New Opportunities in Urban Transport Data: Methodologies and Applications

MASOUD FADAEI OSHYANI

New Opportunities in Urban Transport Data:
Methodologies and Applications

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i transportvetenskap fredagen den 11 december 2015 klockan 13.00 i sal L1, Kungl Tekniska högskolan, Drottning Kristinasväg 30, Stockholm.

Cover illustration: Maryam Fadaei

به نام او

به یاد پدرم

تقدیم به مادرم

In the name of God
To the loving memory of my father and
To my beloved mother

"Have patience. All things are difficult before they become easy."
Saadi (1210-1291)

Abstract

The deployment of Information and Communication Technologies (ICT) is growing in transportation which may contribute to a more efficient and effective service. The data acquired from ICT based systems could be used for many purposes such as statistical analysis and behavior learning and inference. This dissertation addresses the question of how transportation data that was collected for a specific application can be used for other applications. This thesis consists of five separate papers, each addressing a subset of the topic.

The first paper estimates a route choice model using sparse GPS data. This paper demonstrates the feasibility of an Indirect Inference based estimator in a model with random link costs, allowing for a natural correlation structure across paths, where the full choice set is considered.

The second paper presents an estimator for the mean speed and travel time at network level based on indirect inference when the data are spatially and temporally sparse.

The third paper proposes an evaluation framework which outlines a systematic process to quantify and assess the impacts of public transport preferential measures on service users and providers in monetary terms, using public transport data sources.

In the fourth and fifth papers, a methodology is developed and implemented for integrating different prediction models and data sources while satisfying practical requirements related to the generation of real-time information. Then the performance of the proposed prediction method is compared with the prediction accuracy obtained by the currently deployed methods.

Sammanfattning

Användandet av informations- och kommunikationsteknologier (eng. ICT) ökar inom transportområdet, vilket kan bidra till ökad effektivitet. Insamlad data från system med ICT skulle kunna användas för många ändamål såsom statistisk analys, beteendeinlärning och inferens. Denna avhandling tar upp frågan om huruvida transportdata insamlat för en viss tillämpning kan användas för andra. Avhandlingen innehåller fem forskningsartiklar, var och en inriktar sig på sin del av ämnet.

De två första uppsatserna fokuserar på konsistenta estimatorer för hastighet på länkar och ruttval. För många olika tillämpningar är det viktigt att förutsäga en observerad rutts fortsättning, och, givet att det är glest med data, att även avgöra var individen (eller fordonet) har varit. Att skatta den upplevda restiden (och nyttan) av en vald rutt är ett statistiskt svårt skattningsproblem av flera olika skäl. För det första är valmängden ofta mycket stor. För det andra kan det vara viktigt att ta hänsyn till korrelationen mellan de (generaliserade) kostnaderna för olika rutter och därigenom tillåta realistiska ersättningsmönster. För det tredje, på grund av överväganden gällande teknik och den personliga integriteten, kan data vara temporalt och spatialt gles och med endast partiellt observerade rutter. Slutligen, kan det finnas mätfel av fordonens position. Vi utvecklar en estimator för upplevd nytta av en rutt (i den första artikeln) samt för restid på länkar i vägnätverket (i den andra artikeln).

I den tredje artikeln föreslås ett ramverk för utvärdering innefattande en systematisk process som kvantifierar och bedömer inverkan av preferensstyrmedel inom kollektivtrafiken på tjänsteanvändare och leverantörer.

I fjärde och femte artikeln utvecklas och implementeras en metodologi för att integrera olika prediktiva modeller och data-källor i beaktande av praktiska krav kopplade till skapandet av realtidsinformation. Den resulterande prediktionsmetoden jämförs med de metoder som i nuläget används i Stockholm och Brisbane.

Acknowledgements

It would not have been possible to write this thesis without the help, love and support of my family. Above all, I would like to thank my beloved wife, Mina, who has been a source of encouragement, endless love and support. For her patient encouragement when I lost my vision and for being my inspiration and my guardian angel. But most of all, thank you for being my best friend. I owe you everything.

My mother, sister and brother who never stop believing in me and for their constant support at each and every step throughout the way. I would also like to thank all my friends in both Iran and Sweden, particularly Masih.

I would like to appreciate my supervisors Anders Karlström, Oded Cats and Marcus Sundberg for their support, encouragement and great contribution.

I also wish to say thanks to the following:

Storstockholms Lokaltrafik (SL), special thanks to Azhar Al-Mudhaffar, Johan Nordgren and Sofia Rahlén for their contribution in data collection and their supports.

Smart Transport Research Centre at Queensland University of Technology. Specially Ashish Bhaskar for generously welcoming me to spend amazing time as a guest researcher with his group. Thanks my dear friends there: Hasti Tajtehranifard, Neema Nassir and Jake Withehead for hosting me and their supports during my stay. Also Mark Hickman and Mahmoud Mesbah for fantastic discussions that we had for improving my papers.

Centre for Transport Studies at Stockholm. Particularly its former and current directors, Jonas Eliasson and Maria Börjesson for supporting and financing my study.

All the members of the Department of Transport Science create a very friendly atmosphere. Specialy Per Olsson, Susanne Jarl, Eva Petersson, Shiva Habibi, Oskar Blom Västberg, Siamak Baradaran, Jens West, Daniel Jonsson, Erik Jenelius and Haris N. Koutsopoulos. Thank you all for making such

a fantastic environment to study and work in.

List of appended papers

Paper 1. Oshyani, Masoud Fadaei, Marcus Sundberg, and Anders Karlström. "Estimating flexible route choice models using sparse data", 15th International IEEE Conference on Intelligent Transportation Systems, 2012, pp.1215,1220.

Paper 2. Oshyani, Masoud Fadaei, Marcus Sundberg, and Anders Karlström. "Consistently Estimating Link Speed Using Sparse GPS Data with Measured Errors", Procedia-Social and Behavioral Sciences 111, 2014, pp. 1227-1236.

