Heliyon 6 (2020) e03729

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Visualizing public transit system operation with GTFS data: A case study of Calgary, Canada

Postsavee Prommaharaj [a], Santi Phithakkitnukoon [a,*], Merkebe Getachew Demissie [b], Lina Kattan [b], Carlo Ratti [c]

[a] Department of Computer Engineering, Chiang Mai University, Chiang Mai, Thailand
[b] Department of Civil Engineering, University of Calgary, Calgary, Canada
[c] SENSEable City Laboratory, Massachusetts Institute of Technology, MA, USA

ARTICLE INFO

Keywords:
Systems engineering
Data analysis
Data analytics
Data visualization
Big data
Computer-aided engineering
Information technology
Data mining
Information management
Smart city
Urban informatics
Intelligent transportation systems
Visual analytics
GTFS data
Public transit

ABSTRACT

Public transportation agencies are one of the industries that generate a large volume of data on a high frequency and velocity basis. The General Transit Feed Specification (GTFS) is one of the datasets these agencies generate and share openly with the public. GTFS feeds contain data for scheduled transit service including stop and route locations, and schedules information. This paper aims to demonstrate the potential of GTFS data, specifically, the paper describes the development of a GTFS data visualization tool that displays spatial and temporal patterns of transit services from which qualitative information and insights can be gained. In this paper, GTFS data from Calgary Transit was used as a case study. Previous studies focused on the development of visualization tools that display transit movement, or static graphical representation of transit operation. However, there is still a need for a dynamic interactive visualization tool that can also measures and displays transit system operation geographically and statistically. This study builds on the previous investigations and further develops a new public transit system operation visualization tool (called PubtraVis) with six visualization modules that reflect on different transit system operational characteristics; mobility, speed, flow, density, headway, and analysis. The user can evaluate two modules side by side for comparative analysis. The analysis module provides an insightful statistical summary and similarity measure and clustering results based on the transit operation characteristics. PubtraVis was tested with real-world users through a user experience study from which it was found to be useful and easy to start using. PubtraVis can be a useful tool to demonstrate the dynamism of transit vehicles from the entire transit network at a glance, and can be used to facilitate communication between transit operators, city authorities, and the general public regarding the public transit planning and operation.

1. Introduction

The previous gap in our understanding of the temporal and spatial variations of transport demand and supply was primarily the result of data limitations [1]. Measurements of the transport demand and supply has traditionally been made via surveys. However, because surveys typically involve in-person interviews and demand a high workload, this method of data collection is costly and limited. In recent years, as increasing volumes of datasets related to the use of transportation system and human mobility are produced from various sources and become available, new opportunities arise for data-driven analysis especially data visualization, which transforms these datasets into appropriate visual representations that can lead to improvements in the planning, management and operations of transportation services [2, 3].

Data related to transportation are characterized by their spatial and temporal features. Andrienko et al. [4] summarized the transportation (traffic) data into three categories: spatial event data, trajectories of moving objects, and spatial time series. Of the three traffic data types, trajectories are the most explored data in transportation analysis [5]. Trajectories describe positions of individual moving objects (e.g., vehicle, pedestrian) at predefined events such as time interval, distance interval, or sensor change. A disaggregate analysis of trajectories would produce huge difficulties in modeling the transportation system. Normally all transport models make use of zones for aggregate computations on groups of vehicles, locations, and passengers. For example, the trip

* Corresponding author.
E-mail address: santi@eng.cmu.ac.th (S. Phithakkitnukoon).

https://doi.org/10.1016/j.heliyon.2020.e03729
Received 31 July 2019; Received in revised form 24 February 2020; Accepted 30 March 2020

starts and ends of trajectories can be aggregated and summarized in Origin-Destination (O-D) matrices that contain information about the spatial and temporal distributions of trips between different zones in an urban area.

Traffic data in general and trajectory data in particular have received widespread attention in the visualization community [3, 5]. Mainly, previous studies based on mobile phone users trajectories [6, 7, 8] and taxis trajectories [9, 10] have made significant progress in extracting mobility patterns in the form of O-D flows. Visual transformation of these O-D flows are performed and presented in the form of region-based heatmaps [6, 7, 9, 11]; and line-based visualization [12, 13]. Zhou et al. [14] developed a visual exploration system to enhance O-D flow representation which reduces visual clutter of flow map and enhanced correlations of O-D flows. Taxi trajectory data have been also explored to test a visual analysis system which interactively evaluate traffic conditions [15] and to demonstrate a visual query model that supports complex spatio-temporal queries over O-D data [2]. A number of other visualization and analysis methods have been developed to improve the transportation planning, operations, and services. Some of the applications includes the study of accessibility of urban structures [16]; visually study joint traffic situations in multiple separate locations [17]; and urban crowd flow [18, 19, 20].

Public transit system is also increasingly being augmented with a range of information and communication technologies (ICT) that makes them smarter, safer, more efficient, and integrated. The rapid development of ICT and subsequent ICT-enabled transport services has enabled transit agencies to generate streams of data on a high frequency basis. These data are obtained from on-board sensors and data collection points introduced by Automated Vehicle Location (AVL), Automated Passenger Counting (APC) and Automated Fare Collection (AFC) systems. Many transit agencies also generate a General Transit Feed Specification (GTFS) data and made them publicly available. The majority of previous studies utilized the AFC, APC and AVL data to capture the origin and destination flows of transit users [21, 22]; and to analyze transit performance measures for operations and planning [23, 24, 25, 26, 27, 28]. However, recent studies have started developing open architecture platforms for transit data demonstration, analysis and visualization. For example, Kurkcu et al. [29] applied a bus trajectory data to develop a web-based tool to acquire, store, process and visualize bus trajectory data. The AFC data have been also explored to develop a web-based application to monitor and visualize the performance of bus fleets [30].

In this study, we capitalize on the untapped potentials of GTFS data. GTFS feeds contain data for scheduled transit service including stop and route locations, and schedules information. The broad adaptation of GTFS by transit agencies has made it a de facto standard, which can provide unprecedented insight into many different aspects of transit services and operations. GTFS is the least explored transit data. Only a handful of studies have used GTFS data for visualization, such as web-based movement of buses [31], static visualization (graphical representation) of transit operation measures (e.g., transit coverage, waiting time at stops, and daily transit demand fluctuation) using an existing GIS tool (i.e., ArcGIS) [32], and a tool for preprocessing and visualizing GTFS feeds as a simple movement of public transits on a map [33]. In Trains of Data [34] project, a visualization that runs in a 2D map was developed to display the number of transit users in the network and the time trains run behind schedule. Barry and Card [35] applied data captured from GTFS and realtime train locations to develop a line-based visualization technique to represent the Boston subway trajectories. The location and timing of running trains are also represented in a map. However, there is still a need for a 'dynamic' visualization (animation) that can also measure and display transit system operation in terms of both geographical and statistical forms, so that the public transit system operation (from the scheduled transit service's point of view) can be observed and assessed for transit system design refinement. This study aims at filling in this gap by introducing a development process of a public transit operation visualizer (called PubraVis) that measures and dynamically visualizes public transit operation from six perspectives (visualization modules); mobility, speed, flow, density, and headway as well as analysis module that summarizes and shows insightful analytical information. Although some of the previous work have touched on a similar aspect of transit vehicle mobility such as [31] and [33], to the best of our knowledge, this is the first attempt in building a complete GFTS visualizer with these six modules.

The rest of the paper is structured as follows. Section 2 discusses the requirements and data description. Section 3 describes the development of our visual analytics tool whose analytics features are explained in Section 4. A demonstration of how the tool works is in Section 5. The tool was evaluated through a user experience study, which is described in Section 6. Finally, Section 7 concludes the work presented with a summary and future direction.

2. Requirements and data descriptions

The city of Calgary is used as a case study to demonstrate our developed visual analytics tool. The 2016 population estimation showed that Calgary had 1.23 million inhabitants with a total area of 825.3 km^2. The city of Calgary owns the public transit system in Calgary, which is operated by the Calgary Transit that provides bus and Light Rail Transit (LRT) services. The population projection report from the Alberta government indicates that Calgary's population is expected to reach 2.5 million by 2050 [36]. This growth generates more travel that will add burden to the transportation system. During the past decades, Calgary Transit has experienced unprecedented ridership growth that leads to the continual expansion of the LRT system with the latest extension, West LRT, commencing its service in 2012 [37]. In 2015, ridership of Calgary transit reached 110 million. Figure 1 shows the Calgary Transit system map. It consists of two light rail transit (LRT) routes (red and blue) where each route operates in two lines, 433 bus routes (grey), over 6,000 stops, and the total transit routes cover 4,369 km [38].

2.1. Requirements gathering

At the center of our methodology is the user-centered approach. Such an approach ensures the involvement of end-users from the initial research cycle onwards and guarantees the iterative testing of methodologies and solutions so that they can be adjusted to and aligned with real-life and context-sensitive needs. In this study, the development of the visualization tool was done iteratively through repeated communication of a local transit agency (Calgary Transit), transit users, and transportation professionals. We analyzed the current practices of transit

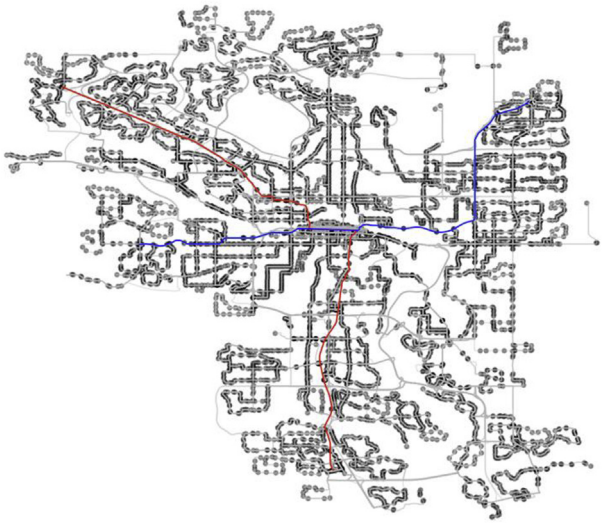

Figure 1. Calgary Transit system map.

agencies with regard to translating large volume transit related data into insights. One of the transit agency's main interests lies in the development of a data visualization tool for monitoring transit operation and performance. The following are two potential analysis scenarios extracted from the discussion with a transit agency and transportation professionals.

