1 Introduction

With the growing adoption of automatic data collection systems (ADCS), transit providers can now collect volumes of data on their operations and their customers' behavior. The use of automated fare collection (AFC) systems, by collecting fares electronically and storing the data digitally, permits urban public transit providers to collect fine-resolution data on how their customers interact with the network. These data can show, for every transaction a customer performs, where and when that transaction occurred. If customers are uniquely identified, it is then possible to determine their behavior over one or many days, including the origins and destinations (OD) of their trips.

By using ADCS data to infer origins, destinations, and journeys, transit providers and researchers can avoid the need to gather information about usage using costly, time-consuming, and often inaccurate surveys (Riegel, 2013). Due to their nature, these surveys are limited in both sample size and frequency, whereas (ADCS) data have the potential to provide information at a near daily frequency. In the USA an example of the potential use of such information is the federally required Title VI and environmental justice (EJ) reporting. Title VI of the Civil Rights Act of 1964 prevents agencies receiving federal funding from having a disparate impact with regards to race, ethnicity, or national origin. In complying with this law, large transit agencies must report regularly on how their service is provided to populations with different demographics. There is a growing critique of the inaccuracies inherent to the required methodology (Bills, 2013; Karner & Golub, 2015; Karner & Niemeier, 2013), based on outdated data collection methods which have not kept pace with the availability of large, passenger-level data sets from ADCS.

A distinction must be made between early implementations of these technologies, which were designed with one task in mind such as AFC or automatic vehicle location (AVL) systems (hereafter legacy systems), and a second generation of ADCS designed with a holistic view of data collection and warehousing. Legacy systems require the synthesis of disparate data sources in order to produce useful information such as origins and destination whereas newer systems will have such synthesis built into data collection.

This thesis builds upon recent work in the synthesis of passenger-centric public transit information and primarily updates the work of Gordon (2013), to infer origins, destinations, and full journeys in London, to a fully open[1] transit network: the Massachusetts Bay Transportation Authority's (MBTA) rapid transit and bus network. Additional algorithms were developed to prepare bus and train vehicle location data for passenger origin and destination inference, as well as arrival time inference on a rail. The inference algorithm was performed on a month's worth of data for April 2014. From this inferred origin–destination (OD) information users' home locations are inferred, and their usage is aggregated to geographic units to demonstrate an

[1] An open transit network is one in which fare payment only occurs at the origin of a trip, at boarding or gate entry, rather than a closed system which requires fare payment at entry and exit.

alternative methodology for Title VI and EJ analysis which betters reflects passenger outcomes.

The remainder of this chapter provides the motivation for this case study, an overview of the state of the art of OD inference, the objectives and methodology for this study, and outlines the rest of this thesis.

1.1 Motivation

1.1.1 State of the Art: Using ADCS to Infer Travel Behaviour

Pelletier, Trépanier, & Morency (2011) review the use of smart card AFC data including OD inference methods for tactical and strategic transit planning. Since their review, the state of the art has moved in a number of directions discussed in the sections below: from improved methodologies, to inferring activities from OD, or using alternative massive passively collected data sets to infer travel behaviour.

Robinson et al (2014) review data collection errors for AFC and AVL systems. They discuss how to isolate faulty data collection units through peer comparison and present methods for error handling and correction with a particular focus on bus or light rail systems that require fare transactions upon both boarding and alighting.

Activity Inference

Moving beyond inferring the origin and destination of a trip, researchers have developed methodologies to elucidate trip purpose from the user and trip characteristics as well as land use characteristics. Lee and Hickman (2014) infer home-based trip purposes (work, university, and other) for bus passengers in the Minneapolis/St. Paul Metropolitan Area in Minnesota using an OD inferred from AFC and GTFS[2] schedules (Nassir, Khani, Lee, Noh, & Hickman, 2011). They assume users start their first journey of the day at home and used start time, activity duration and location as criteria in activity inference.

