# Highlights

## Predicting the probability distribution of bus travel time to measure the reliability of public transport services

Léa Ricard,Guy Desaulniers,Andrea Lodi,Louis-Martin Rousseau

- We propose probabilistic models for travel time density prediction
- We compare two types of probabilistic models to a Random Forests model
- We proposed a simulation to approximate the delay tolerance from travel time density
- Probabilistic models prevail in generating accurate estimates of the delay tolerance

# Predicting the probability distribution of bus travel time to measure the reliability of public transport services

Léa Ricard[a,*], Guy Desaulniers[b], Andrea Lodi[b] and Louis-Martin Rousseau[b]

[a]Department of Computer Science, Université de Montréal, Montréal, Canada
[b]Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal, Canada

ARTICLE INFO

*Keywords*:
Long-term prediction
Travel time
Probabilistic model
Public transport
Reliability

ABSTRACT

An important aspect of the quality of a public transport service is its reliability, which is defined as the invariability of the service attributes. In order to measure the reliability during the service planning phase, a key piece of information is the long-term prediction of the density of the travel time, which conveys the uncertainty of travel times. This work empirically compares probabilistic models for the prediction of the conditional probability density function (PDF) of the travel time and proposes a simulation framework taking as input the latter distributions to approximate the expected secondary delays, a measure of the reliability of public transport services. Two types of probabilistic models, namely similarity-based density estimation models and a smoothed Logistic Regression for probabilistic classification model, are compared on a dataset of more than 41,000 trips and 50 bus routes of the city of Montréal. A similarity-based density estimation model using a $k$ Nearest Neighbors method and a Log-Logistic distribution predicted the best estimate of the true conditional PDF of the travel time and generated the most accurate approximations of the expected secondary delays on this dataset. This model reduced the mean squared error of the expected secondary delay by approximately 9% compared to the benchmark model, namely a Random Forests. This result highlights the added value of modeling the conditional PDF of the travel time with probabilistic models.

## 1. Introduction

In order to increase the ridership and attract new users, public transport agencies put increasing emphasis on improving the quality of the service they provide and particularly its regularity, also referred to as reliability (Ma, Ferreira and Mesbah, 2014). Studies show that a majority of passengers put more value on a reduction of the travel time (TT) variability than on a reduction of TT itself (Bates, Polak, Jones and Cook, 2001). Reliability can be addressed at different levels, either during the strategic planning, the tactical planning or the operational planning stages and during operations. At the strategic planning level, adding reserved lanes for buses can increase the reliability of the service, while during operations, bus holding is a popular solution to alleviate risks of bus bunching. The latter consists of holding a bus at key locations along a bus trip if it is running ahead of time. However, service reliability is rarely taken into account at the tactical and operational planning levels, when the detailed planning of the service is computed (van Oort, 2011). The network design, the frequencies and/or timetables of buses, the vehicle schedules and the crew schedules are built during these stages, among other things (Desaulniers and Hickman, 2007). This work aims at providing tools to measure, and eventually improve, the reliability of one of the output of the service planning phase, namely vehicle schedules. These schedules are defined as a sequence of timetabled trips and waiting times starting and ending at the same depot, such that each travel is either a timetabled trip or a deadhead trip (e.g., between a depot and a terminal or between two terminals). A deadhead trip between two terminals enables the connection of two timetabled trips ending and starting at different terminals.

To assess the reliability of a vehicle schedule, Kramkowski, Kliewer and Meier (2009) introduced the concept of delay tolerance, a term reused in the works of Amberg, Amberg and Kliewer (2019); van Kooten Niekerk (2018), among others. This concept is based on primary and secondary delays that we distinguish below. On the one hand, a primary delay (or exogenous delay) is a deviation from the planned duration of a timetabled trip caused by a disruption (e.g., bus bunching) or variability during operation. This type of delay cannot be avoided by scheduling decisions.