Scenario 1. Overviewing transit service

A transit agency named Transit-Y introduced modifications to 25 routes in the northwest and southwest quadrants of a city. Transit-Y wants to hold several open houses to get an idea of how the route changes would impact rider's ability to get where they need to go. During a previous public engagement process, Transit-Y employed a print version of transit maps and schedules. Even though, the transit system maps outline all the different routes in the city (i.e., where they start, travel, and connect to), the printed maps may not be the ideal way of conveying information to the transit users. Transit-Y is seeking a visual tool to facilitate communication between the transit agency and transit users and non-users, advisory committees, county departments, transit staff, and various stakeholders. For instance, the tool could demonstrate the dynamism of transit vehicles from the entire transit network at a glance – i.e., it could show route changes, service reductions or service extensions; it could enable users to explore information from different perspectives and levels of abstraction; and the user could evaluate two transit services side by side for comparative analysis, and so on.

Scenario 2. Monitoring transit operation and performance

In the last decade, most of the previous efforts of the Transit-Y were directed towards data generation and collection, and data aggregation, such as development of processes and platforms for combining transit data from multiple sources (active and passive data collection channels). The focus for the Transit-Y is shifting towards capturing value from the data that has been gathered and developing actual application of data-driven insights. More specifically, Transit-Y has been exploring how to use GTFS data to grow and improve transit service performance. Transit-Y is in search of an interactive tool for visualizing different characteristics of its transit system operation with the analytics capabilities. Transit-Y wishes to evaluate the characteristics of the transit network when new schedule plan is introduced. The specifics of this scenario include analytics that automate the characterization of routes similarities, identifying crowded stations as a result of new scheduling plan (service changes), and so on.

The sheer volume of available transit data has grown exponentially over the last decade, and we analyzed how these transit agencies would like to turn this flood of raw data into insights. We follow an approach which leads to a development of an interactive visualization tool that fits with realities and needs. The aforementioned examples of analysis scenarios showed us the type of tasks that are relevant to the transit agencies, which were also used in defining our visualization tool modules. Thus, this study aims at filling in this gap by introducing a development process of a public transit operation visualizer (PubraVis) that measures and dynamically visualizes public transit operation from six perspectives (visualization modules); mobility, speed, flow, density, and headway as well as analysis module that summarizes and shows insightful analytical information.

2.2. General Transit Feed Specification (GTFS) data

This study is carried out using GTFS static, which is the most common standardized format for public transit schedules and geographic information. The developed visualization tool is based on the Calgary Transit GTFS data between April 2016 to June 2016. The GTFS defines a common format for public transportation schedules and associated geographic information, which was first conceived by Portland's TriMet transit agency in 2005 [39]. A GTFS feed is a collection of at least six, and up to 13 CSV files representing routes, stops, stop times, trips calendar, and so on. Figure 2 shows an example of a GTFS feed and the relationship between the aforementioned files and their entities. For example, a trip.txt and route.txt files are related through route_id. Currently, GTFS feeds allow public transit agencies to publish their transit data and share them with the general public and application developers. GTFS only includes scheduled operations, but Williams et al. [40] recently developed GTFS data for semi-formal transit systems by adopting the guidelines from the Google Developers website [41]. Recently Google has introduced GTFS Realtime, an extension to GTFS static, which is a feed specification that allows public transportation agencies to provide real-time updates about their fleet to application developers. GTFS Realtime feed needs to be updated regularly and these updates include trips updates, service alerts, and vehicle positions. However, accessing GTFS Realtime data is a lot more difficult than accessing GTFS static.

In addition, the geographic information system (GIS) shapefile of Calgary is used as background map, which contains the information of administrative divisions of Calgary and city boundaries.

2.3. Data preprocessing

Raw GTFS data needs to be preprocessed in order to generate a suitable input for visualization. There were mainly three steps in our data

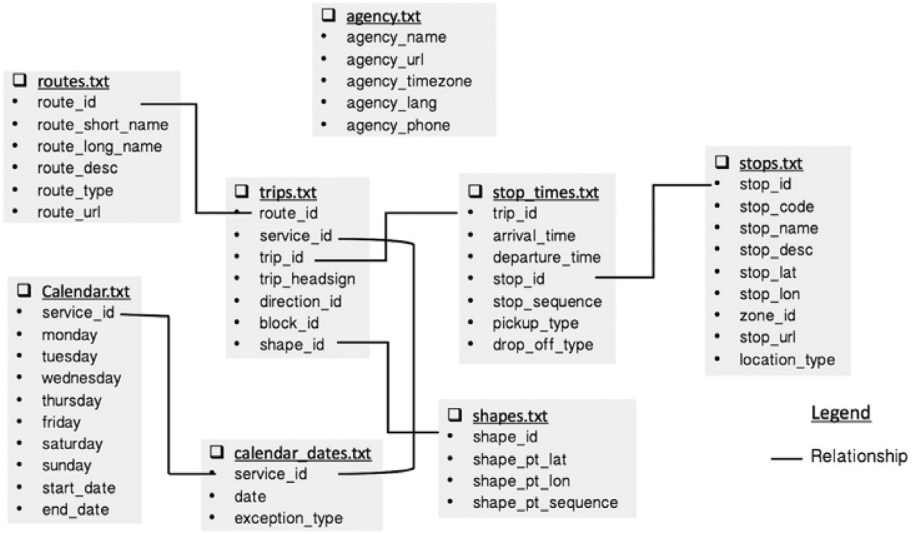

Figure 2. GTFS feed file and entity relationship.

Figure 3. Procedure of GTFS data preprocessing, from raw data to visualization.

preprocessing that includes data cleaning, reorganization, extraction, and filtering (Figure 3).

2.4. Raw data

Our raw data from the Calgary Transit was 755 MB in total, consisting of eight files; agency.txt, calendar.txt, calendar_dates.txt, routes.txt, shapes.txt, stop_times.txt, stops.txt, and trips.txt.

2.5. Data cleaning

There were some errors in the raw data, such as incorrect and missing records which we corrected or removed accordingly. The missing records were incomplete attributes, for example, missing shape_id records in trips.txt and missing latitude and longitude records of stations (stops.txt) which we removed from our database. Example of incorrect records were attribute values that were obviously out of range, such as station's geo-location value (stops.txt) were not in the Calgary Transit area for which we removed from our data. Another example is a missing geo-location between a trip for which we interpolated the missing value based on the adjacent geo-locations. In total, the removed errors accounted for about 3% of the whole data.

2.6. Data reorganization

After cleaning the data, we merged some files together and added an ID list to link them up to create an initial structure file (ISF), so that it can be accessed easily by the visualizer. For example, route.txt, trips.txt, stop_times.txt, and stops.txt were merged into an ISF for visualizing the transit mobility (vehicle movement). An ISF can be faster accessed than GTFS files as it links all required files into one file.

2.7. Data extraction

PubtraVis offers six visualization modules that reflect on different perspectives of the transit system operation. These modules are: mobility, speed, flow, density, headway, and analysis. Mobility module shows the movement of the transit vehicles (buses and trains) in the system as well as number of trips made. Speed module displays the speed level of each trip vehicle in the system. Flow module illustrates the direction of transit flow in each station throughout the system. Density module displays transit vehicle density distribution in the system. Headway module illustrates the average interval of time between vehicles moving in the same direction on the same route for each station throughout the system.

Required information was extracted to feed into each visualization module. Mobility module uses the ISF (described in Section E) while Speed module uses the same ISF with an additional attribute of space mean speed value (s), which can be calculated as in Eq. (1).

$$s = \frac{|P_i - P_{i+1}|}{t} \tag{1}$$

where P_i is the current position and P_{i+1} the next position, and t is the elapsed time. The numerator ($|.|$) is simply the geographical distance between the current and next positions of the vehicle.

For flow module, the system calculates the flow direction (i.e., angle (θ) where the vehicle is headed) according to the current and next positions of the vehicle, as shown in Eq. (2).

$$\theta = \sin^{-1}\left(\frac{y_{i+1} - y_i}{|P_i - P_{i+1}|}\right) \tag{2}$$

where y_i and y_{i+1} are the longitude of the current and next positions, respectively. The numerator is the difference in the longitude directions while the denominator is the distance between the current and next positions. The flow direction is calculated and accumulated at each station.

For density module, the system calculates the accumulative number of vehicles per grid area (2km by 2km) over a 15-minute period.

For headway module, the system calculates the route level headway. A headway is the time between successive arrivals of a transit vehicle from one route at a specific stop. Headway is calculated hourly for each station. For example, a route that runs transit vehicle four times for 1 h

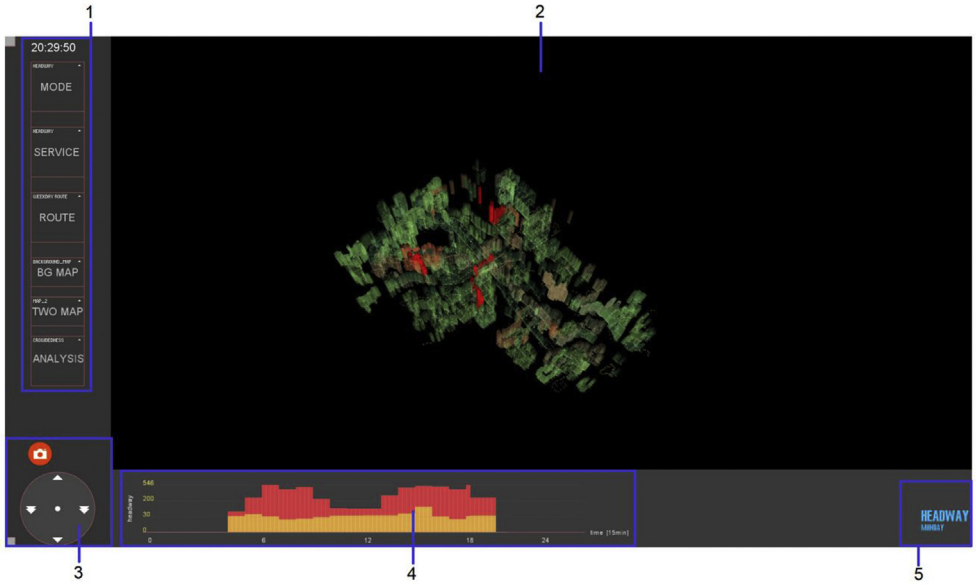

Figure 4. A screenshot of PubtraVis interface.

