Towards an AVL-based Demand Estimation Model

Morriea-Matias , Luis; Cats, Oded

Publication date
2016

Document Version
Final published version

Published in
Transportation Research Record

Citation (APA)
Morriea-Matias , L., & Cats, O. (2016). Towards an AVL-based Demand Estimation Model. *Transportation Research Record*, *2544*, 141–149. http://10.3141/2544-16

Towards an AVL-based Demand Estimation Model

Luis Moreira-Matias (cor. author)
Research Scientist, NEC Laboratories Europe
Kurfürsten-Anlage 36, 69115 Heidelberg, Germany
phone: 0049-6221-4342261
luis.matias[at]neclab.eu

Oded Cats
Assistant Professor, Dep. Transport and Planning
Delft University of Technology, 2600 GA Delft, The Netherlands
o.cats[at]tudelft.nl

A Paper Submitted for Presentation at the 2016 Annual Meeting
of the Transportation Research Board and Publication in the
Transportation Research Record

4958 Words + 6 figure(s) (1500 words) + 1 table(s) (250 words) = 6708 words

November 4, 2015

ABSTRACT
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 is often available for the entire public transport fleet for monitoring purposes. However, the potential value of AVL in estimating passenger volumes has been overlooked. In this study, we examine whether AVL data can be used as a standalone source for estimating on-board bus loads. The modeling approach is to infer maximum passenger load stop from the timetable and then construct the load profile by reverse engineering through a local constrained regression of dwell times as function of passengers flows. In order to test and demonstrate the potential value of the proposed method, a proof of concept was performed by conducting unsupervised experiments on one month AVL data collected from two bus lines in Dublin. 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.

INTRODUCTION

Understanding passenger demand is key for the effective planning and provision of public transport services. Over the last decades, mass transit operators worldwide relied on passenger surveys to understand their mobility needs and adjust their planning and operations accordingly (1, 2, 3). 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. By observing current service attributes, service management could adapt the service to better respond to passenger travel needs. The implementation of such measures require information on passenger flows in order to assess the expected effects of such measures. For example, when deciding whether to allocate an additional vehicle to reduce on-board 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 (APC) and automated fare collection (AFC), the data collected by those systems is often incomplete and hinders 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, in order 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 is 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 checking in and out when boarding and alighting each vehicle, requiring excessive big data analytics and a large number of behavioral assumptions in order 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, on-board) (5) or passengers travel demand at the network level (origin-destination matrix) (6, 7, 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 based on smartcard transactions, often use Automatic 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 is often available for the entire public transport fleet for monitoring purposes. AVL technologies are more well-established and their installation cost has reduced significantly over the years when compared with APC (9). AVL data has been extensively used for studying the determinants of running times, dwell times and headways. In particular, a large number of studies estimated the determinants of dwell time and in particular the relation between boarding and alighting passenger flows on dwell time based on a combination of AVL and APC data (e.g. (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 one to estimate and/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 our knowledge, none of the previous studies suggested using AVL for estimating passenger flows.

In this study, we examine whether AVL data can be used as a standalone source for estimating real-world passenger loads. The modeling approach is to infer maximum passenger load stop from the timetable and then construct the load profile by reverse engineering through a local constrained regression of dwell times as function of passengers flows. A series of machine learning methods and principles are applied in order to estimated boarding and alighting flows based on actual dwell times and the planned schedule. The resulting framework is denominated as *DemandLOCkeR* - Demand Estimation through LOcal Constrained Regression.

The remainder of the paper is structured as follows: Section 2 presents the method proposed in this study and the related estimation procedure. Section 3 describes the case study and data which were selected for testing the feasibility and performance of the proposed method. Section 4 presents the experimental setup along with the results of the application. In Section 5 we conclude with a discussion on the implications and limitations of this study and outline potential directions for future work.

