# VISUALIZING BUS SCHEDULE ADHERENCE AND PASSENGER LOAD THROUGH MAREY GRAPHS

**Eric Mai**
**Analyst, Berkeley Transportation Systems**
**2150 Shattuck Ave.**
**Suite #200**
**Berkeley, CA 94704-1345**
**+1-510-984-1474, mai@bt-systems.com**

**Mark Backman**
**Senior Software Engineer, Berkeley Transportation Systems**
**2150 Shattuck Ave.**
**Suite #200**
**Berkeley, CA 94704-1345**
**+1-510-984-1469, mark@bt-systems.com**

**Rob Hranac**
**Vice President, Berkeley Transportation Systems**
**2150 Shattuck Ave.**
**Suite #200**
**Berkeley, CA 94704-1345**
**+1-510-290-5496, rob@bt-systems.com**

## ABSTRACT

The original Marey graph, published in 1885, has become a frequent example of innovative design in data visualization. It plots a French train timetable on a time-space diagram, intuitively depicting the paths of trains throughout the day. These graphs continue to be used in transit-related applications such as the Google Transit Feed Specification (GTFS) Schedule Viewer. This paper repurposes the original Marey graph for use in transit performance measurement by adding schedule adherence and passenger load information. APC data preprocessing steps are described and technological issues related to the development of the visualization are discussed. Finally, this paper demonstrates how the Marey graph enables quick visual identification of vehicle performance trends across space and time.

*Keywords: transit performance measurement, performance measurement system, visualization, transit data, APC, GTFS, Marey graph*

# INTRODUCTION

Due to the growing ubiquity of GPS technology, data on transit vehicle performance continues to become more common. However, even with the use of aggregate measures of performance, this data can easily become overwhelming without clear and effective means of visualization. The Marey graph is a useful tool for distilling the schedule of a transit system into a single image, however it is still primarily used for viewing schedules, not measures of performance.

The first Marey graph was published in 1885 as a visual depiction of train schedules (see Figure 1) [1]. This view of the transit system is particularly valuable because it is both data-rich and unified in presentation. It is data-rich in that many attributes of the system are depicted at once: vehicle speeds, dwell times, directions of travel, service frequency, and stop spacing (to name a few). It is unified in the sense that all of this data is encoded into the features of the lines, making it all available in one location. These characteristics make Marey graphs an efficient and valuable tool in transit planning and operations activities.

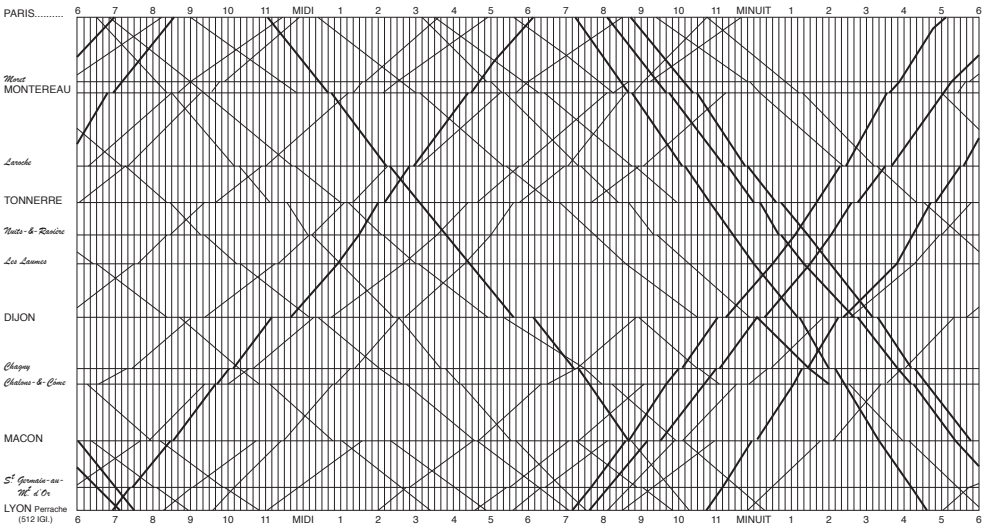


**Figure 1: Original Marey graph depicting a French train schedule [1]**

This paper applies the ideas presented previously in [2] to build Marey graphs that augment the original by depicting delivered service and passenger load in addition to transit schedule data. By plotting these three types of data concurrently on the same graph, it is possible to explore several lines of inquiry on the state of the transit system through a single visualization.