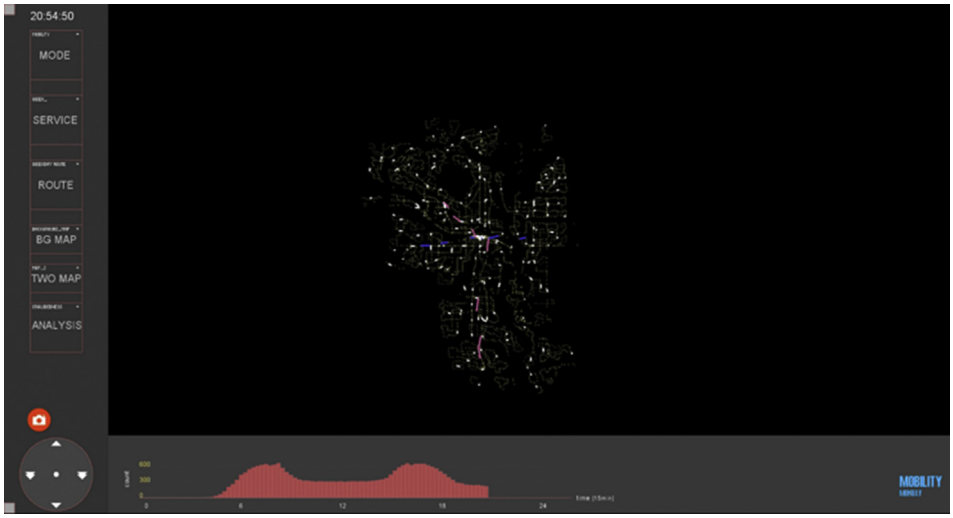

Figure 5. A snapshot of the mobility module.

has an average headway of 15 min. Transit stations that served multiple routes were visualized multiple times so that headway values were included for each route.

3. Visualization

There are two main elements of the visualization; graphics and user interaction.

3.1. Graphics

The system visualizes the transit system operation in the perspectives of the system's mobility, speed, flow, density, and headway, using the data extracted and ready for each module accordingly. A snapshot of PubtraVis is shown in Figure 4 that includes four main display areas. Area 1 includes a digital clock on the top and a drop-down list of six main menus; MODE, SERVICE, ROUTE, BG MAP, TWO MAP, and ANALYSIS. The MODE drop-down menu allows the user to choose one of the visualization modules in which different information is provided. SERVICE menu allows the user to select one of the transit services run on different days throughout the week to observe. ROUTE menu allows the user to select a particular route of the transit services to observe. BG MAP menu is used to either turn on/off the background map. TWO MAP menu is to turn on/off the two-map mode i.e., displaying two maps for comparison. ANALYSIS menu allows the user to view analytics results. Area 2 is the main display window that shows the graphics and animation according to the user's selection. Area 3 consists of a screenshot taking button (camera icon) and camera view controller for viewing the graphics from different perspectives. Area 4 displays a bar chart carrying statistical information of the displayed transit operation measures such as the distribution of transit trips in one day, with different colors showing the hourly minimum, maximum and average values. Area 5 shows the name of the selected module.

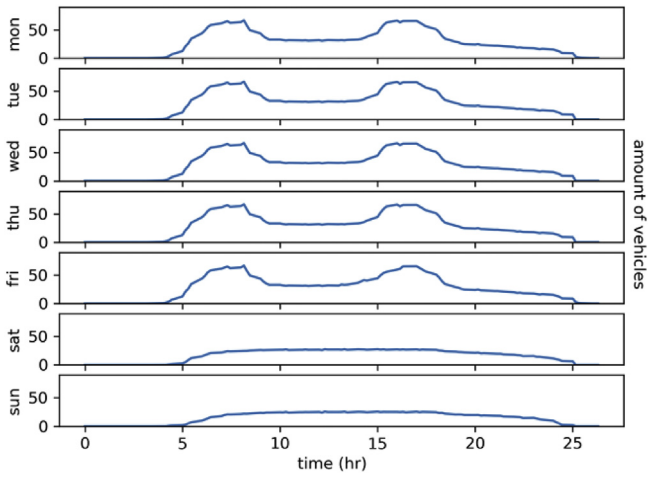

Figure 6. Number of vehicles throughout the week that can be observed from the mobility module.

In the mobility module, the system displays the movement of the public transit including buses and trains over the course of one day of the selected service schedule, so that the user can observe the overall movement of the transit system. Bus movement is shown using a white line connecting the current and previous stops to create an illusion of a moving vehicle, while the train service lines are displayed with red and blue lines for the Red Line (69 Street/Saddletowne) and Blue Line (Tuscany/Somerset - Bridlewood), respectively. A snapshot of the mobility module is shown in Figure 5. The number of operating vehicles accumulated every 15 min throughout the day is shown in the red graph in area 4.

Figure 6 summarizes what the user is able observe from the mobility module throughout all seven days of the week from which two distinct patterns are revealed i.e., weekday's and weekend's, intuitively. There are two peaks, one in the morning hours and the other in the afternoon for a weekday's mobility, while on a weekend there is a more consistent mobility from morning to the evening hours.

Public transit travel speed is the fundamental indicator for the transit operation dynamic monitoring, traveler's information service, as well as the transit service evaluation [42]. With the GTFS data, the system is capable of visualizing the speed of every vehicle in the system throughout the day of the selected transit service. The user can observe the speed statistically, temporarily, and geographically. For each vehicle in the transit system, speed is calculated using Eq. (1). A snapshot of the speed module is shown in Figure 7. A graph of varying speed values over the course of the day, including the minimum (in green), maximum (in red), and average (in yellow) is displayed in area 4 at the bottom part of the interface. Moreover, there are speed dashboard and average speed display at the lower right corner for the user to observe the instance values. The main display uses different blue shades to represent a range of speed graphically, so that the user can observe the geographical distribution of the speed throughout the transit system. The speed module can be used to indicate the low speed transit section which can be good

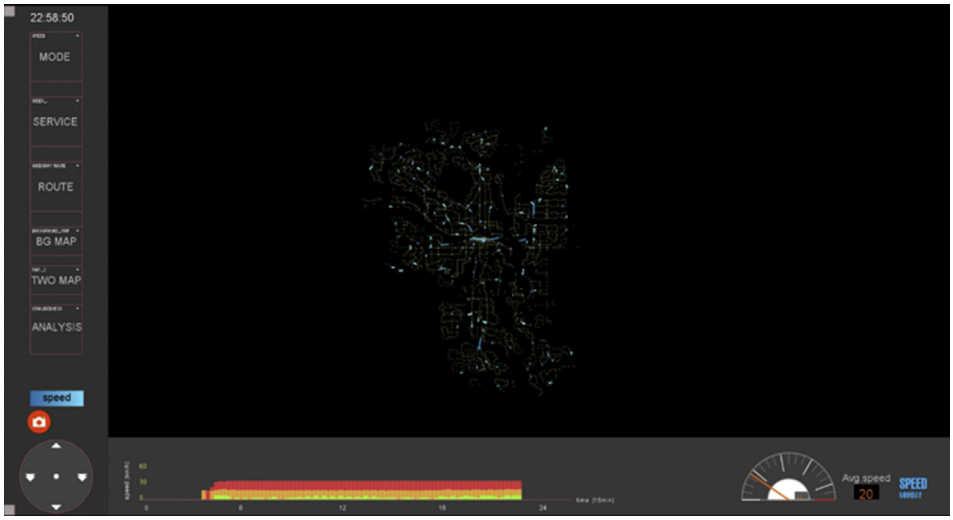

Figure 7. A snapshot of the speed module.

candidate for future transit priority measures such as queue jumps, transit signal priority, and so on.

The user can observe the speed graph of any day of the week. Seemingly, the speed level remains consistent throughout the week as summarized in Figure 8. The value exceeds the 24-hour mark because the scheduled operation goes beyond midnight each day of the week. Apparently, the operations on Tuesday, Friday, and Sunday carry on longer after the midnight than other days.

In addition to speed, there is also a flow of transportation that is a critical factor in the design of public transport. Therefore, in the flow module, the tool visualizes the direction of departure at the transit station. It shows the instant maximum outflow of each station, which collectively indicates the overall flow direction of the transit system. The flow direction is calculated using Eq. (2). The direction with the maximum flow over a 15-minute period at each station is visualized with a color line pointing to one of the four main directions i.e., north (red), south (green), west (yellow), and east (blue). Every 15 min if there is a change in maximum flow direction, the flow line will change from its previous direction color to black and turn to the new direction. Once the flow line points to the new direction, then its color is changed to that direction's color respectively. A snapshot of the flow module is shown in Figure 9. To allow the statistical aspect of the flow to be observed, there are graphs showing the number of flows (vehicles) in four main directions throughout the day in the corresponding color bars (West - yellow; East – blue; North – orange; and South – mint colors). Graphs are updated every 15 min simultaneously with the main graphic display. The flow module could provide insights regarding the intensity of directional transit flow. The user can observe the flow for any given day of the week. As summarized in Figure 10, the flow level or the number of vehicles in four directions throughout the week align with the overall mobility level observed in the mobility mode. Flow in each direction reveals a similar pattern. A distinction can only be made between the weekday's and weekend's flows, while flow level remains similar across the four directions.

The number of public transit vehicles in a given area or vehicle density is another important factor in designing efficient public transit system as traffic congestion continues growing in urban areas. Figure 8 shows a snapshot of the density module in which the density is represented by a 3D box with various colors. The height and the color indicate the density level. Taller and warmer color of the box imply higher level of density (color ranges from greenish to reddish). For example, Figure 11 shows a peak in transit vehicle density in Calgary downtown. In the panel below the main graphic display, there is a graph showing the maximum, minimum, and average density levels throughout the day, in red, yellow, and green color, respectively. There is also an instant average density value (i.e., the number of transit vehicles per 2km-by-2km area at a given time interval) showing in a box at the bottom right corner of the tool. Figure 12 summarizes all density graphs that can be observed in the panel below the main display. The average density tends to be much more stable compared to the maximum density value that exhibits a similar pattern with the overall mobility trends (Figure 6).

Headway is an important concept used in transit planning and modelling. Headway is a measurement of interarrival times of transit vehicles in the system. A lower headway implies a more frequent service, or less waiting time for the passengers. If a bus arrives at the bus stop every half hour, then the bus service has the headway of 30 min. If passengers all turn up to the bus stop randomly without considering the schedule, then the average passenger would be expected to be waiting there for a period of time equal to half of the headway – i.e., 15 min. The transit capacity and quality of service manual provides a useful table on different headway classes and what they mean for waiting passengers [43]. For example, some headway class can mean that passengers do not need schedules, or service is unattractive to choose riders. Being able to measure and observe the distribution of the headways throughout the system is thus very essential.

In the headway module (Figure 13), a higher headway value is represented with taller 3D box with warmer color (color ranges from

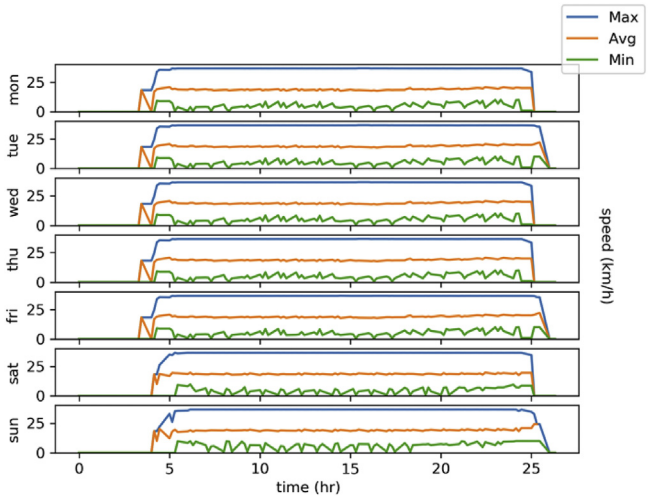

Figure 8. Speed level throughout the week that can be observed from the speed module.