Paper 3. Oshyani, Masoud Fadaei, and Oded Cats. "Evaluating the Impacts and Benefits of Public Transport Preferential Measures", *submmited.*

Paper 4. Oshyani, Masoud Fadaei, and Oded Cats. "Real-time bus departure time predictions: vehicle trajectory and countdown display analysis", 17th International IEEE Conference on Intelligent Transportation Systems, 2014, pp. 2556-2561.

Paper 5. Oshyani, Masoud Fadaei, Oded Cats, and Ashish Bhaskar. "A hybrid scheme for real-time prediction of bus trajectories", *submmited.*

Other papers

The following papers are not included in this thesis due to an overlap in content or that their content goes beyond the scope of this thesis.

Paper 6. Oshyani, Masoud Fadaei, and Oded Cats. "Rolling horizon predictions of bus trajectories", In 1st International Conference on Engineering and Applied Sciences Optimization, OPT-i 2014; Kos Island; Greece; 4 June 2014 through 6 June 2014, pp. 875-886. 2014.

Paper 7. Jonsson, Daniel, Anders Karlström, Masoud Fadaei Oshyani, and Per Olsson. "Reconciling user benefit and time-geography-based individual accessibility measures", Environment and Planning B: Planning and Design 41 (2014): 1031-1043.

Paper 8. Oshyani, Masoud Fadaei, and Oded Cats. "An empirical evaluation of measures to improve bus service reliability: Performance metrics and a case study in Stockholm", In CASPT Conference 2015 in Rotterdam, the Netherlands, 2015.

Paper 9. Oshyani, Masoud Fadaei, and Oded Cats. "Evaluating the performance and benefits of bus priority, operation and control measures," In Transportation Research Board 95th Annual Meeting, (2016), *accepted.*

Declaration of contribution

Paper 1. The idea of the paper I was jointly proposed in the discussion among Masoud Fadaei Oshyani, Anders Karlström and Marcus Sundberg. The first author was responsible for methodology implementation and main contributor in the methodology development, analysis of results and writing.

Paper 2. The idea of the paper II was jointly proposed in the discussion among Masoud Fadaei Oshyani, Anders Karlström and Marcus Sundberg. The first author was responsible for methodology implementation and main contributor in the methodology development, analysis of results and writing.

Paper 3. The idea of the paper V was jointly proposed in the discussion among Masoud Fadaei Oshyani and Oded Cats. The methodology was developed in discussions among Masoud Fadaei Oshyani and Oded Cats. The first author was responsible for methodology implementation and main contributor in the analysis of results and writing.

Paper 4. The idea of the paper IV was jointly proposed in the discussion among Masoud Fadaei Oshyani and Oded Cats. The methodology was developed in discussions among Masoud Fadaei Oshyani and Oded Cats. The first author was responsible for methodology implementation and main contributor in the analysis of results and writing.

Paper 5. The idea of the paper III was jointly proposed in the discussion among Masoud Fadaei Oshyani and Oded Cats. The methodology was developed in discussions among the coauthors. The first author was responsible for methodology implementation and main contributor in the analysis of results and writing.

Contents

1 Introduction

The necessity of transportation in our daily life cannot be disregarded these days. In particular, in large cities people are forced not to always use their private transport facilities, but rather use public transport systems. Due to growing demand for inner-city travel and mixture of different transport modes (i.e. private cars, buses, trams, bicycles and pedestrian) in urban areas, traffic congestion is a growing concern in many cities around the world. High dense urban areas and financial constraints increase importance of investment on new efficient urban traffic management (UTM) systems rather than conventional approach of more road building.

The development of urban traffic management systems requires high quality traffic information in real-time. Availability of exhaustive and accurate data is the first initial prerequisite of a UTM system. Data collection is a fundamental concept in transportation science and particularly in UTM. The traditional paper-based surveys have been substituted by telephone, mail and web-based ones through emergence of new communication systems. These data are often used to investigate traveler behavior considering their socio-demographic data. In addition, manual in-site data collection techniques (e.g. counting passengers, measuring traffic flow and bus dwell time) are replaced by more advanced ones through appearance of new measuring and positioning technologies.

Generally, two automated data collection categories are introduced in the urban transport field and enhanced through technology development. (i) Infrastructure-based technologies, where the data collector is installed on a field infrastructure such as loop detector, automatic number plate recognition (ANPR) cameras and (ii) In vehicle-based technologies a vehicle is equipped by a data collector device (e.g. Global Positioning System (GPS) devices). Although Infrastructure-based collecting data is useful, their limited coverage and expensive costs of implementation and maintenance make them sometimes difficult in an extensive implementation. Hence, in the context of UTM the application, vehicle-based technologies are fast growing due to its competitiveness.

Rapid and reliable communication systems are the second requirement of UTM, in order to enable monitoring and management of the traffic situation, for instance to take proactive decision in the case of disruption. Moreover, existence of such a system enables the management authorities to distribute

real-time information regarding current system situation and potential solution in case of service interruption or congestion (e.g. alternative route). The most common used communication systems in UTM are land line telephones, mobile-based (e.g. GPRS) and asymmetric digital subscriber line (ADSL). All of these communication systems are located under an umbrella term, Information and Communications Technologies (ICT).

The last part of UTM is a set of methodologies which retrieve required information in an adequate time and accuracy; such as, optimized traffic signal plans, fault and incident detection and prediction, in-vehicle traffic information (e.g. route guidance) and evaluation of transport policies.

There are widespread efforts ongoing in order to develop new methodologies or to enhance existing ones to improve UTM. Deployment of such UTM systems results in more efficient and accurate service. Besides, the acquired information could be used for many other purposes such as statistical analysis and system behavior learning and inference.

