Toward a Demand Estimation Model Based on Automated Vehicle Location

Article in Transportation Research Record Journal of the Transportation Research Board · January 2016
DOI: 10.3141/2544-16

CITATIONS
0

READS
23

2 authors:

Luis Moreira-Matias
NEC Laboratories Europe
36 PUBLICATIONS 165 CITATIONS

Oded Cats
Delft University of Technology
68 PUBLICATIONS 274 CITATIONS

Some of the authors of this publication are also working on these related projects:

WEpods: The first autonomous vehicle on Dutch public roads View project

ADAPT-IT View project

All content following this page was uploaded by Oded Cats on 15 December 2016.

The user has requested enhancement of the downloaded file. All in-text references underlined in blue are added to the original document and are linked to publications on ResearchGate, letting you access and read them immediately.
Toward a Demand Estimation Model Based on Automated Vehicle Location

Luis Moreira-Matias and Oded Cats

The rapid increase in automated data collection in the public transport industry facilitates the adjustment of operational planning and real-time operations based on the prevailing traffic and demand conditions. In contrast to automated passenger counts systems, automated vehicle location (AVL) data are often available for the entire public transport fleet for monitoring purposes. However, the potential value of AVL data in estimating passenger volumes has been overlooked. This study examined whether AVL data could be used as a stand-alone source for estimating onboard bus loads. The modeling approach infers maximum passenger load stop from the timetable and then constructs the load profile by reverse engineering through a local constrained regression of dwell times as a function of passengers flows. To test and demonstrate the potential value of the proposed method, a proof of concept was performed by conducting unsupervised experiments on 1 month of AVL data collected from two bus lines in Dublin, Ireland. The results suggest that this method can potentially estimate passenger loads in real time in the absence of their direct measurement and can easily be introduced by public transport operators.

Understanding passenger demand is key for the effective planning and provision of public transport services. Over the past decades, mass transit operators worldwide relied on passenger surveys to understand their mobility needs and adjust their planning and operations accordingly (1–3). The rapid increase in automated data collection in the public transport industry facilitates the adjustment of operational planning and real-time operations on the basis of prevailing traffic and demand conditions. By observing current service attributes, service management could adapt the service to respond better to passenger travel needs. The implementation of such measures requires information on passenger flows to assess the expected effects of such measures. For example, when deciding whether to allocate an additional vehicle to reduce onboard congestion, information on the number of passengers on board is essential to assess the impacts of this decision.

Even though public transport systems are increasingly equipped with automated passenger counts (APCs) and automated fare collection (AFC), the data collected by those systems are often incomplete and hinder the estimation of the overall demand profile. This shortcoming stems from the fact that these systems and their deployments were designed to support tactical planning and managing concessions rather than support real-time information on passenger flows. In particular, to save costs, the common practice is to install APC systems only on a small subset of the fleet. While this is sufficient for obtaining a robust estimation of overall demand patterns, it prohibits the real-time estimation of passenger loads for individual trips. Furthermore, APC is only seldom transmitted in real time. Instead, data collected by the APC equipment are downloaded on a daily or weekly basis at the depot. Similarly, while AFC constitutes a promising source of information on travel patterns (4), it is typically owned by a public agency that is responsible for the offline distribution of ticket revenues. In addition to the data availability, privacy concerns, and ownership issues, most systems do not require passengers to check in and out when they board and alight from each vehicle, thus requiring excessive big data analytics and a large number of behavioral assumptions to infer route choice at the individual traveler level to estimate passenger flows.

Passenger demand estimation may refer to passenger flows at the vehicle run level (boarding, alighting, onboard) (5) or passengers travel demand at the network level (origin–destination matrix) (6–8). The latter can potentially support demand estimation for strategic planning purposes. Studies that try to infer the details of the travel itinerary undertaken by each individual, on the basis of smart card transactions, often use automated vehicle location (AVL) data as a complementary source of information for attaining the respective time stamps (4). Other data collection technologies that have been deployed to estimate passenger counts include vehicle weight sensors (9) and video surveillance (10). Researchers pointed out technical deficiencies that reduce the accuracy and reliability of such systems and restrict their widespread deployment.

