MetroViz: Visual Analysis of Public Transportation Data

Fan Du, Joshua Brul, Peter Enns, Varun Manjunatha and Yoav Segev

Abstract—Understanding the quality and usage of public transportation resources is important for schedule optimization and resource allocation. Ridership and adherence are two main dimensions for evaluating the quality of service [1, 5]. Using Automatic Vehicle Location (AVL), Automatic Passenger Count (APC), and Global Positioning System (GPS) data, ridership data and adherence data of public transportation can be collected. In this paper, we discuss the development of a visualization tool for exploring public transportation data. We present the map view and the route view to help users locate stops in the context of geography and routes information. To convey ridership and adherence information over years, we present a calendar view inspired by the calendar we use in our daily life. The calendar provides an overview of the data, and users can interactively select specific days to explore the trip details in the trip and stop components of the calendar view. In the usability testing, four researchers from the Center for Advanced Transportation Technology (CATT), and six Computer Science PhD students from the University of Maryland took part in a study in which they used our tool to explore the bus transit data of Blacksburg, Virginia. The participants were able to use our tool to explore bus stops, bus routes, ridership and adherence data of bus trips.

Index Terms—Public transportation, visual analysis, spatio-temporal data

1 INTRODUCTION

Public transportation modes follow scheduled routes and timetables, and allow passengers to board and alight at specific stops or stations. Most research regarding AVL, APC and GPS data focus on developing algorithms to predict the travel time [6, 13, 10], to automatically optimize the bus scheduling [3, 16], and to estimate the quality of service [1, 5]. To the best of our knowledge, there is no existing visualization tool that could aid service providers to interactively explore those data and gain insights using the advantages of their domain knowledge.

There are three technical challenges that we believe are critical to visually analyze the ridership and adherence data of public transportation over time. The first challenge is how to design a general data model that can be applied to different public transportation modes. The second challenge is how to design the interface to show the stops / stations, routes, trips, ridership and adherence simultaneously. The third challenge is how to process the large-scale time series data and develop an interactive visual analysis tool.

To address these challenges, we have developed MetroViz, an interactive visual analysis tool that helps users explore public transportation data, and evaluate how the public transportation is serving the public. We model the public transportation data into three levels: 1) the stop / station level, 2) the route level and 3) the trip level. As most public transportation modes follow scheduled routes and timetables, and allow passengers to board and alight at specific stops or stations, our model as well as our visual analysis tool can be applied to different public transportation modes. Based on the three level model, our visual design consists of three views, including the map view, the route view and calendar view. We present the map view and the route to help users locate stops in the context of geography and routes information. To convey ridership and adherence information over years, we present...
a calendar view inspired by the calendar we use in our daily life. The calendar provides an overview of the data, and users can interactively select specific days to explore the trip details in the trip and stop components of the calendar view. Benefiting from the overview-to-detail design of the three views, our tool iteratively and progressively visualizes a small portion of the large-scale data at a time, which accelerates the query response and enables users to explore the data interactively.

In the usability testing, four researchers from the Center for Advanced Transportation Technology (CATT), and six Computer Science PhD students from the University of Maryland took part in a study in which they used our tool to explore the bus transit data of Blacksburg, Virginia. The participants were able to use our tool to explore bus stops, bus routes, ridership and adherence data of bus trips.

To the best of our knowledge, our work is the first to help users visually analyze public transportation data, as well as to allow users to interactively explore the large-scale time series ridership and adherence data over time. Our work presents three technical contributions.

- A general data model that divides the public transportation data into three levels. This model can be applied to different public transportation modes.
- An overview-to-detail interface consists of three views that visualize stops/stations, routes, trips, ridership and adherence data simultaneously.
- An iterative and progressive visual analysis that accelerates the query response and enables users to explore the data interactively.

Following a review of related work, we describe in detail MetroViz’s system design, data processing and visual design of the overview-to-detail interface followed by a description of the usability testing to evaluate the effectiveness of our design.