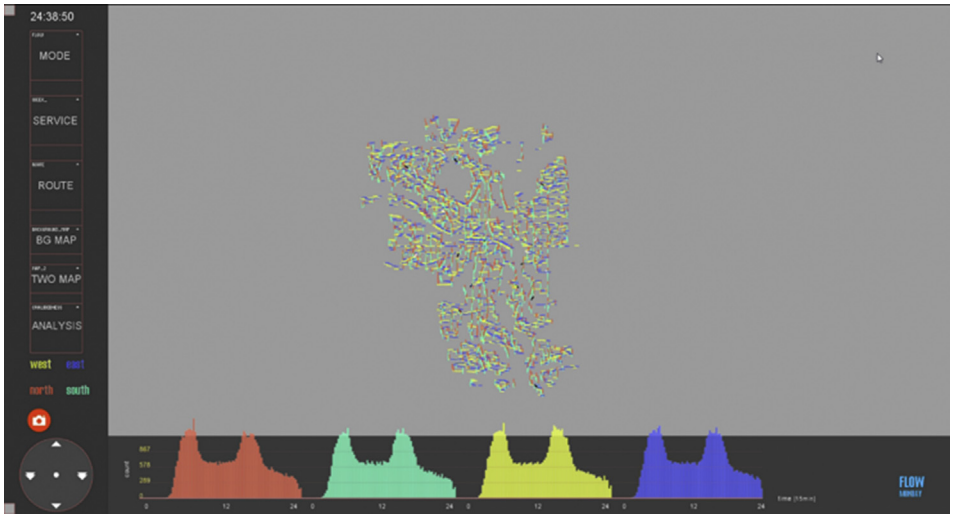

Figure 9. A snapshot of the flor module, showing the direction of maximum flow at each station.

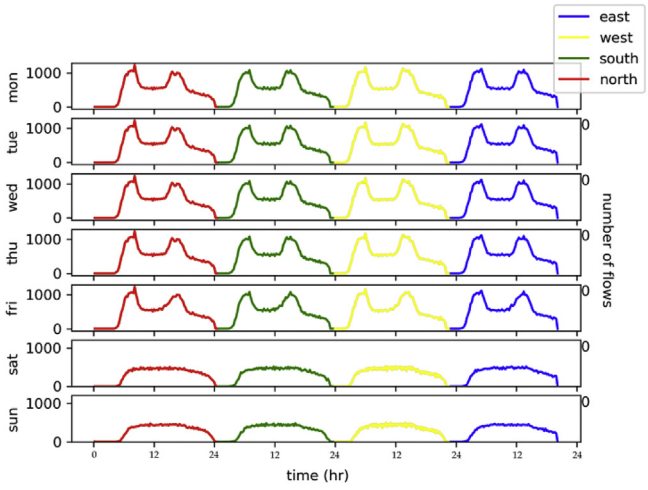

Figure 10. Flow level in different directions throughout the week that can be observed from the flow module.

greenish to reddish). The maximum (red), minimum (green), and average (yellow) headways throughout the day are displayed in a graph below the main display. The user can observe the distribution of the headways both temporal and spatial aspects. Figure 14 summarizes all headway graphs that can be observed by the user. Intuitively, two distinct patterns are observed; weekday and weekend. Average headway drops during the morning (7:00–9:00) and afternoon (16:00–18:00) periods on a weekday, which means that a shorter waiting time can be expected during those hours compared to other periods of the day. This is due to a more frequent transit service during those hours. On the other hand, a weekend's average headway is much more stable throughout the day.

3.2. User interaction

PubtraVis not only measures and visualizes the transit system operation but also allows the user to interact with it by selecting different visualization modules and features to facilitate the user's analysis.

The mobility module is the default module. Other modules can be selected using MODE menu in the drop-down list, except for the analysis module that has its own drop-down menu as it has its sub-list of operations. The user can select a particular operating service and route to be viewed from the lists under the SERVICE and ROUTE menus. Figure 15

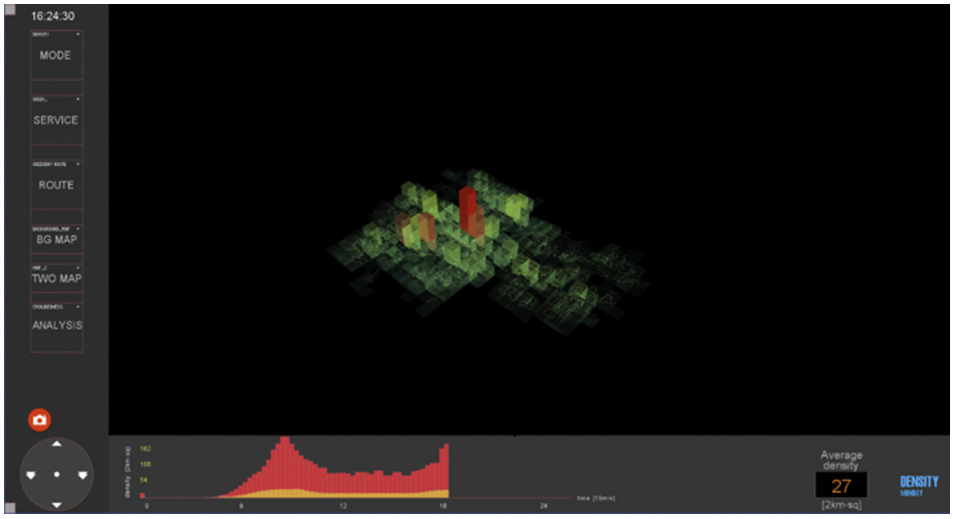

Figure 11. A snapshot of the density module.

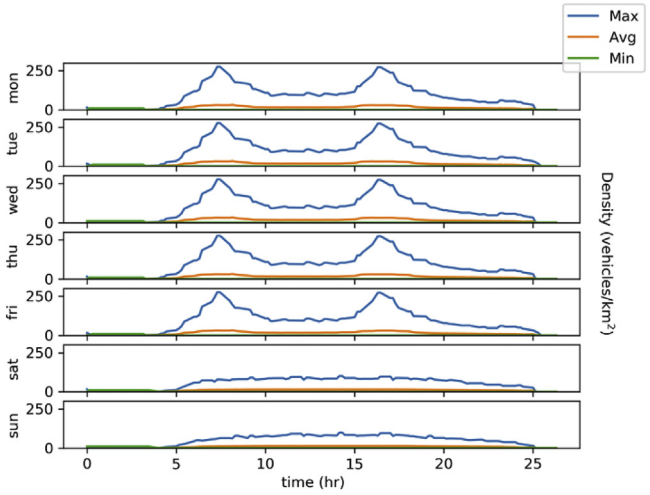

Figure 12. Density values throughout the week that can be observed from the density module.

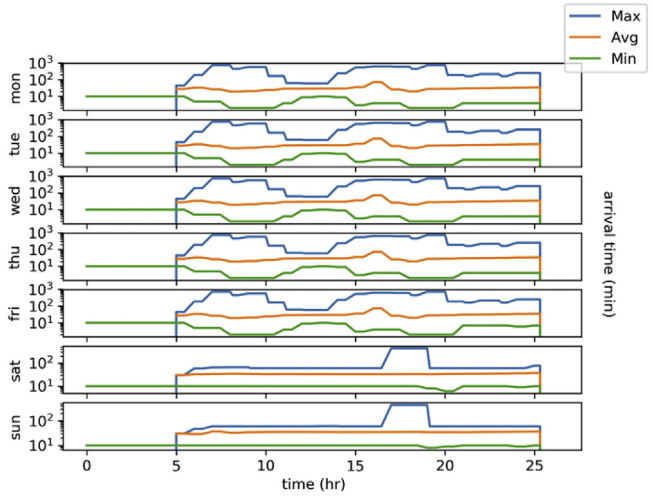

Figure 14. Headway values throughout the week that can be observed from the headway module.

shows an example of viewing Saturday service and route #291–20424, which is the train (LRT) service.

The user can also choose to overlay the graphics on a Calgary map to observe the transit operation in relation to spatial characteristics (e.g., rivers, main roads) by turning on/off the background map using the BG MAP menu. An example of using background is shown in Figure 16.

Comparison between different transit system operation measures is possible and essential for preliminary analysis. To compare two transit operation measures, the tool allows the user to view two different modules at the same time on two separate displays side by side, using the TWO MAP menu. The displayed measures are synchronous. The user can also select to view the background map in either or both displayed measures. Viewing particular services and routes is separately independent. An example of using the TWO MAP option is shown in Figure 17.

Moreover, the user can adjust the camera view by using the camera view controller (bottom left corner) that allows the user to rotate the graphics in 3D space, including zooming in and out, and reset the graphics to the 2D default display (by clicking the middle circle). The controller can be used separately for the TWO MAP option.

With the TWO MAP option, viewing the results of two different GTFS feeds such as from two different periods is also possible for a comparative analysis of transit performance between different periods of operation. This can further facilitate analysis of route change, service reduction or service extension between the two operational periods. The user can evaluate two transit services side by side and explore information from different perspectives between the GTFS feed periods.

4. Analytics

In addition to visualizing different characteristics of transit system operation, the analytics is another fundamental part of a visual analytics tool. PubtraVis provides an analytics functionality via its analysis module, which includes top lists, similarity, and clustering sub-modules.

4.1. Top stations

Besides being able to observe dynamism of transit system operation values over space and time, a basic value sorting can provide an insightful information. The *top lists* sub-module thus provides two lists; most crowded stations and longest waiting time stations, which can be useful for transit system evaluation and planning.

