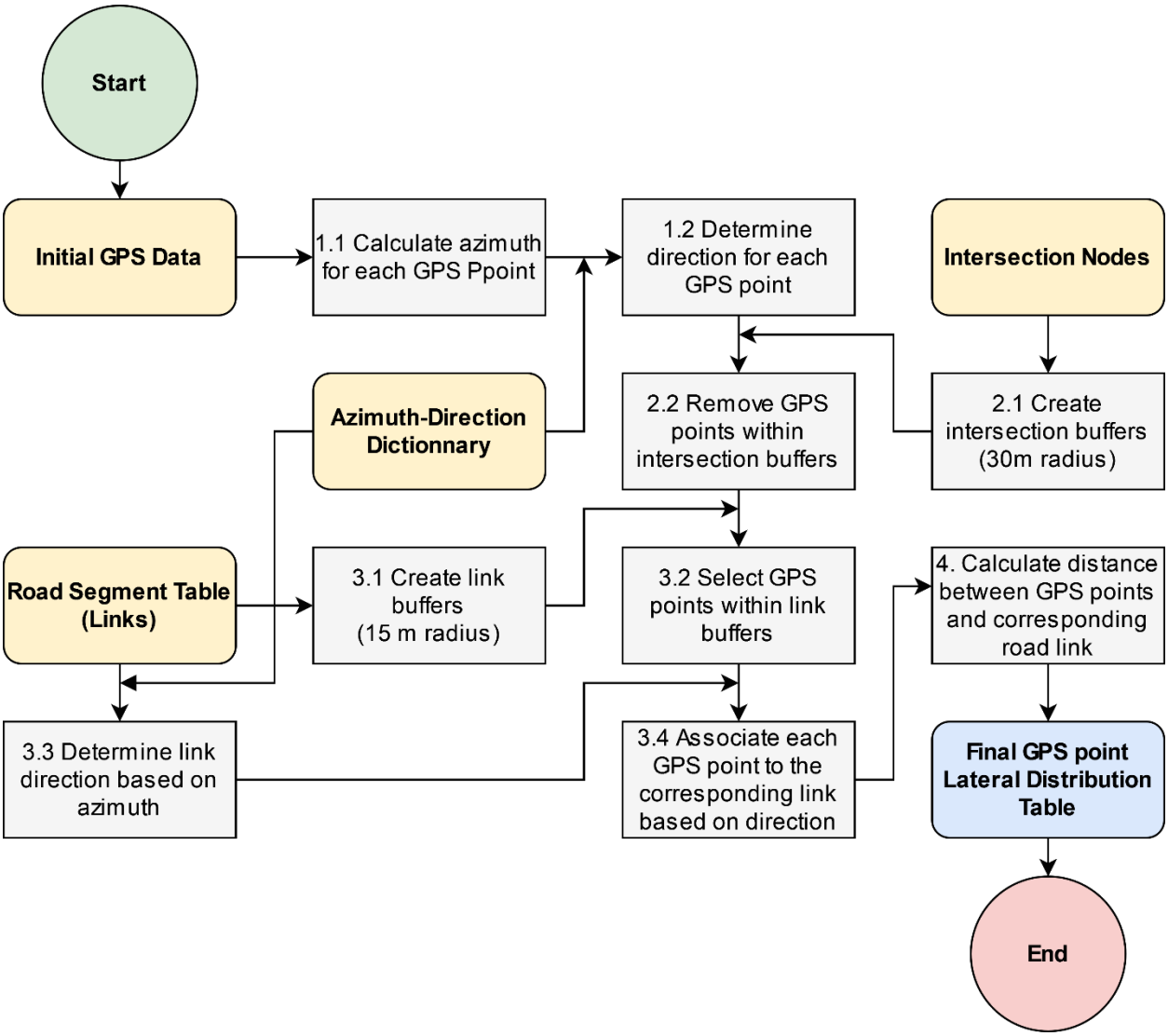
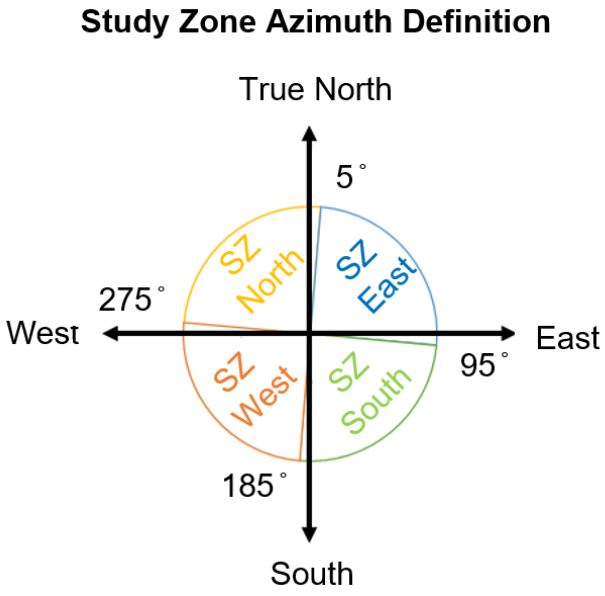


Figure 4-1. GPS Trajectory Points GIS Treatment Diagram

Secondly, intersection buffers were used to remove GPS points that fall within the vicinity of intersections. Given that this study aims to determine the mid-block road segment number of lanes, the GPS trajectory points in the vicinity of intersections were removed since the number of lanes near an intersection is sometimes different to allow for upstream dedicated turning lanes

or downstream insertion lanes. Following visual inspection of the road network, a buffer size of 30-meter radius with respect to the intersection centres was used to filter GPS trajectory points within intersection areas. This ensured that the remaining GPS trajectory points correspond to travel within the road segment.



Azimuth angle range (degrees)			Direction
275	-	4	North
5	-	94	East
95	-	184	South
185	-	274	West

SZ: Study Zone

Figure 4-2. Azimuth-Direction Correspondence

Thirdly, to remove noisy GPS trajectory points, a road segment (link) buffer of 15-meter radius was created to select GPS points associated to each link through nearest neighbour analysis. This buffer size was selected to ensure that the GPS points' lateral distribution profile with the respect to the directional link is captured entirely while minimizing the number of outliers. This was validated in the following steps of the analysis by examining all lateral distance distribution histograms and kernel density estimators. The link direction was also obtained based on the azimuth to add an extra criterion when selecting the nearest neighbour and ensure that every GPS point is associated to the correct directional link.

Fourthly, the shortest distance between each GPS point and the associated directional link is calculated and serves in the following step develop a number of lanes prediction model. This distance corresponds to the length of the perpendicular line, d_i, between the GPS location point and the directional link as seen in Figure 4-3. It was the main variable carried to the next modelling step.

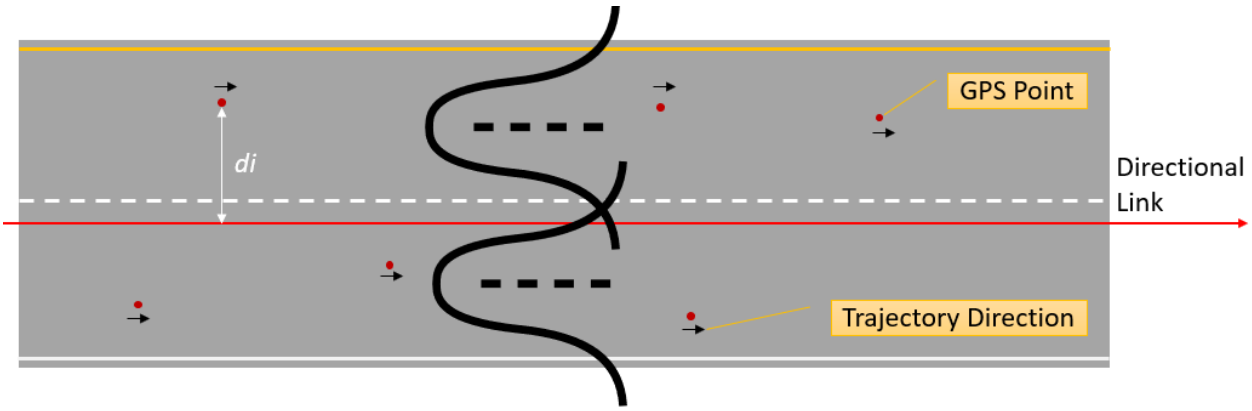


Figure 4-3. Distance from Point to Link Calculation

The location of the directional link with respect to the actual road segment is approximate since it is based on a simple street centreline shapefile. Moreover, for bidirectional road segments, the links for both directions are superimposed. This was considered while determining road segment buffer size.

4.3.2 Road Segment Number of Lanes Prediction Model

Following spatial analysis, the second part of the method consisted of creating the number of lanes prediction model. Assuming that road segments with different numbers of lanes have different GPS trajectory data characteristics (such as spatial distribution pattern and number of points), roads with different numbers of lanes are seen as distinct categories and the question is formulated as a number of lanes classification problem. A classification model was calibrated using input variables derived from GPS trajectory points data to output the number of lanes for

each road segment. For each road segment, input variables were compiled following the spatial analysis part and were used to train the model to predict the number of lanes as a categorical variable. The two main GPS trajectory points descriptors, used to derive input variables to the classification tree model, were the lateral distance d_i and the number of points per directional road segment. A frequency histogram and a kernel density estimator were fitted to the distance variable to visualize the distribution with respect to the reference line (directional link) and determine model parameters. First, it was observed that for some of the links, sample size was too low and resulted in unstable and unmeaningful distributions. Following inspection of the kernel density estimators and frequency histogram, the sample size was limited to a minimum of 500 GPS points per directional link to produce stable results in terms of distribution shape. Road segments with fewer GPS point observations were removed.

Based on the observed distributions and preliminary tests and aiming to create variables that reflect the lateral distribution of GPS trajectory points with respect to the link, distance percentiles, d_{ipc}, were calculated for different percentiles, i, of 5%, 10%, 15%, 20%, 80%, 85%, 90%, and 95%. To standardize these values and render them comparable across different links, new variables were created by calculating the variable D_p defined as the lateral distance containing a proportion, p, of the GPS points data. D_p is calculated using lateral distance percentiles to ensure that this new variable is centered around the median distance value. The following are the lateral distance variables that were calculated:

$$D_{90} = d_{95pc} - d_{5pc}$$

$$D_{80} = d_{90pc} - d_{10pc}$$

$$D_{70} = d_{85pc} - d_{15pc}$$

$$D_{60} = d_{80pc} - d_{20pc}$$

For example, D_{60} corresponds to the difference between the 80th percentile distance and 20th percentile distance, therefore it contains 60% of the GPS points data. A visual illustration is provided in Figure 4-4. The number of GPS points per link and the standard deviation of lateral distance per link were also calculated to be tested in the model specification.

Following the creation of the variables for each road segment, supervised machine learning classification methods were tested. In fact, classification tree analysis was carried out to determine if it can create an accurate model that can be used for prediction. This method is a good option when ground truth data is available for the learning step. Moreover, it is non-parametric and does not require prior knowledge of the distribution of each variable. Another

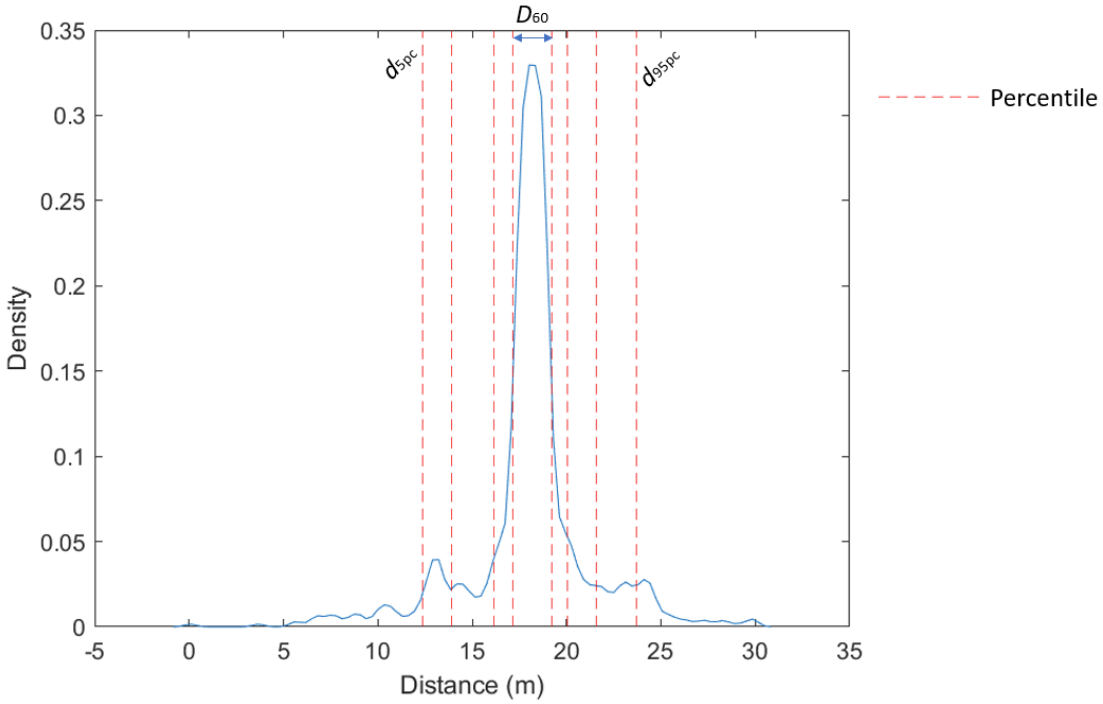


Figure 4-4. Example of Percentile Visualization

advantage of this method compared to other machine learning techniques such as neural networks classification is its transparency which makes the model easy to interpret (Ian et al., 2017).

To ensure protection against overfitting, model validation was carried out using a 5-fold cross-validation. This validation method divides the dataset randomly into five groups. At each step, one of the five groups is held out to be used for validation while the other four groups are used to train the model. Once the model is specified, it is used to make predictions on the group that was held out. For a 5-fold cross validation, this process is repeated five times.

4.3.3 Data

Three main input datasets are used: 1) GPS trajectory points, 2) Modelled directional road network (links and nodes), and 3) Google maps and Street View. GPS data was collected during the spring of 2014 in Quebec City, Canada. It was collected for 21 days by 2000 voluntary users through the Mon Trajet smartphone app, made available by the Municipality. Each point is described by the following attributes: map matched X and Y coordinates, trip ID, speed, and timestamp (Year-Month-Day-Hour-Minute-Second). Following the preprocessing steps, 245 links were selected as the experimental data to model number of lanes, which included 120 000 GPS points (excluding GPS points within the intersection buffers). This study area was selected based on its urban setting since it is in the city centre where more GPS trajectories were available. Figure 4-5 presents a sample of the study area where part a shows the raw GPS trajectory points, and part b shows the processed GPS trajectory points for the same road corridor following spatial analysis steps 1 to 3. In part b of the figure, GPS trajectory points are colored differently depending on the link to which they were associated.

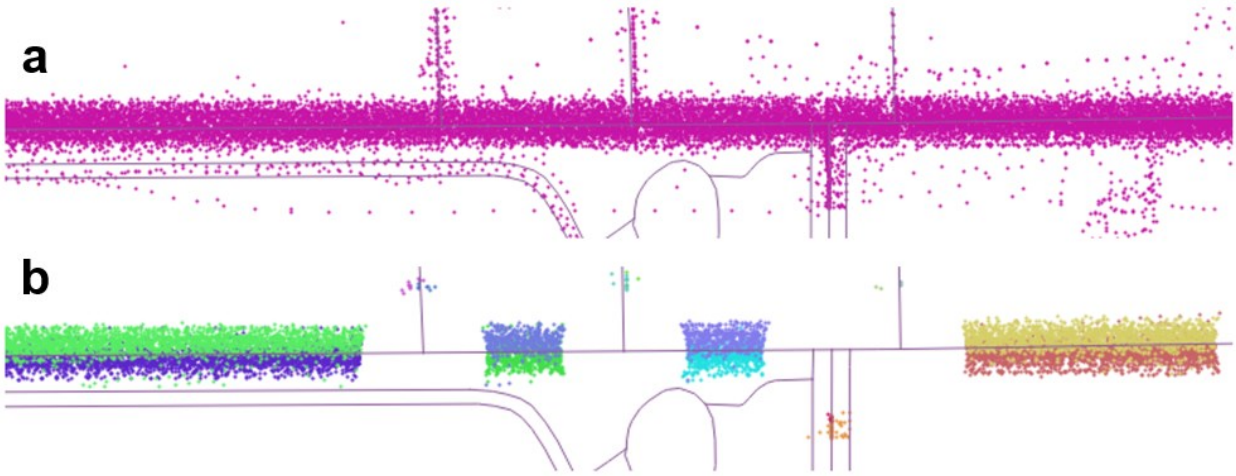


Figure 4-5. Sample of GPS Points Data in Study Area - Before and After Spatial Processing

The directional road network was created using an initial road centreline shapefile which was converted in a network model compatible the EMME transport modelling software to obtain directional links and augmented using the same GPS trajectory data to ensure that road topology and connectivity are valid. Each link is defined by an origin and a destination node. The possible number of lanes per directional link was one, two, or three lanes, for which the ground truth was manually extracted using Google Maps and Street View.