2   Research Outline

This thesis aims to develop new novel methodologies to obtain valuable transportation information from existing big data sources beyond the application that they initially have been developed for. Two big urban transport data sources are mainly used:

1. Sparse GPS data acquired from Taxi to allocate the nearest vehicle to the customer.

In this context, ICT facilitates management of the system by reporting the location of all taxis, their occupancy status (hired/free) and the derivers updated work plan, constantly. The control center then uses this information to optimize allocation process of available taxis to the existing demand.

2. Automatic Vehicle Location (AVL) and Automatic Passenger Count (APC) data acquired from public transport network to improve service planning and operation management.

Many transit agencies have implemented AVL and APC systems to have more advanced operation management over the transportation network. The primary application of these systems is in enhancement communication, fleet

management and provision of traveler information. Availability of these data provides a great advantage which is the usage of massive, accurate and auto-processed data instead of manually collected data and improper simulated data.

Using the aforementioned big data sources above, the following research areas are considered:

1. Route Choice Estimation

2. Travel Time Estimation

3. Public Transportation Policy Evaluation

4. Bus Departure Time Prediction

In the following sections, each area is explained.

2.1 New urban traffic data

Implementation of advanced technologies from various fields, such as mobile communication and image processing, has introduced new data sources in urban transportation. Availability of the mobile phones data (particularly GSM network) provides unique opportunity to retrieve individual origin-destination and travel time information. City video cameras output transform to information regarding aggregated flows and speed using image processing techniques. This information could be enhanced by license plate recognition to provide individual travel time information. Moreover, several new data collection methods are developed based on existing sensor technologies, such as Infrared detectors (e.g. APC), Radar detectors, Induction loop detector (Electromagnetic), Bluetooth sensor and Ultrasonic sensors. These detectors generally detect passing time of any mobile object (individual or vehicle) and gather information regarding flow, travel time (or arrival time) and vehicle type (for details see [41]). In addition increasing availability of data using the Geographical Positioning System (GPS), new methods and algorithms are being developed that are tailor-made for the new data sources to address specific problems.

GPS data have previously attracted attention in the transportation science literature (see, e.g. [12], [16], [30], [27] and [29]). GPS technology allows for

collection of large data sets regarding mobility and route choices of individuals [21]. Lou et al. [25] have previously addressed methods and algorithms associated with the use of sparse GPS data.

Using GPS data has its own restrictions. For example, receiver's noise and clock errors can cause inaccuracy in data. Depending on the number of available satellites, the atmospheric conditions and the local environment, the GPS devices may result in inaccurate location or missing data. Another restriction is that the data is saved in a series of GPS points, and to be able to infer the travelled path it is often necessary to apply data processing such as map-matching. Quddus et al. [33] provided an overview of map-matching methods. Deployment of map-matching methods results in an estimation for actual traversed path. When the sparse GPS data is considered as collected travel data, the traversed path is unknown even between two consecutive points. In addition, GPS data has its own measurements error.

In the public transportation sector, many service providers started to use new technologies in respond to increasing traffic congestion and resulting passenger demands for more reliable service. AVL and APC are two commonly used systems. Operators were primary interested in capability of these systems in provision of immediate information which is essential for real-time operations monitoring and control [43]. Okunieff's TCRP synthesis [42] gives a comprehensive picture of AVL systems and review of AVL's history. Through fast development in communication systems and storage technologies, AVL and APC systems became enabled to gather very large amounts of operating data which is required for performance analysis and management. New urban public transport data sources, congregated as a consequence of widespread application of AVL and APC systems, can be used for a wide variety of applications. Hickman [44] categorized these applications into five main areas: (i) Passenger information, (ii) Transit operation monitoring and control, (iii) Service planning, (iv) Air quality improvements and (v) Safety and security enhancement. In this thesis, the two former are points of interest.

2.2 Route choice

There could be many different purposes underlying a trip such as commuting to work, school and shopping. For analyzing the trips and modeling the travel behavior of passengers within a network, parameters of high interest should be studied, among which are route choice, purpose, destination, travel time,

travel mode. Considering the Route Choice only, one may be able to infer commuters' preferences for travel routes consisting of elements such as, for instance, distance, road type, travel time, cost, or number of traffic lights. Other potentially affecting elements such as gender, age and income may be taken into consideration as individual characteristics. Route choice models are estimated to analyze the travelers' behavior regarding choosing a route in a given network.

The better understanding of route choice could be helpful for predicting behavior under different circumstances. For instance, in a congestion charging system, routes which are taken by travelers crossing the charging gates should accept a more costly trip compared to a previous situation of not having charges. Route choice models could be applied to analyze the congestion charging scenario in order to predict forthcoming travelers' behavior. Furthermore, the route choice models are used in the Vehicle Routing Problem (VRP), where a number of vehicles should find their routes to distribute their products among several destinations.

In route choice research so far, traditional data collection methods include telephone, mail and more recently web-based surveys ([47], [48] and [49]). Through these methods travelers explain their taken paths. The advent of passive monitoring of route choice has prompted different authors to compare different means of data collection (i.e. comparing conventional and the new ones). There is a lot of literature comparing data collected by conventional survey methods to GPS data. (See [50]; [51]). A common point in almost all literature reviews is that passive monitoring methods have many benefits compared to the conventional surveys. For instance, collected data is directly accessible in electronic format. Additionally, trip data could more easily be collected repeatedly for several days of trip (see [52], and [53] for detailed discussions).

The routes chosen by a traveler are determined by an observed sequence of links. In this format, a route choice can be defined as a discrete choice. Ben-Akiva and Bierlaire [45] give a comprehensive description of discrete choice methods in route choice modeling. The choices may be considered from two different approaches. First, travelers are assumed to select their route among paths (Path-based). Second, travelers are assumed to select links at each intersection (Link-based).

The path-based approach using discrete choice methods are elaborated in a large proportion of the literatures. A main problem with path-based approach

5

is that the universal choice set is very large, and often unknown. Different formulations have been used to address this problem, such as the C-Logit approach [46]. The second approach is to use link specific errors, such that the errors are associated to links defining paths. The link costs are assumed random and the summation of link costs gives the cost of the paths [23]. This is the approach implemented in this study as well where law frequency GPS points are used.