METHODOLOGY

Analysis Approach

The approach adopted in this study (*DemandLOCkeR*) for passenger demand estimation relies solely on AVL data involves reverse engineering where the relation 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 our analysis are 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 Fig. 1. The methodology for estimating bus load profiles using AVL data consists of five steps: (A) extracting high-level demand information from the planned timetable, assuming that they were designed based on a max load point method; (B) decomposing real-time dwell times and regressing them based on 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 which 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 Section 3.4); (D) constraining and fitting the results obtained in the previous step based on the actual dwell times and an incremental bandwidth (defined by domain constrains which force a fitting of the regression outputs within the range of admissible loads, given/known each vehicle's capacity) that uses only the most recent dwell time records to obtain realistic load profiles, and; (E) 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/minimum admissible values for the loads on every stop given the load prediction for the immediate previous one - as adequately

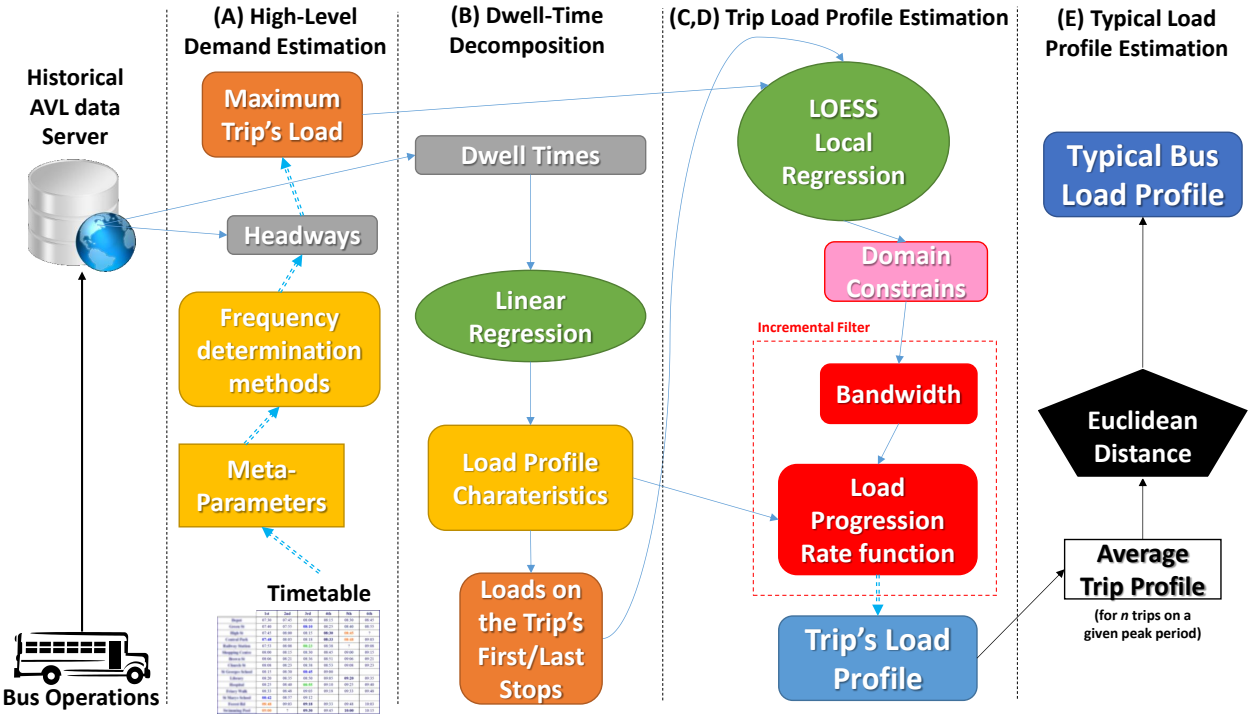


FIGURE 1 : Analysis framework – from data to load profile estimations.

described in Section 3.5). 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, we 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 based on passenger surveys and direct observations (1, 2, 3). There are two different ways of determining such frequencies: (i) stop-based and (ii) route-based. The latter one 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 j of length τ (e.g. $\tau = 60$ minutes), i.e. f_j as follows

$$f_j = \max\left(\frac{o_{s,j}^{\max}}{o_j^d}, f_j^{\min}\right), \forall j \tag{1}$$

where $o_{s,j}^{\max} = \max o_{s,j}, \forall s \in S$ stands for the average/measured on-board occupancy when departing from stop s during time period j for a certain line and S is the set of all stops except for the last stop on the respective line. o_j^d is the desired occupancy for the same time period and $f_j^{\min}$ is the minimum frequency defined by policy makers. In order to extract information on the demand pattern, the following set of assumptions is made:

Assumption 1 *The entire fleet has an* ***equal capacity*** *of ς passengers;*

Assumption 2 *o_j^d is defined by a pre-defined constant value $0 < \delta < 1$ (i.e. percentage-wise definition) for each route and period j, i.e. $o_j^d = \delta \cdot \varsigma$;*

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

Assumption 4 *The first term in Eq.* (1) *is binding. In other words, the frequency needed in order to satisfy the load-desired occupancy ratio exceeds the minimum policy frequency.*

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

Based on these assumptions, it is possible to re-write Eq. 1 as follows

$$o_{s,j}^{\max} = \varsigma \cdot \delta \cdot f_j = \varsigma \cdot \delta \cdot \frac{3600}{\bar{h_j^p}} \tag{2}$$

where $\bar{h_j^p}$ denotes the average planned headway during period j (in seconds). Let $l_m(j,t)$ 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 p. Based on the above relation between max load point and headway, the maximum load of a specific bus trip $k \in K_j$, $o_{s,k}^{\max}$, can be estimated based on observed headways derived from AVL:

$$o_{s,k}^{\max} = \varsigma \cdot \delta \cdot f_j = \varsigma \cdot \delta \cdot \frac{3600}{\bar{h_k}} \tag{3}$$

where K_j is the set of bus trips that operate on a given line during period j and $\bar{h_k}$ is the average observed headway calculated as

$$\bar{h_k} = \sum_{s \in S} \frac{h_{s,k-1} + h_{s,k}}{2|S|} \tag{4}$$

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

$$s_k^{\max} = \arg\max_{s \in S} o_{s,k}, \forall k \in K_j \tag{5}$$

However, the passenger loads upon departing from each stop along trip k, $o_{s,k}$, are unknown. In the following section, these values are estimated based on the dwell times available from AVL data.

Decomposing Dwell Times

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

$$d_{k,s} = \gamma + \max(\alpha \cdot a_{k,s}, \beta \cdot b_{k,s}) + c_{k,s} + \epsilon \tag{6}$$

where α and β are the average alighting and boarding time per passenger, respectively, and, $a_{k,s}$ and $b_{k,s}$ represent the number of alighting and boarding passengers. γ is the fixed delay due to door opening and closing times and ϵ is an error term caused by variations in driver and passenger behavior that is assumed to be distributed $\epsilon \sim N(0, \sigma^2)$. $c_{k,s}$ is the additional dwell time due to on-board crowding and interactions between passengers in crowded situations. In line with the formulation of Weidmann (15), the delay due to on-board crowding can be expressed as a penalty that prolongs the constant dwell time delay:

$$d_{k,s} = \max(\alpha \cdot a_{k,s}, \beta \cdot b_{k,s}) + (\gamma \cdot (1 + e_{k,s})^2) \tag{7}$$

where $e_{k,s}$ is the friction element defined as

$$e_{k,s} = \begin{cases} \left(\max(\alpha \cdot a_{k,s}, \beta \cdot b_{k,s}) - \varsigma \cdot \delta\right) \cdot 1/100 & \text{if } \max(\alpha \cdot a_{k,s}, \beta \cdot b_{k,s}) \geq \varsigma \cdot \delta, \forall i \in j \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The relation between on-board occupancy of trip k upon departure from stop s to past boarding and alighting flows is

$$o_{s,k} = \sum_{y=1}^{s} (b_{k,y} - a_{k,y}) \tag{9}$$

In order to reduce the degrees of freedom that characterize the load profile estimation problem, the following assumption is made based on empirical observations:

Assumption 5 *There are no alightings on the first stops of a route neither boardings on the last ones.*

The notion of *first* and *last* stops of a given route can be defined percentage-wise by introducing the two following user-defined parameters: $0 < \varphi_f << 1$ and $0 < \varphi_l << 1$, respectively. This assumption implies that $e_{k,s} = 0$ for the first and last stops. The dwell time for the first stops is then reduced to $d_{k,s} = \beta \cdot b_{k,s} + \gamma$, whereas the dwell time for the last stops is simplified into $d_{k,s} = \alpha \cdot a_{k,s} + \gamma$. By applying linear regression models with a constrained solution space (i.e. $2 < \beta, \beta < 10$) using the well-known least squares as objective function, α, β and γ can be estimated. The constant delay, γ , 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/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 preformed using Local Regression, namely, Local Scatterplot Smoothing (LOESS) (16). In order to apply the LOESS estimation method, support samples should be

provided to the regression analysis. In our context, these samples are the values of $o_{s,k}, \forall s \in S$. Following the discussion in the previous section, the values of $o_{s,k}$ 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, the eqs. (3,4) provide a way to compute the maximum load. However, this is not sufficient to compute the maximum load point.