The user can choose to view the most crowded stations list on a particular day of the week through the analysis module. A list of most crowded stations list will then be displayed on the main display, as shown in Figure 18. This example is a list of the Monday's most crowded stations that includes ranking, stations, number of vehicles per hour, and time

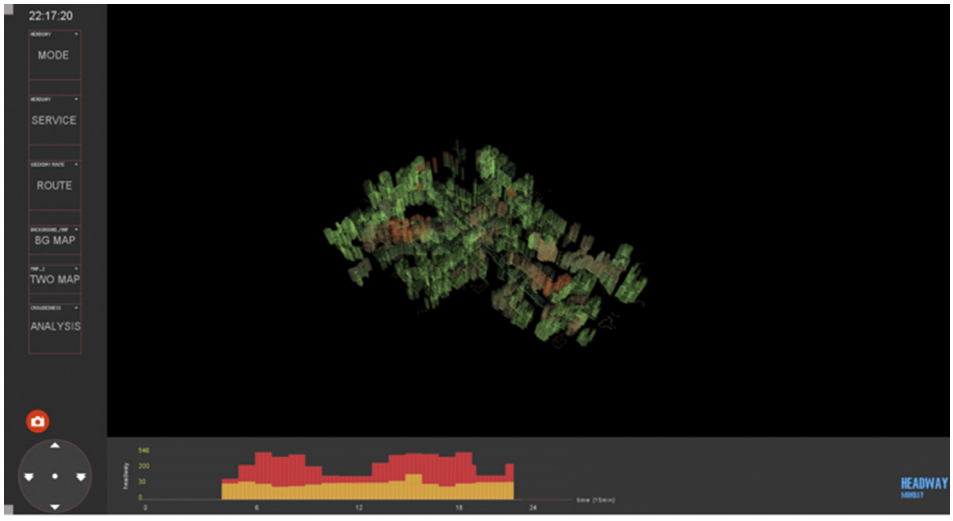

Figure 13. A snapshot of the headway module, showing vehicle interarrival times.

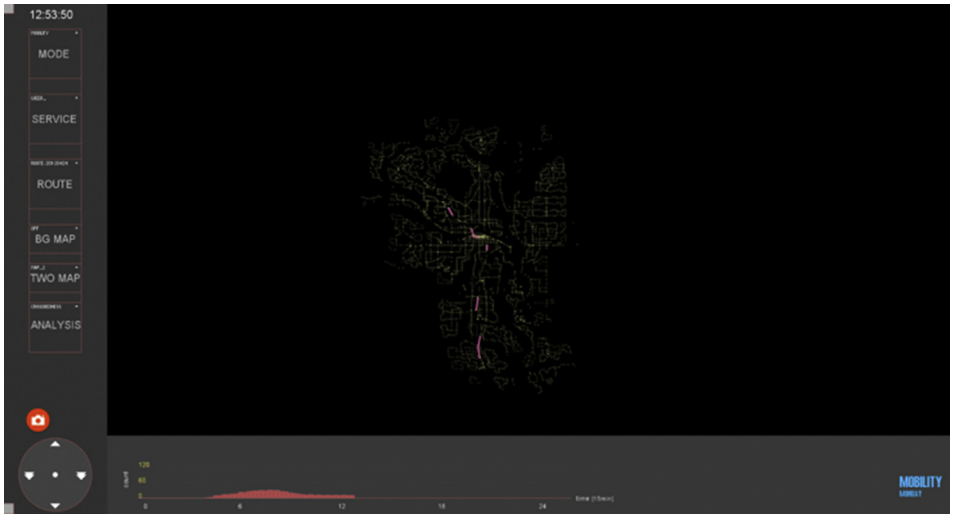

Figure 15. Example of viewing a specific service and route: Saturday service and route #201–20424 which is a train service.

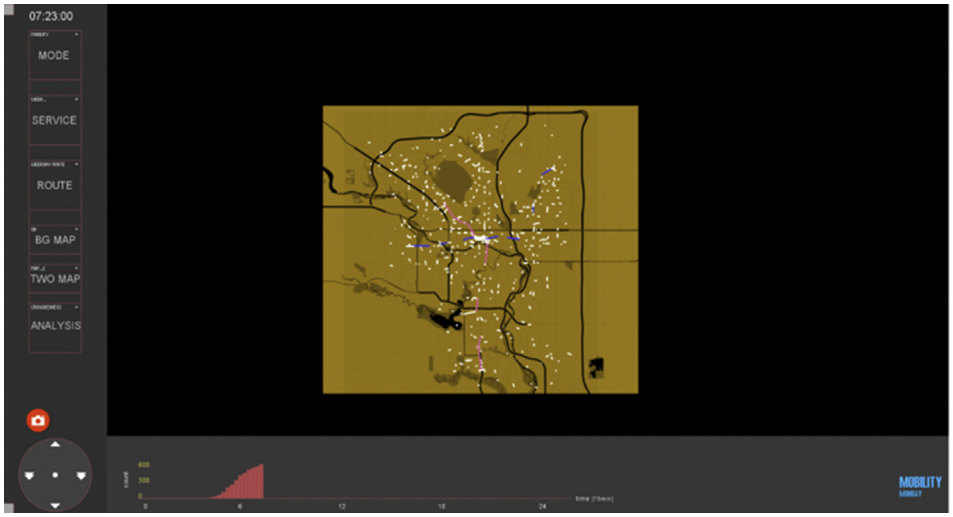

Figure 16. Example of using background map.

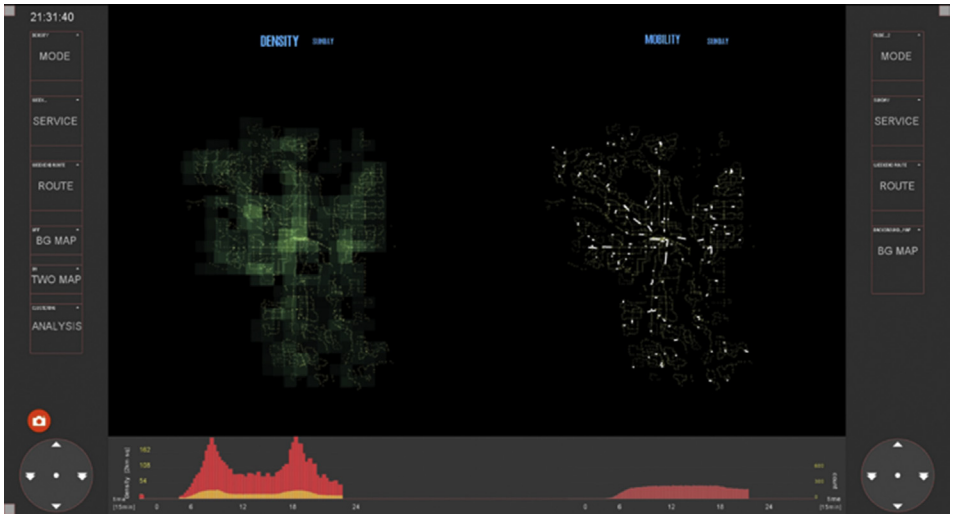

Figure 17. Example of using TWO MAP option, comparing between the density and mobility modules.

Figure 18. A snapshot of the most crowded stations list in the analysis module.

slot. The user can select to view the location of any station on the list by clicking on the station ID. The row will then be highlighted and a marker will be flashing to indicate the station location on the map. In this example, the stations #9978 and #9979 are most crowded with 52 vehicles per hour during 7:00am–8:00am for Monday operation. Moreover, the user can view beyond the top 20 stations shown on the first page by clicking the arrow icon on upper left corner of the list to see more ranked stations (a total of 6,163 stations).

We have observed that the ranking does not change across the weekdays, but varies from a weekday to Saturday and to Sunday. Lists of the top 10 stations for Monday (a representative weekday), Saturday, and Sunday are shown in Tables 1, 2, and 3, respectively.

The results in Tables 1, 2, and 3 show the transit stations ranked from highest to lowest based on the number of transit vehicle departures per hour value. Crowding in public transport can greatly affect passengers' travel experience and therefore affect route and mode choice [44]. The majority of previous studies which do incorporate public transport crowding applied stated preference and revealed preference experiments for crowding valuation [44]. Congestion and bus stop crowding are not usually incorporated in public transport modeling. However, congestion and crowding at the transit stop influence passenger's perceived in-vehicle times [45]; and in extreme cases passengers would not be boarding the transit vehicle, which can result in additional costs for passengers and operators [46]. For instance, the top three ranked stations in Table 1 are positioned along the WB 6 Ave SW route, which accommodates more than 50 bus departures from each station between 7 a.m. and 8 a.m. on a typical weekday. These results improve our understanding of current bus traffic along a given corridor, where transit vehicles can contribute to congestion at least as much as private vehicles in some corridors if the number of transit vehicles is high [47]. The analytical result provides first insights into how crowding in public transport can be assessed from the perspective of delay (due to overcrowding) caused by transit station to buses, their passengers and to other traffic.

4.2. Route similarity

For transit system planning and management, it is important to be able to identify routes with similar and different operational characteristics so that suitable operational workloads can be assigned accordingly. Thus, the analysis module provides the *similarity* sub-module that help the user to identify similar routes to a selected one based on their operational characteristics.

Table 1. Monday's 10 most crowded stations.

Rank	Station ID	Vehicles/hour	Time slot
1	9978	52	7:00–8:00
2	9979	52	7:00– – 8:00
3	9980	50	7:00 – 8:00
4	8418	44	7:00 – 8:00
5	6651	44	7:00 – 8:00
6	4586	44	7:00 – 8:00
7	5579	37	17:00 – 18:00
8	6658	34	16:00 – 17:00
9	7442	34	16:00 – 17:00
10	5579	33	7:00 – 8:00

Table 2. Saturday's 10 most crowded stations.

Rank	Station ID	Vehicles/hour	Time slot
1	5762	16	18:00 – 19:00
2	5579	16	9:00 – 10:00
3	5579	16	19:00 – 20:00
4	5579	16	17:00 – 18:00
5	5579	16	15:00 – 16:00
6	5579	16	13:00 – 14:00
7	5579	16	11:00 – 12:00
8	5192	15	10:00 – 11:00
9	6950	15	18:00 – 19:00
10	6950	15	10:00 – 11:00

Table 3. Sunday's 10 most crowded stations.

Rank	Station ID	Vehicles/hour	Time slot
1	5762	16	18:00 – 19:00
2	5192	15	10:00 – 11:00
3	6950	15	18:00 – 19:00
4	6950	15	10:00 – 11:00
5	5762	15	10:00 – 11:00
6	5192	15	18:00 – 19:00
7	9299	14	23:00 – 24:00
8	9299	14	13:00 – 14:00
9	5579	14	11:00 – 12:00
10	5192	14	9:00 – 10:00

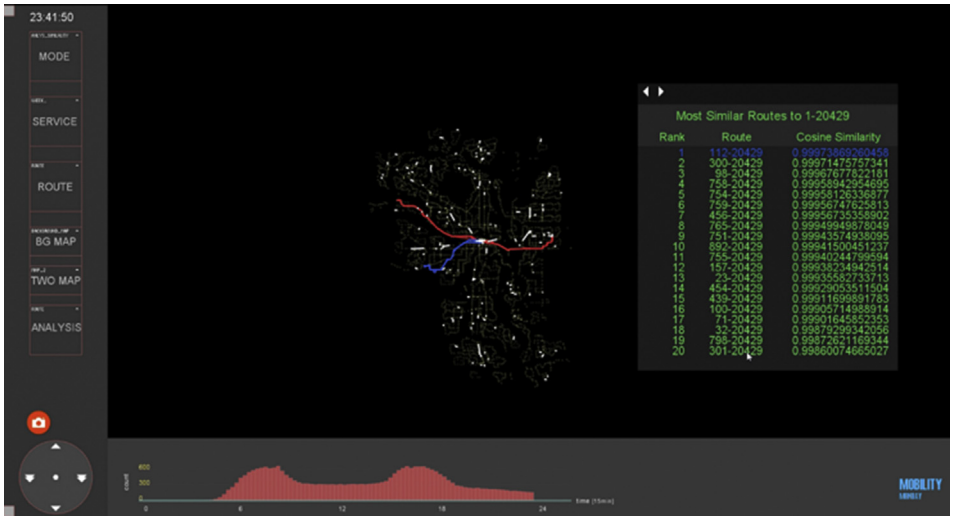

Figure 19. A snapshot of the similarity sub-module in the analysis module. This sample shows the most similar routes to the route #1–20429 (highlighted in red) ranked by the cosine similarity values.