2 RELATED WORKS

2.1 Analysis of Public Transportation Data

Automatic Vehicle Location (AVL), Automatic Passenger Count (APC) and Global Positioning System (GPS) have been used to collect ridership and adherence data—the two main factors for evaluating the quality of service. Rancic et al. [11] present a tracking system using AVL data to analyze city bus transit traffic. Chen [2] introduces a model to simulate bus operation and passenger demand based on AVL and APC data. Camus et al. [1] propose a way to estimation the quality of service based on AVL data. Hammerle et al. [5] analyze the AVL and APC data of Chicago Transit Authority to estimate service reliability. Mai et al. [8] extend the Marey graph by adding schedule adherence and passenger load information to measure transit performance. Kimpel et al. [3] discuss efforts of using the TriMet APC and AVL data to improve buses scheduling. Yu and Yang [16] develop a dynamic holding strategy to optimize the holding strategy.

Public transportation data can also be used to predict the travel time and optimize the route choosing. Lee et al. [6] introduce a real-time travel time prediction method based on multiple samples of similar historical trajectory. Tiesyte and Jensen [13] propose a method to predict the future movement of a vehicle based on the identification of the most similar historical trajectory. Predic et al. [10] use real-time AVL data and historical data to predict bus motion and bus arrival time. Nguyen et al. [9] consider buses as moving objects, and use temporal maps to represent the movements of buses in spatio-temporal domain to help passengers choose appropriate routes. Liu et al. [7] propose a bus trip planning system to help passengers choose the most appropriate lines and transfers, based on traffic data.

However, most of the research regarding AVL, APC and GPS data focus on developing algorithms to predict the travel time, to automatically optimize the bus scheduling, and to estimate the quality of service. To the best of our knowledge, there is no existing visual analysis tool that could aid service providers to interactively explore those data and gain insights using the advantages of their domain knowledge.

2.2 Spatio-temporal Data Visualization

Transportation data is a kind of spatio-temporal data. Public transportation modes follow scheduled routes and timetables, and allow passengers to board and alight at specific stops or stations. A lot of research regarding spatio-temporal data visualization focus on trajectory drawing. Tominski et al. [14] use a hybrid 2D/3D display to show the trajectories and associated attributes in their spatio-temporal context. Scheepens et al. [12] improve density maps to help explore trajectories using multiple density fields. Some research use graphs to visualize the spatio-temporal data. Cui et al. [3] propose an edge-clustering method to reduce edge crossings and visualize geometry graphs. Guo [4] develops a visualization framework to interactively explore large-scale spatio flows. Wang and Chi [15] introduce a focus+context method to visualize metro map on small displaying area of mobile devices.

However, trajectories cannot effectively show data in both the spatio-temporal context and the ridership-adherence context simultaneously. In our design, we choose to use an overview-to-detail interface to show stops/stations, routes and trips information on different levels of visualization.

3 SYSTEM OVERVIEW

MetroViz is organized into three primary views: map view (Fig. 1), route view (Fig. 1) and calendar view (Fig. 2). Those views are designed visualize several dimensions of public transportation datasets at important levels of detail, and allow the user to zoom in on the data they wish to explore and filter out extraneous details.

To begin, the user is presented with a broad overview of the transportation system’s organization in space, and routes with two coordinated views (map and route). In the map view, stops are visualized as blue circles superimposed on a Google map. Mousing over a stop highlights it on both the map and route views and displays the distribution of fare types collected at the stop in a donut glyph around the stop. The route view also displays stops as blue circles. For each route, stops are displayed in linear order that they occur in the actual bus route, followed by the route’s name. Mousing over a stop triggers the same response that it would in the map view. Mousing over a route name highlights every stop in the route in both views.

Route view provides two controls to filter the data in desirable ways. Clicking on a stop displays all of the data collected at that stop in the calendar view. Similarly, clicking on a route’s name displays all of the data collected on a specific route in the calendar view. Calendars in the calendar view are organized in a familiar format. The calendar displays either adherence or ridership for every day where data was collected. The type of data displayed on the calendar can be controlled with a drop down menu on the control panel.

The user may drill down even further from the calendar by clicking on a specific day. If the calendar was generated by clicking on a stop, the stop component (Fig. 2) of the calendar view is displayed. Similarly, if the calendar was generated by clicking on a route, the trip component (Fig. 4) of the calendar view is displayed. Both of these views provide very fine grained details about each stop or route throughout the day.