The identification of the maximum load point $s_k^{\max}$ for a particular k without any passenger-based data is a difficult task. Therefore we restrict our investigation 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 (furthermost upstream) bus stop which experienced the largest dwell time, $d_{k,\hat{s}}$, on a given trip k. It can be computed as

$$\hat{s}_k = \arg\max_{s \in S} o_{s,k} \tag{10}$$

Using these dwell times, we propose to compute the maximum load point of a given trip k, $s_k^{\max}$, as follows

$$s_k^{\max} = \begin{cases} \min_{\grave{s} \in \grave{S}} \grave{s} \text{ if } & o_{s,k} < \chi \\ \hat{s}_k & \text{otherwise.} \end{cases} \tag{11}$$

where $\grave{s} \in \grave{S} : \sum_{y=1}^{\grave{s}} o_{s,k} \geq \sum_{y=1}^{|S|} {}^{o_{s,k}}\!/_2, \grave{S} \subseteq S$. This definition implies that the max 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, χ.

By following these computations, we obtain a set of loads which we denominate as *support vector*. This set contains the known load values which we can use while estimating the remaining loads. The definitions made by the Assumption 5 and eqs. (3,4) imply that the load profile follows a parabola-like function - where its maximum is located at $s_k^{\max}$. However, this pattern may not prevail for every single trip.

LOESS is a regression method which combines linear/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 which 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/last stops using a simple linear function followed by a parabolic 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 λ. Usually, the LOESS requires a large amount of data to obtain accurate fits for the target function. LOESS 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 $o_{s,k}$, using the LOESS method based on the support vector. Our 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 ($0 < o_{s,k} < \varsigma, \forall s, k$).

Fitting the Dwell-Times to the Load Profile using Incremental Filters

After estimating a constrained $o_{s,k}$ using the abovementioned procedure, we need to keep adjusting their results using the dwell times available from AVL data records. To this end, we employ an

incremental filter. This filter is defined stop-by-stop by 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 $o^+_{s,k}$ and $o^-_{s,k}$, whose can be defined as:

$$\begin{aligned} o^-_{s,k} &= o_{s-1,k} - {}^{d_{s,k}}\!/_{\alpha} \\ o^+_{s,k} &= o_{s-1,k} + {}^{d_{s,k}}\!/_{\beta} \end{aligned} \tag{12}$$

(2) a progression rate function, $\rho_{s,k}$, to decompose the loading time into boarding and alighting times, defined as:

$$\rho_{s,k} = \begin{cases} 1 & \text{if } s = \lceil \varphi_f \cdot |S| \rceil \\ 0 & \text{if } s = |S| - \lceil \varphi_l \cdot |S| \rceil \\ \rho_{s-1,k} - \frac{1}{\varphi_l - \varphi_f} & \text{otherwise.} \end{cases} \tag{13}$$

where φ_f, φ_l denote the ratio of stops which are considered first/last stops on the route where it is assumed the absence of friction (i.e. $e_{k,s} = 0$) for those stops, as well as the absence of alightings/boardings for this set of first/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 to the load estimation function. Consequently, the updated on-board load estimation is obtained as follows.

$$\widetilde{o_{s,k}} = \begin{cases} o^-_{s,k} + (o^+_{s,k} - o^-_{s,k}) \cdot \rho_{s,k} + \left[1 - \frac{o_{s,k}}{o_{s-1,k}}\right] \cdot \frac{o^+_{s,k} - o^-_{s,k}}{2} \text{ if} \\ \qquad \lceil \varphi_f \cdot |S| \rceil < s < |S| - \lceil \varphi_l \cdot |S| \rceil \wedge s \neq s^{\max}_k, \\ o_{s,k} \qquad \text{otherwise.} \end{cases} \tag{14}$$

By conducting this procedure, we guarantee that reasonable and consistent load values are obtained. Note that the information on the *load trend* is obtained through the local regression method, which results in a constrained local regression framework.