2.3   Travel time

Travel time is a critical aspect of all trips which is usually considered by people in their route choice. It provides a key aspect in transportation planning and appraisal, and hence accurate travel time measurement is of high importance. In addition, travel time is related to other key factors such as congestion and pollution, and also has a significant impact in social cost benefit analysis, both directly and indirectly [55].

Turner et al. mention three factors of interest in the travel time field studied in the 1990s [54]. The first, travel time is a performance factor which represents traffic congestion in the congestion management systems. The second, travel time is a common factor which is defined for all transportation modes. Therefore it could be a meaningful parameter to compare different transport modes and distribute a common funding source among them. The third, travel time could increase involvement in transportation decision by parties which are not expert in the field. It is evidence that the term is simple and easy to understand, yet precise enough for transportation analyses.

Extensive research has been carried out on route choice modeling and link travel time estimation; however a few studies have used GPS data ([34], [17], [57] and [58]). In this context, map-matching is the main prerequisite to infer the traversed path based on GPS data. Inaccuracy in the map-matching and path inference increases when we are dealing with sparse data points. This motivates the approach in which the nature of map-matching methods is acknowledged and then captured in modelling (e.g. route choice and travel time estimation) rather than trying to fix it through various manipulations. This objective will be addressed in this thesis applying an indirect inference approach.

2.4 Preferential strategies in public transportation

Increasing the share of public transport systems in the city transportation is often considered as a cost efficient and environmentally friendly solution in response to a growing demand in this sector. For this purpose, an extensive transit network and adequate fleet size to provide sufficient service to satisfy the demand and potentially attract more travelers from private car network, is needed. Accuracy, reliability and safety are all significant aspects to make public transit system more attractive for travelers.

Policy makers and transport planners design and implement a large range of preferential measures aimed to make public transport a more effective travel alternative vis-a-vis the private car. Preferential measures are a set of instruments which is applied to reduce travel times and improve service reliability. Abkowitz [56] distinguished between three categories of preferential measures: priority, control, and operational. Priority methods are the specific action to prioritize transit vehicles movement over general vehicular traffic by means of introducing bus lanes or transit signal prioritization. Operational methods involve such methods as schedule modification, route restructuring, and driver training that usually requires a longer implementation period. Control methods are applied in real-time and include vehicle holding, short-turning, stop skipping, and speed modification.

Reliability is an important determinant of passengers' satisfaction and level-of-service. An unreliable service causes disutility since travelers need to accommodate the risk of being late at their destination [31]. Furthermore, an unreliable service hinders operators' ability to adhere to its operational planning and may carry logistical and economic consequences (e.g. penalties, reserve buses). This is especially true for bus systems where small discrepancies can propagate along the line due to limited control. A synthesis of evidence from Europe, North America and Australasia by Currie and Wallis [13] concluded that the largest increase in ridership was related to preferential measures that targeted improving reliability.

Results of preferential measures are often reported in terms of ridership and speed changes aimed to promote the transfer of best practices. An overall assessment of their impacts will enable the comparison of different implementations, assess their effectiveness, prioritize alternative measures and provide a sound basis for motivating investments in such measures. Considering the high availability of big travel data sources resulted from the implementation

of AVL, APC and etc. systems, the question is how total benefits gained from implementing new preferential strategies can be evaluated using big transit data sources.

2.5   Bus travel time

Proper travel-time information, especially in situations where unplanned changes occur, is among the most influential factors on travelers' satisfaction with public transport service [10]. Public transport services in general and urban bus services in particular, are subject to inherent sources of uncertainty. Besides implementation of different affecting measures on the service reliability, prediction errors could be decreased by developing more accurate and reliable prediction methods regarding downstream conditions.

Bus travel times are influenced by different inter-dependent stochastic factors such as traffic congestion, intersection delay, passenger demand, driver's behavior and weather conditions. Any deviation in those factors could possibly propagate and as a result, a deterioration in placed in the schedule adherence and service reliability.

Provision travel time information relies on acquisition and transmission of instantaneous data. Such information helps commuters better adjust their travel plan and reduce the effect of service irregularity [32]. In addition, accurate time predictions facilitate proactive management and control strategies for monitoring and avoiding service disruptions [3]. In this context, public transport operators and users have a variety of viewpoints towards service reliability. While the former are mainly concerned about deviations from the timetable, the latter are primarily concern about waiting times [8]. Therefore, according to such perspectives, we may define two sets of evaluation measures: vehicle-based and stop-based. These distinctive perspectives have some further implications on the prediction methods as well as their respective performance measures.

3   Route choice and travel time estimation

The route choice problem can be formulated as a discrete choice problem with the following main challenges.

1. The choice set can be very large and it is infeasible to enumerate in realistic sized applications.

2. The presence of overlapping paths causes high correlation among some alternatives.

Both of these problems have to be addressed in any model estimation. Generally, GPS data is classified into low and high frequencies. If traversed paths between all two consecutive GPS points can be accurately detected, this data is defined as high frequency; otherwise it would be low frequency. Here, we would like to use sparse GPS data to estimate a flexible route choice model. Low frequency GPS data has two obvious challenging characteristics from the route choice modeling point of view which are added to the abovementioned route choice modeling challenges.

3. Sparseness is a problem since the path cannot be fully observed.

4. The sampled points may be sampled with errors.

Any estimator must be able to correct estimates for biases caused by aforementioned circumstances.

Different techniues have been proposed to cope with the first challenge. It is suggested to arrive at statistically consistent estimates by sampling the choice set (see [26], [18] and [20]).

