Data as a resource: real-time predictive analytics for bus bunching

Rahul Nair¹, Eric Bouillet, Yiannis Gkoufas, Oliver Verscheure
{rahul.nair, bouillet, yiannisg, verscheure} @ie.ibm.com
IBM Research

and

Magda Mourad, Farzin Yashar
{magdam, fyashar}@us.ibm.com
IBM Software Group – Industries & Solutions, IBM Corporation

and

Rosie Perez, Joel Perez, Gerald Bryant
{rdp, jpz, gbryan}@miamidade.gov
Miami Dade Transit, Miami, Florida

August 1st, 2014

4235 words + (3 Tables + 12 Figures) * 250 = 7,985 words

For presentation at the 2015 TRB 94th Annual Meeting and for possible publication in the Transportation Research Record

¹ Corresponding author, Ph. +353-18-269-881. Address: IBM Research, 3 Technology Campus, Damastown Industrial Estate, Mulhuddart, Dublin 15, Ireland
ABSTRACT

This paper outlines the results of an eight week pilot program in Miami that provided a real-time system to predict bus bunching along four routes. Bunching refers to the phenomenon of two buses serving the same route arriving at a stop at the same time. Desired service headways are not maintained in such cases leading to lower levels of service for passengers on account of longer wait times and lower utilization of resources for operators. Predictions of when and where bus bunching is likely to occur in the future equip controllers with tools to regularize headways in an anticipative fashion. The ‘best-effort’ prediction system is works in an entirely data-driven fashion and leverages, in part, a historical catalog of bus propagation. Applications in performance management and schedule optimization are also presented.
INTRODUCTION

As Miami Dade Transit (MDT) seeks to improve transit services under constrained resource environments (unfunded needs for bus improvements for existing services total $81.4 million till 2023 (1)), the use of data and novel analytics methods to gain insights into both operational and planning aspects of the system offer significant opportunities to do more with less. This paper outlines the results of an eight week bus bunching pilot program that provided controllers with real-time predictions of bus bunching. Bunching refers to the phenomenon of two buses serving the same route arriving at a stop at the same time. Desired service headways are not maintained in such cases leading to lower levels of service for passengers on account of longer wait times and lower utilization of resources for MDT. Predictions of when bus bunching is likely to occur in the future equip MDT operations with tools to regularize headways in a proactive fashion. Figure 1 shows a typical occurrence of bunching events during a morning peak period along a route.

![Figure 1](image)

Figure 1 Space-time trajectories showing bunching on two consecutive Wednesdays. The aim of the pilot program was to predict when and where bunching will occur (shown in red)

A ‘best-effort’ prediction system is designed that learns from past histories and covers a wide range of operational scenarios. In part, the system leverages operational data from MDT by building historical models of how each bus in the system performs. The data powers the accurate predictions of bus movements using the intuition that “history repeats itself” and bus propagation along the
route in real-time can rely on past observations to predict the future. Analytics on the gathered operational data paints a robust picture on system performance with an aim to arm MDT scheduling and planning divisions with data-driven insights.

Prediction of bus bunching in this work relies on accurate bus arrival prediction, for which several approaches have been proposed. Cathey and Dailey (2) provide a general framework for predictions based on automatic vehicle location (AVL) data, describing a tracker, filter and prediction module. Among the analytical models, artificial neural networks (3), support vector machines (4), Kalman filters and combinations (5), nearest neighbor trajectory method (6, 7), inclusion of covariates such as weather (8), and K-nearest neighbors (9). Yu et al. (10) compare several of these approaches and present support vector machines as providing the most accurate results. Similarly, Sinn et al. (11) present a comparison of several data-driven approaches, finding that the kernel regression model outperforms other approaches. The fused prediction model employed herein leverages their kernel regression model.

