Forecasting bus ridership using a “Blended Approach”

Catherine T. Lawson¹ · Alex Muro¹ · Eric Krans¹

Published online: 3 December 2019
© The Author(s) 2019

Abstract

As sources of “Big Data” continue to grow, transportation planners and researchers seek to utilize these new resources. Given the current dependency on traditional transportation data sources and conventional tools (e.g., spreadsheets and propriety models), how can these new resources be used? This research examines a “blended data” approach, using a web-based, open source platform to assist transit agencies to forecast bus ridership. The platform is capable of incorporating new Big Data sources and traditional data sources, using modern processing techniques and tools, particularly Application Programming Interfaces (APIs). This research demonstrates the use of APIs in a transit demand methodology that yields a robust model for bus ridership. The approach uses the Census Transportation Planning Products data, modified with American Community Survey data, to generate origin–destination tables for bus trips in a designated market area. Microsimulation models use a transit scheduling specification (General Transit Feed Specification) and an open source routing engine (OpenTripPlanner). Local farebox data validates the microsimulation models. Analyses of model output and farebox data for the Atlantic City transit market area, and a scenario analysis of service reduction in the Princeton/Trenton transit market area, illustrate the use of a “blended approach” for bus ridership forecasting.

Keywords Farebox data · General Transit Feed Specification (GTFS) · Census Transportation Planning Products (CTPP) · Application Programming Interface (API)

Introduction

Transit agencies need to plan as efficiently and effectively as possible to compete with emerging mobility options, while continuing to serve those populations most in need of critical household activity travel (e.g., work commute for household members). Transit agencies have

✉ Catherine T. Lawson
lawsonc@albany.edu

Alex Muro
amuro@albany.edu

Eric Krans
ekrans@albany.edu

¹ State University of New York, Albany, 1400 Washington Avenue AS 218, Albany, NY 12222, USA

a long history of using data for transit planning, yet the data ecosystems within these agencies are often constrained by internal and external policies and procurement practices. These challenges include the use of proprietary software, the lack of data sharing agreements, the lack of data standards, and the lack of internal workforce data-handling skills. At the same time, modern data processing opportunities and new forms of data offer many cost-effective approaches to tackle some of these challenges (Lawson et al. 2019). There is a growing interest in identifying opportunities for the use of emerging data sources (often referred to as “Big Data”) in combination with, or in place of, traditional transportation data sources, for transit planning. Erhardt and Dennett (2017) found that Census data has been used in direct competition with Big Data, as well as complementary to it. The emerging data have many characteristics that have been absent from traditional data sources (e.g., continuously produced, site-specific, voluminous), but, at the same time, lack essential socio-demographic information necessary for forecasting travel behavior. Recent efforts are transforming data ecosystems to blend a various data types.

An example of the transit industry transitioning from traditional data to emerging data sources began in 2005, when the Tri-County Metropolitan Transportation District of Oregon (Tri-Met), partnered with Google, to develop an open data scheduling strategy (Lawson 2016a). Their efforts resulted in the creation of the General Transit Feed Specifications (GTFS), a common format for public transportation schedules that includes associated spatial information (GTFS Static Overview 2016). As an open data approach, it made a unique contribution with the generation of static schedule information (e.g., stop location, route geometrics, and stop times) in a standard format (see <https://developers.google.com/transit/gtfs/>). Wong (2013) and Wong et al. (2013) described the many uses of GTFS providing a better understanding transit ridership. Rodnyansky (2018) reviewed uses of GTFS, providing descriptions of methods for accessing GTFS for individual projects.

Another example is the use of archived Intelligent Transportation Systems (ITS) transit data. Iliopoulou and Kepapoglou (2019) reviewed uses of archived ITS transit data including Automated Vehicle Location (AVL), Automatic Fare Collection (AFC), Automatic Passenger Count (APC). The authors found uses of archived ITS transit data included: strategic level planning; transit assignment; network design; tactical level planning; optimal timetabling; origin–destination and transfer inference; and activity modeling. The lack of integration of these data types, the need for advanced computational analysis, and the lack of data sharing policies are challenges for these uses.

Efforts to harness modern processing techniques for transit planning are currently underway. However, many efforts remain as individual research projects rather than adopted into mainstream usage. This research explores the opportunity to use modern processing techniques in small and medium-sized transit agencies. The next section provides a review of new data types and methods for transit planning. The third section provides a description of data ecosystem elements, including tools. The fourth section details the process of estimating bus ridership using a unique, web platform and a number of data sources. The fifth section describes case studies in two cities in New Jersey. This is followed by a discussion of opportunities, limitations, and future research. The final section provides conclusions, recommendations for transit agencies, and considerations for transit data ecosystems for improving ridership-forecasting tools.

Background

One of the first advanced econometric analysis using the vast data resource of archived Intelligent Transportation Systems (ITS) data was conducted using transit operations data from Tri-Met, the regional transit agency for the Portland, Oregon region (Peng 1994). Peng (1994) developed a route-level transit patronage model. The research identified and accounted for three modeling challenges including data inconsistency, simultaneous transit supply and demand effects, and transit line interrelationships.

A number of studies have focused on social and demographic factors that influence transit ridership (Kimpel 2001; McKenzie 2011; Thompson et al. 2012; Lee et al. 2013a; Wang and Woo 2017; Ma et al. 2018). Other studies examined different service factors (Verbas et al. 2013; Vij and Walker 2013; Brown et al. 2013) or land use aspects (Dill et al. 2013; Frei and Mahmassani 2013; Wang and Woo 2017). Liu et al. (2018) focused primarily on accessibility. Table 1 lists research using new forms of transit data, in combination with other traditional data, and applications.

Advances in the use of software platforms and web-interfaces has spawned a number of transit planning tools and new approaches (Sun et al. 2011; Antrim and Barbeau 2013; Owen and Levinson 2017; Liebig et al. 2014; Giraud et al. 2016; Pi et al. 2018). Karner (2018) reviewed the 2012 Federal Transit Administration (FTA) mandated process for evaluating transit projects with respect to equity. Urbanized areas with populations exceeding 200,000 are required to perform a service equity analysis in order to obtain federal funding for major service changes to determine if proposed changes have a disparate impact on minority households, or results in disproportionate impacts on low-income households. Table 2 provides details of recent transit planning tools taking advantage of new data sources and platforms for analyses.

The Federal Transit Administration (FTA) continues to support their Simplified Trips-on-Project Software (STOPS). STOPS is a variation of the traditional Four Step travel demand-forecasting model that uses the Census Transportation Planning Products (CTPP) rather than trip-generation and trip-distribution tables. The transit network now uses GTFS and relies on traditional zone-to-zone roadway times and distances from regional travel models (current and forecast year). The software requires extensive data input for highway supply, travel demand information, and transit supply components (RSG 2015). The skills required include: experience using one or more GIS packages and ability to create GIS layers; an understanding of the travel forecasting methodology; and familiarity with regional transit systems (e.g., different agencies providing services in the area and using their own schedules). RSG (2019) describes the Incremental Mode, a recent advancement that uses recent detailed transit rider surveys, if available. The process divides survey transit trips by the transit share (from a mode choice model calibrated to match CTPP shares) to capture incremental impacts of changes (e.g., transit levels-of-service).

Conveyal (2019) provides guidance on their platform tools, techniques, and instructions for the assembly of necessary data sources. The open source code for their tool is available at Github (see <https://github.com/conveyal>). Hanft et al. (2016) points out that most transit agencies lack the resources to develop comprehensive ridership data and the complex, transit demand models, similar to those used by New York City Transit (NYCT). Understanding the data ecosystem within a transit agency is critical to employing the most efficient and effective approach to forecasting transit ridership.