To quantify the similarity between routes, we used the *cosine similarity*, which is a measure of similarity between two non-zero vectors of an inner product space by calculating the cosine of the angle between them. It is suitable for our route similarity measurement because the cosine similarity is a comparison based on orientation but not magnitude. In our case, we can characterize each route as a vector with features that each describes a characteristic of the route. Two vectors with 0° apart or both have the same orientation will have the maximum cosine similarity of 1, while two vectors with 180° apart or are diametrically opposed to each other will have the minimum cosine similarity of -1. Cosine similarity of 0 is for two vectors that lie 90° or perpendicular to each other. Mathematically, cosine similarity is described by Eq. (3).

$$S(X,Y) = \cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{3}$$

where X_i and Y_i are components of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. In our case, the components of comparing vectors are mobility (i.e., total number of vehicles per hour), speed (i.e., average speed in km per hour), and headway (i.e., waiting time in minutes per vehicle) values of each route. The density value isn't considered here because it is already described by the mobility value. As for the flow value, it isn't used here either because of the flow in each direction is also already described by the mobility value, and the flow direction is constrained by the road network. Although it's not considered here, its' worth exploring as part of future study.

The user can select this similarity sub-module via the analysis module' menu. Figure 19 shows a snapshot of the routes that are most similar to the selected route #1–20429 (highlighted in red) based on the cosine similarity values. The user can view the routes listed on the table by clicking on the route number, then a blue line will appear on the map to determine the route's location. In this example (Figure 19), the user clicks on the most similar route (ranked #1) to the considered route, a blue line shows the trajectory of the route.

The results from clustering of transit routes opens up new possibilities for gaining insight into how GTFS data can be used to identify set of transit routes that exhibit similarity and thus may require similar operational and planning strategies. For instance, in previous studies bus travel time prediction model has been developed based on profile similarity [48]; and transit schedule plan improvement has been suggested based on clustering group of service days with similar behavior [49, 50].

Figure 20. A snapshot of the clustering sub-module in the analysis module. This sample shows a cluster of routes operating on Monday. A box on the right shows the cluster numbers on which the user can click to view each cluster of routes.

4.3. Route clustering

Being able to identity a list of most similar routes to a given one is important. Additionally, in some cases, transit system or urban planners need to further evaluate the overall transit service provision for review, planning, and decision-making, the overall picture of operation is thus required. Clustering of routes based on their operational characteristics can offer an insightful perspective, which can lead to a discovery of different categories or set of route operations.

A challenge in clustering transit routes is the number of clusters, which is unknown beforehand. An algorithm like *k*-means clustering [51] cannot be used satisfactorily here because the number of clusters or *k* must be predefined to proceed with the clustering. A few studies have applied the *k*-means algorithm to cluster public transit operations but suffered from the predefined *k* issue [52, 53]. To remedy the shortcoming of the *k*-means, we adopted an approach known as *X*-means [54], which is a clustering technique that does not require a predefined number of clusters to proceed with the clustering. It automatically estimates the number of clusters or *k* by repetitively partitioning data and keeping the optimal resultant splits to better fit the data, until some criterion is reached. Bayesian Information Criteria (BIC) [55] is used for cluster splitting and so the algorithm essentially boils down optimizing BIC value.

Similarly to the similarity measure, for each route, a vector with three components each represents mobility, speed, and headway characteristics is constructed and used as the route's data point for clustering. With a total of 540 routes, clustering with the *X*-means is performed for each day of the week. The resulting number of clusters are 20, 20, 16, 26, 28, 23, and 26 for Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday, respectively.

The user can view the clustering results via the analysis module menu. Figure 20 shows a snapshot of the clustering sub-module where the user can view each of the route cluster on a selected day of the week. In this shown example, the user choses to view the cluster #12 of Monday's route clustering. Clustered routes are highlighted on the map. All resultant route clusters of each day of the week are shown in Appendix A, Figures A1 – A7, respectively.

The route clustering analysis was performed solely based on information extracted from GTFS data. These results can provide useful insights to improve the operational planning of public transportation services. For instance, Calgary Transit makes four major service changes (schedule plan) within a year. The changes are usually made based on feedback from passengers and drivers, political decisions, changing ridership levels, and new development areas. In this regard, the clustering results can complement the development of new schedule plan by suggesting transit routes that are characterized by similar patterns in terms of mobility, speed and headway features.

5. Demo

For demonstration purposes, a video clip showing how PubtraVis works is available at: https://youtu.be/7LMVtLoDWH0.

6. User experience study

It is important to evaluate the usability of the PubtraVis. So, we put it into the test with the real users by conducting a user experience study. The user experience study is part of the user experience (UX) design, which is a process of creating software products that provide relevant significant experiences to the users based on their behaviors. Therefore, creating a software like PubtraVis requires a user experience study to gather user behavior and preference information.

We randomly recruited subjects (i.e., users) to our study by visiting people who were in transport and information technology related fields. Each user was approached and explained about the study, and then was asked to use the tool and afterwards answer a questionnaire, which was designed according to the Theory of Four Elements of User Experience [56]. On a 5-likert scale, the user was asked for their lelvel of agreement on each of the four statements regarding the user experience with the tool as following.

1) It is easy to use.
2) It is useful.
3) It is easy to start using.
4) It is fun and engaging.

These four statements are intended to evaluate our tool based on the four elements of user experience; (1) usability, (2) value, (3) adoptability, and (4) desirability. Usability is the degree to which the tool is able or fit to be used for the user's intended taks. Value of the tool is an alignment between its features and user needs. Features of the tool should be designed in the way that they support user needs and thereby the tool will be considered valuable. Adoptability is about quality of being able to start using the tool and become a regular user of the tool, which relates to the stage when the user has not yet used the tool. Desirability relates to quality of being worth having and using the tool. Innovative vidual design can largely influence desirability. However, attractive graphics and appealing designs are not the only factors, but the ability to engage the user is also the key for desirability.

Our user experience study was carried out in Chiang Mai, Thailand and Calgary, Canada over a total period of six weeks. A total of 122 subjects were recruited for the study using the convenience sampling [57]. Subjects were in mix of genders, ages, and occupations. They were students, researchers, professors, and transport authorities. Each user was provided with a laptop computer, which had the PubtraVis installed, and a print version of Calgary transit map. The user used the tool freely

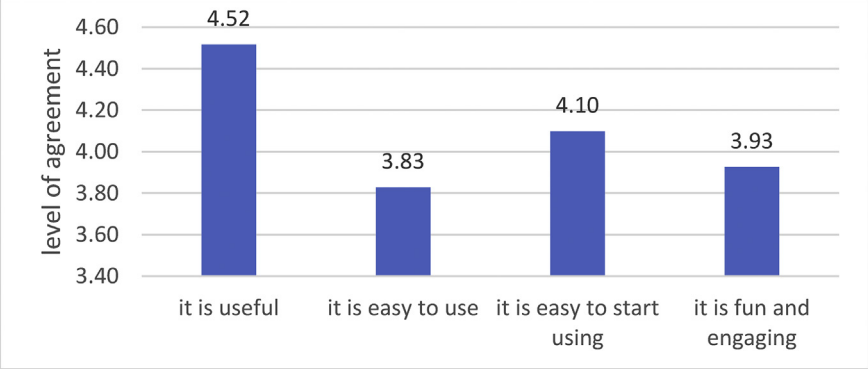

Figure 21. Overall result of the user experience study.

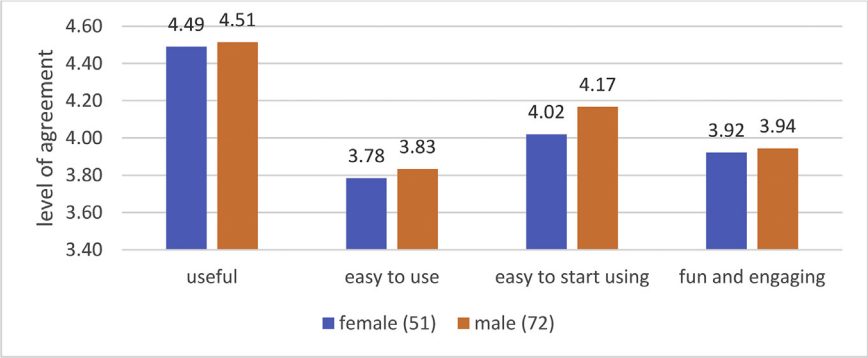

Figure 22. User experience study's result, separated by gender.

while our staff was sitting nearby for answering any questions. For diversity, subjects were recruited from different institutes and organizations. People from five universities and two transport offices participated in our study in Chiang Mai, while the study in Calgary was conducted with people from the University of Calgary, Calgary Transit, city of Calgary, Alberta Transportation, and other active transit users in Calgary, Canada.

6.1. Overall

From all 122 subject responses in the study, Figure 21 shows that the tool was rated the highest for being useful at 4.52, followed by being easy to start using at 4.10, then being fun and engaging at 3.93, and lastly being easy to use at 3.83. Overall, the tool was received well for being useful, however its current version may not be easy to use. This suggests that although the users feel that the tool is useful, the tool still needs to improve on the easy-to-use aspect – possibly through future user-friendly design with creative ways to provide a user instruction.

Some users wandered around figuring out how to go about selecting and viewing features in the tool that they wanted. They spent some time when first used the tool. From conversations with the subject users, they felt that the tool was useful but since they used it for the first time, it required some time to familiarize with the tool. For example, a comment from one of the subjects was "*I like it. I think that tool is useful, especially for people in transport planning and design. It's a little confusing at first when I started using it.*" Another comment was "*Graphics look interesting. I've been working in the Department of Transport for a while but never seen tools making use of GTFS data like this one. It's definitely useful for public transport planning. It would be great to have a user manual to ease learning time.*" These valuable comments suggest that the tool should provide clear instructions or user manuals to improve on "easy to use" as well as "easy to start using" elements. The "fun and engaging" element can also be improved with more interactive features, such as designing attractive graphics, infusing gamification, and potentially social networking functionality.