For each transit run, two lines are drawn: one representing the scheduled service (as in Figure 1), and a second depicting the delivered service (based on APC data). On long time scales, these two lines can be very close together making their order difficult to discern. Thus, in order to tell more easily whether the bus was early or late, the area between the lines is shaded black for late and white for early. Passenger load is shown on both sides in gray as a shadow that increases in width proportional to the number of passengers carried.

APC data from buses on San Diego's #2 (South) route on the morning of Monday, August 2, 2010 was used to construct such a graph. Referring to Figure 2, the following information about the performance of the vehicles can be seen:

- *Trends over time:* Of the five runs shown, the earliest and latest buses had the best on-time performance in general.
- *Trends over space:* The Northern part of the route, before the buses reach Juniper Street, consistently saw lower passenger loads and better schedule adherence than the rest of the route.
- *Trends between runs:* The first and second buses started the route much closer together than they finished. Around I-5, the first bus got ahead of the schedule at the same time as the second bus fell significantly behind.
- *Trends between measures:* As the distance between the first and second buses grew, the second bus began carrying a larger passenger load, falling further behind schedule around the same time.

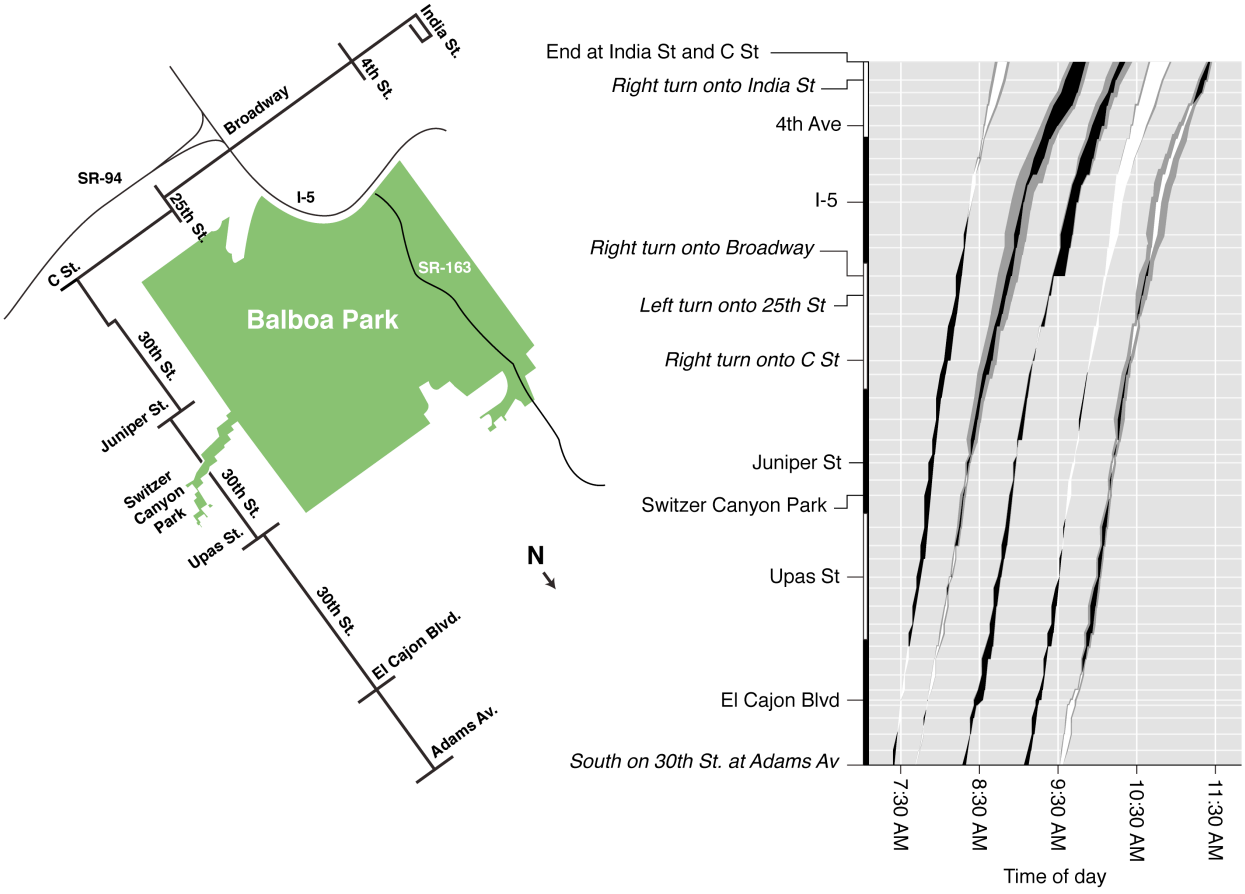


**Figure 2: Example of a Marey graph with APC data integrated (these runs were shifted in time for display purposes)**

# LITERATURE REVIEW

The Marey graph has two specific characteristics that make it an excellent conveyance for information. One is its ability to allow viewers to use their visual operators instead of more demanding logical operators [4]. For example, the time a train spent waiting at a station can be approximated based on the length of a line as opposed to having to subtract time values from a table. Secondly, the Marey graph streamlines the search for information, allowing users to immediately identify trains serving a particular route, or compare speeds between trains without needing to consult a table of arrival times.

The Marey graph is a good candidate to be reimagined using current display technologies. A previous study demonstrated this by combining GIS tools with a Marey graph to visually represent the locations of potential train collisions [5]. In this example, the Marey graphs were modified by techniques such as contour shading, 3D effects, and banding of time and space ranges to highlight data trends.

Other studies envision the potential applications of futuristic Marey graphs. In a paper describing an interactive hardware-software framework to physically couple modeling with real-time simulation, the manipulation of a Marey graph by transit operators was used as a central example to demonstrate the potential of dynamic, interactive visualizations [6]. This paper imagined a transit management center in which Marey graphs were manipulated directly to explore the effects of changes in service.