(Devillaine, Munizaga, & Trépanier, 2012) infer activity types (home, work, education, other) using simple heuristics based on the type of pass, the trip's order (e.g., whether it was the last of the day), the duration of the activity, and the land use of the location of the activity.

Alternative Data Sources

Researchers have developed methodologies to use more ubiquitous data sources to infer travel behaviour. Jiang et al.(2013) offer a comprehensive review of the use of passively collected mobile phone data including challenges and opportunities for that data source.

Montero et al (2015) use Bluetooth data to estimate real-time dynamic OD on a transit network. Their work extends a framework originally designed to predict traffic on roadways, to using a

[2] General Transit Feed Specification: a standard for publishing machine-readable transit schedules

historical OD matrix for Vitoria in Spain updated with counts of passengers equipped with Bluetooth enabled devices. This methodology was validated with simulated rather than actual real-time data.

1.1.2 ADCS Applications in Boston

The MBTA's fare data have been used to analyze the effects of different fare policies. Pincus (2014) determined the impacts of the 2012 fare increase through analyzing months of AFC records. Kamfonik (2013) estimated the added revenue the MBTA's Corporate Pass program provides by analyzing the AFC usage records of Corporate Pass holders. Chow (2014) piloted the use of AFC in a survey of a panel of MBTA customers prompting them to recall their trips based on their transaction history.

Researchers have also used ADCS to improve transit operations. Tribone (2013) used automatically collected track circuit data to identify reasons for delays on the subway and then piloted and evaluated solutions to these. Maltzan (2015) designed real-time control strategies on high-frequency bus routes using real-time AVL data and evaluated the benefits of pilot interventions.

1.1.3 Motivation: Using ADCS to Analyze and Improve Transit Outcomes

Beyond the obvious use of OD matrices for service and network planning, researchers have explored how ADCS can inform social policy. In developing countries where censuses are conducted sporadically, Smith et al. find that indicators calculated from mobile phone Call Detail Records (CDR) correlate well with regional poverty as defined by the Multidimensional Poverty Index (Smith, Mashhadi, & Capra, 2013). In a developed country context, Smith et al. use heavy rail AFC data from the Oyster system in London to calculate proxy indicators that correlate with social indicators derived from census data collected on the order of every 5 years (Smith, Quercia, & Capra, 2013). In contexts with robust censuses, these applications are intended to supplement censuses by providing intermediate diagnostics in between censuses. This can allow for rapid feedback on policy changes without the need for intermediate surveys of target populations or to wait for census results.

In the USA, analysts at New York City Transit (NYCT) have published a number of Title VI and EJ analysis methodologies using ADCS. The agency was the first to determine disparate impacts using statistical methods rather than heuristic rules (Reddy, Chennadu, & Lu, 2010). Fare-impact methodologies were developed to estimate the impacts of the March 2008 and December 2009 proposed fare change using two methods (Hickey, Lu, & Reddy, 2010). Span adjustments and route modifications accompanied the fare increase, and *t*-tests were performed on load factor and travel time distributions to demonstrate no disparate impact (Wang, Lu, & Reddy, 2013).

The federal reporting process has been generally criticized for missing large segments of target populations through analysis of averages aggregated by zones, rather than examining outcomes at the level of individual persons. Furthermore, the comparison of averages between populations

rather than examining different distributions of individuals masks the existence of winners and losers across and within populations (Bills, 2013). This leaves planners without sufficient information for localized decision-making to identify interventions to correct disproportionate impacts.