Kressner et al. (2016) describes the use of passive data as a replacement for travel surveys using public data and cell tower movement data harvested from moving

Table 1 Transit ridership research using archived ITS transit data and emerging data sources

Author(s)	Date	Purpose	Application/Findings
Kimpel	2001	Analysis of socio-economic factors of transit demand	Zero auto-ownership households has major effect on demand for crosstown routes in Portland, Oregon. Growth in population and employment also increase transit patronage
McKenzie	2011	Examination of transit access for areas with high concentrations of blacks, Latinos, and low-income households	Members of black households had better access to transit than Latinos in the Portland, Oregon region
Puchalsky et al.	2012	Development of regional model	Applied in Delaware Valley Regional Planning Commission (DVRPC) nine-county area
Thompson et al.	2012	Analysis of bus ridership	Low-income households and low levels of vehicle access characterized the ridership profile
Lee et al.	2013a	Development of time-varying route-level transit patronage model to analyze ridership	Applied in Minneapolis/St. Paul, Minnesota
Frei and Mahmassani	2013	Examination of disaggregate ridership elasticity estimations applied to large bus transit network	Transit stops located near medical facilities increased ridership in medium and long term. During the day, elasticities were lower for industrial, medical, recreational, and educational areas in Chicago, Illinois
Lee et al.	2013b	Conducted a comparative study of alternative methods for generating route-level mutually exclusive service areas	Applied in Minneapolis/St. Paul, Minnesota
Brown et al.	2013	Examination of ridership on different transit types (rail transit, transit-dependent bus services)	Applied in Atlanta, Georgia
Dill et al.	2013	Development of methodology for predicting transit ridership at stop-level	Applied in Oregon (Tri-Met in Portland, Lane Transit District in Eugene, and Rogue Valley Transit District in Jackson County)
Verbas et al.	2013	Development of multiple scenarios to illustrate ridership with respect to changes in headways	Application designed for Chicago Transit Authority (CTA)
Vij and Walker	2013	Examination of mode share responses to incremental improvements	Without corresponding shifts in individual modality preferences, changes will be smaller than traditional forecasts, using the Bay Area Travel Survey (BATS) and MOBIDRIVE in Karlsruhe, Germany

Table 1 (continued)

Author(s)	Date	Purpose	Application/Findings
Hanft et al.	2016	Development of ridership model that generated 100% of O/D data citywide, using AVL and AFC data	Application for New York City Transit and assisted in service planning. Case studies included neighborhood-level ridership and performance analysis for low-cost re-routes and stop changes and an optimal route split location analysis for poorly performing line
Wang and Woo	2017	Measurement of transit ridership as ratio of transit users that commute to work by transit	Independent variables included socio-economic variables (e.g., race, marital status, income, and employment), physical characteristics (e.g., renter-occupied housing, density, land use, and the distant to be Central Business District (CBD)). Transportation variables included commute mode, travel time, car ownership, and bus stop locations
Wei et al.	2017	Development of a method for evaluating the overall performance of transit services, using a combination of data envelopment analysis (DEA), GIS, and multi-objective spatial optimization techniques	Utah Transit Authority (UTA) applied the methodology in Wasatch Front, Utah
Boisjoly et al.	2018	Assemble of fourteen years of public transit ridership data in 25 North American cities, using a longitudinal multilevel mixed-effect regression model	Study found that vehicle revenue kilometers (VRK) and car ownership (proportion of the population) were important determinants of transit ridership
Liu et al.	2018	Determination of accessibility to transit, using GTFS and CTPP for employment data	Study used First Mile Last Mile (FMLM) compared to Public Transit Accessibility (PTA) measures for the Utah Transit Authority in Salt Lake City, Utah. Measures included Weighted Average Travel Time (WATT), Need for Public Transit Service (NPTS), Public Transit Accessibility Gap (PTAG), Average to Medial WATT ratio (AMWR), and Need for Public Transit Improvement (NPTI)
Ma et al.	2018	Development of a geographically and temporally weighted regression model to predict transit ridership	Independent variables included residential building, places of employment, commercial establishment, service facility, attraction, bus stop, metro station, road, and external station. Coefficients varied by time of day

Table 2 Transit planning tools using emerging data sources and modern approaches

Author(s)	Date	Methodology
Sun et al.	2011	Development of a service-oriented architecture for transit planning using path-finding algorithms. Online geospatial services were able to maintain core functions of itinerary searches unique to individual transit agencies, using Waukesha Metro Transit data, in Wisconsin, comparing the outputs to the existing South-East Wisconsin Transit Trip Planner and route scheduling
Antrim and Barbeau	2013	Transit Boardings Estimation and Simulation Tool (TBEST), funded by the Florida Department of Transportation. Conduct short-term transit ridership forecasting, market analysis and network accessibility analysis in an ArcGIS environment for short-term transit ridership forecasting, market analysis, and network accessibility analysis
Liebig et al.	2014	Exploration of predictive trip planning. Application used smart routing in Smart Cities. The route planning architecture used OpenTripPlanner interface and real-time processing of data from traffic sensors to generate traffic flows, applied in Dublin, Ireland
Owen and Levinson	2017	Development of integrated software framework to facilitate the evaluation of accessibility of public transit. Software included OpenStreetMaps, pedestrian links and residential locations, OpenTripPlanner, PostgreSQL, and PostGIS, to analyze transit travel time and continuous accessibility
RSG	2015	Simplified Trips-on-Project Software (STOPS) uses approach similar to traditional 4-Step Travel Demand model, but replaces trip tables with CTTP origin and destination data
Girand et al.	2016	Development of interface to map load profiles of routes. Web interface visualized maps and various analytics based on route summaries
Hanft et al.	2016	Development of ridership model that generated 100% of O/D data citywide. Application for New York City Transit and assisted in service planning. Case studies included neighborhood-level ridership and performance analysis for low-cost re-routes and stop changes and an optimal route split location analysis for poorly performing line
Conway et al.	2017	Development of an open source tool to provide cumulative opportunities accessibility indicators (number of jobs within 45 min of a location) using optimizations and parallelization with routing algorithms. Uses Monte Carlo methods to develop scenarios in an open source environment with visualizations. Tool uses GTFS, OpenStreetMap (OSM), the TAPTOR algorithm for transit routing, and employs a Monte Carlo procedure
Karner	2018	Development of a transit equity analyses and applied it to the Phoenix, AZ metropolitan region, using publicly available data rather than the data sources employed by the FTA methodology

Table 2 (continued)

Author(s)	Date	Methodology
Pi et al.	2018	Development of a transit data analytics platform that uses APC, AVL, and GTFS. The platform included the Django web framework and an Nginx HTTP server. Both components are open source, and capable of handling featured aspects of service quality (e.g., wait time, stop-skipping frequency, bus bunching occurrences, bus travel time, on-time performance, and bus occupancy levels). As a web application, it includes visualizations of spatial data
Swayne and Miller	2018	Development of access measures on travel time for transit riders, particularly young, entry-level, low-income workers, using the proprietary tool, Remix, to map existing transit networks. Research team modified the network and stop locations, customized the scheduled service, calculated travel times, and created isochrones to simulated young passenger traveling in top five locations to jobs, using ACS and employment data (e.g., LEHD, LODES, LODES Workplace Area Characteristics (WAC) data) within 60 min

vehicles (e.g., AirSage). Recent advances on this methodology include CityCast (see <https://transportfoundry.com/blog/2017/5/26/introducing-citycast>), a web-based software that includes a transit component. The data sources include: the 2010 Decennial Census; the 2012–2016—5-Year ACS Public Use Microdata Sample (ACS PUMS); the 2015 Longitudinal Employer-Household Dynamics, Origin–Destination Employment Statistics, Workplace Area Characteristics (LEHD, LODES, WAC); the 2009 National Household Travel Survey (NHTS); Open Street Maps (OSM); and local GTFS. The tool allows users to look at the various data sources along a selected link. Techniques for blending various types of data provide new ways to increase planning efficiency and effectiveness.