4 DATA PROCESSING

MetroViz is a visual analysis tool designed to visualize public transportation data. In the development, we use bus transit data of Blacksburg, Virginia. The data was collected and given to us by the Blacksburg Department of Transportation. The data consists of three files; each file includes about three years of data from mid 2010 to mid 2013. Uncompressed, the files totaled over three gigabytes worth of data.

The first file consists of the buses adherence to schedule. For every bus route, for every trip on that route, and for every stop on that trip, it shows the scheduled departure time from that stop, as well as the actual departure time from that stop, and the difference in seconds between them.

The second file has data about the number of passengers on the bus at each stop. For every trip and stop, there are data for the number...
of passengers buying a ticket (collected by the ticketing machines installed on the bus), the number of passengers boarding and the number of passenger alighting the bus (both collected by an automatic counting device, and generally less reliable than the fare count).

The third file is a breakdown of the fare counts into categories. Categories include but not limited to: student fare, faculty/staff, full fare, etc.

The first step we took in processing the data, was to write scripts that filter out erroneous records, such as partial records, or records that are inconsistent in any way. After that we used another script to load the files into a SQLite3 database, and use batch SQL queries to align the data of the first file to that of the second file (passengers count, and adherence data). We also loaded the third file (fares breakdown) to the database, in order to easily run aggregated statistics. Once the data was ready and quality assured, we exported it to a different set of tables used in the visualization process.

In order to supply the data to the visualization utility, we wrote a simple python based web server using the CherryPy framework that returns JSON formatted data to several types of HTTP requests with a given set of parameters (such as data types, dates, routes, etc.)

5 VISUAL DESIGN

5.1 Map View and Route View

Map view (Fig. 1) (built on top of Google Maps) is designed to provide an overview the entire Blacksburg bus transit system. Every bus stop in the system appears as a blue dot. Hovering over a stop highlights it in route view (discussed below) and brings up a donut chart summarizing the fare types at that stop.

Map view also includes a search feature to search stops by name. Based on requests from the initial usability study, we added drop-down auto-completion to the search box. Entering a stop will pan/zoom the map to center on the appropriate stop.

The route view (Fig. 1), placed immediately below the map view, is designed to provide a concise overview of all bus stops in the Blacksburg bus transit system. Each stop is represented by a small blue circle, and a single route is represented by an array of touching stops. The name of the route is located to the right of the last stop. The Blacksburg bus transit system has 25 active routes which cannot all be displayed without the use of a scroll-bar. The scroll-bar ensures that the user has a view of any route of his choice, as well as a view of the map.

The initial design of route view had each stop circle increase in size on mouse-over. During the usability test, several participants requested highlighting instead. In the current version, the stop receives a red highlight and appears as the center of focus in the map view with a red highlight and fare information. The numeric position of the stop in the route, and the name of the stop appear as tooltips. If several routes share a stop, they all receive red highlights with position and name tooltips, in their respective routes.

When the user clicks the blue circle representing the individual stop, the calendar view for that particular stop (see section 5.2) is produced to the right of the route view.

When a user clicks the name of the route, the calendar view for that route is generated to the right of the route view.

5.2 Calendar View

The calendar view (Fig. 3) is designed to leverage a coarse visualization of time that most users are familiar with. A year in the calendar view is displayed as a grid with seven columns and enough rows to display 12 months where each cell represents a day. Months are delimited with thick black lines. Multiple years are tiled side by side.

Adherence and ridership are visualized in similar ways. For adherence, the color of a day is determined by the average of the absolute value of the difference between the scheduled arrival time and the actual arrival time. For ridership, the color for each day is determined...
by the average number of passengers who boarded on that day. Less extreme values on each scale (buses arriving close to their scheduled time and low number of passengers) are more blue, and more extreme values (buses arriving very late or early and high numbers of passengers) are more red. Days with missing data are left blank.