A paper by Dix and Ellis is similarly enthusiastic about the potential for interactivity in Marey graphs [7]. In describing the power of interactive visualizations, Marey graphs are imagined with interactive highlighting to allow particular lines in busy graphs to stand out [7]. The inspiration comes from Marey's original train schedule graphic, where certain lines were drawn thicker to aid the viewer in tracing them across overlapping routes (see Figure 1).

# DESIGNING MAREY GRAPHS WITH APC DATA

In considering how best to add APC data to Marey graphs, the means of distribution of the visualizations must first be considered. While physical performance reports are still common, the use of computer-based visualization tools has become widespread. For this reason, any new data visualization effort should take the capabilities of digital media into account. Computers enable direct interaction with the visualization, allowing users to explore the data themselves. In a printed visualization, the scale and range at which to display the data must be fixed. However, using computers, it is possible for users to scroll through the data continuously, zoom and pan smoothly without reloading, highlight individual data points, and turn layers of data on and off. This technology is ideal for the large amount of information conveyed in the Marey graph, while printed Marey graphs can still be valuable for displaying smaller ranges of data.

Of course, as more information is added to the graphs they can become cluttered. The graphs must be designed carefully; otherwise the line representing the vehicle's path can easily become confused with that representing the schedule, especially on long time scales. This means that for

small schedule deviations it may be impossible to tell whether the bus was behind or ahead of schedule. As more data is presented at once, the lines become further compressed, hiding longer schedule deviations. Furthermore, previous efforts [2] have used the space between the APC and schedule lines to display passenger load (in grayscale), however this space disappears as the lines converge.

A strategy to address these challenges follows. This technique involves moving the passenger load information from the area between the lines to a variable width "shadow" outside of the lines (see Figure 3). This has the following effects:

1. It allows for the area between the lines to be used for a different purpose. To clarify whether or not a bus is ahead or behind schedule, this area is repurposed as either being black or white. If the bus is behind schedule, the area is black (and the line representing the bus is on the right). If it is ahead of schedule, the area is white (and the line representing the bus is on the left).
2. The passenger load and schedule adherence become more closely related visually. A thicker border when more passengers were on the bus magnifies the visual impact of large schedule deviations. Thus, times when buses carrying many passengers were very late or early will stand out.
3. This means that the graph is now color-coded into three regions: (1) ahead of schedule, (2) behind schedule, and (3) passenger load (in gray). This allows trends across time and space of each of these measures to be more easily seen. A row, column, or cluster of white, black, or gray (representing times, locations, or time-dependent locations prone to early service, late service, or large loads) is easy to recognize visually.