Gaining advantages from blending data within a transit data ecosystem requires considerations for the legacy systems in place, the ability to ingest newer forms of data, and the willingness of agency leadership to leverage these resources within the agency itself. For example, a number of transit agency now generate GTFS to facilitate the development of mobile applications to serve potential transit riders with accurate scheduling and routing information. At the same time, these agencies lack the ability to utilize GTFS for their own planning purposes after having invested in proprietary software packages for planning. This research examines opportunities for transit agencies to take advantage of blending traditional and emerging data for transit planning purposes. In particular, it describes the development of a low-cost, open-source approach to estimate transit demand, using modern processing methodologies to analyze, visualize, and forecast bus ridership in a web-based format.

Data ecosystem elements and tools

In 2012, New Jersey Department of Transportation (NJDOT), together with New Jersey Transit (NJTransit), sought assistance in leveraging the American Community Survey (ACS) 5-year datasets, to identify relationships between ridership and various sociodemographic factors in order to assist in predicting bus ridership and service needs. The data ecosystem available included ACS, CTPP; GTFS; and farebox data (at the zone level). NJTransit also had recent on-board transit surveys available for this research. The functionality required included the ability to view Census variables of interest for transit planning at the tract level and the ability to add and subtract potential Census tracts for inclusion in customizable market areas. Additionally, the analysis needed to provide route-specific travel characteristic, variations by time of day passenger travel, and visualizations of bus networks for small and medium city bus systems.

Application Programming Interfaces (APIs) for socio-demographic data The Census has been a primary data resource for transportation planning (Lawson 2018a). The decision to change the data collection program to a continuous, monthly survey (e.g., Census long form to ACS) triggered the need for new data practices. The ACS provides timely demographic, housing, social, and economic data, updated every year, across states, communities, and population groups (U.S. Census Bureau 2018). At the same, this continuous data generation burdens transportation planning staff with a constant need to download and manually process in-coming Census data files. Recently, the Census Bureau adopted a modernization strategy for data dissemination: using an Application Programming Interface (API) (see <https://www.census.gov/data/developers/data-sets.html>).

An API makes it possible for a single data source to serve many users using software code over the internet to “call” variables, seamlessly, using a key (a unique string of

alphanumeric characters transmitted used to authenticate the source of a data request). Big Data providers (e.g., Google) use APIs for fast, efficient data delivery. Modern processing leverages APIs in a web environment, opening new avenues for transportation planning. While APIs are routinely used with Big Data, but rarely used with traditional data. Promoting the use of APIs facilitates efforts to blend different data types. Web-based, interactive tools that use APIs, facilitate the creation of web choropleth maps, bar graphs, and tables, by interrogating Census information for specific geographies.

The CTPP is “a set of special tabulations designed by transportation planners using large sample surveys conducted by the Census Bureau” (Census Transportation Planning Products 2015). The CTPP data provides tables of Origin–Destination (O–D) capable of identifying bus riders. CTPP tabulations include three geographies: residence-based tabulations summarizing worker and household characteristics; workplace-based tabulations summarizing worker characteristics; and worker flows between home and work, including travel mode. There is currently no API for the CTPP, requiring the construction of a CTPP API for this research. While the Longitudinal Employer-Household Dynamics (LEHD) also includes home origins and work destinations, it lacks any information on the mode used.

Spatial data Key aspects of transit planning require spatial representations (e.g., route planning, bus stop locations). Smith (2000) pointed out the use of Geographic Information Systems (GIS) on the internet would benefit transit planning. General Transit Feed Specifications (GTFS) has gained popularity as an aid for individuals who want to plan transit trips using their mobile device (e.g., smartphone apps). However, it remains an underused resource within transit agencies with respect to enhancing their own transit planning tools.

A number of recent advancements in geographic information science (e.g., modern processing techniques developed for Netflix and Facebook using open source code) provide web-based platforms with the capabilities to meet the special needs of transit planning (see Lawson et al. 2019). Modern processing using leaflet (<http://leafletjs.com/>) and D3.js (<http://d3js.org/>), both open source software, facilitate the creation of interactive maps organized by Census tract geographies. To accommodate the spatial component of transportation planning, this research combines GIS mapping strategies and data visualizations, using GTFS routes as “backbones” to define market areas. Open source GeoJSON files, rather than proprietary GIS software, allow for easy implementation of specific geographies, based on Census tracts adjacent to GTFS routes. The web-tool automatically appends Census tracts containing bus stops on particular GTFS routes, when market areas add new GTFS routes. Pointing and clicking on a Census tract on a computer screen adds it to a market area. The GTFS routes that define the market area are also included on the maps for reference, or as filters for some of the various data visualizations.

Farebox data In transit systems where agencies have invested in fare collection equipment, as each passenger enters a bus, the specific vendor software interface records the data in real-time. Aggregating the data provides financial information for a variety of needs (e.g., revenue by routes, network totals). However, if the original per passenger information is not processed, or retained, only the aggregate information remains. In addition, when the system only requires “tap-in” be recorded (but does not record a “tap-out”), the data retained only contains stop-specific origins, but no destination information. If transit agencies have fare zones, estimated destinations are derivable based on the fare paid. The farebox data is incorporated in the tool suite to allow users to see the output of the model runs in comparison of the farebox data.

Bus ridership estimation using modern processing

In order to estimate bus ridership, traditionally, planners rely on local travel surveys, on-board transit surveys, and traditional Census data. This research uses an API, developed for the CTPP data, to generate O–D tables (Lawson 2016b). The CTPP trip tables are modified using regression equations developed from ACS data. Then, a routing engine using scheduling constraints, defined in available GTFS data, microsimulates bus ridership for specific NJTransit market areas. The microsimulations are validated using farebox data. This approach generates numerous trip tables, calibrated using various demographic variables, to identify changes in ridership in response to different transit planning scenarios (see Fig. 1).

The API CTPP tool extracts origin (home) and destination (work) information for bus riders directly from CTPP tabulations by Census tract. Census data only provides information on the morning commute, based on the ACS questionnaire. In order to model PM peak ridership, departure times from the work location, rely upon a basic assumption that a return trip back home is expected 8 h after the AM trip (e.g., the 8-h workday). Any commute trips after the morning peak are captured in a full day time period, also with the expectation the return trip home will occur 8 h from the time of departure. Using an 8-h workday assumption, transit trip commute tables are constructed using home origins and work destinations from the CTPP.

CTPP bus ridership reflects responses to the transit network that was available at the time of the Census data collected (e.g., 2006–2010 ACS 5-year estimates). However, to forecast potential ridership for current routes, new routes, or route adjustments, it is necessary to take into account the underlying factors (e.g., socio-demographic variables) that drive transit demand (e.g., zero-car households). The ACS API and GeoJSON Census tract geography files generate Census tracts, transportation-related variables, and household characteristics, for each tract using an open source, web-based platform. For example, as illustrated in Fig. 2, in Atlantic City, New Jersey, Census Tract 34,001,010,600 has 6.25% zero-car households (127 households). Colors differentiate current transit routes, with bus stops illustrated on the routes as circles, based on information available in the GTFS files. Transit planners can add or subtract tracts, based on particular goals, to assemble unique market area for analysis.