The real-time estimation of passenger loads requires a scalable approach that could be applied in real time for the entire public transport fleet. In contrast to APC systems, AVL data are often available for the entire public transport fleet for monitoring purposes. AVL technologies are more well established and their installation cost has been reduced significantly over the years when compared with that of APC (9). AVL data have been extensively used for studying the determinants of running times, dwell times, and headways. Many studies have estimated the determinants of dwell time and, in particular, the relation between boarding and alighting passenger flows on dwell time on the basis of a combination of AVL and APC data (11, 12). The results reported in these studies provide insights on the formulation of the dwell time function and its underlying assumptions. Some researchers explored the fusion of AVL and APC by using the APC data as a complement to the AVL data to estimate or predict the travel time variability (13, 14). However, the potential value of AVL in estimating passenger volumes has been overlooked and, to the best of the authors’ knowledge, none of the previous studies suggested using AVL data for estimating passenger flows.

This study examined whether AVL data can be used as a stand-alone source for estimating real-world passenger loads. The modeling approach was to infer maximum passenger load stop from the timetable and then to construct the load profile by reverse engineering.
through a local constrained regression of dwell times as a function of passengers flows. A series of machine learning methods and principles was applied to estimated boarding and alighting flows based on actual dwell times and the planned schedule. The resulting framework was denominated as DemandLOCkeR (demand estimation through local constrained regression). The paper is structured as follows: The first section presents the method proposed in this study and the related estimation procedure. The second section describes the case study and data that were selected for testing the feasibility and performance of the proposed method. The third section presents the experimental setup along with the results of the application. The final section concludes with a discussion on the implications and limitations of this study and outlines potential directions for future work.

METHODOLOGY

Analysis Approach

The approach adopted in this study (DemandLOCkeR) for passenger demand estimation relies solely on AVL data and involves reverse engineering where the relationship between dwell times and passenger flows is exploited to construct an estimated load profile. By deploying a local constrained regression technique and supervised machine learning techniques, bus loads are visualized for a given time period. Given the high uncertainty that is inherent to the bus operation environment and the respective passenger demand fluctuations, the output of this analysis is an estimated load profile that aims to illustrate a likely load profile that can be assumed to prevail without any claim for exact estimates or measurements. The authors are not aware of any previous attempt to construct load profiles based solely on AVL data.

The analysis framework deployed in this paper is illustrated in Figure 1. The methodology for estimating bus load profiles by using AVL data consists of five steps: (a) extracting high-level demand information from the planned timetable, assuming that the design was based on a maximum load point method; (b) decomposing real-time dwell times and regressing them on the basis of load profile and dwell time function assumptions; (c) estimating the shape of the load profile by using a local regression technique (the local regression is a method that divides the solution space into different folds where, within each one of them, the load function is approximated by a linear function, as described in the section on load profile estimation using constrained local regression); (d) constraining and fitting the results obtained in the previous step on the basis of the actual dwell times and an incremental bandwidth (defined by domain constrains that force a fitting of the regression outputs within the range of admissible loads, given each vehicle’s capacity) that uses only the most recent dwell time records to obtain realistic load profiles; and (e) obtaining the output of this process, is the typical load profile for each short-term period, by minimizing the Euclidean distance and using the law of large numbers (it ends up on making a reasonable use of the dwell times to set maximum and minimum admissible values for the loads on every stop given the load prediction for the immediate previous one, as described in the section on fitting the dwell times to the load profile using incremental filters). The following sections detail the implementation of each of these steps.

Computing the High-Level Demand Profiles

The purpose of this initial step is to deduce information on the demand profile from the provisioned service frequency. By leveraging on the observed frequency, one can then explore headway variations (obtained from the AVL data) to infer the shape of the demand profile, as explained in the description of subsequent steps of this framework.