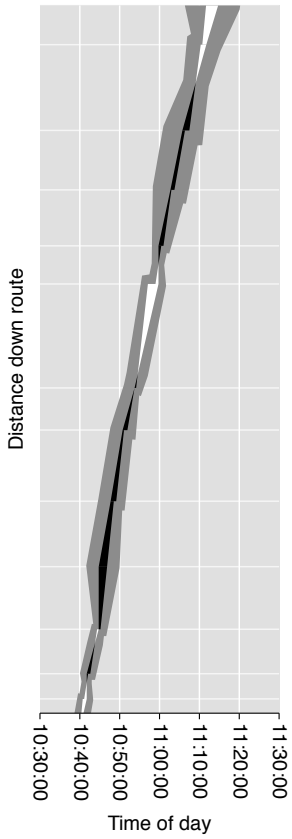


**Figure 3: Example Marey graph with borders indicating passenger load**

The result of applying this technique to San Diego APC data can be seen in Figure 3, a magnification of the fifth bus in Figure 2. The bus enters the figure slightly ahead of schedule, with a moderate passenger load. It remains close to the schedule until the third stop shown when its dwell time pushes it behind schedule for the following four stops. It also picks up more passengers during this stretch, particularly at the fourth stop. By the seventh stop, the bus is back on time. It gets ahead of schedule at the eighth stop, where it also lightens its passenger load some. The bus carried its largest passenger load around the tenth and eleventh stops where it was slightly behind schedule.

# DATA PREPARATION

Before the Marey graphs can be generated, the APC data needs to be preprocessed to fill in gaps. The APC devices used here work in such a way that a data point is created only when the transit vehicle opens its doors. It is common practice for operators of transit vehicles to drive past stops

where no one is waiting and no passengers have requested a stop, leading to missing data. Gaps in the data are overcome by imputing the missing data points. In this case, imputation is based on the available data surrounding the missing point. Missing values are filled in based on a weighting formula which considers the vehicle's schedule adherence immediately before and after the gap. There are definite bounds on the effects of imputation. For example, the travel time between two stops for which data exists will be accurate regardless of the percentage of stops in between which are imputed.

Other data issues include incomplete records due to buses completing only a portion of their scheduled route. Records such as these are discarded. Additionally, the measured passenger count may be inaccurate at times. In most cases, passenger movements are captured by dual light beams, which measure interruptions to detect passengers and their direction of movement (boarding or alighting). When this measurement system malfunctions, the passenger count may get stuck or become unreasonably large.

The goal of this study is to develop a system that automates the generation of Marey graphs such as those shown in Figures 2 and 3 from schedule data (in GTFS format) and APC data *for multiple runs matching a user-specified route*. To facilitate this functionality, precise rules for which runs can be shown on the same graph must be established. This is done through the concept of service patterns [2]. A service pattern is a grouping of trips that share the exact same stops in the same order. Many transit routes (the level of abstraction at which passengers typically interact with a transit system) have multiple service patterns (e.g., express patterns, alternate termination patterns, etc.). This distinction between service pattern and route is essential as attempting to compare individual runs that follow different service patterns on the same Marey graph can lead to mismatches on the vertical axis.

To prepare the data to be plotted, a three step process is followed: (1) Gaps in the APC data are imputed and incomplete records are discarded, (2) The corresponding GTFS data is processed to extract service patterns and other information (such as distance between stops) from it, and (3) The measured APC trajectories are labeled with the service pattern that they belong to. GTFS data is organized by route and trip, and APC data may only contain location, timestamp, vehicle ID, and passenger count, so pairing APC data and schedule data is nontrivial [2].

# PRACTICAL CONSTRUCTION ISSUES

A fine level of detail is required to design this visualization for an environment in which users can zoom in closely without loss of resolution or data distortion. Since the schedule adherence is plotted simply from the data as in other plots, we turn our focus to the borders representing passenger load. The borders are not meant to be a measurable interpretation of the passenger load, but rather a comparative tool to judge the relative passenger load along a route and between runs. As such, there is some freedom in how they are defined.

To plot the passenger load, the schedule adherence borders should be extended out to generate new shapes. A similar effect could be achieved by drawing lines of different weights along the borders of the black and white schedule adherence regions, however this would distort the data.

This is because the lines, if drawn directly between the data points, will overlap the black and white regions. In fact, if lines of differing weights are used, large passenger loads would distort the figure such that the schedule deviation would appear better than it really was. Times when large numbers of passengers were riding a late or early bus (an important case for the transit operator to identify) would become more difficult to identify.

There are several other ways in which the borders could be defined. They could be parallel to the APC and schedule lines, shifted out perpendicularly by the magnitude of the passenger load. This approach has the advantage of maintaining the slope of the schedule or APC data (corresponding to speed) in the representation of passenger load. However, it results in cusps when the passenger load changes dramatically, and these cusps cannot be eliminated without changing the angles of the lines; which would eliminate the advantage of this approach.