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

Finding a Typical Load Profile

Instead of fitting each individual bus trip load, we propose estimating the typical passenger load within a short time window. We thus calculate the mean load value for each bus stop and compute the Euclidean Distance between the average load profile and each individual trip load. Finally, we select as typical trip from the sample which is most similar to the average load profile - the trip with the minimum Euclidean distance.

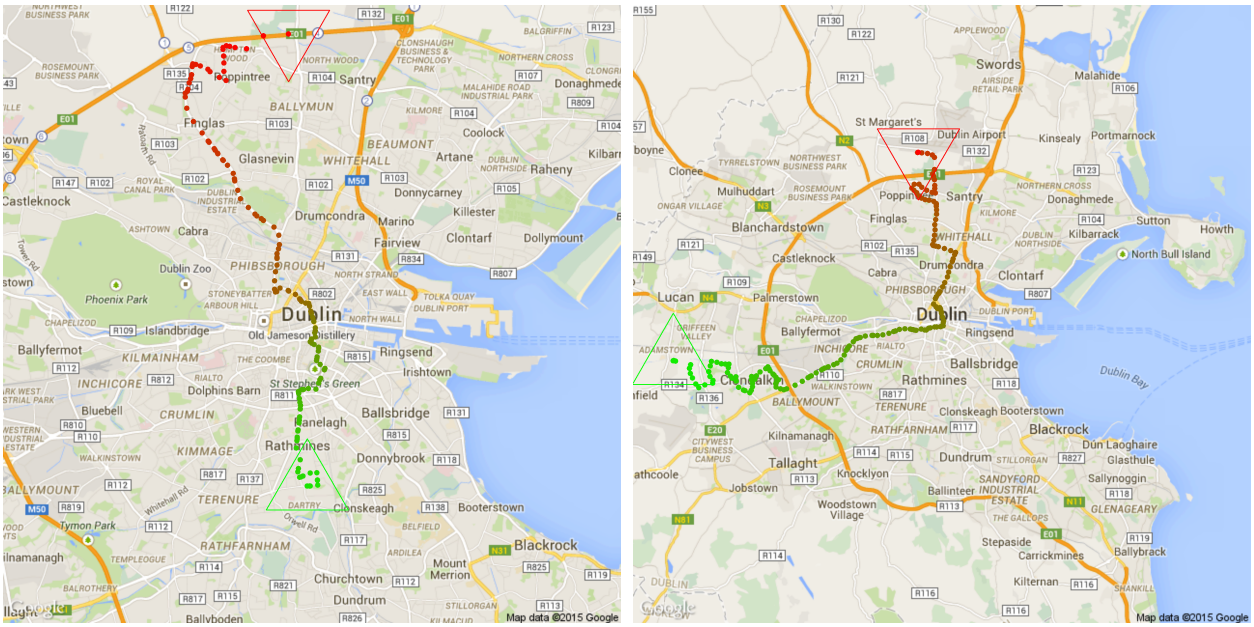
APPLICATION

Case Study and data description

The abovementioned methodology was evaluated using 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 includes also heavy and light rail services. The AVL dataset available for this study was collected on a continuous manner through one month period (January 2013) for 120 bus routes. In addition, the dataset also includes the scheduled time points per route.

AVL data is transmitted by each bus vehicle with 15-second intervals. It includes WGS84 coordinates, timestamp, 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 or not. However, it contains neither information regarding the trip's direction nor a unique ID to identify each individual trip. Moreover, the dataset contains a considerable amount of noise. To tackle such issues, the following data preparation activities were performed: 1) identify the route's direction of each trip through a binary clustering procedure; 2) exclude trips with incomplete or inconsistent data; 3) assign each trip a unique ID using the departure date, the original assignment ID, the trip's line and direction; 3) match this data with the existing schedule time points; 4) exclude trips which were not possible to match with the existing schedule due to data inconsistencies (e.g. deviations from the planned mapped route due to data noise). This process results with a dataset which 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 with 15 seconds steps (since data is collected every 15 seconds, we obtain a non-observed dwell time for some stops).