Service frequencies are determined by operators on the basis of passenger surveys and direct observations (1–3). There are two ways of determining such frequencies: (a) stop-based and (b) route-based. The latter approach requires information on the demand for each stop...
along the route. Conversely, the stop-based approach is based on the ratio between the passenger load at the maximum load point and the desired occupancy specified for a given period of time (which should ideally be characterized by a uniform bus frequency). Formally, it is possible to determine the desired frequency for a given period of length $\tau$ (e.g., $\tau = 60$ min), that is, $f_j$, as follows:

$$f_j = \max \left( \frac{o_{\text{min}}}{o^j}, f^\text{min} \right) \quad \forall j$$  (1)

where

$$o_{\text{min}} = \max_{s \in S} o_s, \quad \forall s \in S = \text{average or measured onboard occupancy when departing from stop } s \text{ during time period } j \text{ for a certain line,}$$

$$S = \text{set of all stops except for the last stop on the respective line},$$

$$o^j = \text{desired occupancy for the same time period,}$$

$$f^\text{min} = \text{minimum frequency defined by policy makers.}$$

To extract information on the demand pattern, the following set of assumptions is made:

Assumption 1. The entire fleet has an equal capacity of $\zeta$ passengers;

Assumption 2. The variable $o^j$ is defined by a predefined constant value $0 < \delta < 1$ (i.e., percentage-wise definition) for each route and period $j$ (i.e., $o^j = \delta \cdot \zeta$);

Assumption 3. The operator determined the frequency on the basis of the maximum load point method where the maximum expected load for a given trip is considered a constant value for a certain time of the year scheduling (typically a season); and

Assumption 4. The first term in Equation 1 is binding. In other words, the frequency needed to satisfy the load–desired occupancy ratio exceeds the minimum policy frequency.

Assumption 3 does not require that the operator has information on passenger demand at each stop. Operators often know the busiest stop along each route and then manually collect data on this particular stop. Moreover, even if the operator does not consciously determine the frequency on the basis of stop-based counts, the frequency is often the outcome of allocating just sufficient capacity to cater for the most heavily used line segment.

On the basis of these assumptions, it is possible to rewrite Equation 1 as follows:

$$o_{\text{min}} = \zeta \cdot \delta \cdot f_j = \zeta \cdot \delta \cdot \frac{3,600}{h_j^p}$$  (2)

where $h_j^p$ denotes the average planned headway during period $j$ (in seconds). Let $l_{sk}(j, t)$ be the maximum bus load of a given trip $t$ during the period $j$. The planned headway is inferred from the data by calculating the average difference between the scheduled departure times within the period $j$. Based on the above relationship between max load point and headway, the maximum load of a specific bus trip, $k \in K'$, $a_{sk}$, can be estimated on the basis of observed headways derived from AVL:

$$o_{\text{max}} = \zeta \cdot \delta \cdot f_j = \zeta \cdot \delta \cdot \frac{3,600}{h_l}$$  (3)

where $K'$ is the set of bus trips that operate on a given line during period $j$ and $h_l$ is the average observed headway calculated as

$$h_l = \frac{\sum_{j \in J} h_{sk} + h_{sk+1}}{2|S|}$$  (4)

where $h_{sk}$ is the observed headway between trips $k$ and $k + 1$. The maximum load point can now be determined by

$$s_{\text{max}} = \operatorname{argmax}_{s \in \text{set}} a_{sk}, \quad \forall k \in K_j$$  (5)

However, the passenger loads on departing from each stop along trip $k$, $a_{sk}$, are unknown. In the following section, these values are estimated on the basis of the dwell times available from AVL data.