For a given road segment, it is important to note that the number of lanes available for traffic can vary spatially and temporally. The presence of lanes dedicated to transit vehicles or high-occupancy vehicles at a fixed schedule on concerned road segments reduces temporally the number of lanes available to general traffic. This is also the case for lanes that are used for parking at fixed schedules. Throughout a road segment, the number of lanes can also change spatially. For example, it is common to see a higher number of lanes at the two extremities of a road segment to allow for traffic insertion and for dedicated turning lanes. The complex nature of traffic lanes can be seen in Figure 4-6 where a reserved bus lane (highlighted in green) is present

at a fixed schedule and the number of lanes at the intersection level is different (usually greater) than the mid-block number of lanes to accommodate turning movement flows. This paper examines the mid-block number of lanes and does not consider reserved lanes.

Figure 4-6. Example of a Complex Road Geometry

4.4 Results

With the proposed steps and parameters, it was possible to extract GPS points for road segments and associate each point to the correct directional link based on the trajectory direction. The sample size filter limited the number of analyzed directional links included in the analysis to 43 links. The buffer sizes were also validated based on the frequency distributions of GPS points' lateral distance with respect to the link since the entire distribution profile is captured. This can also be noted in Figure 4-7 which presents the kernel density estimator fitted to the lateral distance variable distribution for six different links of varying number of lanes. The sample size, N, and the D_{60} values are also presented for each link.

In addition to the distribution profile of lateral distance, the figure also demonstrates the significant difference in the distribution profile between links having one, two, and three lanes. Through observation, it was possible to identify that road segments with fewer lanes have lower values of N and smaller D_{60} values. This can be explained by the fact that roads with fewer users are designed to have fewer lanes, and GPS points are concentrated in a narrower area.

Model specification was carried out to determine the best model and best predictors for the number of lanes. Given the relatively low number of road segments, a 5-fold cross-validation method was performed to avoid overfitting the data (which signifies that approximately 96 000 GPS points are used to specify the model and 24 000 points to validate the prediction). The highest classification prediction

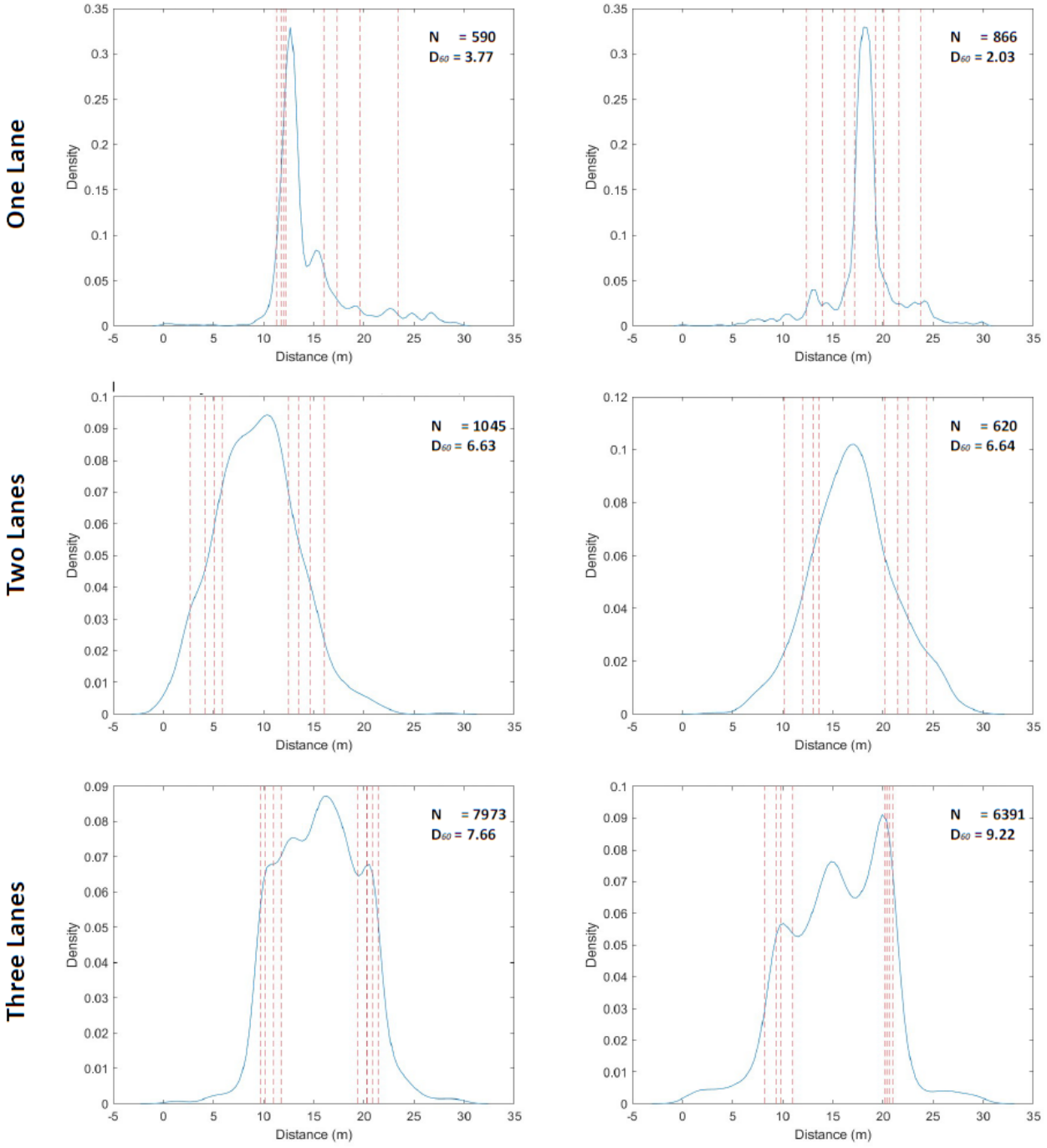


Figure 4-7. Sample Kernel Density Estimator of Lateral Distance for One, Two, and Three Lanes

accuracy was found using a decision tree classifier at 91% using two predictors, the sample size N and D_{60}. The optimizable decision tree classifier tested iteratively different numbers of splits and different split criteria to reach the minimum classification parameters and error.

Figure 4-8 presents a plot of the two selected predictors, showing a clear delimitation between the predictor values for roads with one, two, or three lanes. Moreover, the optimized decision tree is presented with the three split levels and values in Figure 4-9. Ensemble classifiers, such as boosted trees, bagged trees, and subspace discriminant were also tested to improve prediction accuracy and the best accuracy was using the subspace discriminant ensemble classifier at 91%. Given that the optimized decision tree was able to predict at the same accuracy level it was selected as the best model in this case since it is simpler to visualize and interpret.

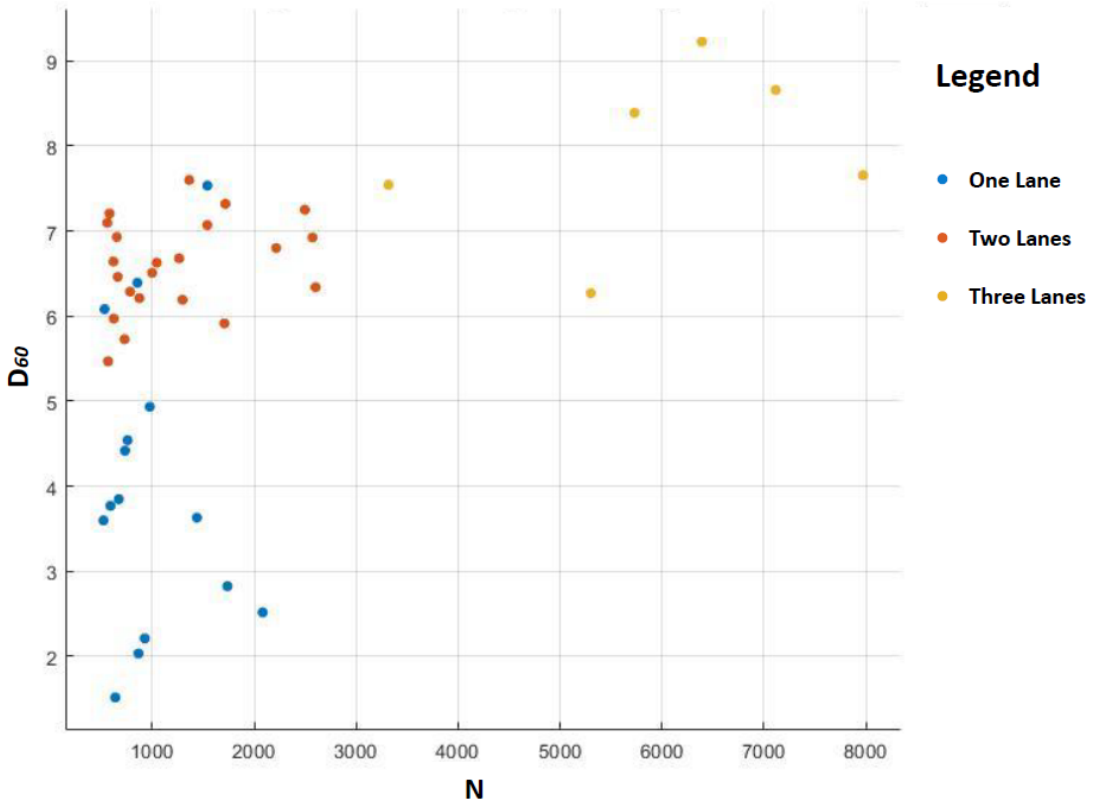


Figure 4-8. D60 vs. Sample Size (N)

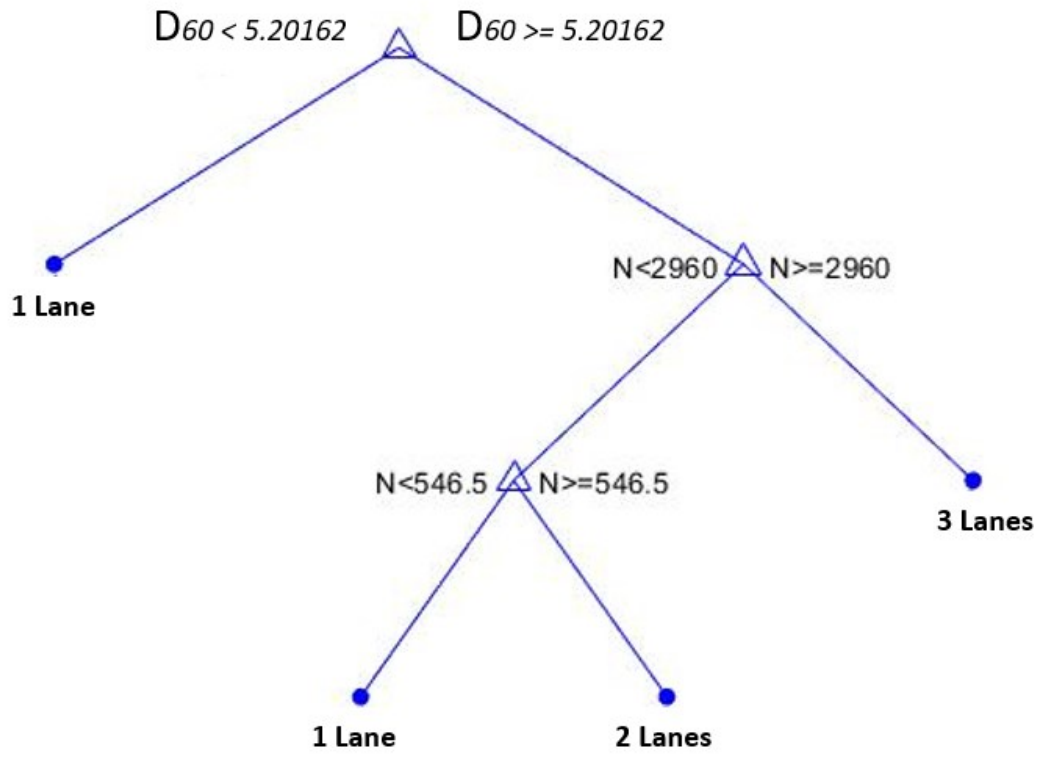


Figure 4-9. Selected Classification Decision Tree

4.5 Discussion

The proposed methodology predicts the number of lanes per road segment based on the number of GPS points associated to the link and the difference between the 80th and 20th percentile distance, representing a lateral distance measure centered around the median lateral distance.

Given that the best prediction model was obtained using only two variables, an optimized decision tree classifier was sufficient to reach a good model accuracy (91%). However, adding new variables will require retesting ensemble classifier methods to verify if they are able to improve prediction accuracy. Moreover, to use this model, the sample size would need to be translated into relative terms or to be specified with respect to the sample size corresponding to a new dataset. The main hypothesis behind using the sample size as a variable is that for a given

period of data collection where we assume a representative sample, it is expected to have a larger number of observations for road segments with a larger number of lanes since they generally have a higher traffic flow.

The spatial analysis and model specification steps were limited by the experimental data available. During the study, it was found that some of the GPS points were map matched in their raw form which signifies that they were snapped to a road centreline at a step prior to accessing the data. Given that this study examines the lateral distribution of raw GPS points with respect to the road link, map matching has a negative impact on data quality. It was also noted that some links had a low number of GPS points, which resulted in unstable lateral distance distribution profiles.

Ideally, larger datasets of uniquely raw GPS points need to be used to have a larger coverage to obtain more realistic distributions and potentially create more predictor variables. The objective is to have more GPS points per link, not necessarily more links as it will also become more complex to obtain the ground truth information. An increase in the number of points per link will also increase the probability of having better coverage for different times of the day, which enables model specification for different time periods to detect the change in the number of traffic lanes temporally.

Although some studies have proposed the extraction of the number of lanes using satellite and street-level imagery with a relative high accuracy, they are not without limitations (Kasmi et al., 2018; Máttyus et al., 2016; Nieroda et al., 2022). In fact, the high cost of imagery data collection is an important limitation that is overcome in this study since GPS data is currently being crowd

sensed by location-based applications through smartphones. Furthermore this study considers 43 road links for model specification which is a larger sample than the work by Tang et al. (2014) which only considers 6 road segments for the analysis.

Comparing this study to some studies using GPS data to extract the number of lanes, the prediction accuracy significantly exceeds the 60% accuracy obtained in the study by L. Zhang et al. (2010). In addition, the method proposed in the current study provides more accurate results and a simpler procedure than the studies by Y. Chen and Krumm (2010) and Arman and Tampere (2020) to obtain the number of lanes for integration in large-scale transport models.

4.6 Conclusion

This study proposes a method to predict the number of lanes per road segment using crowd sensed GPS trajectory data as an input in addition to a simple geographic representation of the road network. The proposed framework is composed of two main steps: to predict the number of lanes of road segments using GPS trajectory data while aiming to keep the cost low and to obtain high prediction accuracy.

The first step is a spatial analysis process to filter and prepare the GPS trajectory data for variable creation. Due to the noise inherent to GPS trajectory, it was crucial to ensure that raw GPS data points were filtered using buffers. This is also necessary to account for the specificities in road design and for the discrepancies in the road network geographic representation. This step also served to produce variables necessary to derive the predictors for the following step. The two main variables were the number of GPS points per road segment and the lateral distance between each point and the reference line representing the road segment. The second step is

the training and validation of a machine learning method using classification tree analysis and ensemble learning. Standardized predictors were derived from the lateral distance variables to ensure that the values are comparable across different road segments.

This study was able to develop a road segment number of lanes prediction model using GPS trajectory point data with an accuracy of 91% using a decision tree classifier and two predictors. This prediction accuracy is higher than prediction results obtained by previous research. This finding demonstrates that it is possible to extract the number of lanes available for general traffic by using crowd-sensed GPS trajectory data. This will facilitate road transport network model development and update. The proposed method was demonstrated using a case study in Quebec City, Canada.

However, the work is not without limitations and can be further developed by having a larger temporal sample coverage to enable the prediction of the number of lanes for different periods allowing the detection of dynamic reserved lanes or parking lanes. This study used manually collected ground truth data which limited the size of the study area, network coverage for model development and validation will be increased in future works by collecting more ground truth data or obtaining this information from another source. Moreover, it is possible to explore adding land use variables that might be correlated with the number of lanes and help in improving the prediction model's accuracy. The potential of this method can also be maximized by automating a procedure that can use GPS trajectory points and other basic input files to create a road network containing the number of lanes per road segment.