For demonstration purposes, we choose to test our method on data from two high frequency routes (140 and 13), respectively. The selection criteria were the small amount of missing data (i.e. $< 10\%$), the high share of trips during peak hours and its 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 which connects the northern neighborhood of *Poppintree*, which lies close to the city outskirts, to the southern neighborhood of *Dartry*. Fig. 2 illustrates the route maps and Table 1 summarizes 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 (Table 1), presumably due to demand variations caused by the irregular passenger flows in the airport which is highly influenced by flight departure and arrival times. The planned headway during the analysis periods ranges between 10 and 30 minutes. Fig. 3 presents the headway distributions of these routes. It is evident that both lines exhibit large headway variations due to 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.



(a) Route 140. (b) Route 13.

FIGURE 2 : Route's definition illustration using R package `[RGoogleMaps]`.

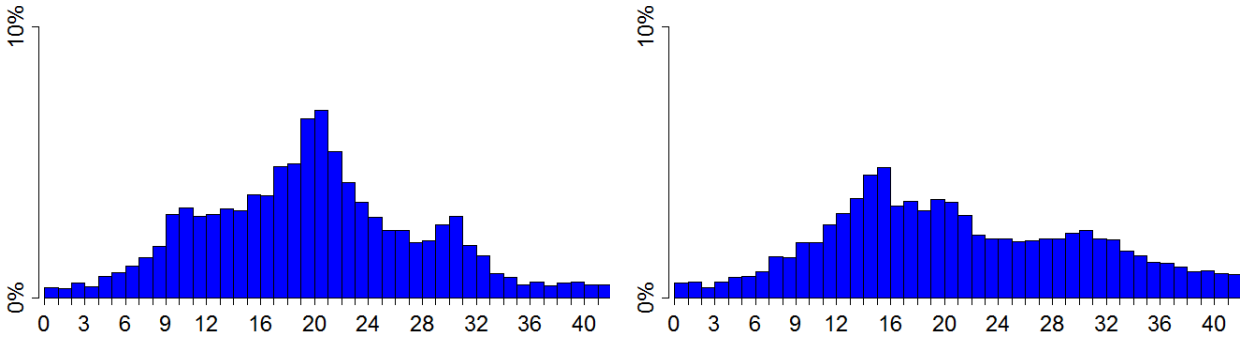


FIGURE 3 : Headway distribution of route 140 (left-side) and 13 (on right-side). Planned headys on peak hours range between 10-30 and and 10-20, respectively. Times in minutes.

TABLE 1 : Descriptive statistics for each route considered. Dwell Times (DwT) in seconds.

Route	Nr. Stops	Total Trips	Daily Mean	Daily Std. Dev.	Route Length
140	45	1320	43	12	18km
13	87	926	30	7	32km
Route	**Max. DwT**	**Mean DwT**	**Std. Dev. DwT**	**Missing Data**	
140	660	11.02	37.49	9.01%	
13	1305	10.02	59.43	15.88%	

Implementation

All the experiments were conducted using the R Software (17). The dwell times were computed by using the midpoint of the registered interval (e.g. if a dwell time of 30 seconds is recorded, it may in fact range between 30 and 45 seconds and thus a dwell time of 37.5 seconds is considered in the analysis). The analysis method involves the specification of six parameters: $\{\delta, \varsigma, \varphi_f, \varphi_l, \chi, \lambda\}$. 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 $\varsigma = 100$. φ_f, φ_l are used to define the concept of *first/last* stops. Their value was set to $\varphi_f = \varphi_l = 10\%$ based on empirical observations. χ 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 seconds). λ 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 with dwell time function coefficients estimates of $\alpha = 3$, $\beta = 4$ and $\gamma = 10$, all in seconds. These values are consistent with dwell time estimates reported in the literature and recommended by the (18).

Results

Fig. 4 illustrates an example of how our framework performs over a single trip on route 140. Note that 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. Figs. 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