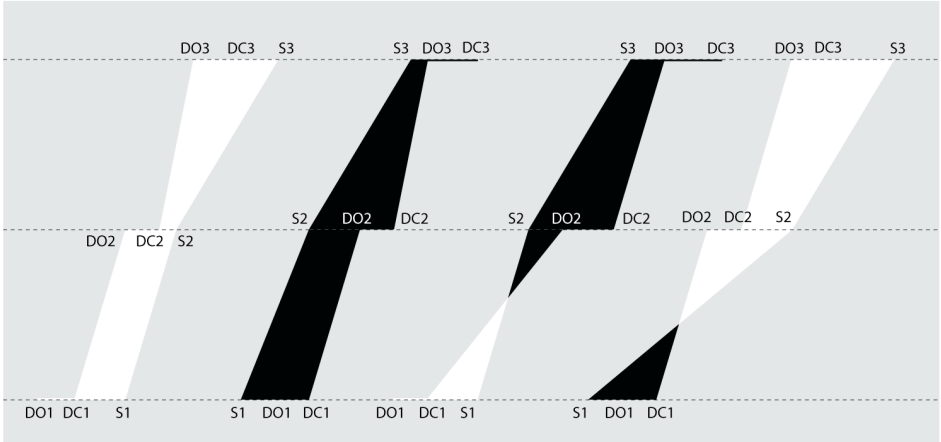


**Figure 4: Four types of bus performance along a link. *DO* = door open, *DC* = door close, and *S* = scheduled times.**

Instead, an angle bisection method is chosen. First, we specify that the widths of the borders will be even on both sides of the schedule adherence curve. Their width will be determined at every stop (the finest level of detail available) and will directly correlate to the proportion of the passenger load at that stop compared to the maximum observed passenger load. Since the interior of the Marey graph (the schedule deviation portion) alternates between black and white depending on the adherence of the bus, the borders will be drawn in gray.

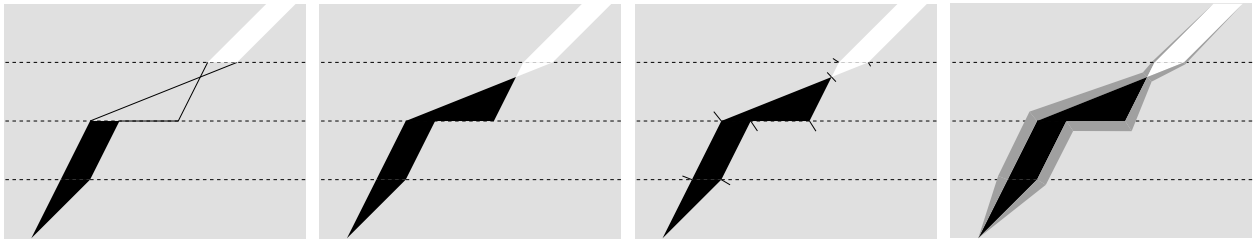

**Figure 5: Applying passenger-load borders to a Marey graph**

These figures are constructed by looping through the data run-by-run and link-by-link. Each stop has time points corresponding to the door opening, the door close, and the scheduled arrival time.

At each stop record, the nature of the following link must be determined. Either the bus was ahead of schedule for the entire link, behind schedule for the entire link, started out ahead and fell behind, or started out behind and made up time (see Figure 4). The vertices of the borders must be determined slightly differently for each of these cases.

Figure 5 depicts how the borders are constructed. In the first panel, links between stops where the bus did not change the sign of its schedule adherence (it was either early or late on the entire link) are filled in according to the APC and schedule data as before, either black or white as appropriate. In the second panel, the links on which the bus crossed the schedule are considered. To draw these links, the intersection between the two lines is found and two triangles representing the schedule adherence are drawn. In the third panel, the vertices of the quadrilaterals that will become the borders are found. The completed depiction of the link is shown in the fourth panel.

To define the borders (as illustrated in panels 3 and 4 of Figure 5), each of the angles exterior to the shapes created by the schedule adherence are bisected. Those bisecting lines are then extended out a distance proportional to the passenger load at that stop, and then connected to the adjacent points. Since there is no corresponding data point in the cases where the APC line intersects the schedule line, care must be taken to preserve continuity of the lines. A good effect can be achieved by extending lines from either side of the vertex, bisecting both angles. The lines should be as long as the average of the passenger load at the two nearest data points, weighted by distance to them.