Decomposing Dwell Times

By assuming simultaneous boarding and alighting passenger flows, it is possible to express the dwell time of trip $k$ at stop $s$, $d_{sk}$, by using the following linear expression:

$$d_{sk} = \gamma + \max(a_{sk} \cdot b_{sk}) + c_{ks} + \epsilon$$  (6)

where

$$\alpha$$ and $\beta$ = average alighting and boarding times per passenger, respectively;

$a_{sk}$ and $b_{sk}$ = number of alighting and boarding passengers, respectively;

$\gamma$ = fixed delay caused by door opening and closing times;

$\epsilon$ = error term caused by variations in driver and passenger behavior that is assumed to be distributed $\epsilon \sim N(0, \sigma^2)$; and

$c_{ks}$ = additional dwell time caused by onboard crowding and interactions between passengers in crowded situations.

In line with the formulation of Weidmann (15), the delay caused by onboard crowding can be expressed as a penalty that prolongs the constant dwell time delay:

$$e_{sk} = \begin{cases} \left( \max(\alpha \cdot a_{sk}, \beta \cdot b_{sk}) - \zeta \cdot \delta \right) \cdot \frac{1}{100} & \text{if } \max(\alpha \cdot a_{sk}, \beta \cdot b_{sk}) > \zeta \cdot \delta, \quad \forall i \in j \\ 0 & \text{otherwise} \end{cases}$$  (7)

The relationship between onboard occupancy of trip $k$ on departure from stop $s$ to past boarding and alighting flows is

$$a_{sk} = \sum_{y \in y} (b_{sk} - a_{ks})$$  (8)

where $y$ is a scalar to refer to bus stops.

To reduce the degrees of freedom that characterize the load profile estimation problem, the following assumption is made on the basis of empirical observations:

Assumption 5. There are no alightings on the first stops of a route, nor are there any boardings on the last stops.
The notion of first and last stops of a given route can be defined percentage-wise by introducing the two following user-defined parameters: 0 < \phi_r < 1 and 0 < \phi_l < 1, respectively. This assumption implies that \epsilon_{ls} = 0 for the first and last stops. The dwell time for the first stops is then reduced to \hat{d}_{s,k} = \beta \cdot b_{l,s} + \gamma, whereas the dwell time for the last stops is simplified into \hat{d}_{s,k} = \alpha \cdot a_{s,k} + \gamma. By applying linear regression models with a constrained solution space (i.e., 2 < \beta < 10) using the well-known least squares as the objective function, the variables \alpha, \beta, and \gamma can be estimated. The constant delay, \gamma, can be taken as the average value of the constants resulting of the two linear regression processes. The number of boarding and alighting passengers for the first and last stops can then be obtained. These estimations will be further used as support vectors to estimate the entire load profile for a given trip, together with the maximum load and the maximum load point of a given trip. This process is detailed in the subsequent section.

Load Profile Estimation Using Constrained Local Regression

The load profile estimation is performed by using local regression, namely, local scatterplot smoothing (LOESS) (16). To apply the LOESS estimation method, support samples should be provided to the regression analysis. In this context, these samples are the values of \epsilon_{ls}, \forall s \in S. Following the discussion in the previous section, the values of \epsilon_{ls} for the first and last stops are known. However, this is not sufficient for estimating the entire load profile. In addition to the support samples, Equations 3 and 4 provide a way to compute the maximum load. However, it is not sufficient to compute the maximum load point. The identification of the maximum load point \hat{s}_k^{\text{max}} for a particular \hat{k} without any passenger-based data is a difficult task. Therefore, the investigation is restricted to understanding the demand for each route for the typical load within a given time period rather than estimating the exact values for each individual trip. Let \hat{s}_k denote the first (furthest upstream) bus stop that experienced the largest dwell time, \hat{d}_{s,k}, on a given trip \hat{k}. It can be computed as

\[
\hat{s}_k = \arg\max_{s \in S} o_{s,k}
\]

(9)

By using these dwell times, the maximum load point of a given trip \hat{k}, \hat{s}_k^{\text{max}}, can be computed as follows:

\[
\hat{s}_k^{\text{max}} = \begin{cases} 
\min \hat{s}_k & \text{if } o_{s,k} < \chi \\
\hat{s}_k & \text{otherwise}
\end{cases}
\]

(10)

where \hat{s}_k: \Sigma_{s,k} = \Sigma_{s,k} / 2, S \subseteq S and \hat{s} and \hat{S} represent subsets of s and S, respectively. This definition implies that the maximum load point is identified as the stop up to which the accumulated dwell time exceeds half of the dwell time for the entire trip or, alternatively, the earliest stop at which the dwell time exceeds a user-defined threshold, \chi.