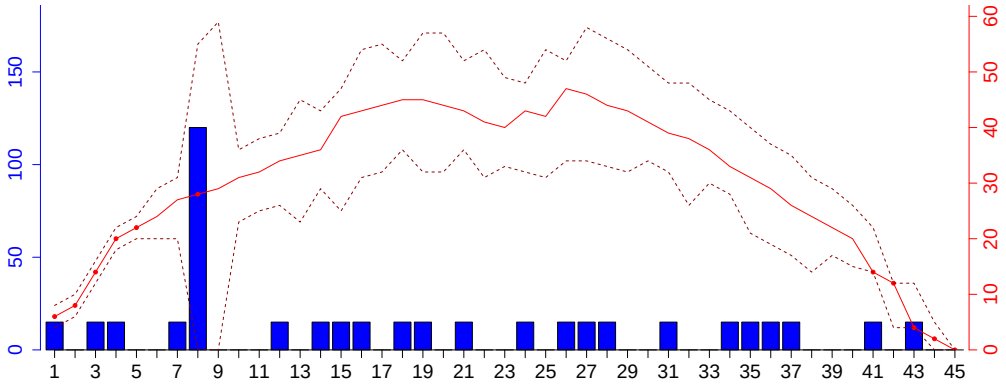


FIGURE 4 : Estimated load profile for a selected trip. The blue bars show the dwell time recorded at each stop. The red line is the estimated load for this particular trip running on route 140. The dashed lines define an interval for the expected load variation for each stop.

(φ_f, φ_l) which are used for estimating the dwell time coefficients. In contrast, the estimation results are sensitive to the desired occupancy value (δ) 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. We therefore focus in our interpretation on the first-order derivative of the load profile and how it evolves rather than 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. Fig. 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 which differ between the morning and evening peak periods. Furthermore, the estimated load profiles provide insights into how a bus route preforms in terms of the number of trips and trip segments that are expected to carry passenger volumes that exceed the desired on-board occupancy (e.g. 50 passengers in this experiment).

Obviousily, the low granularity of the data in this case study (15 seconds) as well as the absence of any information regarding the stops (e.g. nearby/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 - as the computed dwells may not always correspond to the real ones. However, this methodology attemp to model the **typical** demand behavior. Consequently, such rare events are naturally prunned 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 eq. 13 about the progression rate poses a big issue in case the route demand behavior follow 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, big interface hubs, etc.).

CONCLUSION

This paper reports an explorative study into the feasibility of estimating passenger loads based solely on AVL data. The methodology proposed in this study consists of a sequence of steps which 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 based on

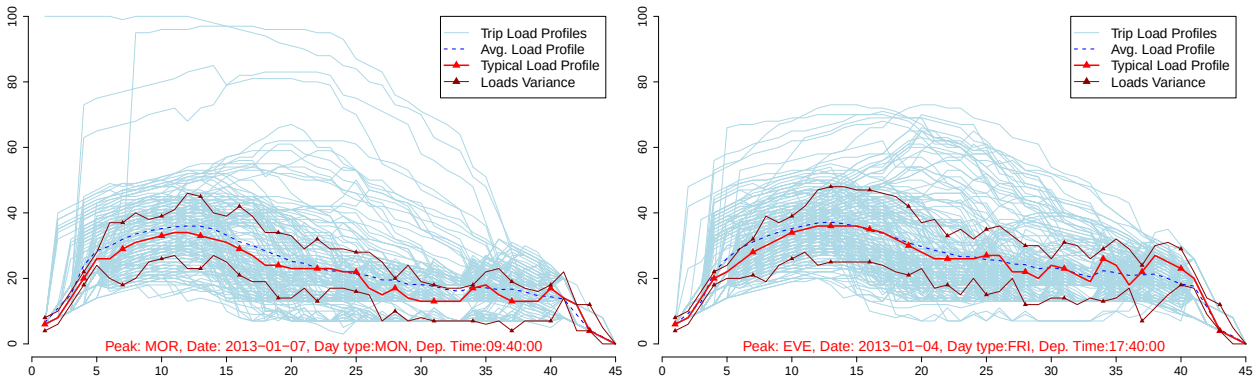


FIGURE 5 : Load Profiles generated for route 140 (morning/evening peak on left/right-sides).