In order to capture the second challenge, the first approach is path-based models using discrete choice methods which is well discussed in the literature (see e.g., [5], [6], and [7]). The second approach defines a route as a sequence of links [14]. Link costs are assumed to be random, and the path cost is given by the summation of link costs ([15] and [23]). Overlapping paths are sharing the same random components and consequently cause a natural correlation structure between paths. Moreover, the approach allows for statistically consistent estimation with consideration of the full choice set. Hence, challenges 1 and 2 are addressed in our modeling approach.

A method is proposed to estimate a route choice model with random link costs. The method is based on indirect inference approach as an alternative of doing hard computations to find the maximum likelihood estimate.

Indirect inference is a simulation based estimation method ([19], and [22]). The methodology is useful even when the likelihood functions are intractable or even impossible to specify. As a simulation-based method, the main pre-

9

requisite of the indirect inference approach is that it should be possible to simulate data from the model of interest for different values of the parameters involved. The main characteristic of the indirect inference method is the use of an auxiliary model to build a criterion function. The indirect inference is aiming to find parameters for the model of interest such that the simulated and observed data look the same from the auxiliary model's perspective. The indirect inference approach intends to estimate a model, which is beneficial in practice, but difficult to estimate.

The GPS points do not usually match directly to a network. Some method are hence deployed to map the reported noisy GPS points on the network. Moreover, these methods have to appropriately estimate the traversed path even though GPS points are sparse. The implimented map-matching method in this thesis is described through the following steps.

1. Take the network data and the given set of GPS points.

2. Identify the origin and destination of the trip.

3. Calculate the summed distance between GPS points and potential paths.

4. Find the path with minimum distance to GPS points and return as the matched path.

Due to sparseness in our dataset, in practice, our map matching method at the forth step usually returns more than one choice as the matched path. In other words, there are usually a number of sub-paths connecting two consecutive points resulting in having different paths with the same total distance to the set of GPS points.

A solution to deal with this problem is to find the actual locations of the GPS points on a path based on the point time stamps and speeds on the links; then, calculating the distances between these locations and their reported positions in the dataset. The choice with minimum distance is detected as a final matched path. The actual location for a GPS point on a given path is defined as the location on the path that the traveler would be there if he/she took that path. Since the information regarding the links speed values is still unknown, we substitute the link speeds in our data set with the speed limits as the average speed values on the links. Although speed limit is not an accurate estimate for the average speed, it could be practically helpful in the indirect inference approach. In other word, since we apply our map-matching method

10

for both real and simulated data in the same way, the error introduced in map-matching will be corrected for by the indirect inference based estimator.

The indirect inference method is applied as a structured procedure to estimate a model with random link costs, where the likelihood function is difficult to evaluate. Rather, we make use of the much simpler likelihood functions related to the auxiliary model. Estimation is performed through simulation. By ensuring that the same data transformations (map-matching) are performed on both real and simulated data, the proposed estimator is able to compensate for the introduced bias such transformations. The main conclusion is that indirect inference is a useful option in the tool box for route choice estimation which can be used for estimating observed path using low frequency GPS sampling data with measurement errors.

Until now, we have developed a method for estimating route choice models when data is spatially and temporally sparse. Following the indirect inference approach we develop an estimator for link travel times, and show how we can jointly estimate both parameters of the route choice model, and link travel times.

One of the main characteristics for the route choice model is travel time. Typically, we want to determine the effect of travel time on route choice, by estimating the corresponding route choice parameter. As link travel times have to be inferred from link lengths and estimated link speeds, a correlation is introduced between the estimates of the route choice parameter and the link speeds. Therefore, we propose a method to simultaneously estimate these two correlated parameters. In other words, a robust estimate for route choice model could lead to accurately estimate the link speed. Thus, we present an iterative method consisting of two sub-estimators. The first sub-estimator takes a given value for route choice parameter and estimates the speed values and the second one estimates route choice parameter given the speeds for the links in the network. This process will continue until we reach the final estimate values for speeds. Both sub-estimators are constructed based on the indirect inference approach.

Due to the large number of links in a typical network and the associated computational cost, it is practically impossible to estimate the average speed on all the links. Thus, we categorize the links into a number of different classes according to their characteristics (i.e. speed limits) and only speed parameters need to be estimated for each link class.

4 Evaluation of preferential strategies

Although there is much interest in implementing public transport preferential measures, evaluation of their impacts remains as a big challenge due to lack of a systematic analysis framework. Such impacts could be either investigated by conducting a before-after comparison of public transport performances or by simulating public transport system. Simulation is often used for studying the effects of real-time control strategies such as public transport signal priority [11], stop skipping [35], holding [8] and short-turning [36].

The majority of previous studies have only considered vehicle-based performance metrics, and effects on passenger travel time (stop-based perspective) have received less attention. Even when both vehicle and passenger travel times are considered, only changes in selected performance metrics are investigated rather than focusing on the benefits of the implemented measures, preventing the overall assessment of investment costs. A new evaluation framework is presented in this thesis in which all the aforementioned shortcomings are fulfilled.

In short, travel time reduction, reliability benefits and operational costs should be precisely evaluated. The proposed approach is to classify measures according to their effects on vehicle time components. Among such measures are link-related (e.g. bus lanes, signal priority, elevated crossing), stop-related (e.g. docking guidance, boarding procedure) and operations and control (e.g. stopping pattern, holding strategy). Specifically, these classes of measures are expected to have impact on the running time between stops, dwelling time at stops or on both.

The proposed evaluation framework tries to systematically quantify and assess the impacts of preferential measures on both service users and operators in monetary terms. In the first step, the change in vehicle-related performance is analyzed with the help of AVL data by considering service speed and reliability metrics. As the next step, the operators' and passengers' benefits are calculated from such change. Any improvement in vehicle performance could possibly result in a reduction of operational costs and passenger travel time. For measuring the overall passengers' time gain or loss, some information is required about passenger demands. This information is normally retrieved via APC or smart cards. The third step is to separately monitor changes in passenger travel time components (walking time, waiting time and in-vehicle time) by checking vehicle performance metrics and passenger demand patterns. The

nominal and perceived travel time components accumulated to reach the total and perceived passenger travel time, respectively. Fourth, both the operators' and passengers' benefits are converted into monetary terms by accounting for operational cost factors and passengers' value of time. Such cost combination gives out the overall benefits attributed to the introduction of the preferential measures.