A significant component of bus travel time with impact on bus bunching is dwell time. This is the time a bus spends serving passengers and can be highly variable in practice. While passenger loading levels directly influence dwell times (13) and passenger data can be used directly in predicting downstream arrival times (5), other cases such as merging of buses into general traffic lanes can also have an impact (12). In contrast to these approaches, the base prediction system employed uses a ‘best-effort’ fusion approach that is applicable in a wide range of operational scenarios. This system is described in greater detail in (14) and the focus of this paper is on the prediction of bus bunching.

Several works have aimed at alleviation of bus bunching under different operational conditions and response strategies (15, 16, 17, and 19). Some works aim to determine spatiotemporal characteristics and causality of bunching (18). Predictive models that provide the opportunity to anticipatively correct collapsing headways have not been previously proposed.

The main contributions of the paper are in (a) describing a real-time prediction system of bus bunching, (b) demonstrating its applicability via a real-world deployment in Miami-Dade Transit, and (c) highlighting key performance indicators, including comprehensive accuracy assessments and performance metrics related to planning and scheduling.

During the pilot operations, the system collected 2.6 million position updates in real-time along four routes and generated 74,015 alerts representing 9,696 separate bunching events. Each of these events were validated a posteriori, and spatiotemporal characteristics of observed bunching analyzed to provide insights into capacity constraints and system operations. Bus bunching prediction accuracy is shown to range between 68% and 80% and was primarily dependent on the polling frequency of the AVL technology used, the time horizon to prediction, and data coverage available for the prediction engine.
From a planning perspective, the operational data allows a rich characterization of service characteristics, its variability and pinpoints bottlenecks with the system. A comprehensive comparison of scheduled services and how they are realized operationally is performed. For example, each scheduled revenue-hour of service is shown to take 66 minutes on the average to provide. This does not include non-revenue time like deadheading trips and includes data on buses that took less time that the scheduled travel times. While each scheduled revenue hour of service for Route 119 took 64 minutes on the average, the same statistic for Route 120 was 71 minutes. Buses along this corridor spent more time stationary away from stops with non-stop related dwell times accounting for 6.7% and 5.1% of runtimes for Routes 119 and 120 respectively.

The paper is organized as follows. The next section describes the high-level system description followed by major implementation considerations. The results from the Miami Dade pilot are described next with key performance indicators for operations and planning. The last section concludes with a summary of ongoing work and conclusions.

**SYSTEM DESCRIPTION**

The real-time prediction system relies on two types of data. Infrastructure data that describe the inventory of stops, routes, lines, along with production timetables for all scheduled services is used to determine the network configuration. Spatial route configurations are then determined dynamically each day, and the production timetable information loaded. The second data source is the real-time vehicle position data that provides the spatial location of the bus along the service reference that is used to correlate the position update with a specific service. Figure 2 shows the major components of the system.

Figure 2 Overview of real-time components

Once the real-time bus position is received, rigorous quality checks, including on-route/off-route analysis through map-matching, de-noising and consistency checks are performed. The spatial position is then converted to an offset metric, which is a linear representation of the bus route. The offset measures the distance of the bus from the start terminal. A collection of bus offsets along
with the time stamps of each position creates a path in space and time, and is referred herein as a *trajectory*. Trajectories are stored in a database and used to build a historical catalog of bus movements that are then used in one of the predictive models \((11, 14)\).

The prediction engine is a ‘best-effort’ system that is designed to operate under a wide variety of operational scenarios. The three sources for the prediction are based on past historical performance (how did this bus do in the past?), current observed conditions (how are other buses doing?) and a reference timetable speed from the schedule. Each prediction model generates an estimated time of arrival (ETA) for all downstream bus stops and a confidence score that is computed by several factors, but mainly related to the quantity of data used to make the prediction, the age of the data available to make the prediction.

A fusion engine then merges the three different predictions to yield a final single prediction. The fusion uses the confidence scores to balance the contribution of each component prediction. This purpose of the fusion engine is to provide accurate predictive capabilities under a wide range of operational scenarios. For example, a low frequency route may not have a sufficient historical catalog to power the historical prediction component. The corresponding weights used by the fusion for this component is therefore reduced. Similarly, for rural routes, there may be insufficient probes to provide coverage on the entirety of the route leading to lower confidence scores for the observed ETA component.