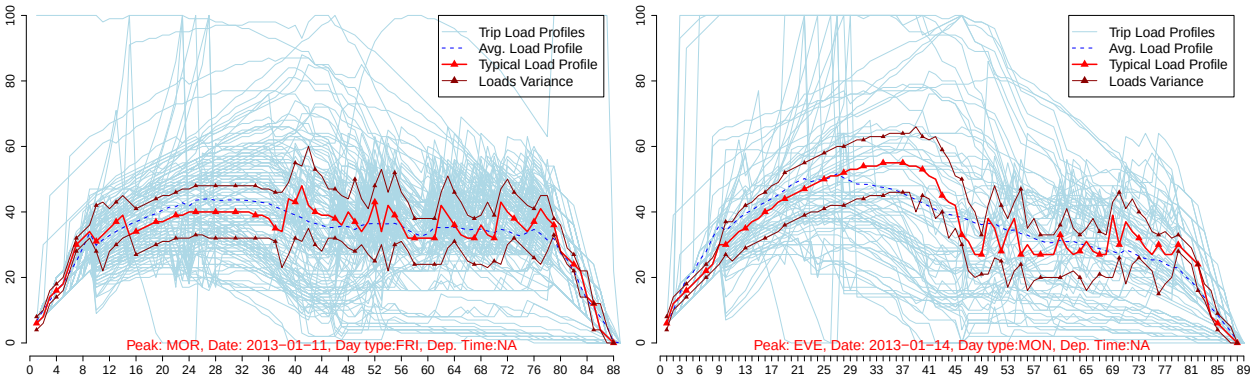


FIGURE 6 : Load Profiles generated for route 13 (morning/evening peak on left/right-sides).

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/AFC data. To the best of the authors knowledge, this 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 on-board passenger loads based on the method proposed in this study could be used for determining whether service frequency or vehicle capacity are 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 based on 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 currently explore the possibility of testing the method for a system where such data is 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 in to the max load point selection.

REFERENCES

[1] Richardson, A., E. Ampt, and A. Meyburg, *Survey methods for transport planning*. Eucalyptus Press Melbourne, 1995.

[2] Vuchic, V., *Urban Transit: Operations, Planning, and Economics*. Wiley, 2005.

[3] Ceder, A., *Public transit planning and operation: theory, modeling and practice*. Elsevier, Butterworth-Heinemann, 2007.

[4] Pelletier, M., M. Trepanier, and C. Morency, Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, Vol. 19, No. 4, 2011, pp. 557 – 568.

[5] Rahbee, A. and D. Czerwinski, Using entry-only automatic fare collection data to estimate rail transit passenger flows at CTA. In *Proceedings of the 2002 Transport Chicago Conference*, 2002.

[6] Trépanier, M., N. Tranchant, and R. Chapleau, Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, Vol. 11, No. 1, 2007, pp. 1–14.

[7] Lee, S. G. and M. D. Hickman, Travel pattern analysis using smart card data of regular users. In *Proceedings of the 90th Annual Meeting of the Transportation Research Board*, 2011.

[8] Wang, W., J. Attanucci, and N. Wilson, Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, Vol. 14, No. 4, 2011, p. 131.

[9] Nielsen, B., L. Frolich, O. Nielsen, and D. Filges, Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A: Transport Science*, Vol. 10, No. 6, 2014, pp. 502–517.

[10] Chen, C., Y. Chang, T. Chen, and D. Wang, People Counting System for Getting In/Out of a Bus Based on Video Processing. In *Intelligent Systems Design and Applications, 2008. ISDA '08. Eighth International Conference on*, 2008, Vol. 3, pp. 565–569.

[11] Dueker, K., T. Kimpel, J. Strathman, and S. Callas, Determinants of bus dwell time. *Journal of Public Transportation*, Vol. 7, No. 1, 2004, pp. 21–40.

[12] Tirachini, A., Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers. *Transportmetrica A: Transport Science*, Vol. 9, No. 1, 2013, pp. 28–49.

[13] Shalaby, A. and A. Farhan, Bus travel time prediction model for dynamic operations control and passenger information systems. *The 82nd Annual Meeting of the Transportation Research Board*, 2003.

[14] Furth, P., B. Hemily, T. Muller, and J. Strathman, *Uses of archived AVL-APC data to improve transit performance and management: Review and potential*. Transportation Research Board, 2003.

[15] Weidmann, U., *Der Fahrgastwechsel im öffentlichen Personenverkehr (In German)*. Ph.D. thesis, Diss. Techn. Wiss. ETH Zürich, Nr. 10630, 1994. Ref.: Heinrich Brändli; Korref.: Adolf Müller-Hellmann, 1994.

[16] Cleveland, W., Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, Vol. 74, No. 368, 1979, pp. 829–836.

[17] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[18] TCRP, *Transit Capacity and Quality of Service Manual*, Vol. 100. Transportation Research Board, 2003.