5 Bus travel time prediction

Several methods were introduced for bus travel time prediction in the last two decades, all targeting at fast and accurate prediction under various circumstances. However, the challenge still remains to choose the "best model" performing well on a wide range of transportation network systems. Besides, the existence of such model is not yet verified in the literature [28]. Hence, development of a fusion method for jointly having the advantage of different models and using different data sources is highly desirable.

Recently, hybrid models have emerged as a promising prediction approach in the context of freeway traffic ([37], [38], [40] and [24]). However, we are not aware of any such model that has been developed and applied for real-time prediction of bus travel times. Although hybrid models perform well, there is still a high tendency towards using simple models in the most recent advanced travel information system [28]. For one reason, simple prediction models are easily adopted to real-time information system providers in practice [9]. Another reason is the computational cost and inherent complexity of hybrid models presented in the literature. In simple words, a suitable model for real-time information needs to be fast, reasonably simple, capable of handling noisy data from various sources, scalable and robust towards unexpected disruptions.

Although noticeable research efforts are made to develop suitable bus travel time prediction methods, it is not yet clear if these methods could be transferred over different transport networks. Also, the added-value for operators and passengers is neither verified within the new method nor compared with the currently deployed methods. Another issue is to prove if the performance of prediction methods could be generalized beyond the single line level and transferred beyond a certain operational context.

A hybrid model is presented as a linear fusion of three prediction methods: Scheduled, Instantaneous and Historical. The main source of information for

scheduled based predictor is a static time-dependent schedule. Instantaneous predictor takes the last few observations for generating prediction. These observations are selected in a way to minimize the prediction error. Historical predictor mines historical data based on certain selection criteria that should be designed to maximize prediction accuracy. The contribution of each predicting component and the associated parameters should be then determined.

Automated Vehicle Location (AVL) data is considered as a data source for the two latter components. Then, prediction is performed on a rolling horizon basis with each prediction including departure time predictions for all the downstream stops. Finally, the proposed hybrid model is evaluated from two different perspectives: operators' and passengers'.

6 Discussion and recommendations

6.1 Paper 1 and 2

We have proposed a methodology approach to estimate route choice models and link travel times using sparse GPS data. The link-based approach is used, where it is assumed that the individual is choosing the shortest path. While route choice models are traditionally estimated using observations of chosen paths, the estimator proposed in these papers enables the estimation of flexible route choice models using sparse GPS data, with only partially observed paths. In the context of route choice, most closely related to our work are [16] and [30], who also consider GPS data for model estimation. Yet, our approach allows for estimation of models with a flexible correlation structure, full choice set consideration, and corrections for potential biases due to data manipulations, such as map-matching.

The proposed indirect inference estimator for the estimation of flexible route choice models also addresses the problem of sparse GPS data with measurement errors. This approach is implemented in paper 2, to estimate both speed and the route choice model simultaneously by using iteration between an estimator of speed and the estimator of the route choice model.

The main conclusion is that indirect inference is a useful option for estimating route choice models and speed values on all the links in a network using low frequency GPS sampling data with measurement errors. The Monte Carlo evidence shows that, applying the indirect inference approach to route choice and speed estimation is a worthwhile solution.

14

Outlook for future work

- The advent of passive monitoring of route choices has provided different data collection methods [29], and that the estimation methods proposed in this paper can equally be further applied to data sets collected using other technologies.

- The assumption of known distribution of GPS-errors is crucial feature of the current approach. We also tested our method with a simulated data set to verify that the method is working satisfactorily. Both these issues will be addressed in future work.

6.2 Paper 3

This paper presents a systematic evaluation framework and a detailed The evaluation framework introduced here assesses the impacts of a combination of public transport preferential measures by considering both passengers and operators benefits. The proposed framework uses automated and passive data collection to ease performance monitoring and post-implementation assessment without imposing the need for any additional data collection.

The overall assessment for impacts of considered preferential measures enables the comparison of different implementations, evaluation of their effectiveness, prioritization of alternative measures and provision of a sound basis for motivating investments in such measures.

Outlook for future work

- In this study, evaluation of preferential measures is restricted to primary and secondary implications on passenger travel times and operational costs for a specific line respectively. At the network-level, the induced demand to the improved line due to route choice or model choice can be assessed as well as variations in accessibility due to frequent changes in stopping patterns. In some cases, non-users may experience difficulties such as prolonged travel times due to the prioritization of public transport. These issues could be considered in the further extension of proposed framework.

- As an another guide for future research, the inclusion of indirect and long-term effects such as economic activity, externalities and land-use development [39] in the evaluation framework should be investigated.

6.3   Paper 4 and 5

The advantages of three prediction methods are well-integrated through development of a hybrid prediction model. This model linearly fuses schedule-based, instantaneous-based and historical-based predictors. The weight for each prediction method is specified by a linear regression based heuristic algorithm to minimize its prediction error.

Application of the hybrid method has been examined for a set of routes together as well as for each route separately. The results prove that in a separate application, more accuracy is attained than in the joint estimation, however the imposed model complexity is not justified.

The hybrid method is applied to five bus lines in Stockholm and Brisbane. The gained accuracy is then assessed by comparing its performance to alternative prediction methods. Assessment results show noticeable improvement for the hybrid method over timetable and delay conservation prediction method. Although the added-value of the hybrid method is different when applied to different bus routes and time periods, the results prove the transferability of the model over a variety of line layouts, passenger demands and operational scenarios.

The real-time application of the proposed method was confirmed through model validation. According to the results, the model parameters could be estimated only once and then applied to other cases while still yielding high accuracy.