The historical ETA component leverages operational data from the bus routes to build accurate historical models of how buses perform. This historical information predicts bus movements using the intuition that “history repeats itself” and bus propagation along the route in real-time can rely on past observations to predict the future. Figure 3 shows sample historical trajectory catalogs used for this predictive model for two different routes. Such catalogs are built for each journey pattern separately and stored in memory for real-time prediction. In real-time, once a bus position is received, the most relevant past histories are referenced from the catalog, and a kernel weighted average taken to determine the most likely progression into the future \((14)\).

The observed ETA component leverages all buses are probes to compute experienced travel and dwell times. These quantities from all downstream buses are factored in the observed ETA in real time. The reference travel time is also considered and factored in for downstream stops.

A bus bunching awareness and prediction module then combines predictions to generate alerts seen by transit operators. The bus bunching alert includes information on where, when and which buses will be involved in a bunch, along with a confidence score. Each time the prediction is updated, the alert is generated, updated, or deleted. Figure 4 shows a sample set of bunching alerts (including all updates) on the space-time reference for a day of operations along a single route. Bunching occurs when trajectories meet. The figure also shows slowdowns during the middle portion of the day.
Figure 3 Sample trajectory data used by prediction models for Route 119 Northbound (top) and Route 120 Northbound (bottom)
IMPLEMENTATION AND KEY CONSIDERATIONS

The prediction system was deployed for four lines of Route H (108), Route L (112), Route S (119), and Route Beach MAX (120) and was operational from January 18th, 2014 through March 18th, 2014 for a period of 8 weeks. Along with a 6 week test period prior to the operational release, a total of 2.6 million bus position updates were processed in real-time from Dec 4th through March 18th, representing close to 20,000 bus runs along the 4 routes over this period. Alerts generated in real-time were available via a browser interface in the control room.

Data on real-time bus positions was made available through an XML feed with all active buses during revenue operations. The infrastructure data along with time table information was made available through the public GTFS feed. The real-time data from MDT was sourced from two systems. The WiFi modem equipped buses provided bus position updates every 20 seconds. The older CAD AVL system provided information at a lower polling frequency of a minimum of two minutes, but in practice polling frequency for the CAD AVL buses was found to be longer. The data coverage for the two systems was therefore different depending on the on-board devices.

Polling frequency of on-board devices

A critical aspect of the prediction system is the polling frequency, the time interval between successive bus position updates, since the prediction computations are made when each bus position arrives. The desired polling frequency of 20 seconds provides sufficient granularity to capture dynamics of the buses as they progress. Longer polling rates, such as those received every two minutes from the CAD AVL system, do not provide complete information for the entirety of
the route. Since in MDT, part of the fleet is equipped with such buses, data coverage can vary based on the number of buses using the CAD AVL system. In particular, routes 119 and 120 had better coverage and better historical models, while routes 108 and 112 had fewer high quality trajectories that were available for the predictive models and subsequent analysis. The system learns over a period time and improves its prediction power as more data becomes available. The data coverage for the duration of the pilot was not complete, since not all bus journeys reported their position in real-time. The consequence of having multiple polling frequencies is that coverage for different routes is different and the available data for the prediction also varies impacting the quality of predictions for certain routes.

**METRICS OF BUNCHING**

A total of 74,015 messages were generated over the course of the pilot representing 9,696 unique bunching events. A message can be of three types. A *create* message is generated with a new bus bunching event is predicted.

When new positions are received, the predictions are updated and subsequently an *update* message is issued that updates the information on the alert. For example, events become more certain when they are in the near term, and as a result confidence scores improve. In total 54,631 update messages were sent. When bunching events are no longer predicted, either due to spacing out of the buses, or due to arrival at the terminal station, a *delete* message is sent to cancel ongoing event.