In the initial usability study, calendar view displayed years horizontally. Several users complained that this counterintuitive and at their suggestion, calendar view was redesigned vertically. Also brought to our attention was the fact that the original color scheme (green to red) was arguably more intuitive, so it would be unacceptable for an individual who was color blind. Additionally, the original color scheme had too many colors, which made it hard to distinguish one from another. We corrected this by changing the color scheme to use a spectrum safe for individuals who are colorblind with only 5 colors.

5.2.1 Trip Component

Trip component (Fig. 4) of calendar view visualizes adherence and ridership for each trip of a route that occurred on a specific day. Data is displayed on a rectangular grid with rows representing trips and columns representing stops. Each cell contains n squares where n is the number of passengers who boarded at that stop on that trip. The squares are colored according to adherence. Adherence is mapped to a spectrum from dark blue (bus arrived on time) to dark red (bus was very early or late). One of the most important goals of MetroViz was to be able to display the stops, routes, and trips where the most passengers were affected by a bus with poor adherence. Trip component achieves this by attracting the user’s eye to large red areas on the grid.

Originally, the trip component displayed stops as circles with different sizes. Our first usability study revealed this to be unsuitable because it was difficult to compare the sizes of circles. Additionally, the original color scheme was inconsistent with the rest of the system and hard to use because it was a continuous spectrum. We addressed this problem by using the same color scheme as the calendar component and displaying the passengers as discrete quantities. Unfortunately, even the improved trip component suffers from the quality of the Blacksburg bus system dataset. Many stops report no passengers boarding while others report unrealistic numbers of passengers boarding (sometimes as high as 160).

5.2.2 Stop Component

Stop component (Fig. 5) of calendar view visualizes ridership and adherence of a single stop for a single day. One of the major challenges of stop component was displaying adherence data in a sensible and intuitive fashion at this level of detail. Unlike the other high-level views, we wanted to capture the distinction between earliness and lateness (a bus leaving 5 minutes earlier than expected may result in a significantly longer wait for the next bus; a bus leaving 5 minutes later than expected is only a 5 minutes wait.) This ruled out the strategy of taking the sum of the absolute value of the deltas between scheduled and actual arrival time (as we did for calendar view and trip component). Averaging the deltas is also unsuitable - a bus that was 10 minutes early and another bus that was 10 minutes late would average out to be exactly on time. In addition, we wanted a display that would simultaneously show the number of passenger boardings to visualize the number of people affected by non-adherence.

Stop component went through a number of iterations. The initial design showed number of boardings and either earliness (in green) or lateness (in red) for every stop that day at a particular location. A variant of this design used the same color scheme, but bucketed stops by hour (the values shown for each hour are the average values for all stops that occurred during that hour) for a more predictable use of space and consistent spacing. Bucketing also had the effect of sometimes showing early (green) and late (red) bars for the same hour.

In the initial usability tests, we also tested versions of the above bar charts with adherence and ridership displayed on separate charts. In addition, there were line chart versions of all of the bar charts.

Most of the initial usability test participants disliked the design choices in the initial bar charts and the line chart versions were universally disliked. The test participants also universally preferred bucketing the stops by hour instead of displaying every stop individually.

The most common complaints were a dislike of different types of data (number of passengers and minutes early / late) displayed on the same scale and the display of earliness and lateness as positive values on the chart.

We had several suggestions that earliness be displayed as a negative value and lateness be displayed as a positive one. Our first re-design kept the green and red bars separate and simply mirrored the earliness bar over the x-axis. Later, we realized that we could reduce space by stacking the early and late bars on top of each other and with a clear origin line, display the stacked bars as a single color.

This display strategy also has the benefit that the total length of each bar is the same length of a bar displaying the sum of the absolute value
of the deltas between scheduled and actual arrival time which helps
maintains consistency with calendar and trip component’s adherence
ranking system. A ridership chart is also shown directly below the
adherence chart to help visualize the number of people affected by
non-adherence.

Finally, we also investigated the possibility of using radar charts
to display population or adherence levels as stop time naturally mapped
to a clock-like display. However, such a chart makes comparisons
difficult and in usability tests was generally rejected in favor of the
simple bar chart.