ACS regressions The first step in the prediction of bus riders is the examination statistically significant correlations in the ACS 5-year data with the Bus to Work (*bus_to_wor*) variable. This step requires a correlation matrix, generated using a statistical software package (e.g., SPSS). Regression models use these variables, based on the assumption of a linear relationship between the dependent variable (*bus_to_wor*) and the set of independent variables. The regression models are run in SPSS, or Geoda (an open source spatial

Fig. 1 Flow of the estimation process, beginning with the generation of estimated trips from the CTPP trip tables, modified by the ACS, converted into individual bus trips in the OTP microsimulation, and finally, validated using the farebox data

Fig. 2 Percentages and counts of zero-car households for Tract 34,001,010,600

statistics tool available at <https://geodacenter.github.io/>). A regression model fits a straight line to a set of observed data and provides the statistical significance of the included variables.

$$Y = a + bX_1 + bX_2 + bX_3 \dots$$

The regression model produces a number of parameters and model fitting indicators, such as the coefficient of determination (R squared). The R squared is defined as the percent of the variation of the dependent variable (bus_to_wor) explained by a set of independent variables. Therefore, the higher the R squared, the more explanatory power the regression model provides.

The regression model output also provides a constant (intercept) which is the average value of the dependent variable when the independent variables equal zero. The slope coefficients indicate the average change in the dependent variable with a one-unit change in the independent variable. For the purposes of this modelling effort, statistical significance is defined as a *p* value of < .05 or a *t*-value > 2.5.

The number of bus riders predicted by the regression is divided by the actual ACS ridership count extracted from each Census tract, to produce an ACS Regression Ratio. The result is the ratio of predicted riders compared to ACS count of riders.

$$\text{RegressionModelRiders} / \text{ACSRiders} = \text{ACSRegressionRatio}$$

Next, bus commute trip in the CTPP, is multiplied by the ACS Regression Ratio, to improve accuracy of the calculated bus ridership numbers for the trip table.

$$\text{TripTableInput} = \text{CTPP} * \text{ACSRegressionRatio}$$

OTP routing microsimulation To model bus passenger behavior, this research uses an approximation of how bus riders behave. For example, when individuals want to know what bus lines are available for a particular trip, they can access stop, scheduling, and routing information using a mobile app on a smartphone, or at an information kiosk. These information resources use algorithms to provide potential transit riders guidance for planning their trip. OpenTripPlanner (OTP), an open-source routing engine, with a core server-side Java component capable of generating itineraries for travelers across modes (e.g., combining transit, pedestrian, bicycle, auto). OTP uses OpenStreetMap (OSM) and GTFS data and exists as a service accessed through an API or by using JavaScript client libraries (OpenTripPlanner, n.d.). OTP uses the pedestrian information to “walk” the synthetic bus rider to the bus stop. (Additional information on the OTP routing engine available at <https://github.com/opentripplanner/OpenTripPlanner/tree/master/src/main/java/org/opentripplanner/routing/algorithm>).

GTFS data for a particular market area (e.g., geographic area with specific Census tracts designated by local transit planners) is loaded into a route planning API that uses OTP. The process generates a request, using each row in the trip table, generated from the CTPP data, and calibrated with ACS Regression Ratio. Each row in the origin–destination (O–D) table is treated as a “synthetic bus rider.” Each synthetic bus rider is algorithmically plotted throughout the market area Census tracts, placed spatially in close proximity to bus stops in the GTFS data (using a one mile radius to ensure the ability to capture at least one stop location). The synthetic bus riders are then taken on their synthetic bus trip in the form of a microsimulated trip, using OTP as a routing engine. In essence, the synthetic bus riders “take a trip” based on the GTFS schedule, as if they are really riding a designated bus, using their smartphones or a kiosk, to navigate their way to work on the bus. OTP returns the three fastest travel-time routes from the origin point (bus stop) to the destination point (bus stop) by departure time. The API randomly chooses one of these three possible (plausible) routes. As part of the processing, the API returns boarding and alighting times. The times are binned into hours for validation purposes. The original departure times, provided in the ACS data in minutes, are also binned to match the binned data in the CTPP data. Departure times are randomly assigned to the synthetic bus riders from these bins. Each trip in the trip tables is placed into its corresponding hour time-bin, and run through the microsimulation. All the details about each trip generated during the process are saved as “legs and trips” data. The process generates an entire population of synthetic bus riders for each market area.

Modeling process The modeling process contains a number of options (e.g., time ranges AM Peak (6:00 AM to 10 AM); PM Peak (3:00 PM to 7:00 PM); and Full Day [see Fig. 3]). Either the model type interface allows the user to either use the CTPP for origins and destinations directly, or market area regression coefficients generated as described above. The model uses origins and destinations either from the bus stops in the GTFS, or

Fig. 3 Interface for setting modeling parameters after generating the trip table

locations extracted from the on-board surveys. Finally, the model can use both the current population and employment from the ACS, or the local forecasts from a regional provider (e.g., the Metropolitan Planning Organization (MPO)). The choice of parameters depends on the type of analysis undertaken.

Validation with Farebox data The farebox data is processed by fare zone and compared to the trip destinations predicted during the modeling process. The tools allow the user to filter the farebox data by route, by time of day, and by the three time period aggregates (AM Peak, PM Peak and Full Day).

In summary, the processing of the entire market area uses a trip table of Census tract to Census tract flows, given an origin and destination, running through the OTP routing engine. The microsimulation process aggregates each trip leg assigned to a bus route into market area output, calculating route-level ridership by time of day in a web-based dashboard. Open source code for the transit demand modeling tool is available at <https://github.com/availabs/transitModeler>. Researchers and practitioners are welcome to make modifications and advancements based on the open source code and use the code with their own databases.

Case studies

Below are three examples that demonstrate uses of the tools for day-to-day planning. The first example focuses on what will happen to ridership patterns, using base year ridership, if there is a projected 10% reduction in population in a particular Census tract in the Atlantic City, New Jersey market area. The second is a model run for the Atlantic City market area, using the farebox data to validate individual routes and overall total ridership. The third examines the impacts on the Princeton/Trenton market area, and routes individually, with and without a new route.

Atlantic City: projected population reduction Atlantic City, a small city on the southeastern New Jersey coastline with a population of approximately 40,000 people. The transit market area, however, serves a population of more than 700,000 and a labor force of nearly 370,000. Approximately 4% of the labor force use the bus to commute to work. NJTransit

operates twenty-one bus routes in Atlantic City. The variables for the Atlantic City analysis include bus to work, households with zero vehicles available, employment in the arts sector, and employment density (a special tabulation created by dividing total employment in the Census tract, by the total area). For the 110 Census tracts in the market area, 60.8% of the dependent variable, *bus_to_wor* is explained by the independent variables, *car_0*, *arts*, and *emp_den* (based on the R squared). All of the independent variable coefficients are statistically significant, using a .05 threshold.

Table 3 displays the values for Census tract 34,001,012,200. The Atlantic City Regression Model parameters is as follows:

$$\text{bus_to_wor} = -41.505 + (0.230 \times (\text{car_0_hous})) + (0.163 \times (\text{arts})) + (0.019 \times (\text{emp_den}))$$

Applying the values from the ACS data produces the following:

$$\text{bus}_{\text{to_wor}} = -41.505 + (0.230 \times (196)) + (0.163 \times (991)) + (0.019 \times (2251))$$

The number of riders in Census tract 34,001,012,200 predicted by the Atlantic City Regression is 208. The Regression Ratio of predicted riders to ACS riders is .54, and is applied to the CTPP data.

$$208/388 = 0.54$$

The resulting trip table depicted in Table 4 displays the number of bus trips from the origin point (Census tract 3,400,101,220) to each corresponding work Census tract.

What would be the expected impacts on bus ridership for tracts where jobs are located if Census tract 34,001,012,200 experiences a 10% reduction in population in the next year? Table 5 displays the ridership impacts for each of the Census tracts expected to receive bus commuters.