Outlook for future work

- The prediction model parameters are homogeneously used throughout the route. By dynamically updating the weights for individual prediction, the prediction method could be further improved. This will allow for self-learning dynamically altering the contribution of each predictor based on its recent performance.

- The proposed hybrid method is applied to provide real-time information for predicting downstream vehicle trajectories and next bus arrival time.

Evaluation is separately done to assess operators' and passengers' gained benefits. These two perspectives could be verified by integrating different prediction methods for vehicle-based and stop-based viewpoint.

- The hybrid method proposed in this thesis can fuse various predictors and be useful in many different applications. This method can be further developed to predict total journey time by estimating expected headways and thus deducing the transfer times. Moreover, the possibility of embedding such prediction methods in a decision support system should be investigated to ease the implementation of proactive fleet management and control strategies.

Bibliography

[1] Oshyani, M. F., Sundberg, M., & Karlström, A. (2012). Estimating flexible route choice models using sparse data. In Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on (pp. 1215-1220). IEEE.

[2] Oshyani, M. F., Sundberg, M., & Karlström, A. (2014). Consistently estimating link speed using sparse GPS data with measured errors. Procedia-Social and Behavioral Sciences, 111, 829-838.

[3] Oshyani, M. F., & Cats, O. (2014). Rolling horizon predictions of bus trajectories. In 1st International Conference on Engineering and Applied Sciences Optimization, OPT-i 2014; Kos Island; Greece; 4 June 2014 through 6 June 2014 (pp. 875-886).

[4] Oshyani, M. F., & Cats, O. (2014). Real-time bus departure time predictions: vehicle trajectory and countdown display analysis. In Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on (pp. 2556-2561). IEEE.

[5] Ben-Akiva, M. and Ramming, S., (1998). Lecture notes: Discrete choice models of traveler behavior in networks. Prepared for Advanced Methods for Planning and Management of Transportation Networks, Capri, Italy.

[6] Ben-Akiva, M. and Bierlaire, M., (2003). Discrete choice models with applications to departure time and route choice, in R. Hall (ed.). Handbook of Transportation Science, 2nd edition, Operations Research and Management Science, Kluwer, pp. 7-38. ISBN:1-4020-7246-5.

[7] Cascetta, E., Nuzzolo, A., Russo, F. and Vitetta, A., (1996). A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks, in J. B.

Lesort (ed.). Proceedings of the 13th International Symposium on Transportation and Traffic Theory, Lyon, France.

[8] Cats, O., Larijani, A., Koutsopoulos, H., & Burghout, W. (2011). Impacts of holding control strategies on transit performance: Bus simulation model analysis. Transportation Research Record: Journal of the Transportation Research Board, (2216), 51-58.

[9] Cats, O., & Loutos, G. (2015). Real-time bus arrival information system: an empirical evaluation. Journal of Intelligent Transportation Systems, (ahead-of-print), 1-14.

[10] Cats, O., Abenoza, R. F., Liu, C., & Susilo, Y. O. (2015). Identifying priority areas based on a thirteen years evolution of satisfaction with public transport and its determinants. In 94th Annual Meeting Transportation Research Board, Washington, USA, 11-15 January 2015; Authors version. TRB.

[11] Chandrasekar, P., Long Cheu, R., & Chin, H. C. (2002). Simulation evaluation of route-based control of bus operations. Journal of Transportation Engineering, 128(6), 519-527.

[12] Chen, W., Yu, M., Li, Z., Chen, Y., Chao, J., (2003). Tight integration of digital map and in-vehicle positioning unit for car navigation in urban areas. Wuhan University Journal of Natural Sciences, 8(2):551-556.

[13] Currie G. & Wallis I. (2008). Effective ways to grow urban bus markets - a synthesis of evidence. Journal of Transport Geography 16(6), 419-429.

[14] Dial, R.B., (1971). A probabilistic multipath traffic assignment model which obviates path enumeration. Transportation Research, 5(2):83-111.

[15] Fosgerau, M., Frejinger, E. and Karlström, A., (2011). A logit model for the choice among infinitely many routes in a network. Technical report, Center for Transport Studies, Royal Institute of Technology KTH, Stockholm, Sweden.

[16] Frejinger, E. and Bierlaire, M., (2008). Route choice modeling with network-free data. Transportation Research Part C, 16:187-198.

[17] Frejinger, E., & Bierlaire, M. (2007). Capturing correlation with subnetworks in route choice models. Transportation Research Part B: Methodological, 41(3), 363-378.

[18] Frejinger, E., Bierlaire, M., Ben-Akiva, M. (2009), Sampling of alternatives for route choice modeling, Transportation Research Part B: Methodological, 43(10):984-994.

[19] Gourieroux, C., Monfort, A. and Renault, E., (1993). Indirect inference. Journal of Applied Econometrics, 8(S1):S85-S118.

[20] Guevara, C.A., (2010). Endogeneity and sampling of alternatives in spatial choice models. PhD Dissertation, MIT, USA.

[21] Jan, O., Horowitz, A.J., Peng, Z.R., (2000). Using Global Positioning System Data to Understand Variations in Path Choice. Transportation Research Record: Journal of the Transportation Research Board, 1725(1):37-44.

[22] Keane, M., & Smith, A. A. (2003). Generalized indirect inference for discrete choice models. Department of Economics, Yale University (manuscript).

[23] Karlström. A., Sundberg, M., Wang, Q., (2011). Consistently estimating flexible route choice models using an MNL lens. International Choice Modelling Conference, Leeds, UK.

[24] Li, C. S., & Chen, M. C. "A data mining based approach for travel time prediction in freeway with non-recurrent congestion." Neurocomputing 133 (2014): 74-83.

[25] Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., (2009). Map-matching for low-sampling-rate GPS trajectories. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09). ACM, New York, NY, USA, pp. 352-361.

[26] McFadden, D. (1978). Modelling the choice of residential location (pp. 75-96). Institute of Transportation Studies, University of California.

[27] Miller, H. J. (2005). A measurement theory for time geography. Geographical analysis, 37(1), 17-45.