6 Evaluation
6.1 Methodology
We decided that the best method for evaluating the performance of
the MetroViz system was to conduct Usability Testing on several test
subjects. There were no hypotheses to test, nor could we define dependent
or independent variables, so a controlled experiment would not be
suitable. Although domain experts were amongst our test subjects, the
short duration of this project meant that a long term case study was
infeasible. Usability Testing is a frequently used technique to evaluate
user’s efficiency in utilizing a tool. A Usability Test involves a training
session followed by a set of pre-designed tasks that the test subject
is asked to perform. We encourage participants to “think out loud”
and give verbal feedback which we record during the test session. We
observe if the participant has any undue difficulty in performing the
task. Many of the participants we tested were extremely enthusiastic
in providing suggestions and feedback.

Our usability test had four distinct steps. First, we gave a demonstra-
tion of MetroViz individually to the subject. Then, we collected
some information from the participant - the goal here was to find out
if they had any prior domain knowledge on visualizing transportation
data. After this, we gave the subject a set of tasks with seven different
themes. We would assist the users verbally if they had extreme diffi-
culty in performing a task. Fourthly and finally, we subjected the tester
data. After this, we gave the subject a set of tasks with seven different
themes. We also sought opinions on the color schemes we used, and whether
there were any features that the users would like added to the system.
We also took note of any part of the system that the user found exces-
sively difficult or frustrating.

The list of tasks for Usability Testing and the feedback question-
naire and provided in the appendix.

Initial usability tests consisted of three members of the CATT lab.
Follow-up usability tests are pending.

6.2 Results
We gained some valuable insights upon performing two rounds of us-
ability testing. The first round of usability testing was performed with
three members of the CATT lab.

The second round of usability testing, unlike the previous round
was performed with both domain experts and non-experts. Our test
subjects concurred on several points of feedback and this will be valu-
able for future revisions to our software. The calendar view scored
high on the Likert Scale, but opinions were divided on the other views.
Some users liked the color scheme (red for more passengers and non-
adherence, blue for fewer passengers and adherence), while others
preferred a different color scheme, or a reversal of the color scheme.
Some of the suggestions are trivial to implement, but other recommenda-
dations, although simple in concept, are tricky to implement.

Some of the recommendations made were : R1: Disable hovers
upon clicking any of the stops in the route view and map view, until
the user clicks on an empty space.

R2: Disable clicking while the data is loading.

R3: The trip component is extremely large, especially if a partic-
ular route has many trips per day, necessitating a lot of up and down
scrolling. We could divide the trip component into morning, after-
noon, evening and night trips to reduce the size of the trip component.
Similarly, if the route had several stops (for example, one of the routes
had as many as 80 stops), a lot of left-right scrolling is needed.

R4: Provide month and day-of-the-week labels in the calendar view.

R5: Ensure all graphs and plots have titles and axis labels.
7 Discussion and Future Work

Although 3 gigabytes is not exceptionally large, MetroViz suffers from many of the problems that plague "Big Data" - long load times and incomplete and inaccurate data.

Calendar view handles missing data very well - a day without adherence or stop data is displayed as an unobtrusive blank square. Map view virtually ignores missing data - as along as a single fare count by fare type is available, the donut charts for each stop can be rendered.

Unfortunately, trip component does not handle missing data as well - during usage we often ran into problems where selecting a single day would bring up a very large number of trips - but with no ridership data available resulting in a large amount of wasted space.

Despite stacking the early and late bars, stop component charts required a significant amount of space to get precise readings on adherence and ridership levels - this made simultaneously comparing multiple days impractical on standard screen sizes. Either larger screen sizes or dynamically resizable charts would be needed to mitigate this shortcoming. However, the ability to visualize earliness, lateness and total adherence in a single chart may be worth the cost of extra screen space.

The long load times are somewhat unavoidable - calendar view requires the complete dataset for a stop / route over the entire time span available and preloading data would result in an excessively long startup time. However, a faster server backend may help mitigate some of the problem.

Aggregate selection in calendar view (multiple days / months) and more comprehensive filters would allow for more precise and faster exploration of the data.

Finally, higher quality data would be ideal and would allow for a better evaluation of trip component. Map view, route view and stop component may not be ideally implemented but work together reasonably well. Trip component requires further study.

References