Two buses are considered to be bunched when the time headway is below a certain threshold. For this pilot, several thresholds were evaluated, and the final threshold used was 60 seconds. Figure 5 shows unique bunching events for the four routes by direction. Looking at the intensity of bunching events through the course of a day, it is predominant during the PM peak periods for all the routes. In terms of differences across routes, Route 119 was the main route for which the maximum number of bunching events were generated.

Figure 6 shows the spatial dimension of bunching with respect to where along the route bunching is likely to occur. While the stochastic characteristics of travel time experienced along a route suggest that bunching should occur during the latter part of the route, along some routes, capacity considerations that impact dispatch policies can lead to bunching early on as well as shown the Route 119 Northbound direction.
Figure 5 Unique bunching events by hour of day for the duration of the pilot
Figure 6 Bunching locations along Route 119 by hour of day (Southbound - top, Northbound - bottom). The y-axis ranges from 0 to 40 events for each hour and the direction of travel is from left to right.
The early bunching indicates that the cause of early bunching in the northbound direction are not due to variability in travel times or boarding patterns, but due to other systemic factors. The main driving factor is the capacity constraints the downtown terminal that cause compressed recovery times.

In terms of prediction horizons, Table 1 shows the how far ahead of time the events were generated. Only create events are considered as the first occurrence of bunching. The table shows a large number of events were predicted within 10 minutes especially for Route 119.

Table 1 Prediction horizon of events

<table>
<thead>
<tr>
<th>Route</th>
<th>&lt;10 min.</th>
<th>10-20 min.</th>
<th>20-30 min.</th>
<th>30-40 min.</th>
<th>40-50 min.</th>
<th>50-60 min.</th>
<th>&gt;60 min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>160</td>
<td>30</td>
<td>13</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>112</td>
<td>983</td>
<td>62</td>
<td>42</td>
<td>17</td>
<td>14</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>119</td>
<td>5081</td>
<td>759</td>
<td>438</td>
<td>257</td>
<td>198</td>
<td>146</td>
<td>190</td>
</tr>
<tr>
<td>120</td>
<td>940</td>
<td>174</td>
<td>52</td>
<td>45</td>
<td>26</td>
<td>16</td>
<td>11</td>
</tr>
</tbody>
</table>

The reason for the large number of on-going events is that, once buses bunch they tend to persist in that state till the end of their run. Under the current solution, bunching alerts are generated if buses that bunch, then space out and then bunch again. The prediction quality of these events are reported next.

ACCURACY MEASURES

Accuracy of the entire solution is investigated in two stages. First, the system prediction of bus arrival times at each stop are studied and compared to “actual” arrivals of buses. Actual arrival of buses are inferred by looking at the position data after the fact, to see if in fact the prediction made ahead of time had actually been accurate. At the second step, the bunching prediction accuracy is quantified to show if bunching occurred as predicted or not.

ETA prediction accuracy

Root Mean Squared Error (RMSE) is used as a metric of accuracy, where prediction error = predicted arrival at a stop – actual arrival at the stop. Table 2 shows the RMSE computed for each route by time horizon. On the average, the prediction of a bus arrival time 30 minutes into the future has an error or roughly 2-4 minutes, depending on the route, while predicting 60 minutes out the average error is in the range of 3-7 minutes. Routes for while there is better WiFi data, mainly 119 and 120, show lower RMSE values than those dependent on the older CAD AVL buses that provide lower polling frequency.
Table 2 ETA Prediction RMSE in minutes by time horizon