By following these computations, a set of loads can be obtained, which is denominated as support vector. This set contains the known load values that can be used while estimating the remaining loads. The definitions made by Assumption 5 and Equations 3 and 4 imply that the load profile follows a parabola-like function, where its maximum is located at \hat{s}_k^{\text{max}}. However, this pattern may not prevail for every single trip.

The LOESS method is a regression method that combines linear and nonlinear regression methods in a simple fashion. Instead of trying to fit a function globally (i.e., for all bus stops), it does so locally by fitting models to localized subsets of data to build up a function that can describe the deterministic part of the variation in the data, point by point (i.e., stop by stop). In simple terms, it fits segments of the data (e.g., first and last stops using a simple linear function followed by a parabola shape around the maximum load point). The partitioning of the data is determined by deploying a nearest neighbors algorithm, where the neighborhood concept is given by a bandwidth-type user-defined parameter denoted by \lambda. Usually, the LOESS requires a large amount of data to obtain accurate fits for the target function. The LOESS method is applied in this study for estimating the local shape parameters of each passenger load profile.

The deterministic part of the function is fitted using the dwell times. The first step of the load profile estimation procedure is to fit a possible function to describe \epsilon_{ls}, by using the LOESS method based on the support vector. The interest lies in the first-order derivatives (e.g., is the load going up or down in the next stop). The regression output is constrained to the possible range of load values (0 < \epsilon_{ls} < \zeta, \forall s, k).

Fitting the Dwell Times to the Load Profile Using Incremental Filters

After estimating a constrained \epsilon_{ls} by using the above mentioned procedure, one needs to keep adjusting the results by using the dwell times available from AVL data records. To this end, an incremental filter is employed. This filter is defined stop by stop using the load prediction obtained for the last stop. It is composed of two components:

1. A bandwidth defining the maximum and minimum admissible load values denoted by \epsilon_{ls}^{\alpha} and \epsilon_{ls}^{\omega}, which can be defined as

\[
\epsilon_{ls}^{\alpha} = \epsilon_{l,s} - \frac{d_{l,s}}{\alpha}
\]

and

\[
\epsilon_{ls}^{\omega} = \epsilon_{l,s} + \frac{d_{l,s}}{\beta}
\]

(11)

2. A progression rate function, \rho_{s,k}, to decompose the loading time into boarding and alighting times, defined as

\[
\rho_{s,k} = \begin{cases} 
1 & \text{if } s = [\varphi_s \cdot |S|] \\
0 & \text{if } s = |S| - [\varphi_s \cdot |S|] \\
\rho_{s,k} = \frac{1}{\varphi_s - \varphi_f} & \text{otherwise}
\end{cases}
\]

(12)

where \varphi_s, \varphi_f denote the ratio of stops that are considered first and last stops on the route where the absence of friction (i.e., \epsilon_{ls} = 0) for those stops is assumed, as well as the absence of alightings and boardings for this set of first and last stops, respectively. The progression rate is thus one for the first stops and zero for the last stops and diminishes in between. This function originates from empirical observations and the assumption that the ratio between the number of boarding and alighting passengers is negatively correlated with
the distance from the origin stop on a given route. It is then used to update the load estimation function. Consequently, the updated onboard load estimation is obtained as follows:

\[
\omega_{i,k} = \begin{cases} 
\omega_{i,k} + (\omega_{i,k} - \omega_{i,k}) \cdot \rho_{i,k} \cdot \left[1 - \frac{\omega_{i,k}}{\omega_{i,k}}\right] \frac{\omega_{i,k} - \omega_{i,k}}{2} 
\end{cases} 
\]

\[\text{if } [\varphi_i \cdot |S|] < s < [\varphi_i \cdot |S|] \land s \neq s_{i}^{\text{min}} \]

\[\text{otherwise} \]

By conducting this procedure, it is guaranteed that reasonable and consistent load values are obtained. The information on the load trend is obtained through the local regression framework.