Atlantic City: market area and route-specific validation This example illustrates the use of farebox data to validate overall market area bus ridership, and route-specific ridership. Table 6 displays a model run using an AM peak ridership estimation and farebox data for the twelve routes in the Atlantic City market area. There is only a 3.26% difference between the model output and the farebox data for the overall market area total ridership. However, using a Mean Absolute Percentage Error (MAPE), which uses the absolute value of the percentage differences between the forecast and the farebox, divided by the number of cases, indicates nearly a 70% error due to the variation across the routes. The route-specific estimates either over or under estimate ridership, compared to the farebox data. For example, routes 505 and 508 over-estimate ridership compared to the farebox data. This is

Table 3 Equation variables and census tract 34,001,012,200 Data

Equation variable	ACS description	ACS category	Value in census tract 34,001,012,200
<i>bus_to_wor</i>	Journey to works by public transportation by bus or trolley bus	Journey to work	388
<i>car_0</i>	Households, zero vehicles available	Household	196
<i>arts</i>	Employment in the arts sector	Labor force	991
<i>emp_den</i>	Employment/area	Total employment/total area	2251

Table 4 Bus riders from home tract 34,001,012,200

Work census tract	Riders	Trip table output = CTPP* regression ratio (0.54)
34,001,002,400	160	86
34,001,002,300	60	32
34,001,001,400	60	32
34,001,011,900	25	14
34,001,000,400	25	14
34,001,001,100	20	11
34,001,013,201	15	8
34,001,013,302	10	5
34,001,011,702	4	2
Total	379	205

Table 5 Ridership forecast from home Tract 34,001,012,200

Work census tract	Riders	Trip table output = CTPP* regression ratio (0.54)	Forecasted ridership (– 10%)
34,001,002,400	160	86	78
34,001,002,300	60	32	29
34,001,001,400	60	32	29
34,001,011,900	25	14	12
34,001,000,400	25	14	12
34,001,001,100	20	11	10
34,001,013,201	15	8	7
34,001,013,302	10	5	5
34,001,011,702	4	2	2
Total	379	205	184

not surprising as local Jitneys compete for riders on these two routes, suggesting the current methodology is most appropriate for locations with no competing modes.

Another complication with using farebox data to validate bus ridership estimates is the lack of non-work trips in the calculation of riders. A proportional relationship between work and non-work bus trips, developed from on-board surveys could account for those trips in the farebox counts. Another source is the NHTS that includes all trip types by mode. It is likely that non-work transit trips occur outside of the morning and evening peaks, making the full day comparisons more difficult due to non-work trips than the peak periods. Routes 551, 552 and 559 farebox data indicate many more riders than are predicted using the work commute simulation. Future research needs to address cross-town trips (not originating from a home location) and improvements in the allocation process where routes compete for the same bus commuters.

Princeton/Trenton route impacts analysis The Princeton/Trenton market area has approximately 103,000 households and includes the Princeton University campus. NJTransit introduced new route, 655, in the Princeton/Trenton market area, to address a perceived

Table 6 Estimated AM ridership and farebox data for Atlantic City, New Jersey

Route Number	Run 121	Run Farebox	Percent difference (%)
505	2035	1431	29.68
508	1151	530	53.95
502	797	719	9.79
553	618	716	− 15.86
507	509	770	− 51.28
509	479	499	− 4.18
554	464	536	− 15.52
501	382	305	20.16
552	337	601	− 78.34
504	304	185	39.14
559	155	326	− 110.32
551	91	465	− 410.99
Total	7322	7083	3.26

need, but later removed the route due to low patronage. The route impacts analysis uses this real world example to demonstrate how running models with and without a particular route can help explain how bus riders would travel under both conditions.

The regression model for Princeton/Trenton market area is as follows:

$$bus_to_wor = (0.199 \times (car_0_hous)) + (0.24 \times (age25_29))$$

The R squared for this regression specification is 62.3%, indicating that roughly 62% of bus ridership can be explained by zero-car households and individuals in the 25–29 year old age range, with 69 cases. The regression model specifications are sensitive to the particular Census tracts aggregated for each market area, and thus, no single equation applies across all jurisdictions. In the case of Princeton/Trenton, the absence of a vehicle, and being in the 25–29 age group, were the only statistically significant independent variables.

This analysis requires running two different models for the Princeton/Trenton market area. The two models runs (with and without Route 655) are compared to farebox data. Run 119 includes Route 655; Run 120 excludes Route 655. The GTFS tools make it easy to add a new route and modify an existing route. Options available include: the first departure time; the last departure; headway; idle time; runtime; route distance; and number of buses on the route (see Fig. 4).

As indicated in Table 7, Run 119 estimates 80 AM peak riders on Route 655, while the farebox data shows an average of 47 riders. Run 119, therefore, overestimates AM Peak ridership on Route 655 by 33 riders.

When Route 655 is removed (Run 120), 32 of the 80 riders estimated in Run 119 were unable to be routed. These synthetic bus riders, accounted for in the trip table, could not find service in the microsimulation. This possibly indicates the existence of latent demand served by route 655, but unserved by the transit network without Route 655. The remaining 48 riders found their way onto the existing service network.

The modeling process produces visualizations depicting estimated boarding and alightings using the CTPP trip tables developed at the Census tract level as origins and destinations. Figure 5. displays a visualization of the stop-level boardings for Run 120. This feature can also be toggled to display the alightings.

Fig. 4 GTFS tools for route creation and modifications

Table 7 Princeton/Trenton estimated AM peak ridership, farebox data, with/without Route 655

Route number	Run 119	Run 120	Run Farebox	Percent change Run 119/Fare- box (%)	Percent change Run 120/Fare- box (%)	Percent change Run 119/120 (%)
606	1259	1294	634	49.64	51.00	-2.78
600	397	409	329	17.13	19.56	-3.02
609	325	329	723	-122.46	-119.76	-1.23
603	316	313	384	-21.52	-22.68	0.95
613	244	221	413	-69.26	-86.88	9.43
605	195	205	93	52.31	54.63	-5.13
619	119	131	209	-75.63	-59.54	-10.08
655	80	0	47	41.25	N/A	N/A
612	21	22	41	-95.24	-86.36	-4.76
Total	2956	2924	2873	2.81	1.74	1.08

Run 119 overestimates AM Peak ridership on Route 655 by almost exactly the same amount as the number of total network riders missing from Run 120, when Route 655 is removed. This example of the route 655 demonstrates that this model shows promise in estimating latent demand; that it is capable of locating potential riders in a market area unserved by the transit network. The 80 riders on Route 655, as estimated by Run 119, are a collection of both latent demand ridership (by 32 riders) and ridership that is served by the transit network (by 48 riders).

In summary, Run 119 illustrated that 48 riders were either randomly placed close enough to route 655 to find their way onto Route 655 through microsimulation, or they are located in the Route 655 commute-shed, but did not appear in the farebox data as

Fig. 5 Visualization of boardings from Model Run 120

“actual” 655 riders due to previously formed habits of commuting. Again, although there are differences on a route to farebox analysis, the overall differences for the market area are small.

Discussion and future research

While it is possible for transit researchers to incorporate archived ITS transit data in individual analyses, transportation planners have found many challenges trying to take advantage of emerging data sources. Sun et al. (2011) note that the majority of transit trip planners are proprietary vendor systems, making it difficult to take advantage of advancements in geospatial information and web technologies. Open source software, in contrast, has source code that is available for modification, or enhancement, by anyone. This openness provides opportunities for additional progress towards more cost-effective and efficient approaches, while providing feedback on these features and improvements to the original open source software creators. Open source allows planning agencies to make updates to the software either in house or through a third party and to receive the benefits of all future updates as they are made by other agencies.