<table>
<thead>
<tr>
<th>Route</th>
<th>Time horizon</th>
<th>&lt;10 min.</th>
<th>10-20 min.</th>
<th>20-30 min.</th>
<th>30-40 min.</th>
<th>40-50 min.</th>
<th>50-60 min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td></td>
<td>1.71</td>
<td>2.4</td>
<td>2.53</td>
<td>3.77</td>
<td>4.54</td>
<td>6.8</td>
</tr>
<tr>
<td>112</td>
<td></td>
<td>2.26</td>
<td>3.55</td>
<td>4.43</td>
<td>4.91</td>
<td>5.51</td>
<td>7.37</td>
</tr>
<tr>
<td>119</td>
<td></td>
<td>1.89</td>
<td>2.73</td>
<td>3.49</td>
<td>4.07</td>
<td>4.56</td>
<td>5.18</td>
</tr>
<tr>
<td>120</td>
<td></td>
<td>1.83</td>
<td>2.49</td>
<td>2.74</td>
<td>3.15</td>
<td>3.63</td>
<td>3.67</td>
</tr>
</tbody>
</table>

The error distributions for each route are also illustrative as they show if there are systematic biases in the predictions. For routes with greater histories the prediction distributions are centered on the mean. Sample distributions for Route 119 are shown in box-plot diagrams in Figure 7.
Figure 7 Sample ETA Prediction error distribution by time horizon (blue boxes represent ranges between the 25th and 75th percentile, the “whiskers” represent three standard deviations, and the red crosses mark outliers). Errors computed over 0.9 million and 0.87 million ETA predictions made for the top (Route 119 southbound) and bottom (Route 119 northbound) respectively.

**Bus bunching prediction accuracy**

Bunching prediction accuracy is evaluated by comparing if predicted bunching events occurred or not by comparing the bus trajectories after the fact to see if bunching could be captured. Since good quality trajectories can be utilized to evaluate if bunching actually took place or not, a minimum threshold of 20 position updates was used to evaluate if bunching occurred or not. Three basic outcomes are possible as shown in Figure 8. (a) True positives indicate predictions that happened in reality. (b) False positives indicate predictions that were made but did not occur in reality, (c) False negatives, where the system did not correctly identify bunching that occurred in reality, and (d) no data scenarios, where there was insufficient number of GPS data points to evaluate if bunching actually happened or not. “Reality” is reconstructed from the observations and interpolating to determine arrival at stops.

Note that bunching is determined using the predictive models, so even if position updates were not received, the bus is assumed to be moving according to the predictions. No data entries show the value of using the prediction based approach, since such events would not have been picked up by conventional methods.
Overall the system had a false positive rate of 13% and no false negatives. For roughly 53% of the cases, there was sufficient resolution of data to evaluate if bunching occurred or not. Figure 8 shows an example of where the system got it wrong. A bus position update was received at 10am that indicated a series of bunching downstream that did not end up happening. The next position update for the bus put it on a bunching pattern with the next services – a set of bunching alerts that the system got correct.

DATA INSIGHTS

The main observations and insights from the data observed are presented, highlighting some of the patterns observed. Due to the differing AVL systems, data gaps and long polling frequencies for some buses, depending on the calculation, only subsets of data that meet certain quality thresholds are used. The number of bus runs that were used to determine the statistic are noted where applicable (and denoted by ‘n’).

Figure 9 shows the space-time diagram for a trajectory that is constructed from the bus position updates and compared with its scheduled trajectory, constructed from the timetable. Deviations from schedule at the end terminal station are used to compute the on-time performance values. Deviations from schedule at the start terminal station are used to compute the dispatch time.
performance. The difference of the scheduled time and actual run time is used to evaluate en-route performance.

All attributes shown in Figure 9 where analyzed. On-time performance was founded to be driven largely by dispatch performance, which in turn was driven, for the major routes, by capacity constraints at the terminal station. In other words, delays in start terminal dispatch explained the on-time performance at the end terminal. Dispatch operations were constrained by downtown terminal capacity issues leading to accumulating delays. We now focus on en-route performance and speed profiles to evaluate revenue operations.