Eventually, with the arrival of autonomous vehicles, new data sources may also be available in terms of geotagged imagery data that can be automatically collected and treated by these vehicles during their operation. These processed images may in the future be used to mine road network features at a low cost and high accuracy.

Acknowledgement

The authors would like to acknowledge the generous support of McGill University's Faculty of Engineering and the Vadasz Scholars Program.

Author Contribution

The authors confirm contribution to the paper as follows:

Study conceptualization and design: All authors.

Formal Analysis, Investigation, Writing - Original Draft: Adham Badran

Supervision: Ahmed El-Geneidy and Luis Miranda-Moreno

Interpretation of results and manuscript review and editing: All Authors

Conflict of Interest Statement

The authors declare that there are no financial or non-financial competing interests to report.

References

ARMAN, M. A. & TAMPERE, C. M. J. 2020. Road centreline and lane reconstruction from pervasive GPS tracking on motorways. Procedia Computer Science, 170, 8.

BOUNINI, F., GINGRAS, D., LAPOINTE, V. & POLLART, H. Autonomous Vehicle and Real Time Road Lanes Detection and Tracking. 2015 IEEE Vehicle Power and Propulsion Conference (VPPC), 19-22 Oct. 2015 2015. 1-6.

CHEN, Y. & KRUMM, J. Probabilistic modeling of traffic lanes from GPS traces. Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, 2010. 81-88.

GUO, C., KIDONO, K., MEGURO, J., KOJIMA, Y., OGAWA, M. & NAITO, T. 2016. A Low-Cost Solution for Automatic Lane-Level Map Generation Using Conventional In-Car Sensors. IEEE Transactions on Intelligent Transportation Systems, 17, 2355-2366.

GUO, Y., LI, B., LU, Z. & ZHOU, J. 2021. A novel method for road network mining from floating car data. Geo-spatial Information Science, 16.

IAN, H., FRANK, E., HALL, M. & CHRISTOPHER, J. 2017. Data mining: Practical machine learning tools and techniques—Part II: More advanced machine learning schemes. Morgan Kaufmann, Burlington, MA.

KADALI, B. & VEDAGIRI, P. 2020. Role of number of traffic lanes on pedestrian gap acceptance and risk taking behaviour at uncontrolled crosswalk locations. Journal of Transport & Health, 19, 100950.

KASMI, A., DENIS, D., AUFRERE, R. & CHAPUIS, R. Map Matching and Lanes Number Estimation with Openstreetmap. 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 4-7 Nov. 2018 2018. 2659-2664.

LEICHTER, A. & WERNER, M. 2019. Estimating road segments using natural point correspondences of GPS trajectories. Applied Sciences-Basel, 9, 11.

MÁTTYUS, G., LUO, W. & URTASUN, R. Deeproadmapper: Extracting road topology from aerial images. Proceedings of the IEEE international conference on computer vision, 2017. 3438-3446.

MATTYUS, G., WANG, S., FIDLER, S. & URTASUN, R. Enhancing road maps by parsing aerial images around the world. Proceedings of the IEEE international conference on computer vision, 2015. 1689-1697.

MÁTTYUS, G., WANG, S., FIDLER, S. & URTASUN, R. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 3611-3619.

MERRY, K. & BETTINGER, P. 2019. Smartphone GPS accuracy study in an urban environment. PLOS ONE, 14, e0219890.

MONTRÉAL, V. D. MTL Trajet Study.

NIERODA, B., WOJAKOWSKI, T., SKRUCH, P. & SZELEST, M. A Heatmap-Based Approach for Analyzing Traffic Sign Recognition and Lane Detection Algorithms. 2022 26th International Conference on Methods and Models in Automation and Robotics (MMAR), 22-25 Aug. 2022 2022. 217-221.

TANG, L., GAN, A. & ALLURI, P. 2014. Automatic Extraction of Number of Lanes from Georectified Aerial Images. Transportation Research Record, 2460, 86-96.

TREIBER, M. & KESTING, A. 2013. Traffic flow dynamics. Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg, 983-1000.

ZHANG, C., XIANG, L., LI, S. & WANG, D. 2019. An intersection-first approach for road network generation from crowd-sourced vehicle trajectories. ISPRS International Journal of Geo-Information, 8, 26.

ZHANG, L., THIEMANN, F. & SESTER, M. Integration of GPS traces with road map. Proceedings of the Third International Workshop on Computational Transportation Science, 2010. 17-22.

ZHONGYI, N., LIJUN, X., TIAN, X., BINHUA, S. & YAO, Z. 2018. Incremental road network generation based on vehicle trajectories. ISPRS International Journal of Geo-Information, 7, 19.

Link Between Chapters

Chapter 4 proposed a method to enhance the road network model developed in Chapter 3 by extracting the number of lanes of road segments based on GPS trajectory points related predictors. A two-step procedure composed of a spatial analysis method and a machine learning modelling technique were able to predict the number of lanes at a high accuracy, convenient for large scale transport models. To complement the developed road network model, Chapter 5 proposes a method to determine the road intersection control type based on the same GPS trajectory points dataset. Knowledge of the intersection control type is useful in the transport model development process as it can increase its accuracy by better adapting the volume delay functions used to estimate travel time on the road segments. It will enable the development of volume delay functions per intersection control type to reflect the difference in vehicle dynamics.

Chapter 5 - Inferring Road Intersection Control Type from GPS Data

Inferring Road Intersection Control Type from GPS Data

Adham Badran (adham.badran@mail.mcgill.ca),

Ahmed El-Geneidy (ahmed.elgeneidy@mcgill.ca)

Luis Miranda-Moreno (luis.miranda-moreno@mcgill.ca)

McGill University

5.1 Abstract

Transport modelling requires accurate and usually hard to find intersection control rules. The widespread use of smartphone applications enabled the automatic collection of road network-related data that can contribute to and improve transport modelling. Global Positioning System (GPS) point data collected in Quebec City, Canada, was used to develop a model inferring intersection control type (traffic light, stops on all approaches, or stops on the secondary approach). Data was used to train and validate supervised machine learning classification models. The developed model predicted intersection control types on a validation dataset with a 96% accuracy. This work presents the best predictors for intersection control type.

Keywords**:** GPS, Transport Model, Road Network, Intersection Control, Map Inference

5.2 Questions

Transport Modelling requires large quantities of data, depending on the project size and level of detail. For example, building a mesoscopic or microscopic model for a neighbourhood, requires detailed road geometry, road type, origin-destination transport demand, and intersection control type and traffic light phasing, to name a few. The work by Barceló et al. (2010) presents different

data collection efforts to estimate travel demand, traffic state, and traffic performance. Additionally, Antoniou et al. (2018) discusses the integration of big data and machine learning in transportation.

Depending on the modelling needs and available resources, data is collected by different means and for different sample sizes. Global positioning system (GPS) data is now collected by widespread communication devices such as smartphones. These devices provide their geographic location and a timestamp at a predetermined high-resolution frequency offering new information that can help in determining road network features.

This work develops a method to infer road intersection control type from GPS points. Such information can be of value for transport modelling when the study area is large, and data cannot be collected as efficiently using traditional observation methods.

5.3 Methods

The primary data source consists of GPS trajectory points, collected during the fall of 2014 in Quebec City, Canada. Data was collected during 21 days by 2000 voluntary users through the Mon Trajet phone app, made available by the city. Each trajectory consists of consecutive GPS location points recorded by the app every second. Each point is described by the following attributes: X and Y coordinates, trip ID, instantaneous speed, and timestamp (Year-Month-Day-Hour-Minute-Second). Figure 5-1 is a map of the raw GPS points (226,000 points) inside the study zone, which consists of 81 intersections. The location and control type of all intersections were also obtained from the municipality for model calibration and validation. Four different control types were available: traffic light, all-way stop, east-west stop, north-south stop.

First, the intersection locations within the study area were determined using the road network and a 20-meter buffer was created around each intersection. The buffer size was determined by examining the road geometry and the spacing between intersections. In fact, the selected buffer size was able to capture all vehicles that are passing through any given intersection without having overlapping buffers. However, some buffers were merged for intersections that are very close to each other and operate as one intersection. The GPS data points were then filtered to only keep the points within the intersection buffers. The final sample size was 81,000 GPS points located within the 127 intersection buffers. At this point, all filtered points for a given trip within an intersection were converted into directional lines representing intersection movements. The

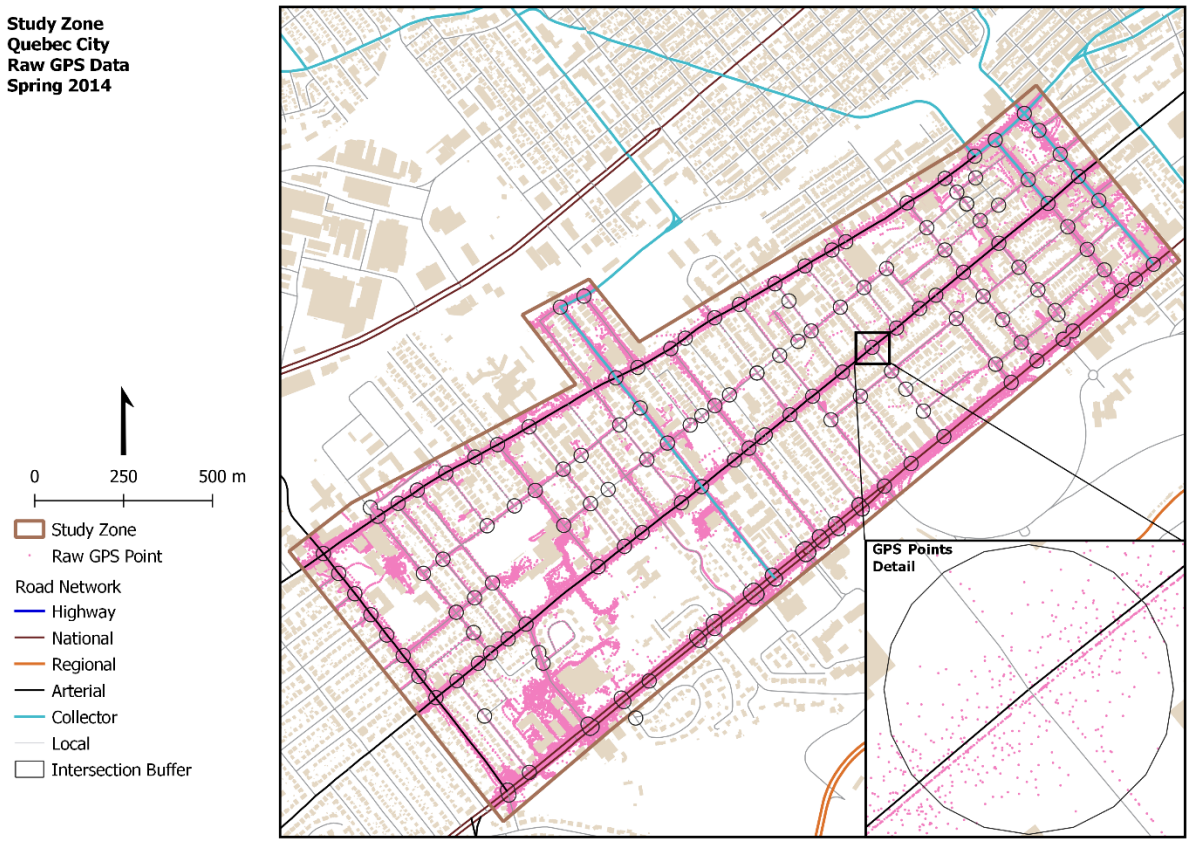


Figure 5-1. Raw GPS Points in Study Zone

intersection movements were then used to determine the inbound and outbound directions for each movement. For a given intersection, trip segment within an intersection buffer area (see Figure 5-2 (a) for the direction definition specific to the study area).

The calculated attributes, inbound direction, and outbound direction were then added to the GPS data points. The intersection control type attribute was also added to the GPS data points to act as the ground truth. For each trip segment within an intersection buffer, the delay (D), in seconds, was calculated using the following equation:

$$D = T_{out} - T_{in}$$

where T_{in} is the time stamp of the first point to enter the buffer area and T_{out} is the timestamp of the last point before exiting the buffer. Following data compilation, the result was a final database containing attributes at the approach level (northern, southern, eastern, or western approach) and at the intersection level. Figure 5-2 (b) illustrates the nomenclature for approaches and movements used in this paper. At the approach level, the following variables were calculated: average speed, standard deviation of speed, minimum speed, maximum speed, trip count, average number of points per trip within the buffer, and average delay. For example, trip count was calculated for each of the four approaches, to know the number of trips that are entering the intersection through each leg. At the intersection level, one speed related variable was calculated: the percentage of points with a speed of less than or equal to 5 km/hr. The developed explanatory variables were based on the expected difference in speed profiles and traffic intensity at intersections of different control types. For example, a traffic light-controlled intersection is expected to serve higher intensity traffic conditions than an all stop intersection.

Therefore, the trip count variable can be significant in differentiating between these two control types. Moreover, at an all-stop intersection, the approach speed is expected to be very low for all the vehicles, while at a traffic light-controlled intersection, some vehicles may not need to decelerate if their approach has a green light. This is expected to be reflected in the different speed variables. Other data disaggregation levels that are expected to show significant difference per intersection control type are specific times of day were traffic performance is impacted, such as peak periods, and specific turning movements, where distinct movement speed profiles may be an indication of a specific control type.

Data processing and manipulations were performed using the FME software, visualizations were produced in QGIS, and model specification and validation were performed in MATLAB. Different model specifications were tested to find the best model to predict intersection control type. Although only the best model specification results are discussed in this paper, the following.

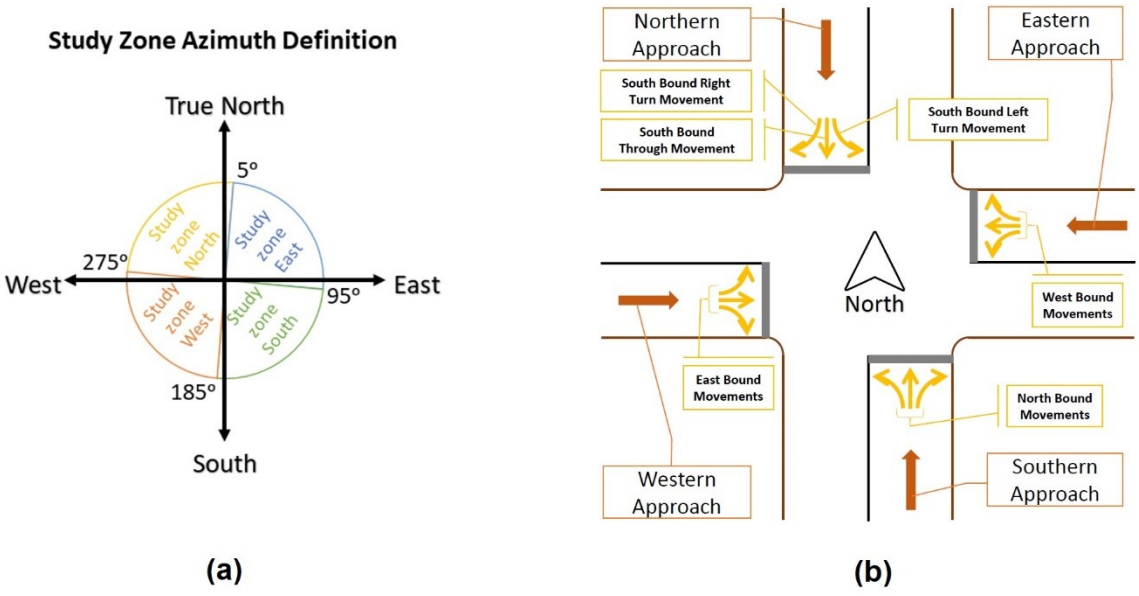


Figure 5-2. Definition of Direction (a), Movements and Approaches (b)

models were tested at the intersection level:

- Speed and count attributes for all week
- Speed and count attributes for workday AM peak period
- Speed and count attributes per approach for all week
- Speed and count attributes per movement for all week
- Delay and count attributes per movement for all week

Two supervised machine learning classification techniques were tested: decision trees and nearest neighbours. The classifiers were trained using 80% of all intersections within the data set. The model was then applied to the remaining 20% of the intersections (validation dataset – 25 intersections) to predict the control type. The model prediction was compared with the ground truth to assess the accuracy and select the best model using the validation dataset.

5.4 Findings

It was found that the best predictors of intersection control type were average speed per approach, standard deviation of speed per approach, maximum speed per approach, trip count per approach, and the percentage of points having a speed lower than or equal to 5 km/h per intersection. Table 5-1 presents the average values of the significant approach-level variables over all the study area intersections. These variables were able to distinguish between the speed and trip count characteristics specific to each control type. For example, the average speed approaching an intersection was a significant indicator in determining if an all stop control, stops at the secondary approach, or a traffic light was present as they have different average speeds.