[28] Mori, U., Mendiburu, A., Álvarez, M., & Lozano, J. A. "A review of travel time estimation and forecasting for Advanced Traveller Information Systems." Transportmetrica A: Transport Science 11, no. 2 (2015): 119-157.

[29] Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting?. Transportation research part c: emerging technologies, 7(2), 149-165.

[30] Newman, J., Chen, J., & Bierlaire, M. (2009). Generating probabilistic path observation from GPS data for route choice modeling. In European Transportation Conference 2009 (No. EPFL-TALK-152440).

[31] Noland, R. B., & Polak, J. W. (2002). Travel time variability: a review of theoretical and empirical issues. Transport Reviews, 22(1), 39-54.

[32] Patnaik, J., Chien, S., & Bladikas, A. Estimation of bus arrival times using APC data. Journal of public transportation 7, no. 1 (2004): 1.

[33] Quddus, M. A., Ochieng, W. Y., Zhao, L., & Noland, R. B. (2003). A general map matching algorithm for transport telematics applications. GPS solutions, 7(3), 157-167.

[34] Schüssler, N., and K. W. Axhausen. Accounting for Route Overlap in Urban and Sub-Urban Route Choice Decisions Derived from GPS Observations. Proc., 12th International Conference on Travel Behavior Research, Jaipur, India, Dec. 2009.

[35] Sun, A., & Hickman, M. (2005). The real-time stop-skipping problem. Journal of Intelligent Transportation Systems, 9(2), 91-109.

[36] Tirachini, A., Cortés, C. E., & Jara-Díaz, S. R. (2011). Optimal design and benefits of a short turning strategy for a bus corridor. Transportation, 38(1), 169-189.

[37] Van Lint, J.W.C. "Online learning solutions for freeway travel time prediction." Intelligent Transportation Systems, IEEE Transactions on 9, no. 1 (2008): 38-47.

[38] Van Hinsbergen, C.P.I.J., van Lint, J.W.C., van Zuylen, H.J. Bayesian committee of neural networks to predict travel times with confidence

intervals. Transportation Research Part C: Emerging Technologies 17, no. 5 (2009): 498-509.

[39] Weisbrod G. and Reno A. (2009). Economic impact of public transportation investments. American Public Transport Association.

[40] Ye, Q., Szeto, W. Y., & Wong, S. C. "Short-term traffic speed forecasting based on data recorded at irregular intervals." Intelligent Transportation Systems, IEEE Transactions on 13, no. 4 (2012): 1727-1737.

[41] Barceló, J., Kuwahara, M., & Miska, M. (2010). Traffic data collection and its standardization (pp. 1-10). Springer New York.

[42] Okunieff, P. E. (1997). TCRP Synthesis of Transit Practice 24: AVL Systems for Bus Transit. Transportation Research Board, National Research Council, Washington, DC, 4-5.

[43] Okunieff, P. E. (1997). TCRP Synthesis of Transit Practice 24: AVL Systems for Bus Transit. Transportation Research Board, National Research Council, Washington, DC, 4-5.

[44] Hickman, M. (2004). Bus automatic vehicle location (AVL) systems. In Assessing the Benefits and Costs of ITS (pp. 59-88). Springer US.

[45] Ben-Akiva, M., & Bierlaire, M. (2003). Discrete choice models with applications to departure time and route choice. Handbook of transportation science, 32.

[46] Cascetta, E., Nuzzolo, A., Russo, F., & Vitetta, A. (1996, July). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In Proceedings of the 13th International Symposium on Transportation and Traffic Theory (pp. 697-711). Oxford, NY, USA: Pergamon.

[47] Ramming, M. S. (2001). Network knowledge and route choice (Doctoral dissertation, Massachusetts Institute of Technology).

[48] Prato, C. G. (2011). Latent factors and route choice behaviour.

[49] Vrtic, M., Schuessler, N., Erath, A., Axhausen, K., Frejinger, E., Stojanovic, J., ... & Maggi, R. (2006). Including travelling costs in the modelling of mobility behaviour (No. TRANSP-OR-REPORT-2006-021).

[50] Jan, O., Horowitz, A., & Peng, Z. R. (2000). Using global positioning system data to understand variations in path choice. Transportation Research Record: Journal of the Transportation Research Board, (1725), 37-44.

[51] Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS) improve trip reporting?. Transportation research part c: emerging technologies, 7(2), 149-165.

[52] Wolf, J., Hallmark, S., Oliveira, M., Guensler, R., & Sarasua, W. (1999). Accuracy issues with route choice data collection by using global positioning system. Transportation Research Record: Journal of the Transportation Research Board, (1660), 66-74.

[53] Zito, R., d'Este, G., & Taylor, M. A. (1995). Global positioning systems in the time domain: How useful a tool for intelligent vehicle-highway systems?. Transportation Research Part C: Emerging Technologies, 3(4), 193-209.

[54] Turner, S. M., Eisele, W. L., Benz, R. J., & Holdener, D. J. (1998). Travel time data collection handbook (No. FHWA-PL-98-035,).

[55] Mishan, E. J., & Quah, E. (2007). Cost-benefit analysis. Routledge.

[56] Abkowitz, M. (1978). Transit service reliability (No. UMTA-MA-06-0049-78-1Final Rpt.).

[57] Quddus, M. A., & Enoch, M. P. (2013). Estimating Link Travel Time from Low-Frequency GPS Data. In Transportation Research Board 92nd Annual Meeting (No. 13-3909).

[58] Zheng, F., & Van Zuylen, H. (2013). Urban link travel time estimation based on sparse probe vehicle data. Transportation Research Part C: Emerging Technologies, 31, 145-157.

[59] Oshyani, M. F., & Cats, O. (2015). An empirical evaluation of measures to improve bus service reliability: Performance metrics and a case study in Stockholm. In CASPT Conference 2015 in Rotterdam, the Netherlands.