**En-route performance**

Once in service, how long does a bus take to complete its trip and how does that compare to how much travel time the schedule allocates? In other words, discounting for delays in dispatching, how do revenue hours of a bus compare with the schedule. To compute the en-route performance, few assumptions were made. Buses that reported less than 30 positions during the entire trip were excluded, as were buses that did not provide updates within 0.62 miles (1 kilometer) of the start and end terminal station. To complete data gaps at start portions of the trip, an average speed of all buses for that route was computed for the first segment and this was employed to compute the start time. To complete data gaps at the end portions of the trip, the last available prediction was employed to determine the arrival time at the end terminal station. Since the prediction is based on historical data, it is equivalent in some ways to the method by which the start gaps are filled.
In total, 2,740 bus runs were used for this analysis. Many completed their journeys in less time than the time allocated by the schedule, but the majority took more time than the scheduled travel times called for. On the average, from an operational sense, it took 9.4% more time to complete service than what the schedules called for. Put in other words, for each revenue hour in the schedule, operations delivered service in 65.66 minutes – on the average. This time includes only revenue hours - i.e. time spent serving customers. Data on deadheaded trips, layovers, and recovery times were not available to this study and are not included. Further, since this is averaged across all buses, it is important to read this value in conjunction with the delay distributions that follow.

Table 3 Aggregate en-route performance by route and expected weekday cost impacts of en-route delays

<table>
<thead>
<tr>
<th>Route</th>
<th>direction</th>
<th># Buses observed</th>
<th>Scheduled revenue time (hrs)</th>
<th>Actual revenue time (hrs)</th>
<th>% rise</th>
<th>Mean delay w.r.t. schedule</th>
<th>#Weekday services</th>
<th>Daily Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>South</td>
<td>28</td>
<td>28.17</td>
<td>32.11</td>
<td>14.0%</td>
<td>11.4</td>
<td>38</td>
<td>$911</td>
</tr>
<tr>
<td>108</td>
<td>North</td>
<td>15</td>
<td>15.27</td>
<td>16.46</td>
<td>7.8%</td>
<td>8.7</td>
<td>38</td>
<td>$699</td>
</tr>
<tr>
<td>112</td>
<td>East</td>
<td>280</td>
<td>376.10</td>
<td>419.03</td>
<td>11.4%</td>
<td>13.4</td>
<td>88</td>
<td>$2,490</td>
</tr>
<tr>
<td>112</td>
<td>West</td>
<td>200</td>
<td>253.25</td>
<td>291.30</td>
<td>15.0%</td>
<td>13.5</td>
<td>87</td>
<td>$2,481</td>
</tr>
<tr>
<td>119</td>
<td>South</td>
<td>705</td>
<td>1186.33</td>
<td>1294.40</td>
<td>9.1%</td>
<td>15.0</td>
<td>94</td>
<td>$2,967</td>
</tr>
<tr>
<td>119</td>
<td>North</td>
<td>917</td>
<td>1566.05</td>
<td>1647.21</td>
<td>5.2%</td>
<td>11.1</td>
<td>89</td>
<td>$2,078</td>
</tr>
<tr>
<td>120</td>
<td>South</td>
<td>203</td>
<td>272.20</td>
<td>304.50</td>
<td>11.9%</td>
<td>15.6</td>
<td>71</td>
<td>$2,331</td>
</tr>
<tr>
<td>120</td>
<td>North</td>
<td>392</td>
<td>484.82</td>
<td>572.11</td>
<td>18.0%</td>
<td>15.9</td>
<td>69</td>
<td>$2,314</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2740</td>
<td>4182.19</td>
<td>4577.12</td>
<td>9.4%</td>
<td>574</td>
<td></td>
<td>$16,271</td>
</tr>
</tbody>
</table>