RSG (2015) points out the extensive data tasks required to run the STOPS program (including GIS skills). The NJTransit tool uses APIs that automatically feed the data

into a web-interface. In addition, while some academic researchers continue to look for more exotic applications for transit planning (Zhang et al. 2018; Wu and Cao 2018), simply applying a modern processing approach (e.g., use of APIs) with blended data for bus ridership forecasting, promises benefits in the near term, as well as longer-term. At the same time, abandoning traditional datasets (losing the critical socio-demographic variables necessary for understanding travel behavior) is a risk associated with using only Big Data sources. By deploying options for blending the traditional datasets, using modern processing techniques, makes it possible to integrate numerous types of data, providing the best of both worlds. The NJTransit project demonstrates the use of blended data for transit planners.

While modern processing has accelerated a number of industries (e.g., entertainment venues such as Netflix), transit has been slow to transform their data ecosystem to reap the benefits of the tools and techniques available. Potential barriers to transformation include institutional barriers within organizations and lack of understanding of benefits by decision-makers. An initial question is how to introduce a new approach. Existing staff members are not likely to have, or be able to gain the requisite computing skills to build a program from scratch. In addition, trying to hire talent with these skills means competing with private industry capable of offering much larger compensation packages. Strategies to reduce these barriers could include leadership at the federal level to offer guidance in how best to find the right type of computing services (e.g., consultants, university programs, internship programs), with an emphasis on open source to share benefits from efforts easily across the transit industry. State Departments of Transportation could also offer support and guidance, including providing direct assistance to interested transit agencies within their state, forming a technical team to address issues as a consortium. University Research Centers are also able to provide research support, however, depending on the terms of their research administration, may or may not be able to provide continued support after the initial research is completed. Consulting firms interested in promoting new uses of platforms and leverage advancements into a larger customer base, are also an option.

Transit agencies need to address hosting options (e.g., in-house, commercial services, university programs) and different levels of technical support, ranging from once or twice a year maintenance visits to aggressive program development to address particular needs (e.g., new functionality that includes bike-share and scooter data for multi-modal accessibility). Web interfaces permit different forms of access, making it possible to have a public-facing site with limited functionality, or access with a password to advanced analytics for transit planning teams. New forms of training for using platform software has advanced rapidly, including embedded video for instruction to click-based learning where the software “teaches” users throughout the entire site, requiring no previous knowledge by users.

There are cost-savings gained through implementing APIs including auto-loading of a variety of data types, and instantaneously conducting analysis from simple queries to advanced machine-learning algorithms. The agile nature of platforms provides benefits across a transit agency as the web interface can be shared with different departments within the agency (e.g., marketing) and with decision-makers. It is also possible to share analyses with outside agencies using a platform approach. For example, transit agencies can share strategies with MPOs and state DOTs for a larger, regional perspective. More forward thinking opportunities could include land use planners as they evaluate the impacts of new commercial or residential developments. Other stakeholders who rely on bus services, including emergency response, evacuation strategies, medical institutions, special generators (e.g., universities, stadiums) could participate in transit planning through specialized designed screens, available as a web-app with options for

running scenarios for particular needs. Opportunities could even interface with customers and log their responses to service changes.

Trip types Given the original focus of this research was to forecast bus commuters using ACS and CTPP data for socio-demographic variable, the current tool lacks the capacity to directly forecast transit trips for other purposes. This complicates validating model outputs with farebox data where non-work trips are the predominant trip type (e.g., mid-day trips). As a result, market area models may underestimate full day ridership, despite often overestimating peak-time ridership. To account for the full range of bus riders, an enhanced methodology needs to include other trip purposes (e.g., shopping, medical). On-board surveys collect all trip purposes useful for inclusion in the modeling process (e.g., factoring a proportion of different types of trips based on ACS characteristics). Future data processing could forecast non-work trips using regression models that create synthetic non-work travelers modified with point-based trip destinations (e.g., landmarks). The NHTS state-level add-on data contain geocoded origins and destinations by trip purpose by mode, and may be a future source for trip types for buses (Lawson 2018b).

Trips in the peak Due to assumptions made in trip table generation regarding an 8-h workday, and the lack of information about work-to-home trips, the microsimulation algorithm shows overly concentrated peaks, compared to farebox data, as well as a PM Peak that generally begins later than farebox data (based on actual passenger loads). The AM and PM Peak settings are currently hard-wired into the demand modeling and analysis tools. Future research could explore alternative data sources (e.g., smart phone apps records associated with transit travel to establish variations in hours of work in log data) to better tie work-to-home department times to farebox collection. Another approach would be to explore hours-of-work details found in public data sources and generating modifications for bus riders from particular industries, based on work locations. For example, the 2017–2018 American Time Use Survey (ATUS) provides information on the percent of workers with a non-day-time schedule by shift and by occupation type (Bureau of Labor 2019).

Census tract geographies Transportation planners often use Transportation Analysis Zones (TAZs) for trip origins and destinations, rather than Census tracts. TAZs are generally smaller geographies and useful for transportation planning purposes. The Census Bureau recently decided to discontinue the formal generation of TAZs for the CTPP (see Lawson (2018a) for further discussion on the issues surrounding TAZs). Going forward, local transportation planners will establish their own TAZs (a number of transportation modelers already have their own unique TAZs). Using Census tracts provides the most generalizable geography at this time and is preferred for generalizable tool suites.

Trip origin geographies The microsimulation algorithm currently distributes synthetic riders randomly throughout each home and work Census tract, using a one mile radius around the GTFS-designated bus stop, to increase the likelihood synthetic riders will find a bus in the OTP processing (which includes pedestrian links). Traditionally, transportation planners have used a smaller radius (e.g., $\frac{1}{4}$ mile or $\frac{1}{2}$ mile) to predict ridership. While the number of bus riders per Census tract would remain the same, having an improved approach to assigning riders to particular bus stops would improve route-specific counts. There are a number of approaches that could be explored for improving bus stop allocations including: using the MicroSoft Building Footprint data (see <https://github.com/microsoft/USBuildingFootprints>), or OSM building footprints (see <https://osmbuildings.org/>), to explicitly identify residential structures within a Census tract. Other approaches to consider include predicting trips with population distributions using parcel data polygons; point-based establishment and employment data; or using smartphone Location-Based Services (LBS) data.

Latent demand The current version of the research tool uses socio-demographic data without the addition of other important factors that influence the decision to ride the bus to work. Future research needs to determine whether different probabilities for individuals in households previously unserved by bus services, to account for the unobservable preferences, or circumstances that still influence bus ridership. In addition, bus service quality and quantity should be included as independent variables, or modeled in the form of simultaneous equations. While many new data types (e.g., GPS traces from smartphones of bus riders) are becoming available, they, unfortunately, lack socio-economic information. Using APIs to blend various data types could improve the predictive capacity of models with new routes, or route modifications.

Disclosure concerns In order to be granted permission from the Census Bureau to use the raw ACS data to develop the CTPP, disclosure concerns are treated with a method referred to as perturbation. This method uses a technique that adds random data when the data is processed. For example, some origins and destinations are randomized from the original raw data. As a result, there is some error purposely embedded in the CTPP data.

Route overlap In dense urban areas with two Census tracts in downtown and a number of buses going between the two tracts, the microsimulation may not be able to distribute the trips as accurately as when there are fewer choices. This issue would arise while attempting to forecast cross-town ridership using a residentially generated AM bus to work trips. Service levels are included in the microsimulation-modeling algorithm. While the overall market area is accurate in the peaks (e.g., 3.26% difference in total for Atlantic City run), there are a number of trips captured in farebox on a specific route, were assigned to a different route during the microsimulation phase. The algorithm is not currently capable of differentiating between two routes competing for the same riders where routes have overlapping Census tracts in common. One approach would be to use a three-stage-least-squares estimation method such as the one developed by Peng (1994) for competing routes.