A higher average speed was observed for approaches that are controlled by traffic lights or that are uncontrolled. In addition, trip count was a good indicator of control type since traffic lights have higher observed trip counts than all stop-controlled intersections, because traffic lights are usually implemented at higher traffic intersections. Intersections with stops on the secondary approaches also have a significantly higher trip count on the main approaches compared to the secondary approaches, which classifies them in their own category. Since the variables were compiled per approach, it was possible to predict on which approaches were the stops located (E-W or N-S). Moreover, standard deviation of speed was found to be a good determinant of control type since it reflects the different classes of variability in speed for different control types. It is seen that stop controlled approaches have a lower standard deviation, because all vehicles are coming to a stop, while traffic light-controlled approaches have a higher standard deviation due to the higher variability in speeds caused by the traffic light colour. Finally, the maximum speed was found to be the highest for traffic light-controlled approaches, followed by uncontrolled approaches, and then stop-controlled approaches, which was significant in discriminating between intersection control types. The higher maximum speed of traffic light-controlled approaches compared to uncontrolled approaches, is that a green light ensures that the driver has the right of way and traffic lights are usually implemented on higher capacity roads that usually have higher posted speeds less traffic calming measures.

The best predictions were obtained using all weekdays data set using the nearest neighbours classifier. The model predicted the intersection control type with the accuracy of 96% for the validation dataset. Figure 5-3 presents a confusion matrix showing the prediction error for the validation intersections using the best model.

Control Type	*Std Dev. of Speed West*	*Std Dev. of Speed South*	*Std Dev. of Speed North*	*Std Dev. of Speed East*	*Max. Speed West*	*Max. Speed South*	*Max. Speed North*	*Max. Speed East*
All-Way Stop	0.55	2.49	1.91	2.20	5.71	15.01	15.65	14.75
E-W Stop	1.94	7.15	6.63	0.90	8.90	33.68	31.91	10.06
N-S Stop	5.71	1.37	0.82	4.21	30.84	3.74	7.80	32.49
Traffic Light	10.56	6.40	6.74	10.27	45.64	28.31	27.76	52.70
Control Type	*Avg. Speed West*	*Avg. Speed South*	*Avg. Speed North*	*Avg. Speed East*	*Trip Count West*	*Trip Count South*	*Trip Count North*	*Trip Count East*
All-Way Stop	4.81	11.60	12.42	11.60	4.95	11.95	10.21	5.58
E-W Stop	6.92	23.30	22.65	9.10	2.26	26.16	34.68	1.58
N-S Stop	22.67	2.52	7.11	23.60	40.47	0.68	0.95	41.79
Traffic Light	26.57	14.91	16.71	28.98	89.58	23.42	29.42	100.89

Table 5-1. Average of Approach Variables' Values per Control Type Over All Intersections

Developing the model based on the AM peak period of workdays reduced the total sample size considerably, resulting in a low prediction accuracy. In addition, introducing the detail of all intersection movements (inbound and outbound direction) in the model, also reduced the model's prediction power.

For projects requiring a higher prediction accuracy, the model can potentially be improved by using a larger sample size to train it. A larger sample size enables the model to have a higher resolution and examine the data patterns in more detail. In addition, since traffic conditions have significantly different characteristics during different times of the day/week, developing a model based on homogeneous temporal characteristics might improve the prediction accuracy if a larger sample is available. Another potential avenue would be to test different model types. In sum, GPS data has a great potential to infer transport network variables for areas where such data is not easily available.

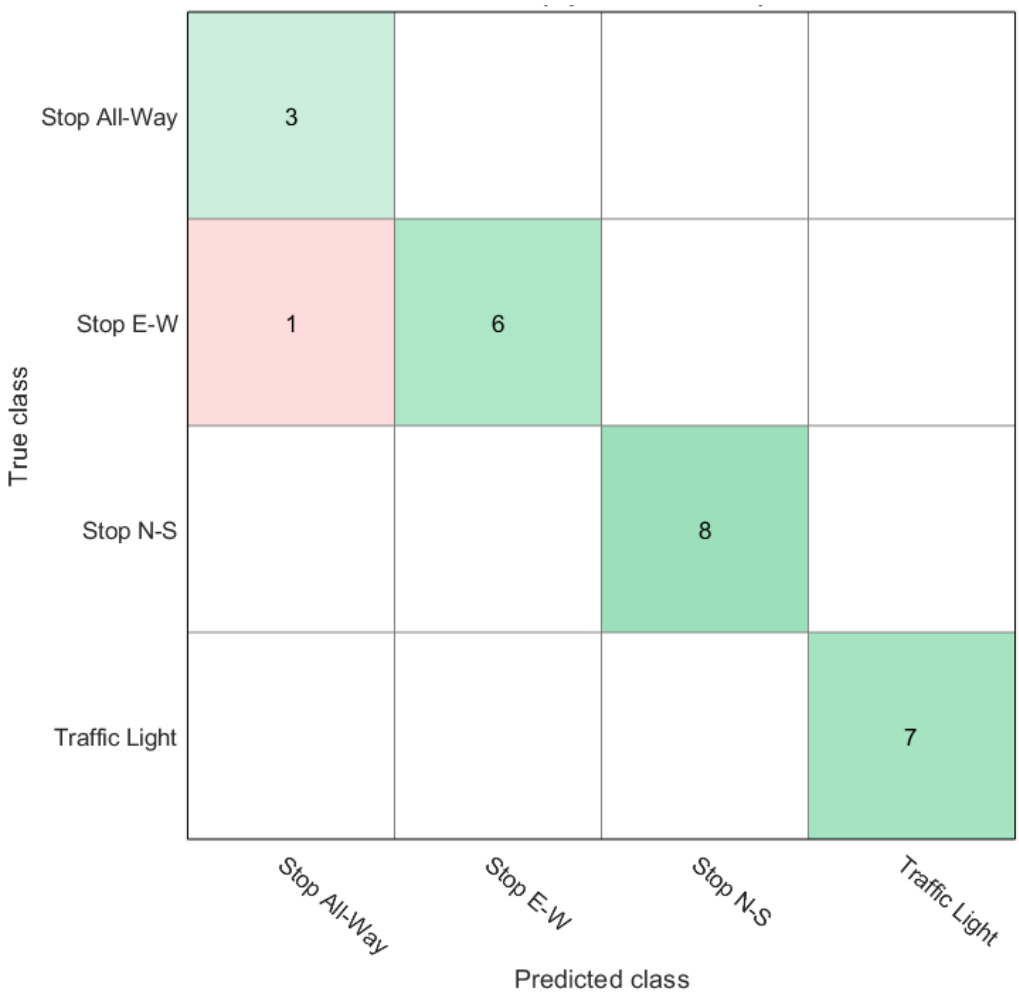


Figure 5-3. Confusion Matrix

References

Antoniou, C., Dimitriou, L., & Pereira, F. (2018). Mobility patterns, big data and transport analytics: tools and applications for modeling: Elsevier. doi:10.1016/C2016-0-03572-6

Barceló, J., Kuwahara, M., & Miska, M. (2010). Traffic data collection and its standardization: Springer. doi:10.1007/978-1-4419-6070-2_1

Link Between Chapters

In Chapter 5, the developed road network model was enhanced by extracting road intersection control type from GPS trajectory data. It is expected that this method can be generalized to a larger road network to predict intersection control type. The preferred method would be to train a model using GPS data obtained from the same region for which the prediction needs to be carried out. The model can further be expanded by including more intersection control types, such as yield controlled approaches.

The contribution of Chapter 5 is built upon in Chapter 6 to classify the road intersections with respect to the control type and develop turn penalty functions per road type, per intersection type, and per turn type. Considering the intersection control type allows to reflect the difference in vehicular dynamics for each type of intersection control. This has been demonstrated in Chapter 6 since vehicle speed-related attributes were found to be good predictors of intersection control type. Finally, Chapter 6 presents a method to develop intersections turning movement penalties based on GPS trajectory data and integrate the findings in large scale transport models.

Chapter 6 - Intersection Movements Delay Modelling Based on Crowd-sensed Global Positioning System Trajectory Data

Intersection Movements Delay Modelling Based on Crowd-sensed Global Positioning System Trajectory Data

Adham Badran [a*], Ahmed El-Geneidy [b], and Luis Miranda-Moreno [a]

[a] Civil Engineering Department, McGill University, 817 Sherbrooke Street West, Montreal, H3A 0C3, Canada

[b] School of Urban Planning, McGill University, 815 rue Sherbrooke West, Montreal, H3A0C2, Canada

[*] Corresponding Author

Contact: adham.badran@mail.mcgill.ca, Department of Civil Engineering, McGill University, 817 Sherbrooke Street West, Montreal H3A 0C3, Canada.

6.1 Abstract

Developing accurate large-scale transportation models, used to guide policy adoption and evaluate infrastructure alternatives or changes in sociodemographic conditions, is data and resource intensive. This research proposes a novel method for modeling intersection movement delay using crowd-sensed Global Positioning System (GPS) data. This is achieved by providing a general definition of turning movements and extracting travel times from GPS trajectory data analysis. Additionally, a straightforward method is proposed to integrate the observed delays per movement type into volume-delay functions. The spatial definition provided for turning movements captured distinct speed profiles per turn type. The significant differences in mean speeds for different turn types highlights the importance of integrating turn penalty functions

based on real observations and underscore the importance of crowd-sensed GPS data. A simple technique is also proposed to integrate the proposed method into the volume-delay functions used in large scale transport models.

Keywords: Intersection Delay Model; Macroscopic Model; Turn Performance Function; Global Positioning System, Transport Planning.

6.2 Introduction

Transport models are decision-making tools used to evaluate current system conditions and propose modifications to it to optimize its performance (Jacyna et al., 2014). They assist in evaluating the impact of policies, sociodemographic changes, and infrastructure projects on the transport system (Wegener et al., 1991). Large-scale transport models, known as macroscopic transport models, consist of three components: i) the supply, a digital representation of the transport network for all modeled transport modes, ii) the transport demand, representing all the trips that need to be made, and iii) the performance, depicting network conditions when the demand is assigned to the transport network reflecting the influence of demand on route choice and traffic conditions (Ortúzar and Willumsen, 2011). Road network performance is usually evaluated by examining travel time delays on road segments and at intersections (Ledezma-Navarro et al., 2018, Sun et al., 2014). Delays at intersections originate from two main sources, traffic signals and turning movements. Turning movement delay at intersections depends on multiple factors, such as the number of approaches, intersection control type, intersection size, number of conflicting movements, traffic intensity, presence of dedicated turning lanes, and traffic signal phasing and timing (in presence of traffic lights)(HCM, 2022). Acquiring data for all

these variables at a regional level is challenging and even more complex to maintain up to date. Due to the complexity of developing such models, some modelers rely on major assumptions regarding turn penalty functions that represent turn movement delays in macroscopic models or use generic penalties that represent turning movement delays with sufficient accuracy. The impact of these inaccuracies is directly reflected in the route choice results since the generalized cost is mostly based on delays or travel times, which can lead to misleading results. This weakness has also been identified by Abedini (2022) who proposed a data-driven method to calibrate more accurate link performance functions.

Recently, Global positioning systems (GPS) trajectory data has been collected by GPS enabled smartphones, creating large databases of GPS trajectories. This emerging data source has the potential to provide high-resolution and high-coverage information about the observed motorist's speed or travel time throughout the road network, offering an opportunity to improve the current macroscopic modelling practice. The objective of this work is to demonstrate the potential of crowd-sensed GPS data to accurately model road intersection turning movement delay, using as a case study dataset from Quebec City, Canada. It also aims to show how such information can be integrated into large-scale simulation models to provide more accurate intersection delay functions. This is achieved through the adoption of a replicable and standardized procedure to calculate the average speed per turning movement. Average speed is selected since large-scale transport models are deterministic and represent an average day. This method is not adapted for use with dynamic traffic assignment models since it does not model turning movement delay as a random variable. The case study examined in this paper examines

the turning movement delays at traffic signal-controlled intersections of arterial-arterial or arterial-collector type roads.

6.3 Literature Review

Intersection delay estimation and modelling, using GPS trajectory data, has been addressed in multiple studies (Jiang and Zhu, 2005, Ko et al., 2008, Strauss and Miranda-Moreno, 2017). These studies can be categorized based on the examined transport mode (car, bus, or bicycle).

Strauss and Miranda-Moreno (2017) conducted a study using crowd-sensed GPS trajectory data in Montreal, Canada to estimate performance measures at signalized intersections. They developed models to relate bicycle intersection delays to predictors such as intersection geometry and built environment. While this work provides detailed steps in GPS data processing, it confines the analysis to the approach and intersection levels without exploring detailed intersection movements. Another study by Gillis et al. (2020) used crowd-sensed cyclist GPS trajectory data to determine road intersection delays. This research focuses on the main cyclist movements across the intersection and emphasizes the importance of having an adequate sampling rate to capture details before and after the intersection. The main limitations of the two studies examining cyclist GPS data are the fact that they do not consider the impact of traffic flow on delay and that they do not propose a standardized method to extract delays at the intersection movement level.

Using real-time bus GPS trajectories, Wang et al. (2016b) proposed a method to predict intersection delays and bus arrival time. This method, designed for real time use, does not explicitly consider intersection movements, making it inapplicable for macroscopic transport

models. Another study by Wang et al. (2016a) uses low-resolution transit bus GPS data to estimate control delays; however, it does not consider turning movements. In addition, using bus GPS data to estimate control delays cannot be used to represent the dynamics of the general population of motorists, as it may be biased due to differences in vehicle characteristics and the presence of bus stops, which can create additional delays.

One of the most used methods to estimate intersection movement delays is proposed by the Highway Capacity Manual (HCM). It combines three models: uniform, random, and overflow delay models. This method can be seen in the work by Leong (2017) and requires the collection of signal phasing and timing information, in addition to intersection configuration. Although this method can yield good results, it requires significant data collection efforts for large-scale models, limiting its suitability to small-scale models.

Other studies have explored the use of passenger vehicle GPS trajectory data to estimate delays while reducing data collection efforts and having a satisfying accuracy level. In fact, a study by Liu et al. (2006) investigated the effect of different GPS trajectory sampling rates on delay estimation quality and the ability to capture the delay. This study focused on reducing the cost of real-time data transmission and does not propose a method to estimate or model intersection movement delays.

In another study, Alkaissi et al. (2021) conducted an experiment by instrumenting a vehicle with a GPS device to record 50 trips through an arterial corridor. Based on speed and acceleration, they were able to determine delays at intersection; however, the study only considered a limited number of trips and did not examine delays from movements at the intersection.

Intersection delay estimation techniques were examined based on a theoretical framework of vehicle dynamics. In a study by Jiang and Zhu (2005), a GPS-equipped vehicle was used to collect trajectory data, proposing a method to calculate the approach delay. The approach delay is defined as the difference between the actual time for the vehicle to pass the intersection and the time it would take to pass the intersection at the driver's desired speed. This delay can be estimated by measuring different various components such as stopped delay, control delay, approach delay, midblock delay, or segment delay. A variation of this technique was explored by Hoeschen et al. (2005). However, these measures remain limited to traffic signal operation applications and only consider delays at the intersection approach level.

Intersection delay is crucial information for assessing intersection control performance and determine the level of service (LOS). Tišljarić et al. (2018) estimated intersection control delays based on GPS trajectory points by locating the first deceleration and stopping points upstream on the intersection. The information was also used to create a queuing profile for the examined intersections. However, this technique was limited to the approach level and the queuing profiles were not compared to ground truth for validation.

When studying delay modelling, understanding the level of detail required depends on the model type and the capabilities available in transport planning and modelling software to be able to produce results that can be integrated to the modelling tool. Macroscopic models integrate intersection movement delays differently depending on the modelling tool used. For example, the Aimsun simulation software divides delay into three different components: link delay functions, turn penalty functions (TPF), and junction delay functions (JDF). TPF and JDF are used for traffic signal-controlled intersections and stop or yield controlled intersections, respectively.

The TPF is also capable of using the programmed signal timing plan to estimate macroscopic level delays based on green time, cycle duration, and equations provided in the Highway Capacity Manual. Although this possibility is interesting, integrating and maintaining all signal timing plans for different time periods and for a whole metropolitan region requires important resources and is generally not feasible.

Other tools used for macroscopic modelling, such as EMME or Visum also offer the possibility to add turn penalties for each possible movement at an intersection. However, the challenge remains in finding the correct values or functions that represent the observed conditions adequately. Due to limited resources, in practice, this usually results in the oversimplification of turn delay modelling by assuming fixed generic values or even by limiting turn modelling to simple turning permissions indicating whether each movement is permitted or prohibited.