The following histograms in Figure 10 and Figure 11 show what fractions of buses took longer (or shorter) than the travel times in the schedule for each route. To interpret these figures, let’s take the example of Route 119 in Figure 10. The southbound services are more likely to take longer than the schedule, while the northbound services shows that a good fraction of buses are likely to be quicker than the schedule time. This indicates high variability in service which has a scheduled run time of a little over one hour. The data shows that in 35% of the cases, the observed buses took a little over an hour. In 10% of the cases, the buses deviated by 30 minutes, showing an actual run time of more than 1.5 hours. In a perfect system, all buses will be clustered around the 0 mark – showing perfect synchrony between schedules and operations.
Figure 10 En-route performance distribution for Route 119
The cost impacts of these differences can be computed by merging this delay measures with the average cost to MDT to provide these services. MDT operating costs were computed as $126.54 per revenue hour and $10.6 per revenue mile (20). Table 3 also shows how delayed service runs cost MDT for the various routes. This cost is only incurred when schedule revenue hours are exceeded by the operational revenue hours. Note that this is not the true cost to MDT (in terms of overtime pay, etc.), but rather the cost associated with increase revenue hours due to operational conditions.

En-route performance is broken down by hour of day for Route 119, which along with Route 120, shows the greatest variability in service. Each time period shows a distinct delay pattern. Travel times increase significantly in the both directions beginning in the mid-day period. The variability of northbound routes is markedly different, since it shows buses are as likely to be fast as slow during the PM Peak period.

Figure 11 En-route performance distribution for Route 120
Speed Profiles

While the on-time, dispatch and en-route measures can be considered aggregate in some sense, as they measure what the bus did in “total”, this section describes the performance at the stop-to-stop level and aims to quantify where the delays accumulate and the variability experienced by services.

The speeds are computed for all data available on weekdays with reasonable quality (more than 30 observations) and data gaps are ignored, i.e. not filled in. The distribution of speeds in mph are measured for each stop pair and the summary of this data below show the mean computed across this distribution, and variability is documented by the 10th and 90th percentile scores. Ignoring the extreme tails of the speed distribution essentially excludes special events, breakdowns and other outlier behavior that could bias the results and not indicative from a planning perspective. The diagrams below therefore summarize the range of “normal” operations that were experienced during this period. The baseline scheduled speeds are also shown as a reference.

It is important to note that some key differences in interpreting the speed information below as compared to the delay statistics presented above. The main difference is that since speed and travel times are inversely proportional (lower speeds lead to higher travel times), portions in the speed profile graphs that have low speeds could have much greater impact on the travel times than what is visually perceptible. For example, in the Lincoln Road portion in Miami Beach of Route 119 where significantly lower speeds are shown in Figure 12, the impact on travel times is significantly greater than what the speed profile graphs suggest.
The main outcome of this analysis is that the scheduled speed closely follows the observed speed for much of the route. In portions of the routes where variability is high, the scheduled speed is conservative to account for this increased uncertainty. Examples of this are the early part of Route 119 Northbound (around the causeway), where high variability in bus travel speeds are observed. When traffic conditions allow for free flow, buses travel can reach more than twice the scheduled speed and as evidenced in the bunching metrics section, lead to greater bunching in the portion of the route. Among the four routes, Route 120 shows the greatest differences between the scheduled speeds and observed speeds, suggesting that the travel times in the schedule can be updated to include modified time profiles from the data.

CONCLUSIONS

A predictive system for detecting bus bunching is described and piloted in Miami-Dade Transit for four routes. The prediction system is entirely data-driven, and leverages operational data to build historical models of bus propagation that provide high accuracy predictions even when polling rates are more than two minutes. For the four routes the prediction error for predictions made one hour out were between 3-7 minutes. Bus bunching predictions had a false positive rate of 13%. Data insights from the operational data were analyzed to demonstrate the potential for bridging the gap between operations and scheduling via schedule refinements. Current research efforts are focused on generating anticipative corrective actions.

ACKNOWLEDGEMENTS

Contributions from IBM Software Group Services, specifically Tom Luzzi, Keith Baldwin, Jun Xia Zhou, and Sanjeev K. Sinha on project management and cloud infrastructure, and software services are acknowledged. From Miami Dade Transit, we wish to thank Andrew Hagewood and Hector Garnica for their assistance with the project.
References