As noted earlier, this calculation is completely unsupervised because the real load values are not known. This situation prohibits the computation of confidence intervals for the predictions, which requires sample standard deviations. To address this limitation, an online procedure was developed to compute a dwell-based load bandwidth that aims to illustrate the uncertainty around the load predictions graphically. It uses a sliding window based on a number of upstream bus stops to assess the range of realistic minimum and maximum loads by using their dwell times (e.g., if \( \alpha = 2 \) and \( d_{i,k} \), then arguably \( a_{i,k} \leq 5 \)).

**Finding a Typical Load Profile**

Instead of fitting each individual bus trip load, this paper proposes estimating the typical passenger load within a short time window. Therefore, the mean load value is calculated for each bus stop and the Euclidean distance between the average load profile and each individual trip load is computed. Finally, a typical trip is selected from the sample that is most similar to the average load profile (i.e., the trip with the minimum Euclidean distance).

**APPLICATION**

**Case Study and Data Description**

The above mentioned methodology was evaluated with AVL data collected from a real-world case study in Dublin, Ireland. Dublin’s urban area has a population of 1.3 million inhabitants. In addition to buses, the public transport network in Dublin also includes heavy and light rail services. The AVL data set available for this study was collected in a continuous manner through a 1-month period (January 2013) for 120 bus routes. In addition, the data set also includes the scheduled time points per route.

AVL data are transmitted by each bus vehicle with 15-s intervals. They include WGS84 coordinates, time stamp, trip ID (which identifies the particular trip assignment that the vehicle is performing, which is recurring), line ID, and a binary value indicating whether the bus is halting at a bus stop. However, the data do not contain information about the trip’s direction or a unique ID to identify each individual trip. Moreover, the data set contains a considerable amount of noise. To tackle such issues, the following data preparation activities were performed: (a) identify the route’s direction of each trip through a binary clustering procedure; (b) exclude trips with incomplete or inconsistent data; (c) assign each trip a unique ID by using the departure date, the original assignment ID, and the trip’s line and direction; (d) match these data with the existing schedule time points; and (e) exclude trips that were not possible to match with the existing schedule because of data inconsistencies (e.g., deviations from the planned mapped route caused by data noise). This process results in a data set that describes the trip trajectory of each route at a stop level and includes the following variables: trip ID, stop ID, latitude and longitude, scheduled arrival and scheduled departure time at stop, actual arrival and departure time at stop, and the observed dwell time. The latter ranges discretely between 0 and 600 s with 15-s steps (since data are collected every 15 s, a nonobserved dwell time is obtained for some stops).

For demonstration purposes, it was decided to test this method on data from two high-frequency routes (140 and 13). The selection criteria were the small amount of missing data (i.e., <10%), the high share of trips during peak hours, and the data’s distinct function in the network. Route 13 connects the airport (north of the city), located in the city’s northwest corner, to Adamstown, a large neighborhood in the westernmost part of the urban area through downtown, serving several transport hubs along its route. Route 140 is a commuter line that connects the northern neighborhood of Poppintree, which lies close to the city outskirts, to the southern neighborhood of Dartry. Figure 2 illustrates the route maps and Tables 1 and 2 summarize information on the number of daily trips, the observed dwell times, and the amount of missing data for these routes. The analysis focuses on the two peak periods, morning (8:00–12:00) and evening (16:00–20:00), which were defined by identifying the periods of the day during which the largest round-trip delays were experienced. Large variations in dwell times are observed on Route 13 (Tables 1 and 2), presumably attributable to demand variations caused by the irregular passenger flows in the airport, which are highly influenced by flight departure and arrival times. The planned headway during the analysis periods ranges between 10 and 30 min. Figure 3 presents the headway distributions of these routes. It is evident that both lines exhibit large headway variations caused by both planning and irregularity in their operations. The irregular demand pattern is arguably also the underlying reason for the highly irregular headways that characterize Route 13.