Scalability The transit demand-modeling tool developed in this research is designed to analyze bus-to-work ridership in small and mid-sized market areas. The tools are not calibrated for more complex transit environments. Future research could test the possibility of modeling bus rider in neighborhoods within larger, urban areas, where trips outside of the neighborhood would be assigned to areas external to the immediate market area, but still within the urban area. These neighborhood tools would need to be calibrated to the larger area, regional, multi-modal models.

Combining transit assignment and latent demand The web-based tool suite was designed to contribute to both assignment (using the OTP microsimulation process) and demand (identification of underlying socio-demographic factors using regression models). The regression models provide coefficients for the statistically significant ACS variables within each market area (e.g., zero vehicle householders taking the bus, 25–29 years of age for Princeton/Trenton). When these coefficients are applied to neighborhoods currently without transit service (but with similar socio-demographic characteristics), this assumption suggests that households with the combination of characteristics would be likely users of the new service, and thus could be used to better understand potential demand. Future tests of this assumption would require the use of back-casting (e.g., creating output from the modeling process for potential routes and then comparing these outputs to behaviors over time on the new routes).

Regression modeling options The regression analysis, run outside of the platform, for the individual market areas, demonstrated a high sensitivity to the Census tract level socio-demographic variables. Over time, it may be necessary to update the regression models (e.g., expansion of employment centers, substantial residential development).

This suggests the need to incorporate the capability to produce the regression, using an open source code within the platform itself (e.g., incorporating open source software such as “R” routines, or developing an open source regression modeling procedure in the tool itself).

Stop-level farebox data The most promising future research should address the use of farebox data at the stop level and the landmarks near the stop to clarify trip purpose. This could reduce the need for traditional on-board surveying to collect origins and destinations, while providing a monitoring and validating data strategy going forward. This improvement would also inform the allocation process to better route trips within a Census tract.

Conclusions

The transformation of transportation planning is already underway with new types of data (e.g., Big Data sources). At the same time, some of the critical variables (e.g., socio-demographic information), are only available in traditional datasets (e.g., Census data). Recent data dissemination strategies (e.g., APIs) being deployed by the Census Bureau will require a “retooling” of the transportation planning industry to take full advantage of the ease and speed these modern processing tools. This research demonstrates a blended approach for bus ridership forecasting that uses both traditional and emerging data through the use of an open-source, web-based platform. The key component to facilitating this strategy is the use of APIs. Moving to an API-centric approach, now common in other applied data science uses (e.g., Netflix and Facebook), could provide transportation planners with a seamless method for future improvements in analysis, visualization, and forecasting. This research demonstrates its usefulness in a bus ridership forecasting application. The Census Bureau is expanding their contributions to data dissemination with APIs. Transportation researchers and planners will benefit most from these investments by increasing their understanding and use of these new applied data science tools.

There is an urgency to move to more agile and easy to use methodologies as bus systems are experiencing more competition for riders (e.g., ride sharing). Modern processing tools and techniques ingest many new sources of data, compared to labor-intensive GIS and manual data input approaches. Overcoming obstacles that discourage transit agencies from considering modern processing begins with an analysis of the data ecosystem currently in place, and determining what next steps would assist in facilitating the integration of data sources internal and external to the agency while maximizing opportunities to provide better service, and to respond more rapidly to an ever-increasing multi-modal environment.

Acknowledgements This research was supported by New Jersey Department of Transportation (NJDOT) (UTRC/RF Grant number 49997-54-24, 75144-05-24), New Jersey Transit (NJTransit), and the Research and Innovative Technology Administration of the U.S. Department of Transportation through the Region 2 University Transportation Research Centers Program. Special thanks to David Vadney, and Joel Tirado for their contributions to this research.

Author's contribution CTL: Literature Search and Review, Statistical Analysis, Manuscript Writing and Editing. AM: Software Design and Code Development. EK: Manuscript Writing and Editing, Graphic Interfaces.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interests.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Antrim, A., Barbeau, S.J.: The many uses of GTFS data—opening the door to transit and multimodal applications. Location-Aware Information Systems Laboratory at the University of South Florida, 4 (2013)
- Boisjoly, G., Grisé, E., Maguire, M., Veillette, M.P., Deboosere, R., Berrebi, E., El-Geneidy, A.: Invest in the ride: a 14 year longitudinal analysis of the determinants of public transport ridership in 25 North American cities. *Transp. Res. A: Policy Pract.* **116**, 434–445 (2018)
- Brown, J., Thompson, G., Bhattacharya, T., Jaroszynski, M.: Understanding transit ridership demand for the multi-destination, multi-modal transit network in Atlanta, Georgia: Lessons for increasing rail transit choice ridership while maintaining transit-dependent bus. A paper presented at the 92nd transportation Research Board Annual Meetings, January 13–17, 2013, in Washington, DC (2013)
- Bureau of Labor: Job Flexibilities and Work Schedules—2017–2018 Data from the American Time Use Survey (2019). <https://www.bls.gov/news.release/pdf/flex2.pdf>
- Census Transportation Planning Products: (2015) Retrieved September 7, 2016, from http://www.fhwa.dot.gov/planning/census_issues/ctpp/
- Conveyal: Conveyal Analysis Documentation: Release v4.5.0 (2019). <https://buildmedia.readthedocs.org/media/pdf/analysis-ui/latest/analysis-ui.pdf>
- Conway, M.W., Byrd, A., van der Linden, M.: Evidence-based transit and land use sketch planning using interactive accessibility methods on combined schedule and headway-based networks. *Transp. Res. Rec.* **2653**, 45–53 (2017). <https://doi.org/10.3141/2653-0>
- Dill, J., Schollossberg, M., Ma, L., Meyer, C.: Predicting transit ridership at the stop level: the role of service and urban form. A Paper Presented at the 92nd Transportation Research Board Meetings, January 13–17, 2013, in Washington, DC (2013)
- Erhardt, G.D., Dennett, A.: Understanding the role and relevance of the census in a changing transportation data landscape. *Applying Census Data for Transportation*, **96** (2017)
- Frei, C., Mahmassani, H.: Riding more frequently: disaggregate ridership elasticity estimation for a large urban bus transit network. A Paper Presented at the 92nd Transportation Research Board Annual Meetings, January 13–17, 2013, in Washington, DC (2013)
- GTFS Static Overview: (2016) Retrieved September 7, 2016, from <https://developers.google.com/transit/gtfs/>
- Giraud, A., Trépanier, M., Morency, C., Légaré, F.: Data fusion of APC, smart card and GTFS to visualize public transit use (No. CIRRELT-2016-54). CIRRELT, Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport=Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (2016)
- Hanft, J., Iyer, S., Levine, B., Reddy, A.: Transforming bus service planning using integrated electronic data sources at NYC transit. *J. Public Transp.* **19**(2), 6 (2016)
- Iliopoulou, C., Kepaptsoglou, K.: Combining ITS and optimization in public transportation planning: state of the art and future research paths (2019)
- Karner, A.: Assessing public transit service equity using route-level accessibility measures and public data. *J. Transp. Geogr.* **67**, 24–32 (2018)
- Kimpel, T.: Time Point-level Analysis of Transit Service Reliability and Passenger Demand (2001). An unpublished dissertation. [<http://www.pdx.edu/sites/www.pdx.edu.cus/files/SR036.pdf>]
- Kressner, J.D., Macfarlane, G., Huntsinger, L., Donnelly, R.: Using passive data to build an agile tour-based model: a case study in Asheville. In: 6th Transportation Research Board Conference on Innovations in Travel Modeling, Denver, CO (2016)
- Lawson, C.T.: Transformative trends in bus data: a bright future ahead. *TR News* **303**, 28 (2016a)
- Lawson, C.T.: Integration of Bus Stop Counts Data with Census Data for Improving Bus Service. Final Report FHWA-NJ-2016-001, published for the New Jersey Department of Transportation (2016b). <https://dspace.njstatelib.org/xmlui/bitstream/handle/10929/40830/t7642016c.pdf?sequence=1&isAllowed=y>
- Lawson, C.T. Applying census data for transportation: 50 years of transportation planning data progress. *Transp. Res. Circ.*, (E-C233) (2018a)