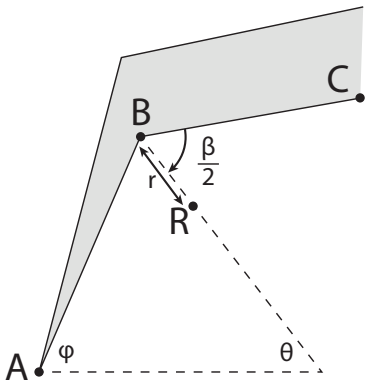


**Figure 6: Angle bisection method to place the border**

Figure 6 demonstrates the angle bisection method for determining the vertices of the quadrilaterals defining the passenger load. First, $\frac{\beta}{2}$ is found by applying the law of cosines to triangle ABC:

$$\frac{\beta}{2} = \frac{\cos^{-1}\left(\frac{AB^2 + BC^2 - AC^2}{2 * AB * BC}\right)}{2}$$

From this and the values of A and B, θ can be found:

$$\vartheta = \pi - \frac{\beta}{2} - \tan^{-1}\left(\frac{B_y - A_y}{B_x - A_x}\right)$$

Which means that the coordinates for R can be given as:

$$R = \left(B_x \pm r\cos(\vartheta), B_y \pm r\sin(\vartheta)\right)$$

The signs of which must be adjusted depending on which side of the curve the border is being applied to. The gray quadrilaterals can then be plotted by connecting consecutive *R* points with the appropriate *door open*, *door close*, or *scheduled time* points depending on the nature of the link (as shown in Figure 4).

# CONCLUSIONS

Since data visualizations are related to the performance of tasks, visualization design should focus on designing for efficient visual tasks. Decisions about how to encode and structure data visually should be based primarily on supporting efficient and accurate performance of the anticipated task for which it will be used.

The addition of APC data to Marey graphs as shown in this paper benefits users of these visualizations in two important ways [4]:

1. *It allows users to use visual operators in place of more demanding logical operators.* The on-time characteristics of the buses are presented visually, making them easy to grasp more quickly. For example, lines that grow close to one another represent bus bunching. The size of the passenger load is presented as a comparative measure designed for visual effect.
2. *It streamlines the search for information.* White, black, and gray regions are immediately visible, allowing patterns across time and space to be seen easily.

Since the passenger load "shadow" is a comparative and not an absolute representation of passenger load, the user could potentially adjust the scale at which it is shown. For example, a Marey graph could be accompanied by a dial which turns up or down the scale of the passenger load. Perhaps it would be turned down when buses are close together and adherence is being considered and turned up it to view periods of low ridership more easily. Furthermore, vehicle capacity information could be added to present passenger load as a ratio, with loadings above 1.0 (indicating the presence of standees) highlighted [3].

Transit performance visualization tools are most effective when combined with good performance measurement practices such as goal setting, integrating performance metric results with agency decision making, etc. [3]. The Marey graphs presented in this paper could serve such

a role for planning and operations tasks by presenting a multifaceted view of the transit system performance in a clear and centralized way.

# REFERENCES


(1) Étienne-Jules Marey. *La méthode graphique*, Paris. 1885

(2) Rob Hranac, Jaimyoung Kwon, Ph.D., Mark Bachmann, Karl Petty, Ph.D. *Using Marey graphs to visualize transit loading and schedule adherence*. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC. 2010

(3) Kittleson; Urbitran; LKC Consulting; MORPACE International; Queensland University. *Transit Cooperative Research Program Report 88: A Guidebook for Developing a Transit Performance-Measurement System.* Transportation Research Board of the National Academies, Washington, D.C. 2003

(4) Stephen Casner. *A Task-Analytic Approach to the Automated Design of Information Graphics.* Technical Report AIP – 82. Learning Research and Development Center. University of Pittsburgh. Pittsburgh, PA. 1989

(5) Darren M. Scott and Micha Pazner. *Image Processing of Space-Time Graphs/ An Example Using the Toronto-Montreal Train Schedule.* University of Western Ontario/Canada. 1992

(6) Daniel Chak. *Enhanced Modeling: Real-Time Simulation and Modeling of Graph Based Problems on Interactive Workbenches*. Master's Thesis. Massachusetts Institute of Technology. 2004

(7) Alan Dix, Geoffrey Ellis. *Starting Simple - adding value to static visualization through simple interaction*. Proceedings of the 2098 International Conference on Advanced Visual Interfaces. 1998