**Implementation**

All the experiments were conducted by using R software (17). The dwell times were computed by using the midpoint of the registered interval (e.g., if a dwell time of 30 s is recorded, it may range between 30 and 45 s and thus a dwell time of 37.5 s is considered in the analysis). The analysis method involves the specification of six parameters: \( \delta, \varsigma, \varphi_i, \varphi_o, \chi, \lambda \). In the absence of information from the public transport planner on their design criterion, a desired occupancy level of 50% of vehicle capacity was assumed, \( \delta = 0.5 \) where \( S\text{var} = 100\% \). The parameters \( \varphi_i \) and \( \varphi_o \) are used to define the concept of first and last stops. Their value was set to \( \varphi_o = \varphi_i = 10\% \) based on empirical observations. The parameter \( \chi \) is the maximum dwell time threshold for identifying the maximum load point. The parameter was specified after testing the results: \{90, 120, 150\}. As the output profiles on both routes did not vary significantly (i.e., <1%), the lowest available value was chosen (90 s). The parameter \( \lambda \) is a user-defined bandwidth parameter and was tested with all the default values for the implementation provided by the built-in R package [stats]. The same procedure was followed when applying the least squares linear regression method and resulted in dwell time function coefficients estimates of \( \alpha = 3, \beta = 4, \) and
FIGURE 2  Route’s definition illustration with R package (RGoogleMaps): (a) Route 140 and (b) Route 13.

TABLE 1  Descriptive Statistics for Each Route Considered: Number of Stops, Total Trips, Daily Mean, Daily Standard Deviation, and Route Length

<table>
<thead>
<tr>
<th>Route</th>
<th>Number of Stops</th>
<th>Total Trips</th>
<th>Daily Mean</th>
<th>Daily SD</th>
<th>Route Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>45</td>
<td>1,320</td>
<td>43</td>
<td>12</td>
<td>18 km</td>
</tr>
<tr>
<td>13</td>
<td>87</td>
<td>926</td>
<td>30</td>
<td>7</td>
<td>32 km</td>
</tr>
</tbody>
</table>

TABLE 2  Descriptive Statistics for Each Route Considered: Maximum Dwell Time, Mean Dwell Time, Standard Deviation Dwell Time, and Missing Data

<table>
<thead>
<tr>
<th>Route</th>
<th>Maximum DwT</th>
<th>Mean DwT</th>
<th>SD DwT</th>
<th>Missing Data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>130</td>
<td>660</td>
<td>11.02</td>
<td>37.49</td>
<td>9.01</td>
</tr>
<tr>
<td>13</td>
<td>1,305</td>
<td>10.02</td>
<td>59.43</td>
<td>15.88</td>
</tr>
</tbody>
</table>

Note: DwT = dwell time.

FIGURE 3  Headway distribution of (a) Route 140 and (b) Route 13. Planned headways during peak hours range between 10 and 30 min and 10 and 20 min, respectively.
γ = 10, all in seconds. These values are consistent with dwell time estimates reported in the literature and recommended by Kittelson and Associates, Inc. (18).

Results

Figure 4 illustrates an example of how the framework performs over a single trip on Route 140. The maximum load point is expected at Stop 26 while Stop 8 experiences the longest dwell time and therefore introduces large variation into the estimation procedure.