The Massachusetts Bay Transportation Authority (MBTA), the transit provider for the Boston Metropolitan Area (Boston), must regularly perform Title VI and environmental justice reporting as a condition of Federal funding (most recently conducted by CTPS (2014)). A 2014 study by Williams, Pollack, & Billingham used American Community Survey (ACS) data from 2011 to examine commute times by race and income for the Boston Metropolitan Area. They found a significant commute time penalty for Black commuters versus White ones across all modes which was most pronounced on the bus (on average an extra 70 hours per year). In an Ordinary Least Squares (OLS) regression, when controlling for income, racial penalties persisted. While the data used for their analysis was from individual surveys, it had been geographically anonymized to such a degree to render a spatial analysis of impacts impossible. Additionally, information about the journey to work was missing important contributors to travel times such as journey distances, the number of transfers between vehicles or modes, and whether the journey required mandatory stops.

1.1.4 Summary

As ADCS have become more prevalent, researchers have explored how these systems can guide social policies. The US federal regulation requiring transit agencies to provide equitable service benefits regardless of color, race, or national origin is an example of such a social goal that can be better informed with these systems. While agencies have begun incorporating these data into Title VI and EJ analyses, the current federal reporting requirements have not kept pace with the availability of data and the ability to perform finer resolution analysis. This thesis explores how inferred OD data can enable finer grained equity analysis.

1.2 Objectives

This thesis demonstrates the feasibility of using a month of inferred origins and destinations from transit ADCS to perform periodic analysis of the spatial variation of transit service as part of ongoing service monitoring and in order to fulfill Federal Title VI reporting requirements. In order to accomplish this goal the following objectives must be met.

Infer boarding and alighting locations and times for bus journey stages in Boston

AVL data at a stop-level resolution are required for OD inference, however the bus AVL system was not designed to specifically record arrival times at stops. An appropriate source of AVL must be selected and processed to synthesize the stop-level input for inferring boarding and alighting locations.

Infer alighting locations and times for rail journey stages in Boston

In a rail system where users' exit information is not collected, alighting locations and arrival

times must be inferred. Rail AVL must be processed to derive stop-level arrival times, and these data must be used in an arrival time inference process to determine the time at which each user arrived at their rail destination.

Infer interchanges between journey stages of any AFC-enabled mode

From the inferred OD, link together stages into journeys as accurately as possible by adapting heuristic parameters for the specific MBTA transit context.

Prepare OD inference process to be run over months

This analysis requires multiple days of data to analyze travel behaviour and transit performance over time. The inference and preprocessing algorithms must be automated to be able to run over multiple days.

Develop methodologies for analyzing spatial variation of transit effectiveness

Assess current equity analysis methods and their critiques in the literature and from these propose new methodologies for analyzing spatial variation of transit effectiveness that better reflect passenger outcomes using ADCS.

Link users to demographic census data

In order to compare user behaviour and experienced service across different demographics and geographies, a link must be made between the fare payment ID and that user's home location. By determining users' home locations, the demographics of their home neighbourhood can be linked to their ID to demonstrate how transit use and experience differs by neighbourhood demographics.

Determine if differences exist in home based journey characteristics across demographics and space

Determine the distance, travel time, speed, and number of transfers for home-based trips for different users and compare the distributions by demographics. Map differences in behaviour and experience. Identify regions where differences are larger and explore causes of poor transit effectiveness.

Propose and evaluate a set of solutions to differences in transit

Based on the analysis performed, determine a set of potential solutions; for example: bus route modification, increased bus frequencies, fast frequent commuter rail service. Evaluate the impacts of these solutions.

1.3 Thesis Organization

The thesis is divided into two parts: the work required to process and infer ODs in the Boston context (Chapter 2) and the subsequent use of this OD to analyze spatial variation in transit effectiveness in Boston (Chapters 3-5). Chapter 3 presents an overview of efforts to quantify transportation equity and presents a history of analyses in Boston. Chapter 4 contains the methodology used to process the inferred OD from Chapter 2 into metrics used for the analysis

of spatial variability in transit outcomes. The results of this analysis are discussed in Chapter 5 and select solutions to identified differences are proposed and evaluated. The final chapter reflects on the study's findings and presents conclusions and recommendations for future research.