6.2. Genders

If separated by gender, there were 72 male and 51 female subjects. Both gender groups perceived the tool quite similarly, as shown in Figure 22 where the subjects in both genders rated the tool being useful the most at 4.51 (male) and 4.49 (female), followed by being easy to start using at 4.17 (male) and 4.02 (female), being fun and engaging at 3.94 (male) and 3.92 (female), and lastly being easy to use at 3.83 (male) and 3.78 (female), which is the same ranking pattern with the overall result (Figure 21). Each aspect was rated closely between the two genders, except for the "easy to start using" aspect for which the female users' rating was significantly higher.

Through our conversions with the subjects, the female users seemed to be more impressed with the tool's graphics than males. One of the comments from female users was "*I love the visual arts of the tool. It should have sound effects too. It feels like I'm playing a computer game.*" Another comment from a female user was "*It looks interesting although I don't quite understand it completely. I think it's not really easy to use. But I'm sure it can be useful.*" One of the commends from a male user was "*Looks like it can be useful for people in transport engineering. It looks attractive with its graphics. It's easy to use. It's interesting to observe different transportation measures and statistics.*" The rating result and comments suggest that the style of the visual arts is attractive to the users, sound effects are something to be considered for our future development of the tool, and user instructions are needed to improve on the "easy to use" aspect.

6.3. Ages

When separated by age, there were eight subjects who were under 20 years old, 46 subjects between 20-29 years old, 26 subjects between 30-39 years old, 28 subjects between 40-49 years old, 11 subjects between 50-59 years old, and 4 subjects over 60 years old. The result is shown in Figure 23. The users who were less than 20 years old rated the tool overall higher than other age groups. This can be interpreted that young

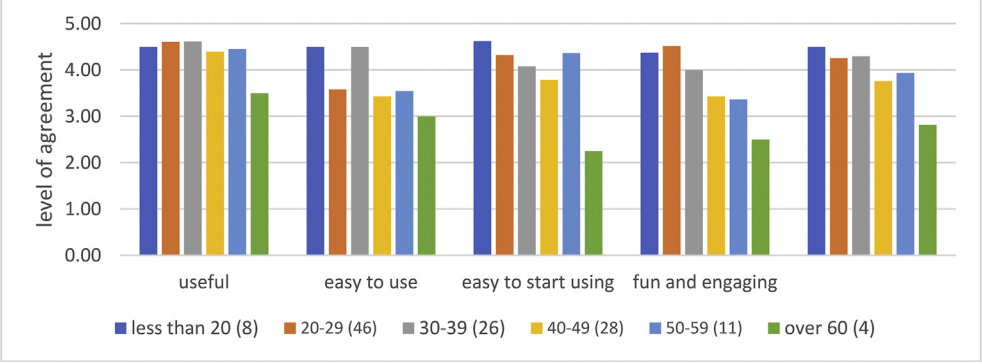

Figure 23. User experience study's result, separated by age.

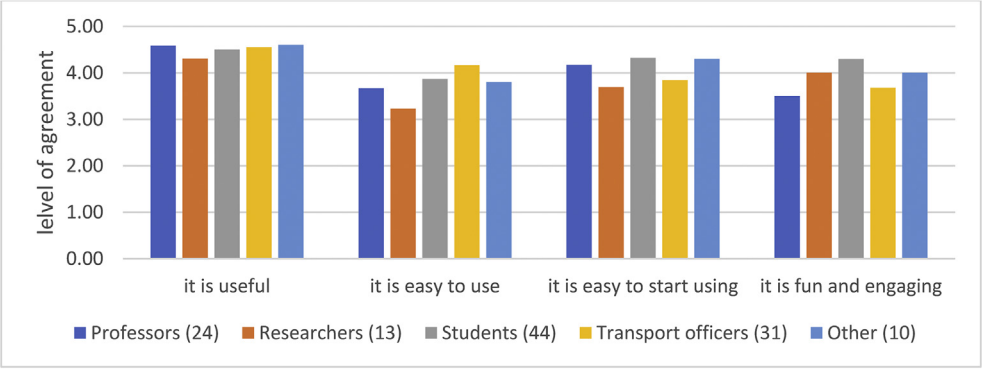

Figure 24. User experience study's result, separated by occupation.

users who are growing up in the ICT era and hence are generally more familiar with information technologies found the developed tool more useable than other age groups in general. They rated the tool highest for "easy to start using" aspect at 4.63, followed by being useful (4.5) and easy to user (4.5), and "fun and engaging" aspect was rated lastly at 4.38 which is still relatively high.

On the other hand, senior users who were over 60 years old rated the tool relatively lower than other age groups in all usability aspects. They rating ranking was "useful" (3.5), followed by "easy to use" (3.0), "fun and engaging" (2.5), and "easy to start using" (2.25). Presumably they were less familiar with the ICT and some features of the tool may not be suitable and useful for them. For example, one of the comments from this age group was "*It looks like a useful tool. But for me, it's difficult to interpret the results shown on the screen. It would be better to also provide the meaning of the results shown. Perhaps it should provide some recommendations.*" Another comment was "*I think it's useful for people who are involved in transportation but not for me.*" This suggests that in our future development, we should consider adding features that can provide result summary/interpretation and recommendations for assisting better transit system design based on the observed operation measures in relation to the provided service schedules.

Users who were in 20–29 years old age group gave the highest rating to "useful" aspect (4.61), followed by "fun and engaging" (4.52), "easy to start using" (4.32), and lastly "easy to use" (3.58). One of the comments was "*The tool is very useful. I like its graphics. I can see myself using it. I suggest to have a user manual or help page.*"

The 30–39 age group highly rated the tool in all four aspects. The tool was rated highly for being useful at 4.62, followed by "easy to use" (4.50), "easy to start using" (4.08), and "fun and engaging" (4.00). One of the comments was "*I can use it in my research. I like the way that I can observe different transport measures separately and individually for specific routes. It would be great if the tool allows the user to export the data.*" This suggests that we should consider adding a feature that enables data export for the user's analysis in forms of spreadsheets and images in future development.

Users who were in 40–49 and 50–59 age groups rated the tool quite similarly. They rated highly for the tool being useful (4.39 and 4.45), followed by being easy to start using (3.79 and 4.36), easy to use (3.43 and 3.55), and lastly fun and engaging (3.43 and 3.36). One of the comments was "*I like the tool. It can be useful for my research and teaching in transport design. It's difficult to read some texts. They should be larger for more readability.*" Another comment was "*It would be more useful for research if it could assist in-depth analysis like simulation with different parameters*". This suggests that features for statistical analysis and simulation can improve the tool's usability.

6.4. Occupations

When separated by occupation, there were 24 college professors, 13 researchers, 44 students who were in the transport and information technology related fields (i.e., transportation engineering, urban planning, geographic information systems, computer science, and media technology), as well as 31 transport department officers who were engineers and planners. There were 10 subjects who were transit vehicle drivers, teachers, and non-transport engineers, grouped into the 'other' occupation category. The result is shown in Figure 24. The tool was highly rated across all occupations regarding its usefulness (above 4.0). Its aspect of being easy to use was rated the lowest among other aspects by researchers and students. On the other hand, professors and transport officers rated the "fun and engaging" as the lowest aspect for them. Most comments suggested that there should be a user manual as they felt that the tool although looked easy to use at first, there were unclear selection buttons. Professors, researchers, and students mostly commented about the graphics and usefulness of the tool, while transport officers focused more to the insights gained from the analysis module's results. Staff who were transit planners and transit engineers at the Calgary Transit were particularly interested in the analysis module's results displayed on the two-map mode, comparing GTFS feeds from two different periods. They suggested additional statistical measurement and display for comparative analysis of different GTFS feeds.

7. Conclusion

Public transportation provides an essential service that is especially important in large cities, where increasing vehicular traffic flows continues to a be a major challenge for urban planners. This article presents a new interactive visual analytics tool called *PubtraVis* that makes use of the GTFS data that carries schedule information to measure and display the public transit system operation in different perspectives through six visualization modules: mobility, speed, flow, density, headway, and analysis. Each module displays both dynamic graphics and statistics of the perspective transit operation measure. Mobility module conveys information regarding the movement of transit vehicles in the system. Speed module displays the variation in speeds throughout the system. Flow module demonstrates and varying directions of the traffic flows in the system. Density module displays the traffic density that changes throughout the system. Headway module illustrates the vehicle inter-arrival times at each transit station in the system. Analysis module displays insightful analytical information including top lists of the most crowded and longest waiting time stations, similarity measures between routes, and route clusters based on the transit operational characteristics. PubtraVis is an interactive tool that allows the user to select to view particular services and routes. It also allows the user to compare two modules side by side where services and routes of each module can be viewed and selected separately for comparison purposes. To our knowledge, PubtraVis is the first interactive GTFS data visual analytics tool that measures and displays public transit operation. We believe that our developed tool can be useful for transit system design and planning. The data visualization can be useful for facilitating communication between transit operators, city authorities and the general public regarding public

transit planning and operation. The tool was designed and developed through repeated communication with a local transit agency, transit users, and transportation professionals by focusing on current practices of transport agencies with regard to translating large volume transit related data into insights. The approach in the development and the tool itself are the main contribution of this work.

PubtraVis has been tested with 122 real-world users from which constructive comments and suggestions were received for our future development. The tool was generally found to be useful and easy to start using. However, the "easy to use" aspect needs improvement. We are continuing to improve the tool in different directions among which we will consider the design that potentially includes user instruction manual, sound effects, result interpretation, recommendation, multi-scenario simulation, comparative analysis, and data export. Another area of future improvement is the use of the GTFS-static data combined with the GTFS Realtime data to develop additional visual analytics modules to perform a variety of real-time strategies and actual system performance evaluations: (i) schedule adherence analysis; (ii) headway regularity analysis; and (iii) speed, delay, and dwell time analysis.

Declarations

Author contribution statement

Postsavee Prommaharaj: Performed the experiments; Analyzed and interpreted the data.

Santi Phithakkitnukoon: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Merkebe Getachew Demissie & Lina Kattan: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Carlo Ratti: Conceived and designed the experiments.

Funding statement

This work is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (RGPIN/03923-2014), the Urban Alliance Chair in Transportation Systems Optimization, and the Alberta Innovate Strategic Grant on Integrated Urban Mobility (G2018000894).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e03729.