In summary, intersection delay was studied by multiple researchers using GPS trajectory collected by different transport modes, such as bicycles, buses, and passenger cars. Depending on the study objective, delay was defined differently in terms of spatial or temporal resolutions (intersection level or approach level) to obtain indicators used for traffic signal control operation and optimization. However, additional work is required to explore crowd sensed GPS data and develop methods that consider delays at the intersection movement level without the knowledge of signal phasing and timing or signal groups. This is essential to model turning movement delays for large-scale models. Therefore, this work proposes a framework and method to extract intersection movement delays for use in large-scale transport models from GPS data, avoiding the use of data that is difficult to obtain or collect.

6.4 Methodology

Definitions

Before describing the theoretical framework and the proposed method, it is important to define a few terms. An intersection turning movement refers to a possible vehicular movement at an intersection, usually described by the direction and the turn type (Board et al., 2022). Intersection turn type refers to the maneuver performed at the intersection, which can be left turn, through movement, or right turn. Although delay and speed are two different concepts, this work interchangeably uses the two words. Since the proposed method needs to be applicable to intersections of different dimensions, speed was calculated instead of delay to eliminate the distance dimension and reduce the bias. This is important for the proposed method, as it includes the upstream segment travel time in the delay (speed) calculation. Calculating a typical delay value for all types of intersections would incorrectly assume that all intersections have the same geometric configurations and upstream road segment length.

To capture the average delay incurred by a vehicle associated with a given turning movement and keeping in mind the macroscopic aspect of the transport model, it was important to have an adequate definition of intersection movements. For each intersection, an intersection zone is defined as the area containing the road intersection in addition to all the upstream and downstream road segments that connect the given intersection to the neighboring intersections (see Figure 6-1).

Moreover, the start and end points for each movement type (left turn, through movement, and right turn) are defined as seen in Figure 6-2. The start point of every movement is the entrance

point of the upstream road segment (LT_{Start}, T_{Start}, RT_{Start}). The movement end point is the point where the vehicle exits the analyzed intersection (LT_{End}, T_{End}, RT_{End}). Defining the start and end point of every movement enables the calculation of length of each of the left, through, and right movements, which are L_{LT}, L_T, and L_{RT}, respectively. This definition makes it possible to differentiate between delays of vehicles performing different movement types. In a similar logic, the traffic flows for each of the movement types are referred to as F_{LT}, F_T, and F_{RT}, representing flows for left turn, through, and right turn movements, respectively. Connecting back to macroscopic models, it becomes possible to adjust turn penalties based on real observations while considering mid-block traffic delays due to traffic propagation associated with the downstream control type and turning movement type.

Proposed Procedure

The method proposed by this work uses GPS trajectory points, traffic counts, and a road network geographic representation to create an integrated database containing, for each intersection movement, the mean 15-min speed and the corresponding 15-min traffic count. Figure 6-3 presents a summarized diagram of the procedure used to create the traffic count-speed database.

The yellow boxes represent input data while the grey rectangles represent data processing steps, and the green cylinder represents the final output database.

The first step consists of spatially filtering the map-matched GPS trajectory data to allow only relevant data points to be kept and reduces the size of the database. This step is required to only keep the required GPS points and avoid working with a large data file. The second step is to

manually select, for each trip segment within the intersection, the first point (LT_{Start}, T_{Start}, RT_{Start}) and the last point (LT_{End}, T_{End}, RT_{End}). Each trip within an intersection zone is visually inspected to verify if its start point and end point are located at an acceptable distance of the theoretical start and end points defined above. This step is carried out manually and is labor intensive given the large number of trips per intersection. At the third step, the trip ends' timestamps and the geographic coordinates are extracted to create a polyline representing the turn movement of each trip segment within the intersection. The fourth step connects the trip ends using the shortest path algorithm over the digital road network. The process allows the elimination of noise caused by the GPS signal when a vehicle is stationary at trajectory points situated between the trip ends. This step is carried out using the Network Analyst Extension of the ArcGIS software which implements Dijkstra's algorithm to find the shortest path. This algorithm was deemed suitable since it was able to correctly connect the first and last points of intersection trajectories. Figure 6-4 presents the raw GPS data in addition to two sample trip segments that were manually selected to be processed into a line using the shortest path algorithm and considered in the delay analysis.

The fifth step consists of using the turning movement trip segment polyline to calculate the intersection movement length and speed.

The following step, each turning movement trip segment is analyzed to determine the movement type (left turn, through movement, or right turn) based on the movement's in and out directions. A movement type-direction correspondence dictionary is used at that step to determine the entering and exiting direction for each trip and associate it to the correct movement type. For

example, a vehicle entering an intersection from the south and exiting from the east is considered a right turn. At the seventh step, mean 15-min speeds are calculated per intersection movement.

The last (eighth) step is an independent treatment of traffic counts carried out to extract and prepare traffic count data to be integrated to the mean 15-min speed table. Therefore, a traffic count database is created containing detailed 15-min traffic counts for all intersections per turning movement. This database is integrated into the mean 15-min speed table based on the intersection ID and the turning movement to create the final 15-min traffic count-speed database. The final database is used to perform exploratory analysis to gain insight into the different movement types.

Integration to Macroscopic Models

To connect with large scale transport models, a method is then proposed to integrate the findings to the volume delay functions used in macroscopic simulation models. Assuming that through movement delays are already included in the link, or road segment, volume delay function, it is possible to express the turn penalty, seen as an additional delay, as a function of through movement travel time T_T. This assumption is applicable since large scale transport models are calibrated based on floating vehicles that drive straight through main road corridors without turning at intersections. This results in link volume delay functions that integrate road segment and intersection delay for through movement only (T_T in Figure 6-2). The following are the proposed left turn and right turn penalty functions based on the observed GPS trajectory data.

(1) $$T_{LT} = T_T + a * T_T = T_T(1 + a)$$

(2) $$T_{RT} = T_T + b * T_T = T_T(1 + b)$$

Where T_{LT} and T_{RT}, are the travel times for the left and right turns, respectively, and parameters *a* and *b* are the speed adjustment ratios for left and right turns respectively. These parameters are calculated using the trajectory length and travel time extracted from the GPS trajectory points. The parameters a and b are calculated as follows:

(3) $$a = 1 - \frac{L_{LT}/T_{LT}}{L_T/T_T}$$

(4) $$b = 1 - \frac{L_{RT}/T_{RT}}{L_T/T_T}$$

For macroscopic models, the adjusted travel time for turning movements at intersections, or turn penalty functions can be considered as follows:

(5) $$TP_{LT} = a * T_T$$

(6) $$TP_{RT} = b * T_T$$

Where TP_{LT} and TP_{RT} are the additional delay incurred for left turning vehicles and right turning vehicles, respectively, with respect to the through movement travel time. The use of these penalties results in the inclusion of all delays incurred at the intersection for all turn types.

Case Study

This study is based on data collected in Quebec City, Canada. Three sources of data were necessary. First, GPS trajectories data was recorded during the spring of 2014 in Quebec City, Canada. It was collected during 21 days by 2,000 voluntary users through the Mon Trajet smartphone app, made available by the Municipality. Each point is described by the following attributes: X and Y coordinates, trip ID, speed, and timestamp (Year-Month-Day-Hour-Minute-

Second). The GPS data had gone through a preliminary round of preparation and map matching. The second data source, used at step number 8 of the methodology, is traffic counts collected and provided by the Municipality of Quebec City. Traffic counts were available for a one-day period per intersection for 15-min time intervals from 7:00 to 10:00 and from 15:00 to 18:00. These periods were selected by the municipality to cover peak traffic periods. Finally, the last data source was a geographic representation of the road network in the form of a shapefile which was obtained from OpenStreetMap (OpenStreetMap, 2023). Figure 6-5 presents the location of the four intersections selected to perform this study. These intersections were selected based on the road type and the control type. These variables are expected to have an influence on intersection movement delay and can be obtained with a reasonable amount of effort for large scale transport models. In this study, traffic light-controlled intersections were selected, and the road type was limited to arterial-arterial or arterial-collector intersections.

A total of 1400 intersection movements were individually examined and 1136 were found to be adequate and selected for further analysis.

6.5 Results

Considering the four intersections that were analyzed in the case study, a total of 1136 trip segments (126 left turns, 153 right turns, 857 through movements) were extracted for the analysis period. The 15-min mean speed was the lowest for left turns at 14 km/h, followed by the right turns at 17 km/h, and through movement at 21 km/hr. Left turns are typically face conflicts with the opposite through traffic, requiring sharing of the green phase (with priority given to the opposite direction). In addition, left turns often conflict with pedestrian and cyclist users who

also have priority over motorists. To mitigate these conflicts, left turn movements are sometimes given a dedicated protected phase depending on traffic control design standards. Both situations contribute to the expectation that left turning movements have often slower travel times with respect to right. Regarding right turns, generally this movement conflicts with cyclists and pedestrians (who have priority), and occasionally conflicts with left turns from the opposite direction, but this is less frequent and less critical. Therefore, right turn delays are expected to fall between left turn delays and through movement delays. Through movement generally do not conflict with other movements (except for right turn on red); however, it's delay depends on the signal timing design based on traffic flows for all movements. Thus, observed speeds for through movements are reasonable since they are expected to be the fastest.

In parallel, the mean traffic count was the lowest for left turns at 33 vehicles per 15 minutes, followed by right turns at 36 vehicles per 15 minutes, and through movement at 77 vehicles per 15 minutes. The final database was used to visualize the frequency distribution of mean 15-minute speeds and 15- minute traffic counts for each intersection movement type, as shown in Figure 6-6.

Further analysis was conducted to examine the relationship between speeds and observed traffic counts. No evident relationship was found between the two variables. Additionally, the mean 15-minute speed is relatively volatile, explained by the fact that speed is affected by the intersection's signal timing, operation mode, and geometric configuration rather than traffic flow. Additionally, traffic counts and GPS trajectories were not collected at the same moment, which is not ideal when comparing relatively fine resolution data.

For this case study, “a” and “b” for traffic light controlled arterial-arterial or arterial-collector intersections are calculated using equations 3 and 4 to be 0.33 and 0.19, respectively. In other words, a left turn movement is 33% slower than a through movement, considering movement definitions in Figure 6-2, and a right turn movement is 19% slower than a through movement. These parameters (a and b) represent an average behavior of the analysis period as estimated using all observations. However, with more data is available, it is possible to recalculate these parameters per peak period or hour of the day to increase the accuracy.

6.6 Discussion

The large-scale aspect of macroscopic transport models, sometimes referred to as strategic level models, can benefit from the availability of new sources of data for calibration. The proposed framework and methodology can process crowd-sensed GPS data to estimate turning movement delays and integrate them to macroscopic models. The proposed solution is a balance between the delay estimation methods proposed by the HCM or by Hoeschen et al. (2005) and Jiang and Zhu (2005), which are data-intensive when the model is very large, and the simplifications imposed to macroscopic models due to the lack of data and resources. Using GPS trajectory data, it was possible to develop a standardized method to extract speed information at the intersection turning movement level. Traditionally, delays were only calculated for operational purposes to design and optimize traffic signal phasing and timing, therefore, research mostly examining approach level delay, which is also used for level of service assessment, as can be seen in the work by Tišljarić et al. (2018).

Using the extracted results, it was possible to determine the frequency distribution of speeds and traffic counts for each of the turning movement types. These distributions can eventually serve to calibrate other stochastic transport models through distribution fitting and sampling variable delays based on the observed mean and variance values. However, for macroscopic transport models, aggregate speed results were used to propose a method to include GPS-based delays to turning movements. In fact, the main finding is that left turn movements for traffic signal-controlled arterial-arterial or arterial-collector intersections have the lowest average speed compared to through movements and right turns. In addition, right turns were also found to have a lower average speed than through movements. This justifies the importance of including turn penalty functions that reflect this difference in observed speeds, which was the motivation of this work.

The proposed method can be applied to a larger sample of intersections, a larger sample of GPS trajectories, and for a variety of road types for better coverage of the road network. The procedure is semi-automated for the moment and will require the automation of some the tasks to make it feasible to treat many trajectories rapidly. This will also allow for the inclusion of more GPS trajectories in the analysis allowing for better temporal coverage.

No clear relationship was found between mean 15-min speeds and 15-min traffic counts. Although this is explained mainly by the intersection control type, which in this study was traffic signal control, the fact that only one day of traffic counts was available per intersection from a different year might contribute to the randomness observed in the speed-flow chart.

This study controlled for intersection control type and road type. Intersection delay can be influenced by additional variables such as the number of available lanes, the presence of dedicated turning lanes, the permission to perform a right turn on red, the number of conflicts, the type of traffic signal (fixed vs. actuated). Obtaining and maintaining these variables up to date at a regional level is challenging. However, if any of them is available, it could be interesting to include it to improve the classification of turning movements and improve the delay prediction.

6.6.1 Limitations

This work explores a new method to use GPS trajectory data to model turn movement delay per road type, movement type, and intersection control type for large-scale transport models. Although it makes use of the emerging availability of GPS trajectory, it is not without limitations. First, the applicability of the proposed method is to deterministic static transport models that aim to represent an average situation to be used for strategic planning and alternative comparison. Therefore, it is not possible to apply this method to dynamic traffic assignments, further analysis would be required to do so. Moreover, the case study examined in this work was limited by the available data. The GPS trajectory data sample, traffic counts availability, and unavailability of ground truth data were all limiting factors. To cover all types of intersection turn types, road types, and control types, a larger road network should be used in addition to a larger GPS trajectory data sample. Moreover, a larger GPS trajectory data temporal coverage will enable the modelling of turning movement delay per time of day to better reflect the variation of travel time during peak and off-peak periods.

6.7 Conclusion

This work emphasizes the need to consider intersection movement delays in macroscopic transport models. It explores the availability of a new data source that can overcome data collection challenges, typical in macroscopic models. It also complements the work done on delay modelling for different transport modes, which focuses on the operational needs. It was found that crowd-sensed GPS data is suitable to estimate intersection movement delays at the intersection movement level. The case study examined traffic signal-controlled arterial-arterial and arterial-collector type intersections. Average speeds were found to be different for left turns, right turns, and through movements, justifying the importance of considering turn penalties. These speeds were then used to propose a method to integrate them back into macroscopic transport models to improve travel time estimation and consequently improve route choice.

The proposed method can be further improved by increasing the automation of the procedure, allowing for the rapid treatment of many GPS trajectories. This, in turn, will increase the sample size of the observations and allow to estimate different turn penalties per peak period or per hour. Moreover, an extension of this work can examine different methods to address the length variable to ensure that no bias is introduced since different road segments can have different lengths, which can in turn influence the calculated turning speed. Furthermore, if more intersection variables are available, such as the number of lanes, the number of conflicts per movement type, the possibility to turn right on red, the presence of dedicated turning lanes, or other intersection control variables, they can be included to classify turning movement to improve turn penalty estimation accuracy.

Acknowledgments

The author would like to acknowledge the generous support of McGill University's Faculty of Engineering and the Vadasz Scholars Program.

References

ABEDINI, M. 2022. A Machine-Learning Framework for Clustering and Calibration of Roadway Performance Models with Application in the Large-Scale Traffic Assignment. M.A.S., University of Toronto (Canada).

ALKAISSI, Z. A., KADEM, A. J. & ALATTAR, E. F. 2021. Travel Time Prediction Models for Major Arterial Road in Baghdad City using Manufactured GPS device. IOP Conference Series: Materials Science and Engineering, 1090, 1-14.

BOARD, T. R., NATIONAL ACADEMIES OF SCIENCES, E. & MEDICINE 2022. Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis, Washington, DC, The National Academies Press.

HCM 2022. Highway Capacity Manual 7th Edition: A Guide for Multimodal Mobility Analysis. Washington, DC: The National Academies Press.

HOESCHEN, B., BULLOCK, D. & SCHLAPPI, M. 2005. Estimating Intersection Control Delay Using Large Data Sets of Travel Time from a Global Positioning System. Transportation Research Record, 1917, 18-27.

JACYNA, M., WASIAK, M., LEWCZUK, K. & KŁODAWSKI, M. 2014. Simulation model of transport system of Poland as a tool for developing sustainable transport. Archives of Transport, Vol. 31, iss. 3, 23-35.

JIANG, Y. & ZHU, K. Q. 2005. Traffic delay studies at signalized intersections with global positioning system devices. ITE Journal, 31.

KO, J., HUNTER, M. & GUENSLER, R. 2008. Measuring Control Delay Components Using Second-by-Second GPS Speed Data. Journal of Transportation Engineering, 134, 338-346.

LEDEZMA-NAVARRO, B., STIPANCIC, J., ANDREOLI, A. & MIRANDA-MORENO, L. 2018. Evaluation of level of service and safety for vehicles and cyclists at signalized intersections.