- Lawson, C.T.: 2018 National household travel survey workshop. *Transp. Res. Circ.*, (E-C238) (2018b)
- Lawson, C.T., Tomchik, P., Muro, A., Krans, E.: Translation software: an alternative to transit data standards. *Transp. Res. Interdiscip. Perspect.* **100028** (2019)
- Lee, S., Hickman, M., Tong, D.: A time-varying route-level transit patronage model. A Paper Presented at the 92nd Transportation Research Board Meetings, January 13–17, 2013, in Washington, DC (2013a)
- Lee, S., Tong, D., Hickman, M.: A comparative study of alternative methods for generating route-level mutually exclusive service areas. A Paper Presented at the 92nd Transportation Research Board Meetings, January 13–17, 2013, in Washington, DC (2013b)
- Liu, C., Porter, R., Zlatkovic, M., Fazzaz, K., Taylor, J.: First and Last Mile Assessment for Transit System (2018). <https://pdfs.semanticscholar.org/d760/ba6f9a0a69a83df7481a71119efbb26ac4d8.pdf>
- Liebig, T., Piatkowski, N., Bockerman, C., Morik, K.: Predictive trip planning-smart routing in smart cities. In: *Extended Database Technology/International Conference on Database Theory (EDBT/ICDT) Workshops* (pp. 331–338) (2014)
- Ma, X., Zhang, J., Ding, C., Wang, Y.: A geographically and temporally weighted regression model to explore the spatiotemporal influence of built environment on transit ridership. *Comput. Environ. Urban Syst.* **70**, 113–124 (2018)
- McKenzie, B.: Transit Access and Labor Market Outcomes across Segregated Neighborhoods. An unpublished dissertation (2011)
- OpenTripPlanner: (no date). <http://www.opentripplanner.org/>
- Owen, A., Levinson, D.M.: Developing a comprehensive US transit accessibility database. In: *Seeing Cities Through Big Data* (pp. 279–290). Springer, Cham (2017)
- Peng, Z.: A Simultaneous Route-level Transit Patronage Model: Demand, Supply and Inter-route Relationship. An unpublished dissertation (1994). <http://www.pdx.edu/sites/www.pdx.edu/files/SR027.pdf>
- Pi, X., Egge, M., Whitmore, J., Silbermann, A., Qian, Z.S.: Understanding transit system performance using AVL-APC data: an analytics platform with case studies for the Pittsburgh Region. *J. Public Transp.* **21**(2), 2 (2018)
- Pulchalsky, C., Joshi, D., Scherr: Development of a regional model based on Google Transit Feed Specification. A Paper Presented at the 13th TRB Planning Application Conference, May 2011, in Reno, NV (2012)
- Rodnyansky, S.: Do it yourself: obtaining updated transit stop and route shapefiles in urban and nonurban areas. *Cityscape* **20**(1), 205–214 (2018)
- RSG: User Guide Simplified Trips-on-Project Software (STOPS). An unpublished report (2015). https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/STOPS_1.50_user_documentation_v5.pdf
- RSG: User Guide Simplified Trips-on-Project Software: Version 2.50. An unpublished report (2019)
- Smith, B.L.: Using geographic information systems and the world wide web for interactive transit-trip itinerary planning. *J. Public Transp.* **3**(2), 3 (2000)
- Sun, D., Peng, Z.R., Shan, X., Chen, W., Zeng, X.: Development of web-based transit trip-planning system based on service-oriented architecture. *Transp. Res. Rec. J. Transp. Res. Board* **2217**, 87–94 (2011)
- Swayne, M., Miller, M.: Innovation on Job Accessibility with General Transit Feed Specification (GTFS) Data. An unpublished report (2018)
- Thompson, G., Brown, J., Bhattacharya, T.: What really matters for increasing transit ridership: understanding the determinants of transit ridership demand in Broward County, Florida. *Urban Stud.* **49**(15), 3327–3345 (2012)
- Census Bureau, U.S.: Understanding and Using American Community Survey Data: What All Data Users Need to Know. U.S. Government Printing Office, Washington, DC (2018)
- Verbas, I., Frei, C., Mahmassani, H., Chan, R.: Stretching resources: sensitivity of optimal bus frequency allocation to stop-level demand elasticities. A Paper Presented at the 92nd Transportation Research Board Annual Meetings, January 13–17, 2013, in Washington, DC (2013)
- Vij, A., Walker, J.: You can lead travelers to the bus stops but you can't make them ride. A Paper Presented at the 92nd Transportation Research Board Meetings, January 13–17, 2013, in Washington, DC (2013)
- Wang, K., Woo, M.: The relationship between transit rich neighborhoods and transit ridership: evidence from the decentralization of poverty. *Appl. Geogr.* **86**, 183–196 (2017)
- Wei, R., Liu, X., Mu, Y., Wang, L., Golub, A., Farber, S.: Evaluating public transit services for operational efficiency and access equity. *J. Transp. Geogr.* **65**, 70–79 (2017)
- Wong, J.: Leveraging the general transit feed specification (GTFS) for efficient transit analysis. A Paper Presented at the 92nd Transportation Research Board Annual Meetings, January 13–17, 2013, in Washington DC (2013)

- Wong, J., Reed, L., Watkins, K., Hammond, R.: One transit data: state of the practice and experiences from participating agencies in the United States. A Paper Presented at the 92nd Transportation Research Board Annual Meetings, January 13–17, 2013, in Washington DC (2013)
- Wu, X., Cao, J.: Exploring satisfaction with arterial BRT in the Twin Cities: a machine learning approach. Presented at the 2018 Annual Transportation Meetings on January 7–11, 2018, in Washington, DC (2018)
- Zhang, J., Ma, X., Ding, C., Wang Y.: Forecasting subway demand in large-scale networks: a deep learning approach. Presented at the 2018 Annual Transportation Meetings on January 7–11, 2018, in Washington, DC (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Catherine T. Lawson (lawsonc@albany.edu) is an associate professor in the Department of Geography & Planning at the University at Albany in New York. She is the Director of the Albany Visualization and Informatics Lab (AVAIL). She received her master's degrees in urban and regional planning and in applied economics and her doctor's degree in urban studies/regional science from Portland State University in Oregon. Her research interests include data science applications for transit, passenger, and freight transportation, travel behavior, water-related transport, archived intelligent transportation systems (ITS) data visualization and hazard mitigation.

Alex Muro received his B.S. in Computer Science and Applied Mathematics from the University at Albany in 2006. He is the lead programmer at the Albany Visualization and Informatics Lab (AVAIL). His research interests include web-based geospatial mapping and analysis and open-source applications for transportation planning and hazard mitigation.

Eric Krans received his B.A. in English and Communications from UMass Amherst in 2003. He is the Program Manager at the Albany Visualization and Informatics Lab (AVAIL). His research interests include transit planning, equity in transit, hazard mitigation, and using data science to assist planners and researchers.