Load profiles were estimated for each bus trip and were then analyzed jointly for each route direction and time period. Figures 5 and 6 present the load profile obtained for each one of the two routes during the morning and evening peak periods. The typical load profile is highlighted in each case. It is evident that the estimated load profiles for individual trips demonstrate considerable variation. Such variations could be expected by service irregularity and demand variations. However, in the absence of ground truth passenger demand data, it was not possible to verify the extent of these variations. However, the variations in load profile estimates mirror the extent of headway variations for both routes. A preliminary sensitivity analysis suggests that the estimation results are robust with respect to the dwell time threshold (χ) and the share of first and last stops (φf, φl) that are used for estimating the dwell time coefficients. In contrast, the estimation results are sensitive to
the desired occupancy value ($\delta$) since it determines the reference load value at the max load point, which is then used when scaling the remaining load profile based on AVL data. This interpretation is therefore focused on the first-order derivative of the load profile and how it evolves rather than on the exact absolute values.

The load profile estimates provide operators and schedule planners a direct visual insight into which stops are subject to large demand variations. Figure 5 suggests that Route 140 has a more uniform (over stops) and stable (over trips) passenger load when compared with Route 13. The latter exhibits several load profile peaks that differ between the morning and evening peak periods. Furthermore, the estimated load profiles provide insights into how a bus route performs with respect to the number of trips and trip segments that are expected to carry passenger volumes that exceed the desired onboard occupancy (e.g., 50 passengers in this experiment).

Obviously, the low granularity of the data in this case study (15 s) as well as the absence of any information about the stops (e.g., nearby and faraway from a signalized intersection) or the special operations conducted during the dwells (e.g., wheelchair boardings) may appear to be major limitations of this framework, because the computed dwells may not always correspond to the real ones. However, this methodology attempts to model the typical demand behavior. Consequently, such rare events are naturally pruned throughout the last step of the framework, where the median profile is considered as reference to select a trip representative of the entire input (statistical) population, even though meta information about the vehicles and the stop’s location could indeed improve the framework robustness to such issues.

Moreover, the assumption introduced in Equation 13 about the progression rate poses a big issue in case the route demand behavior follows a considerably different pattern. Yet this specific issue may be countered by including any other type of high-level prior knowledge of the demand patterns along a specific route (e.g., maximum load points or big interface hubs).

**CONCLUSION**

This paper reports an explorative study into the feasibility of estimating passenger loads solely on the basis of AVL data. The methodology proposed in this study consists of a sequence of steps that involve the identification of the max load point and the corresponding load by reserve engineering the frequency determination methods. Dwell time function coefficients are then estimated on the basis of locally constrained linear regression models. Passenger loads are constructed by applying machine learning algorithms to smoothen the load profile based on actual dwell time records. The typical load profile is then obtained for each time period. The feasibility of the proposed methodology was tested for a case study in Dublin, which demonstrates its potential value.

The proposed method can be integrated into an operation planning software to support operators in designing timetables and allocating resources for improving service reliability. The deployment of such an estimation method can save operators the high costs associated with equipping the bus fleet with APC devices or be useful in case that the operator does not own the fleet or has no access to detailed APC and AFC data. To the best of the authors’ knowledge, this study is the first attempt to uncover the potential of AVL data in providing information on passenger demand.

Public transport service planning involves assessing the impacts of alternative service provisions on travelers. Information on travel demand is therefore essential in supporting authorities and operators in the service planning process. Estimates of onboard passenger loads based on the method proposed in this study could be used for determining whether service frequency or vehicle capacity is adequate and identifying potential for stop consolidation. Furthermore, key performance indicators such as vehicle utilization rate, empty seat running distance, and exceeded load running distance can be approximated on the basis of the estimated load profile (3). These indicators can support service providers in the assessment of service effectiveness across the network.

Further research is needed to validate and improve the proposed method. In particular, the performance of the estimation method should be validated against passenger counts by examining the mean absolute error. The authors are currently exploring the possibility of testing the method for a system in which such data are available. The consideration of different time windows for establishing the typical passenger load will allow examining the possible real-time deployment of the proposed method. Moreover, some of the assumptions made in this paper can be relaxed and based on the operational practice. For example, accounting for mixed fleet operations or introducing fuzzy logic into the max load point selection.
REFERENCES


*The Public Transportation Group peer-reviewed this paper.*