LEONG, L. V. 2017. Delay functions in trip assignment for transport planning process. AIP Conference Proceedings, 1892, 1-8.

LIU, K., YAMAMOTO, T. & MORIKAWA, T. 2006. Estimating delay time at signalized intersections by probe vehicles. Proceedings of ICTTS, 644-655.

OPENSTREETMAP 2023. OpenStreetMap. OpenStreetMap.

ORTÚZAR, J. D. & WILLUMSEN, L. G. 2011. Modelling transport, John wiley & sons.

STRAUSS, J. & MIRANDA-MORENO, L. 2017. Speed, travel time and delay for intersections and road segments in the Montreal network using cyclist Smartphone GPS data. Transportation Research Part D: Transport and Environment, 57, 155-171.

SUN, D. J., LIU, X., NI, A. & PENG, C. 2014. Traffic congestion evaluation method for urban arterials: case study of Changzhou, China. Transportation Research Record, 2461, 9-15.

TIŠLJARIĆ, L., ERDELIĆ, T. & CARIĆ, T. Analysis of Intersection Queue Lengths and Level of Service Using GPS data. 2018 International Symposium ELMAR, 16-19 Sept. 2018 2018. 43-46.

WANG, H., ZHANG, G., ZHANG, Z. & WANG, Y. 2016a. Estimating control delays at signalised intersections using low-resolution transit bus-based global positioning system data. IET Intelligent Transport Systems, 10, 73-78.

WANG, Y., JIANG, C. & REN, H. 2016b. Model of delay prediction for signalized intersection based on GPS data. Proc. AMTIA, 1-8.

WEGENER, M., MACKETT, R. L. & SIMMONDS, D. C. 1991. One city, three models: comparison of land-use/transport policy simulation models for Dortmund. Transport Reviews, 11, 107-129.