References

[1] M. Getachew Demissie, "Combining Datasets from Multiple Sources for Urban and Transportation Planning: Emphasis on Cellular Network Data", University of Coimbra, 2014.
[2] N. Ferreira, J. Poco, H.T. Vo, J. Freire, C.T. Silva, Visual exploration of big spatio-temporal urban data: a study of New York city taxi trips, IEEE Trans. Vis. Comput. Graph., 2013.
[3] W. Chen, F. Guo, F.Y. Wang, A survey of traffic data visualization, IEEE Trans. Intell. Transport. Syst. (2015).
[4] N. Andrienko, G. Andrienko, Visual Analytics of Movement: an Overview of Methods, Tools and Procedures, *Information Visualization*, 2013.
[5] G. Andrienko, N. Andrienko, W. Chen, R. Maciejewski, Y. Zhao, Visual analytics of mobility and transportation: state of the art and further research directions, IEEE Trans. Intell. Transport. Syst. (2017).
[6] M.G. Demissie, S. Phithakkitnukoon, L. Kattan, Trip Distribution Modeling Using Mobile Phone Data: Emphasis on Intra-zonal Trips, *IEEE Transactions on Intelligent Transportation Systems*, 2018.
[7] L. Alexander, S. Jiang, M. Murga, M.C. González, Origin-destination trips by purpose and time of day inferred from mobile phone data, Transport. Res. C Emerg. Technol. 58 (2015) 240–250.
[8] M.G. Demissie, S. Phithakkitnukoon, L. Kattan, Understanding human mobility patterns in a developing country using mobile phone data, Data Sci. J. 18 (1) (2019) 1–13.
[9] A. Lacombe, C. Morency, Modeling taxi trip generation using GPS data: the Montreal case, Transp. Res. Board 95th Annu. Meet. (2016).
[10] C. Yang, E. Gonzales, Modeling taxi trip demand by time of day in New York city, Transp. Res. Rec. J. Transp. Res. Board (2014).
[11] J. Tang, S. Zhang, X. Chen, F. Liu, Y. Zou, Taxi trips distribution modeling based on Entropy-Maximizing theory: a case study in Harbin city—China, Phys. A Stat. Mech. its Appl. (2018).
[12] M.G. Demissie, G.H. de A. Correia, C. Bento, Exploring cellular network handover information for urban mobility analysis, J. Transport Geogr. (2013).
[13] M.G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, C. Bento, Inferring passenger travel demand to improve urban mobility in developing countries using cell phone data: a case study of Senegal, IEEE Trans. Intell. Transport. Syst. (2016).
[14] Z. Zhou, et al., Visual abstraction of large scale geospatial origin-destination movement data, *IEEE Trans. Vis. Comput. Graph.*, 2019.
[15] F. Wang, et al., "A Visual Reasoning Approach for Data-Driven Transport Assessment on Urban Roads,", in: 2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014 - Proceedings, 2015.
[16] F. Kamw, et al., Urban structure accessibility modeling and visualization for joint spatiotemporal constraints, IEEE Trans. Intell. Transport. Syst. (2019).
[17] S. Al-Dohuki, Y. Zhao, F. Kamw, J. Yang, X. Ye, W. Chen, QuteVis: visually studying transportation patterns using multi-sketch query of joint traffic situations, IEEE Comput. Graph. Appl. (2019).
[18] Y. Ma, T. Lin, Z. Cao, C. Li, F. Wang, W. Chen, Mobility viewer: an eulerian approach for studying urban crowd flow, IEEE Trans. Intell. Transport. Syst. (2016).
[19] F. Wu, M. Zhu, X. Zhao, Q. Wang, W. Chen, R. Maciejewski, Visualizing the Time-Varying Crowd Mobility, 2015.
[20] F. Wang, W. Chen, Y. Zhao, T. Gu, S. Gao, H. Bao, Adaptively exploring population mobility patterns in flow visualization, IEEE Trans. Intell. Transport. Syst. (2017).
[21] D. Li, Y. Lin, X. Zhao, H. Song, N. Zou, Estimating a Transit Passenger Trip Origin-Destination Matrix Using Automatic Fare Collection System, 2011.
[22] Y. Ji, R.G. Mishalani, M.R. McCord, Transit passenger origin-destination flow estimation: efficiently combining onboard survey and large automatic passenger count datasets, Transport. Res. C Emerg. Technol. 58 (B) (2015) 178–192.
[23] E. Mazloumi, G. Currie, G. Rose, Using GPS data to gain insight into public transport travel time variability, J. Transport. Eng. 136 (7) (2009) 623–631.
[24] G. Guido, A. Vitale, D. Rogano, Assessing public transport reliability of services connecting the major airport of a low density region by using AVL and GIS technologies, in: EEEIC 2016 - International Conference on Environment and Electrical Engineering, 2016.
[25] M. Berkow, A.M. El-Geneidy, R.L. Bertini, D. Crout, "Beyond generating transit performance Measures : visualizations and statistical analysis using historical data, Transp. Res. Rec. J. Transp. Res. Board (2009).
[26] T.B. Glick, W. Feng, R.L. Bertini, M.A. Figliozzi, Exploring applications of second-generation archived transit data for estimating performance measures and arterial travel speeds, Transp. Res. Rec. J. Transp. Res. Board (2015).
[27] M. Mesbah, G. Currie, C. Lennon, T. Northcott, Spatial and temporal visualization of transit operations performance data at a network level, J. Transport Geogr. (2012).
[28] X. Ma, Y. Wang, Development of a data-driven platform for transit performance measures using smart card and GPS data, J. Transport. Eng. (2014).
[29] A. Kurkcu, F. Miranda, K. Ozbay, C.T. Silva, Data visualization tool for monitoring transit operation and performance, in: 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings, 2017.
[30] A. Anwar, A. Odoni, N. Toh, BusViz: big data for bus fleets, Transp. Res. Rec. J. Transp. Res. Board (2016).
[31] H. Bast, P. Brosi, S. Storandt, Real-time movement visualization of public transit data, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2014, pp. 331–340.
[32] C. Stewart, E. Diab, R. Bertini, A. El-Geneidy, Perspectives on transit: potential benefits of visualizing transit data, Transp. Res. Rec. J. Transp. Res. Board 2544 (1) (2016) 90–101.
[33] N. Kunama, M. Worapan, S. Phithakkitnukoon, M. Demissie, GTFS-VIZ: Tool for preprocessing and visualizing GTFS data, in: UbiComp/ISWC 2017 - Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers, 2017.
[34] K. Kloeckl, X. Chen, C. Sommer, C. Ratti, A. Biderman, Trains of data [Online]. Available: http://senseable.mit.edu/trainsofdata/, 2011. (Accessed 12 June 2018).
[35] M. B. and B. Card, "Boston's Massachusetts Bay Transit Authority (MBTA). .
[36] Statistics Canada, Alberta population projections [Online]. Available: https://open.alberta.ca/dataset/5336155/resource/971cb905-80fb-4f1b-9132-57ae88676151, 2015. (Accessed 2 January 2018).
[37] Calgary Transit, West LRT one year review [Online]. Available: https://www.calgarytransit.com/sites/default/files/reports/west_lrt_one_year_review.pdf, 2014. (Accessed 1 February 2018).
[38] Calgary transit, "statistics for 2017 [Online]. Available: https://www.calgarytransit.com/about-us/facts-and-figures/statistics, 2017. (Accessed 20 January 2018).
[39] B. McHugh, "Pioneering Open Data Standards: the GTFS Story," *beyond Transparency - Open Data And Future Of Civic Innovation*, 2013.
[40] S. Williams, A. White, P. Waiganjo, D. Orwa, J. Klopp, "The digital matatu project: using cell phones to create an open source data for Nairobi's semi-formal bus system, J. Transport Geogr. 49 (2015) 39–51.

[41] Google Developers, Google transit APIs [Online]. Available: https://developers.google.com/transit, 2016. (Accessed 20 January 2016).
[42] K. Karato, N. Sato, T. Hatta, The speed–density relationship: road traffic flow analysis with spatial panel data [Online]. Available: www.shiratori.riec.tohoku.ac.jp/~takita/ARSC2009/Paper/ARSC2009_58.pdf, 2009. (Accessed 3 March 2018).
[43] Transportation Research Board, Transit Capacity and Quality of Service Manual, second ed. [Online]. Available: http://onlinepubs.trb.org/onlinepubs/tcrp/docs/tcrp100/Part3.pdf, 2017. (Accessed 13 April 2018).
[44] M. Batarce, J. Muñoz, J. de D. Ortúzar, S. Raveau, C. Mojica, R.A. Ríos, Valuing crowding in public transport systems using mixed stated/revealed preferences data: the case of santiago, in: Transp. Res. Board 94th Annu. Meet., 2015.
[45] M. Yap, O. Cats, B. van Arem, Crowding valuation in urban tram and bus transportation based on smart card data, Transp. A Transp. Sci. (2018).
[46] O. Cats, J. West, J. Eliasson, A dynamic stochastic model for evaluating congestion and crowding effects in transit systems, Transp. Res. Part B Methodol. (2016).
[47] J. Gibson, I. Baeza, L. Willumsen, Bus-stops, Congestion and Congested Bus-Stops, *Traffic Eng. Control*, 1989.
[48] G. de B, C.R.G. Teresa Cristóbal, Gabino padrón, alexis quesada-arencibia , francisco alayón, "bus travel time prediction model based on profile similarity, Sensors 19 (13) (2019).
[49] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, J. Freire De Sousa, Validating the Coverage of Bus Schedules: A Machine Learning Approach, *Inf. Sci. (Ny).*, 2015.
[50] J. Khiari, L. Moreira-Matias, V. Cerqueira, O. Cats, Automated setting of bus schedule coverage using unsupervised machine learning, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2016.
[51] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, 1973.
[52] R.A. Kadir, Y. Shima, R. Sulaiman, F. Ali, Clustering of public transport operation using K-means,, in: IEEE 3rd International Conference on Big Data Analysis, ICBDA 2018), 2018, pp. 427–432.
[53] X. Fei, O. Gkountouna, Spatiotemporal clustering in urban transportation: a bus route case study in Washington D.C, SIGSPATIAL Spec. 10 (2) (2018) 26–33.
[54] D. Pelleg, A.W. Moore, X-means: extending K-means with efficient estimation of the number of clusters, Proc. Seventeenth Int. Conf. Mach. Learn (2000) table contents.
[55] R.E. Kass, L. Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion, J. Am. Stat. Assoc. (1995).
[56] F. Guo, More than asability: the four elements of user experience, part IV [Online]. Available: http://www.uxmatters.com/mt/archives/2012/04/more-than-usability-the-four-elements-of-user-experience-part-i.php, 2012. (Accessed 21 December 2017).
[57] O.C. Robinson, Sampling in interview-based qualitative research: a theoretical and practical guide, Qual. Res. Psychol. (2014).