Figures

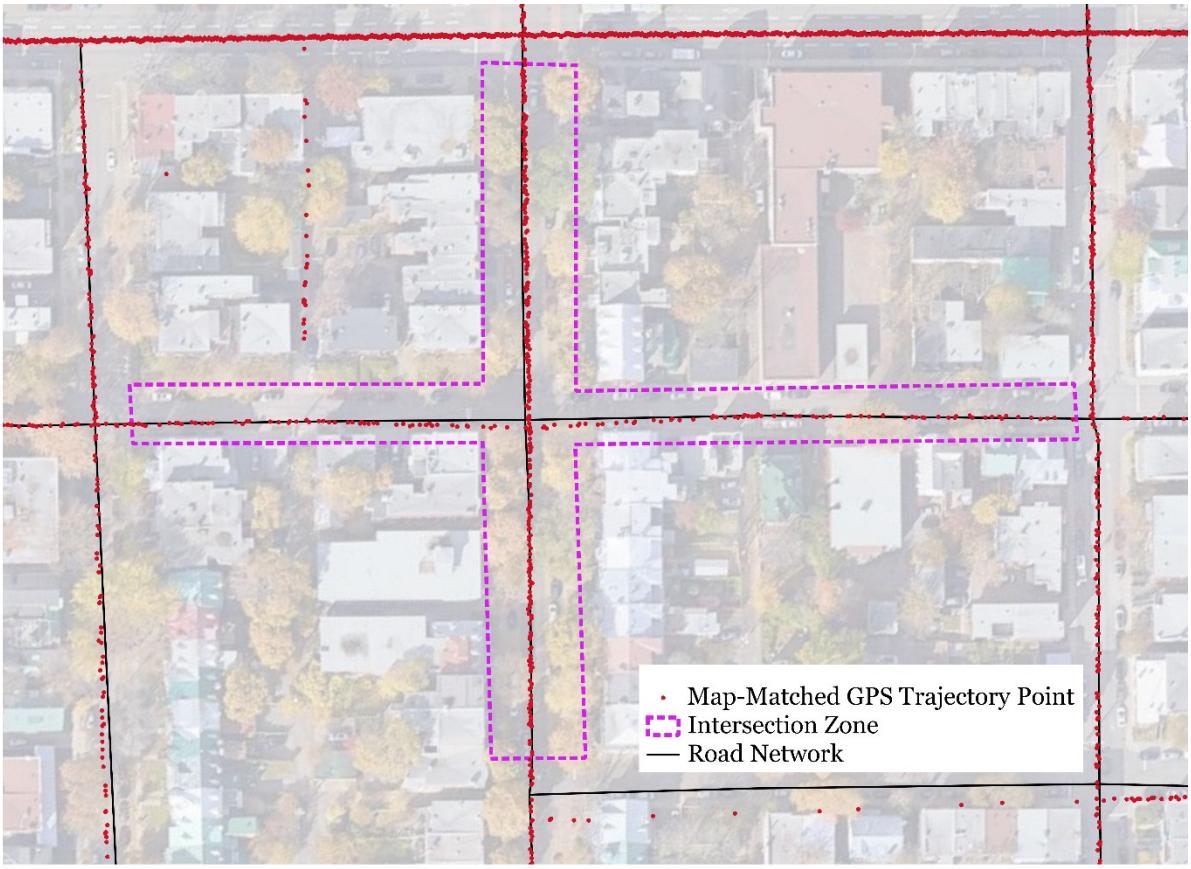


Figure 6-1. Intersection Zone Example

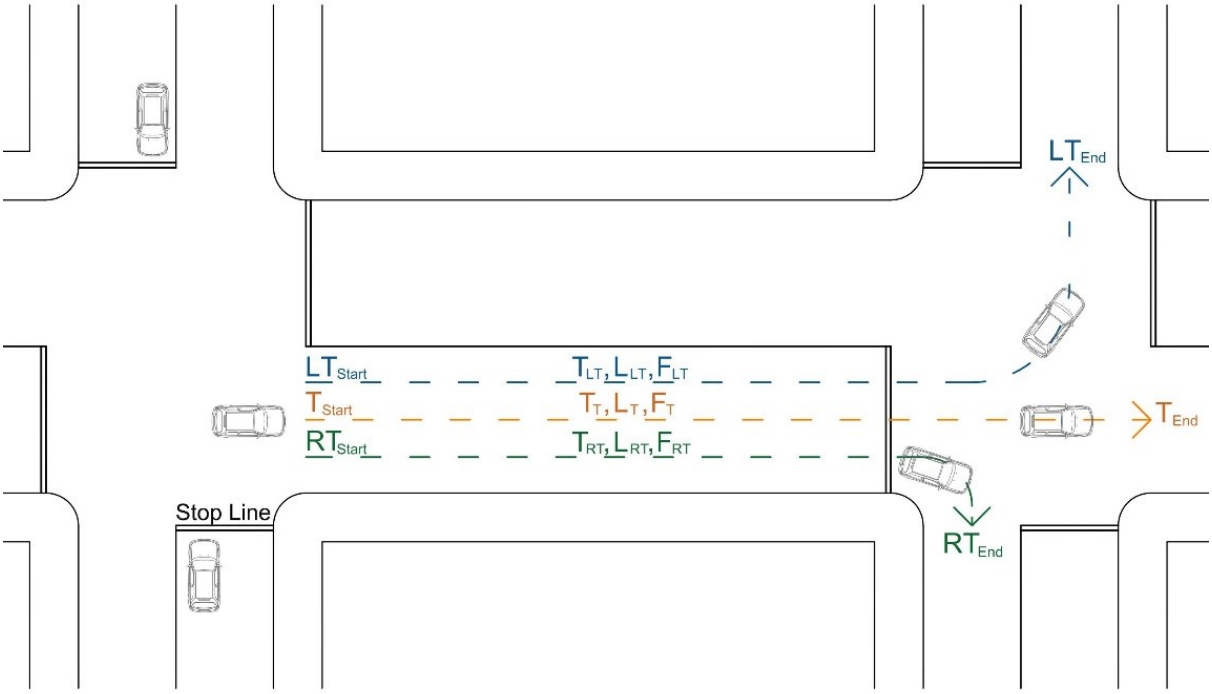


Figure 6-2. Intersection Movement Definitions

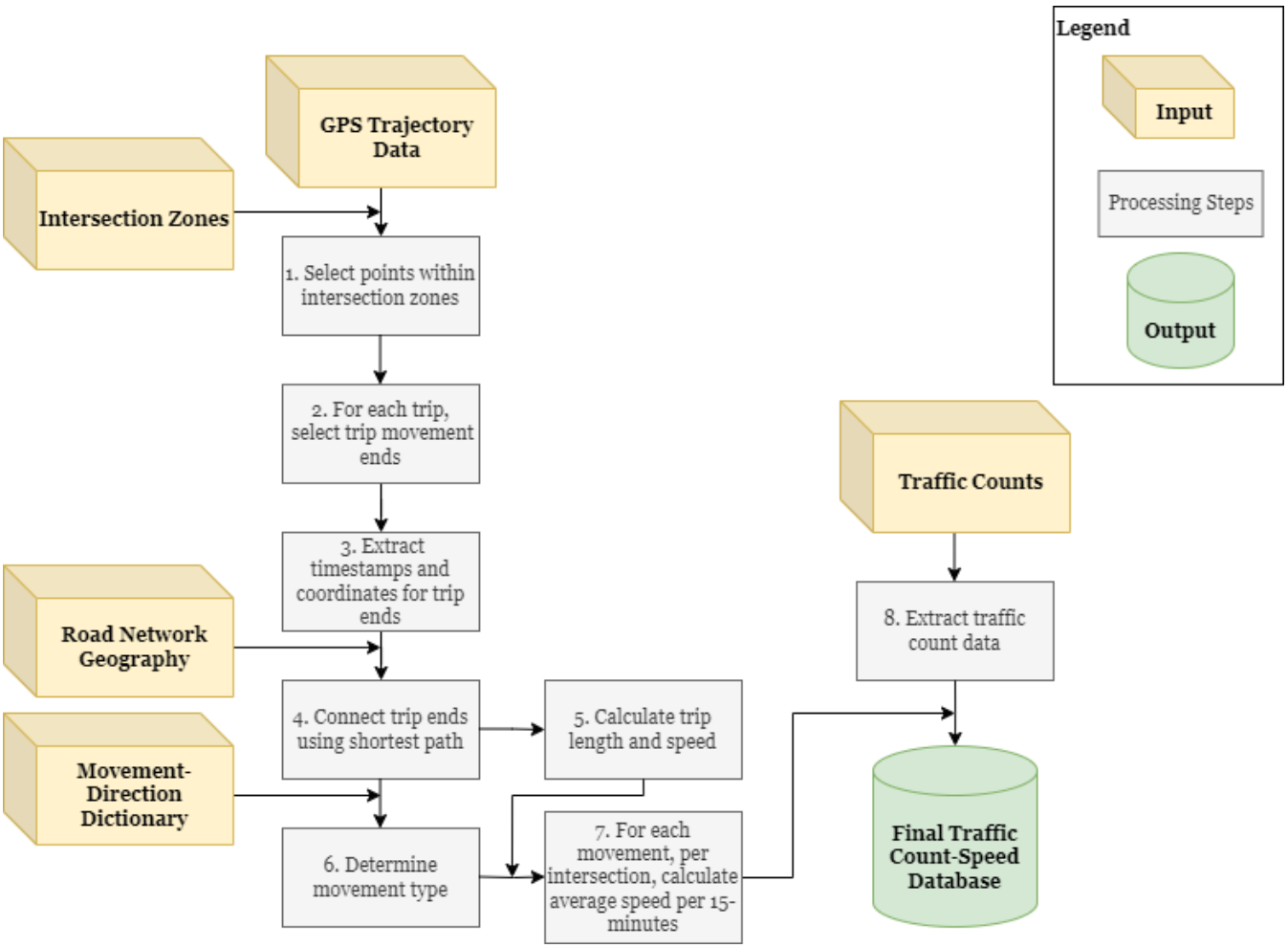


Figure 6-3. Diagram of Database Creation Procedure

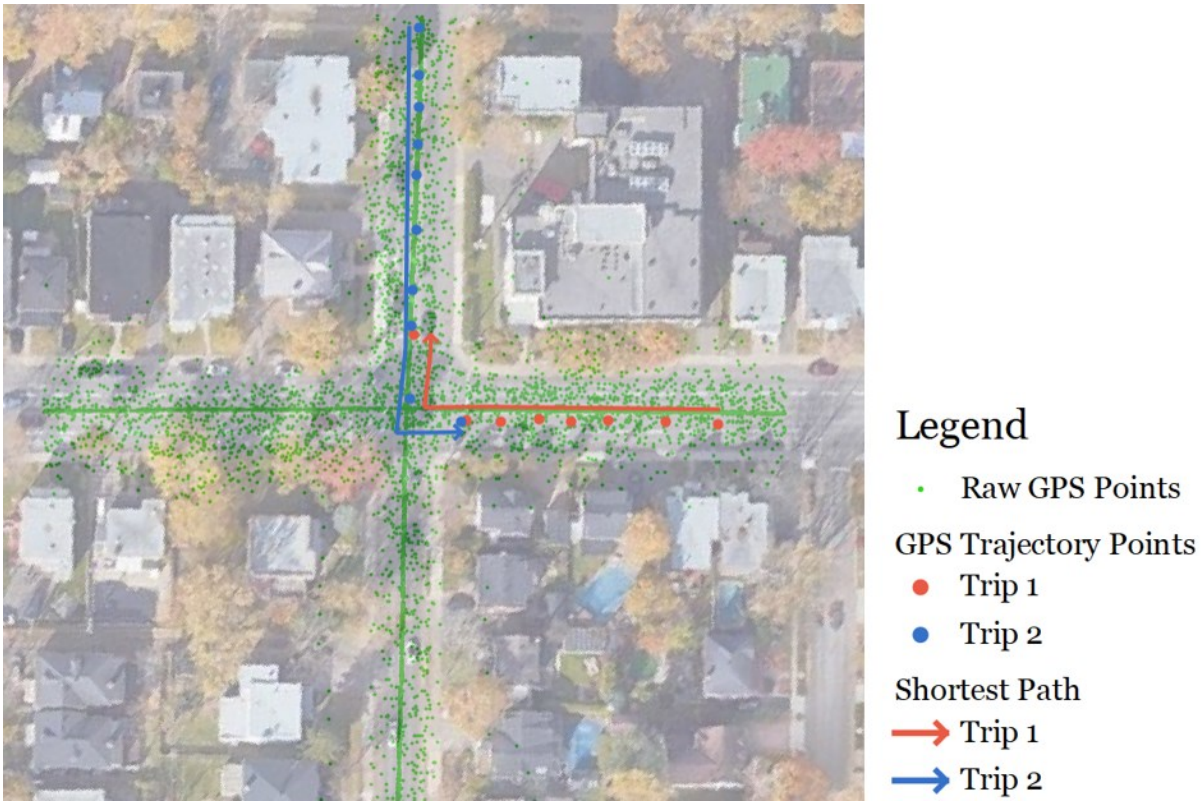


Figure 6-4. Sample GPS Trip Points Converted to Lines

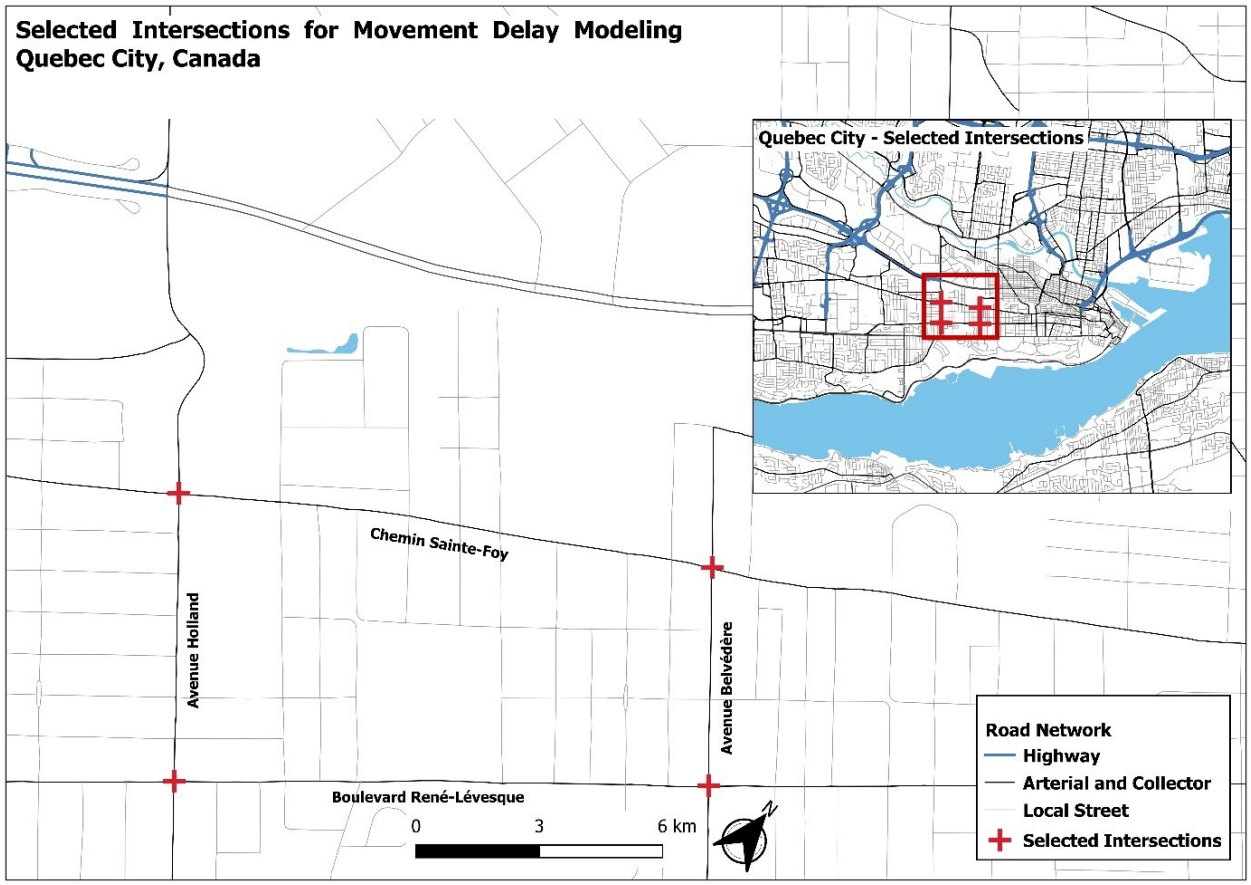


Figure 6-5. Study Location - Selected Intersections

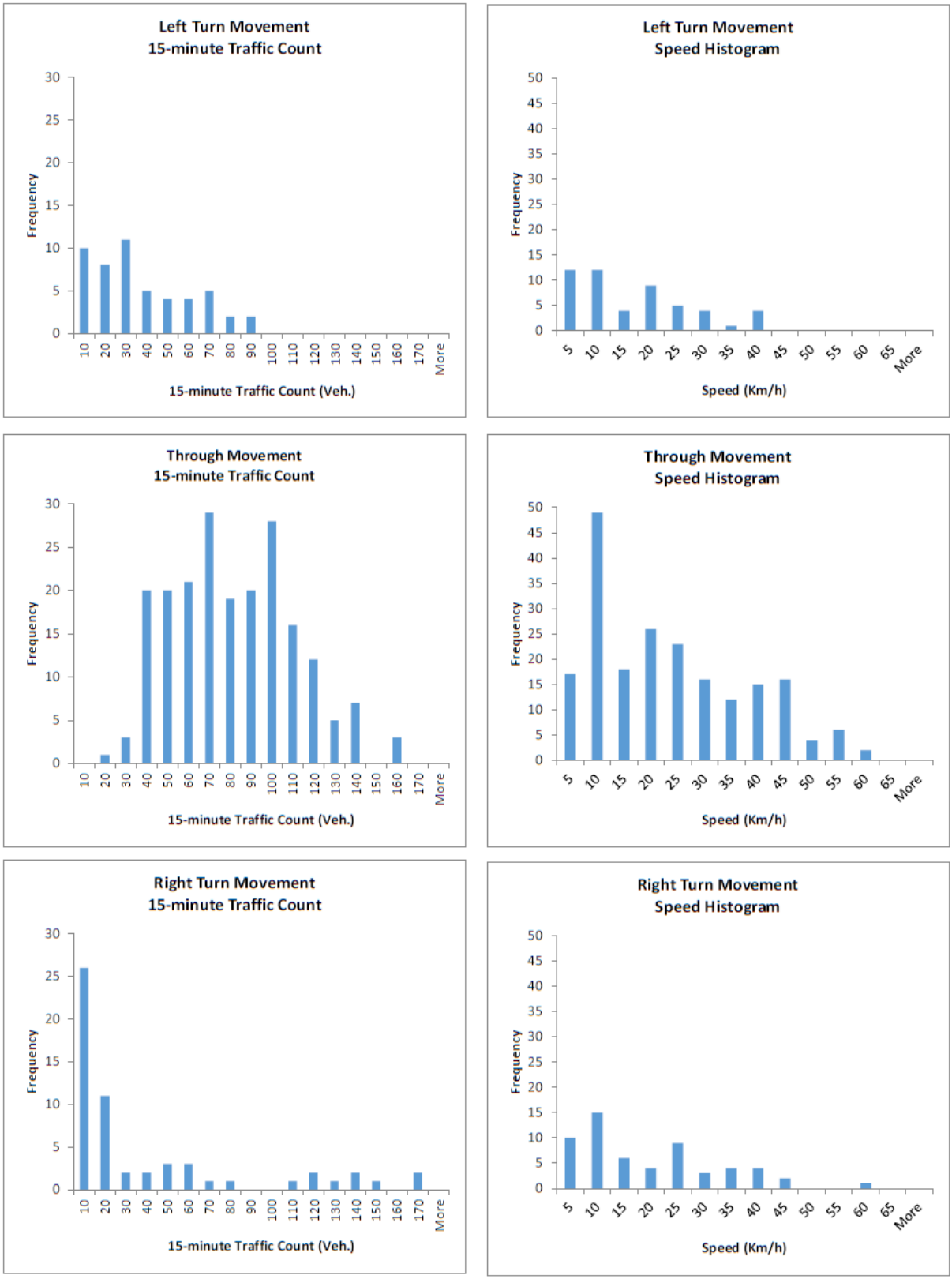


Figure 6-6. Frequency Distributions of Mean 15-min Speeds and 15-min Traffic Counts

Chapter 7 - Discussion

This dissertation addresses the use crowd-sensed GPS data to build on current large scale model development practice by increasing network modelling accuracy while facilitating its' updateability and reducing the required resources by proposing GPS data-driven methodologies to extract road networks attributes including: road segments' number of lanes information and road intersection control type based on GPS trajectory data as well as to calibrate turn delay functions per road type and turn type. This discussion presents a summary of the main findings, limitations, and future work directions.

7.1 Main findings

The comprehensive literature review found that multiple research efforts were carried out to extract road network features from GPS trajectory data and highlighted the different families of techniques: clustering, intersection linking, and track alignment. This review also confirmed the relevance of using GPS trajectory data to extract road network related information and showed that research was mainly carried out in the geography and computer science fields, which aim to achieve different objectives. Although road network extraction accuracy was high (Y. Guo et al., 2021), it was limited to extracting the road centerline and intersection location, therefore not adaptable transport network model development. Some of the past research efforts were not reproducible by reading the publication, therefore, one of the objectives of this research was to provide a detailed documentation of the methods to make it easier to implement and practice-ready. This research builds on current tool capabilities and makes use of available data sources to extract additional network features and improve current network modelling quality.

Figure 1-2 in Chapter 1 presented an overview of the contributions of this research to transport modelling and the relationship between the different developed methods, GPS trajectory data,

and current practice. Currently, transport modelling tools such as EMME and Aimsun use a road centerline shapefile to construct a road network in their proprietary format useable for traffic simulation. Since road centerline shapefiles usually have limited information regarding road directionality, connectivity, number of lanes, delay information, or intersection control type, the created road network using these tools, although in the correct format, require additional treatment to become more accurate and represent the actual road network. Therefore, this research proposes methods able to use crowd sensed GPS data to create a refined network model. First, EMME was used to create a simple road network for the study area. A procedure was then used to extract road directionality and intersection movements from GPS trajectory data. The road network was further refined by using the GPS trajectory data to predict the number of lanes for each road segment and the intersection control type. Based on the intersection control type, turning movement type and road type, a simple method was proposed to integrate turn penalties based on the GPS trajectory data.

Road Direction and Intersection Movements

To improve road network modelling, this research demonstrated the capability of crowd-sensed GPS trajectory data in improving a road network model generated by the EMME transport modelling tool using a road centerline geographic file. Using an existing road centerline representation provided a good road network skeleton and important priori information about road and intersection locations. Based on the frequency and direction of observed GPS trajectories, it was possible to infer road segment directions and turning movement permissions at intersections. The method also accounted for the inaccuracies in GPS data by performing a sensitivity analysis to determine a frequency-based threshold that allows best filter the noise.

For the study area and based on the available GPS trajectories sample, it was possible to extract road directionally at an accuracy of 95%. An additional sensitivity analysis also found that increasing the sample was also able to increase the extraction accuracy to 99% for this case study. For turning movements extraction, the proposed method was found to be 98% when GPS trajectory observations were available. However, for a given GPS trajectory data sample, turning movement extraction was found to be more sensitive to the sample size since there are usually three movements (through, right, and left) for each road segment. Moreover, through movements usually have more observations than right or left turning movements and higher capacity roads such as arterials have more GPS trajectory observations than local streets. Therefore, some movements are not observed in the data simply due to the sample size. This explains why considering all turning movements of the study area, the accuracy was reduced significantly to 68%. The sample size effect was validated by observing that 97% of the wrong predictions correspond to turning movements that are permitted within the ground truth dataset but for which no observation was extracted from the GPS dataset.

Road Segment Number of Lanes

Having obtained an accurate representation of the road network directionality and connectivity, another attribute essential to obtain for the network model was the number of lanes. This enables the estimation of road capacity and the modelling of travel time and route choice. Based on the lateral distribution of GPS trajectory points with respect to the road segment reference line and the number of GPS points per road segment, it was assumed that GPS trajectory points are concentrated around the centerline of each road lane. It was also assumed that GPS trajectory points' inaccuracy is not biased and will be distributed randomly around the real

location of the GPS enabled device. One essential requirement regarding the input GPS trajectory data is that it had to be raw (not map-matched or snapped to a road centerline). It was possible to train a classification learner using a labelled dataset to predict the number of lanes at an accuracy of 91% using decision trees classification.

Road Intersection Control Type

Knowledge of road intersection control types is another essential component of transport models. It can be used to better classify road segments since intersection control type has a significant impact on traffic dynamics. The proposed method was successful in predicting intersection control type at 96% accuracy. The selected predictors also demonstrated how intersection approach level speed characteristics and number of observed trips can be used to predict intersection control type. The application of this method is targeted at large scale transport models where intersection control type information is rarely available and requires important collection efforts to obtain manually. Moreover, intersection control type knowledge is essential prior to modelling intersection turning movement delay since control type has a significant on speeds.

Road Intersection Turn Delay

Having built an accurate road network containing the correct road segment directionality and connectivity, number of lanes, and intersection control type, it was possible to build on that by proposing a method to model turn delay functions per road type, turn type, and intersection control type. The case study presented was for arterial-arterial or arterial-collector road intersections that controlled by traffic lights. Intersection movements were defined in a standardized way to ensure that delay of different movement types for a given approach can be

compared together to find the relative difference in their speeds. This enables the findings to be easily treated and integrated back into large scale macroscopic models. This work assumes that macroscopic models are generally calibrated using floating vehicle data that drive through intersections in a straight movement (through) to cover specific corridors. This signifies that the volume delay functions implicitly represent queuing for the through movement. This logic justified the use of crowd-sensed GPS data to model the difference in congestion that can be observed between through movement and turning movements (left or right) since crowd-sensed data has a better coverage of all the movements that are made at intersections.

It was demonstrated that left turns, right turns and through movements have different observed speed profiles which justifies the importance of making use of the large spatiotemporal coverage of crowd-sensed GPS trajectory data to make travel time estimation more accurate. Traffic speeds were averaged over 15-minute periods and traffic counts were collected for 15-minute periods. It was found that 15-min mean speed was the lowest for left turns at 14 km/h, followed by the right turns at 17km/h, and through movement at 21 km/hr. This can be explained by multiple factors such as traffic light phasing and timing, traffic light operation mode (fixed vs. actuated) intersection movement radius, permission to turn right on red.

In parallel, the mean 15-min traffic count was the lowest for left turns at 33 veh. /15-min, followed by right turns at 36 veh. /15-min, and through movement at 77 veh. /15-min. This information guides the traffic light phasing and timing design and demonstrates that through movement requires, on average, more green time, thus having an influence on observed speeds as presented above.

Given that in practice, travel time calibration is performed mainly for through movements (TRANS, 2014), a method was proposed to efficiently integrate the findings into large scale models by introducing speed adjustment factors, per road type, turn type and intersection control type, that relate turning movement travel time to the through movement travel time. This integration increases the accuracy of travel time calibration, therefore resulting in more accurate route choice and traffic assignment results.

Data Privacy

Given that GPS trajectory data, in its raw form, can be sensitive information if not anonymized and handled adequately, the methods proposed by this research were verified to ensure that it is impossible to identify any individual using the GPS trajectory dataset.

In fact, the GPS trajectory dataset did not contain any personal information and was never merged to any other data source that can identify personal information. Moreover, the proposed methods only process the GPS trajectories on road segments and intersections and extract variables such as speed and direction, that cannot be used to identify any individual. Trips were never examined in their entirety and origins and destinations were not examined at the individual trip level nor at the aggregate level, thus making it impossible to identify any individual.

It should be noted that the proposed methods can be applied using any crowd sensed GPS trajectory data. Depending on the GPS trajectory data source, it is possible to provide additional privacy protection by trimming the first and last 100 meters of each trajectory to ensure that no individual be identified by merging additional data sources.

7.2 Limitations

The main limitation that was faced while extracting network-wide road direction and turning movement rules was the limited sample size per intersection. This limitation was greater for right and left turns that are observed less frequently in the GPS trajectory data. Moreover, intersections movements of lower traffic streets were also less observed due to the sample size. Increasing the sample size by collecting the data over a longer period or accessing a larger number of mobile devices will result in higher accuracy, especially for turning movement modelling. The increase in sample size will also allow modelling the permitted intersection movements by time of day and day of week. This is important in regions where turning movement prohibition policies are scheduled for specific periods to increase traffic fluidity or improve road safety by reducing conflicts during peak traffic periods.

The proposed method to extract the number of lanes was limited by the quality and the quantity of the GPS trajectory. In terms of quality, around 50% of the data appeared to be snapped to a road centerline file. This reduced the GPS trajectory points' distribution quality, which in turn reduces the capacity of the algorithm to distinguish between road segments of different number of lanes. Moreover, the sample size was insufficient in some cases, mostly for local streets, since there were not enough points to obtain a stable distribution of the points with respect to the reference line. Since local streets only have one lane most of the time, this algorithm was still able to classify the one lane streets based on the sample size.

Given that modelling road intersection control type is based mainly on vehicular speed characteristics, increasing the GPS trajectory data sample size might enable the analysis of speed

profiles for specific time periods. This can provide additional insight by reducing heterogeneity in the observations and reducing the variability in speed characteristics that is usually observed throughout the day based on the traffic flows and reflected in traffic light timings. Additionally, targeting the analysis at specific times of the day will filter movements that are prohibited during specific periods to increase traffic fluidity or improve road safety by reducing the number of conflicts.

Regarding turn movement delay modelling, the main limitation was the specification, at a given road intersection, of the turning movement start point and end point for each trip. The uncertainty inherent to GPS trajectory data points in terms of sampling rate and errors in some observed trips made the full automation of the labelling process impossible and had to be done manually. The current study is based on 1136 turning movements that were manually labelled to demonstrate the applicability of the method.

7.3 Future Work Directions

The proposed method for turn delay modelling can be improved by developing an automated method to label the movement start and end points within the GPS trajectory data following the provided definition of turning movements. This is necessary for turn delay modelling and must be done for each trip segment passing through each intersection. This will make it feasible to analyze all intersections' movements more efficiently and take advantage of the large coverage of crowd-sensed data.

Moreover, a larger sample size with sufficient observations at low traffic streets and turning movements can be used to improve the model results. In fact, the sample size depends on the

road network feature that is being modelled or extracted. For example, in some urban areas, the road network segments have a fixed number of lanes available for general traffic and do not vary by time of day or day of the week. In this case, the sample period is not limited to specific times of the day and is usually dictated by the number of trip observations on local streets since they were observed to be lower. If the number of available lanes varies based on a temporal criterion, the sample should be large enough to be representative of each of the road network states. A good rule of thumb is having at least 30 GPS trip trajectory observation per road segment per analysis period. A future work can determine minimum sample size to be recommended per road network feature depending on the analysis period and the road type. In the presence of more data sources and larger data samples, the analysis becomes more complex and additional machine learning techniques such as ensemble learning can be explored to improve modelling accuracy.

Modelling intersection control type can be further developed by using the model predictions to model generic traffic light phasing and timing that can be used for mesoscopic level models, which are large scale models that require more detailed input information such as traffic light programming information.

Moreover, GPS trajectory data can be used to estimate road segment (or link) capacity. The idea would be to develop a method that scales the GPS trajectory sample to represent the entire population and relate that information to the travel time (or speed). Conditions can then be set to identify the link capacity by determining the maximum traffic flow (obtained from the GPS estimated population) just prior to congestion.

Since posted speeds are not necessarily free flow speeds as drivers tend to driver at higher speeds than the limit, another utilization of crowd-sensed GPS data would be to estimate free flow speeds that used in macroscopic model's volume-delay functions. This is relatively to simple to develop and can be done by estimating travel speeds during off-peak times of day. A relationship can also be established between each posted speed and the observed free flow speeds to make the information more generalizable.

Although it was not explored by this research a large sample of crowd-sensed GPS data can be used to extract origin-destination demand travel demand information. This can be done and compared to origin-destination information obtained from traditional household origin-destination surveys to assess if there exists any bias or limitation to this new data source.

Although the developed methods might require some additional refinement and adjustments to make them more generalizable, the presence of a plethora of trajectory data creates a potential for their commercialization through transportation planning and modelling software such as EMME and Aimsun.

It is important to keep in mind the privacy protection concern and take the necessary measures to ensure that no individual can be identified throughout the process.

Chapter 8 - Conclusion and Summary

This dissertation presented the use crowd-sensed GPS data to build on current large scale model development practice by increasing network modelling accuracy while facilitating its' updateability and reducing the required resources.

Using observed GPS trajectories to extract road network topology and connectivity features was achieved at a high accuracy (95%). The road segment number of lanes was also extracted at a high accuracy (91%) and can be further improved by using better quality GPS trajectory data. Having extracted the main road network features, the third objective was also achieved by predicting intersection control types with a high accuracy (96%). This information was necessary to classify intersection movements per road type, intersection control type and turn type. A method to calibrate turn penalty functions using GPS trajectory data was also presented while demonstrating the added value of using this large coverage data source in improving large scale transport model calibration.

With knowledge in spatial and data analysis, and access to GPS trajectory data, it is possible to reproduce the proposed methods for further research or for implementation in practice based on the information provided in the manuscripts. The techniques are transferable to new GPS trajectory samples and new study areas. However, it is important to re-train machine learning models based on the new data to obtain good results. For example, sample size related variables will vary depending on the GPS trajectory data sample.

The proposed methods demonstrate the utility of this new data source in improving road network modelling. It reduces the resources required to perform network modelling and improves quality of the model by serving as a large coverage floating vehicle travel time survey.

This information will help in calibrating more precise transport models and in updating the road network model more frequently and more efficiently.

In addition to achieving high accuracy, these methods build on current modeling tools capabilities by refining its output through using a new data source, spatial analysis, and machine learning. These methods are effective since GPS trajectory data is currently being collected by multiple location-based service providers at a relatively low cost. This reduces the costs associated with collecting satellite imagery data or street level imagery data. Moreover, the proposed methods achieved all the objectives while protecting the privacy of the individuals who were making the trips.

References

Adresses Québec - AQréseau. 2022. https://adressesquebec.gouv.qc.ca/aqreseau.asp

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. 2015a. A comparison and evaluation of map construction algorithms using vehicle tracking data. Geoinformatica, 19(3). 31

Ahmed, M., Karagiorgou, S., Pfoser, D., & Wenk, C. 2015b. Map Construction Algorithms. Springer Publishing Company, Incorporated.

Alexander, L., Jiang, S., Murga, M., & González, M. C. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies, 58. 240-250

Antoniou, C., Dimitriou, L., & Pereira, F. 2018. Mobility patterns, big data and transport analytics: tools and applications for modeling. Elsevier. https://doi.org/10.1016/C2016-0-03572-6

Arman, M. A., & Tampere, C. M. J. 2020. Road centreline and lane reconstruction from pervasive GPS tracking on motorways. Procedia Computer Science, 170. 8

ARTM Faits saillants EOD 2018. (2018). https://www.artm.quebec/faits-saillants-eod-2018/

Banqiao, C., Ding, C., Wenjuan, R., & Guangluan, X. 2020. Extended Classification Course Improves Road Intersection Detection from Low-Frequency GPS Trajectory Data. ISPRS International Journal of Geo-Information, 9(3). 181 (120 pp.)

Barceló, J., Kuwahara, M., & Miska, M. 2010. Traffic data collection and its standardization. Springer. https://doi.org/10.1007/978-1-4419-6070-2_1

Bender, P., Ziegler, J., & Stiller, C. Year. Lanelets: Efficient map representation for autonomous driving. 2014 IEEE Intelligent Vehicles Symposium Proceedings.

Biagioni, J., & Eriksson, J. 2012. Inferring Road Maps from Global Positioning System Traces Survey and Comparative Evaluation. Transportation Research Record(2291). 61-71

Bounini, F., Gingras, D., Lapointe, V., & Pollart, H. Year. Autonomous Vehicle and Real Time Road Lanes Detection and Tracking. 2015 IEEE Vehicle Power and Propulsion Conference (VPPC).

Carpenter, C., Fowler, M., & Adler, T. J. 2012. Generating route-specific origin–destination tables using Bluetooth technology. Transportation research record, 2308(1). 96-102

Chao, P., Hua, W., Mao, R., Xu, J., & Zhou, X. 2022. A survey and quantitative study on map inference algorithms from GPS trajectories. IEEE Transactions on Knowledge and Data Engineering, 34(1). 14

Chen, B. Q., Ding, C. B., Ren, W. J., & Xu, G. L. 2021. Automatically tracking road centerlines from low-frequency gps trajectory data. ISPRS International Journal of Geo-Information, 10(3). 26

Chen, Y., & Krumm, J. Year. Probabilistic modeling of traffic lanes from GPS traces. Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems.

Dabiri, S., & Heaslip, K. 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transportation Research Part C: Emerging Technologies, 86. 360-371

Daigang, L., Junhan, L., & Juntao, L. 2019. Road network extraction from low-frequency trajectories based on a road structure-aware filter. ISPRS International Journal of Geo-Information, 8(9). 17

Demissie, M. G., & Kattan, L. 2022. Estimation of truck origin-destination flows using GPS data. Transportation Research Part E: Logistics and Transportation Review, 159. 102621

Deng, T. 2013. Impacts of Transport Infrastructure on Productivity and Economic Growth: Recent Advances and Research Challenges. Transport Reviews, 33(6). 686-699

Données ouvertes 2022. https://www.ville.quebec.qc.ca/services/donnees-services-ouverts/index.aspx

Dorum, O. H. 2017. Deriving double-digitized road network geometry from probe data. 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.

El-Geneidy, A. M., & Bertini, R. L. Year. Toward validation of freeway loop detector speed measurements using transit probe data. Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749).

Elleuch, W., Wali, A., & Alimi, A. M. 2015. An investigation of parallel road map inference from big GPS traces data. 2015 INNS Conference on Big Data, San Francisco, CA, United states.

Ezzat, M., Sakr, M., Elgohary, R., & Khalifa, M. E. 2018. Building road segments and detecting turns from GPS tracks. Journal of Computational Science, 29. 13

Fan, J., Fu, C., Stewart, K., & Zhang, L. 2019. Using big GPS trajectory data analytics for vehicle miles traveled estimation. Transportation research part C: emerging technologies, 103. 298-307

Fortin, P., Morency, C., & Trépanier, M. 2016. Innovative GTFS Data Application for Transit Network Analysis Using a Graph-Oriented Method. Journal of Public Transportation, 19(4). 18-37

Freytes, L. 2022. Updating Travel Demand O/D Matrices from Old Travel Surveys through More Recent Traffic Counts: a case study in Turin, Politecnico di Torino].

Gately, C. K., Hutyra, L. R., Peterson, S., & Wing, I. S. 2017. Urban emissions hotspots: Quantifying vehicle congestion and air pollution using mobile phone GPS data. Environmental pollution, 229. 496-504

Guo, C., Kidono, K., Meguro, J., Kojima, Y., Ogawa, M., & Naito, T. 2016. A Low-Cost Solution for Automatic Lane-Level Map Generation Using Conventional In-Car Sensors. IEEE Transactions on Intelligent Transportation Systems, 17(8). 2355-2366

Guo, Y., Li, B., Lu, Z., & Zhou, J. 2021. A novel method for road network mining from floating car data. Geo-spatial Information Science. 16

Hashemi, M. 2019. Automatic inference of road and pedestrian networks from spatial-temporal trajectories. IEEE Transactions on Intelligent Transportation Systems, 20(12). 17

Hrabec, D., Hvattum, L., & Hoff, A. 2022. The value of integrated planning for production, inventory, and routing decisions: A systematic review and meta-analysis. International Journal of Production Economics, 248. 108468

Ian, H., Frank, E., Hall, M., & Christopher, J. 2017. Data mining: Practical machine learning tools and techniques—Part II: More advanced machine learning schemes. Morgan Kaufmann, Burlington, MA.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40. 63-74

Jia, Q., & Ruisheng, W. 2016. Road map inference: A segmentation and grouping framework. ISPRS International Journal of Geo-Information, 5(8). 20

Kadali, B., & Vedagiri, P. 2020. Role of number of traffic lanes on pedestrian gap acceptance and risk taking behaviour at uncontrolled crosswalk locations. Journal of Transport & Health, 19. 100950

Kan, Z., Tang, L., Kwan, M.-P., & Zhang, X. 2018. Estimating Vehicle Fuel Consumption and Emissions Using GPS Big Data. International Journal of Environmental Research and Public Health, 15(4). (

Karagiorgou, S., & Pfoser, D. 2012. On vehicle tracking data-based road network generation. 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, United states.

Kasmi, A., Denis, D., Aufrere, R., & Chapuis, R. Year. Map Matching and Lanes Number Estimation with Openstreetmap. 2018 21st International Conference on Intelligent Transportation Systems (ITSC).

Kucharski, R., & Drabicki, A. 2017. Estimating Macroscopic Volume Delay Functions with the Traffic Density Derived from Measured Speeds and Flows. Journal of Advanced Transportation, 2017. 4629792

Lee, K., & Sener, I. N. 2021. Strava Metro data for bicycle monitoring: a literature review. Transport Reviews, 41(1). 27-47

Leichter, A., & Werner, M. 2019. Estimating road segments using natural point correspondences of GPS trajectories. Applied Sciences-Basel, 9(20). 11

Lesani, A., & Miranda-Moreno, L. 2019. Development and Testing of a Real-Time WiFi-Bluetooth System for Pedestrian Network Monitoring, Classification, and Data Extrapolation. IEEE Transactions on Intelligent Transportation Systems, 20(4). 1484-1496

Li, H. F., Kulik, L., Ramamohanarao, K., & Acm. 2016. Automatic generation and validation of road maps from GPS trajectory data sets. 25th ACM International on Conference on Information and Knowledge Management.

Li, J., Pei, X., Wang, X., Yao, D., Zhang, Z., & Yue, Y. 2021. Transportation mode identification with GPS trajectory data and GIS information. Tsinghua Science and Technology, 26(4). 403-416

Lin, C., Zhou, X., Wu, D., & Gong, B. 2019. Estimation of Emissions at Signalized Intersections Using an Improved MOVES Model with GPS Data. International Journal of Environmental Research and Public Health, 16(19). 3647

Liu, J., Zhu, J., Lu, D., Yuan, D., & Azadi, H. 2023. The Effectiveness of Improvement Measures in Road Transport Network Resilience: A Systematic Review and Meta-Analysis. Sustainability, 15(13). 10544

Ma, J., Li, H., Yuan, F., & Bauer, T. 2013. Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data. International Journal of Transportation Science and Technology, 2(3). 183-204

Máttyus, G., Luo, W., & Urtasun, R. Year. Deeproadmapper: Extracting road topology from aerial images. Proceedings of the IEEE international conference on computer vision.

Mattyus, G., Wang, S., Fidler, S., & Urtasun, R. Year. Enhancing road maps by parsing aerial images around the world. Proceedings of the IEEE international conference on computer vision.

Máttyus, G., Wang, S., Fidler, S., & Urtasun, R. Year. Hd maps: Fine-grained road segmentation by parsing ground and aerial images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Mennis, J., & Guo, D. 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. Computers, Environment and Urban Systems, 33(6). 403 - 408

Merry, K., & Bettinger, P. 2019. Smartphone GPS accuracy study in an urban environment. PLOS ONE, 14(7). e0219890

Montréal, V. d. MTL Trajet Study. (

Munizaga, M. A., & Palma, C. 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. Transportation Research Part C: Emerging Technologies, 24. 9-18

Nieroda, B., Wojakowski, T., Skruch, P., & Szelest, M. Year. A Heatmap-Based Approach for Analyzing Traffic Sign Recognition and Lane Detection Algorithms. 2022 26th International Conference on Methods and Models in Automation and Robotics (MMAR).

Ortúzar, J. D., & Willumsen, L. G. 2011. Modelling transport. John wiley & sons.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. 2016. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 5(1). 210

Phondeenana, P., Noomwongs, N., Chantranuwathana, S., & Thitipatanapong, R. 2013. Driving Maneuver Detection System based on GPS Data. 1-6

. *Plan québécois des infrastructures 2022-2032*. (2022). Gouvernement du Québec. https://www.tresor.gouv.qc.ca/budget-de-depenses/archives/budget-de-depenses-2022-2023/

PRISMA. 2015. Transparent reporting of systematic reviews and meta-analyses. PRISMA Ottawa, ON, Canada http://www.prisma-statement.org/

Sinha, K. C., & Labi, S. 2011. Transportation decision making: Principles of project evaluation and programming. John Wiley & Sons.

Stanojevic, R., Abbar, S., Thirumuruganathan, S., Chawla, S., Filali, F., & Aleimat, A. 2018. Robust road map inference through network alignment of trajectories. 2018 SIAM International Conference on Data Mining (SDM), San Diego, CA, United states.

Stipancic, J., Racine, E. B., Labbe, A., Saunier, N., & Miranda-Moreno, L. 2021. Massive GNSS data for road safety analysis: Comparing crash models for several Canadian cities and data sources. Accident Analysis & Prevention, 159. 106232

Sun, Y., & Mobasheri, A. 2017. Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data. International Journal of Environmental Research and Public Health, 14(3). 274

Tang, L., Gan, A., & Alluri, P. 2014. Automatic Extraction of Number of Lanes from Georectified Aerial Images. Transportation Research Record, 2460(1). 86-96

Tantiyanugulchai, S., & Bertini, R. L. Year. Arterial performance measurement using transit buses as probe vehicles. Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems.

TRANS. (2014). *Evolution of the TRANS Regional Travel Demand Forecasting Model*. http://www.ncr-trans-rcn.ca/model/

Treiber, M., & Kesting, A. 2013. Traffic flow dynamics. Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg. 983-1000

Turner, S., Tsapakis, I., & Koeneman, P. 2020. Evaluation of StreetLight Data's Traffic Count Estimates From Mobile Device Data [Tech Report]. (

UN. 2019. *World Urbanization Prospects - The 2018 Revision* (

Wegener, M., Mackett, R. L., & Simmonds, D. C. 1991. One city, three models: comparison of land-use/transport policy simulation models for Dortmund. Transport Reviews, 11(2). 107-129

Xingzhe, X., Wenzhi, L., Aghajan, H., Veelaert, P., & Philips, W. 2016. A novel approach for detecting intersections from GPS traces. Piscataway, NJ, USA.

Xingzhe, X., Wong, K. B. Y., Aghajan, H., Veelaert, P., & Philips, W. 2015. Inferring directed road networks from GPS traces by track alignment. ISPRS International Journal of Geo-Information, 4(4). 26

Yang, H., Cetin, M., & Ma, Q. 2020. *Guidelines for Using StreetLight Data for Planning Tasks* (Tech Report.

Zhang, C., Xiang, L., Li, S., & Wang, D. 2019. An intersection-first approach for road network generation from crowd-sourced vehicle trajectories. ISPRS International Journal of Geo-Information, 8(11). 26

Zhang, L., Thiemann, F., & Sester, M. Year. Integration of GPS traces with road map. Proceedings of the Third International Workshop on Computational Transportation Science.

Zhang, Y. F., Zhang, Z. X., Huang, J. C., She, T. T., Deng, M., Fan, H. C., Xu, P., & Deng, X. S. 2020. A hybrid method to incrementally extract road networks using spatio-temporal trajectory data. ISPRS International Journal of Geo-Information, 9(4). 15

Zhao, Y., Zheng, J., Wong, W., Wang, X., Meng, Y., & Liu, H. 2019. Various methods for queue length and traffic volume estimation using probe vehicle trajectories. Transportation Research Part C: Emerging Technologies, 107. 70 - 91

Zhongyi, N., Lijun, X., Tian, X., Binhua, S., & Yao, Z. 2018. Incremental road network generation based on vehicle trajectories. ISPRS International Journal of Geo-Information, 7(10). 19

Zito, R., & Taylor, M. A. P. 1994. The use of GPS in travel-time surveys. Traffic Engineering and Control, 35. 685-685