*Corresponding author
lea.ricard@umontreal.ca (L. Ricard); guy.desaulniers@gerad.ca (G. Desaulniers); andrea.lodi@polymtl.ca (A. Lodi); louis-martin.rousseau@polymtl.ca (L. Rousseau)
ORCID(s): 0000-0002-2061-1170 (L. Ricard)

Indeed, day-to-day disruptions and delays are considered unavoidable on the day of operation (Amberg et al., 2019; Kramkowski et al., 2009) due to the randomness of incidents and the variation in demand and capacity factors. Bus sharing the road with other road-based vehicles (e.g., cars, bikes and trucks) are likely to have even higher degrees of variability, because they are subject to the same - morning and evening peaks - traffic patterns (Comi, Nuzzolo, Brinchi and Verghini, 2017). On the other hand, a secondary delay (or endogenous delay) occurs when the primary delays of previous trips using the same resource (e.g., vehicle or crew) cannot be absorbed during idle time and thus propagate to the next trip. If a trip starts on time, its secondary delay is null. Otherwise, it is equal to the delay at the departure. Scheduling decisions, that is, the allocation of timetabled trips to resources, can influence the expected secondary delays of timetabled trips. Thus, the delay tolerance of a vehicle schedule is measured by the average expected secondary delay of its timetabled trips.

Secondary delays are stochastic, meaning that the departure of a timetabled trip may be late on a given day and on time on the next day even if the two trips belong to the same vehicle schedule because the secondary delay of a trip depends on the TTs of the previous trips covered in the schedule, which are also stochastic. TT variability is explained by yearly, monthly, day-to-day and hourly variability as well as vehicle-to-vehicle variability (Kumar, Vanajakshi and Subramanian, 2014; Büchel and Corman, 2018; Kieu, Bhaskar and Chung, 2015). The following equations show the dependence between the secondary delay and the TT. Consider a vehicle schedule $s = \{v_1, v_2, ..., v_{m_s}\}$ with $m_s$ trips planned on a given day. For notational conciseness, we denote the trip $v_i$ by $i$ directly in the following. The secondary delay $R_i$ of a trip $i$ is the difference between its actual departure time $D_i$ and its planned departure time $d_i$(assuming that $D_i \geq d_i$), computed as

$$R_i = D_i - d_i. \tag{1}$$

The random variable $D_i$ is a convolution of the previous trip's actual departure time ($D_{i-1}$), actual TT ($T_{i-1}$) and the minimum in-between time between trips $i-1$ and $i$ ($l_{i-1,i}$):

$$D_i = max\{D_{i-1} + T_{i-1} + l_{i-1,i}, d_i\}, \quad i = 2, \dots, m_s \tag{2}$$

$$D_1 = d_1. \tag{3}$$

Precisely, $l_{i-1,i}$ accounts for the deadhead travel between terminals, if the trip $i-1$ ends at a different terminal than the departure terminal of trip $i$, and the minimum break time for drivers. The duration of deadhead travels is stochastic, but for simplicity a fixed value for each pair of terminals is used in the following. This value is given by the operator.

In order to compute $\mathbb{E}(R_i)$, the expected secondary delay of trip $i$, we claim that the expected TT of trips $1, \dots, i-1$ provide insufficient information. To illustrate this, let's have a look at a simple case. Consider trip 2 scheduled to start at $d_2 = 8{:}40$AM and preceded by trip 1 that has started at $D_1 = 8{:}00$AM. Let also $l_{1,2} = 5$ minutes. If the probability that the actual TT of trip 1 is equal to 34 minutes is $P(T_1 = 34 \text{ minutes}) = 0.75$ and the probability that it is equal to 38 minutes is $P(T_1 = 38 \text{ minutes}) = 0.25$, then the expected TT of trip 1 is $\mathbb{E}(T_1) = 35$ minutes. Thus, if we only consider $\mathbb{E}(T_1)$, we get that the expected secondary delay of trip 2 is $\mathbb{E}(R_2) = 0$ minute. However, considering the probability distribution of $T_1$, we get that $R_2 = 0$ minute with a probability of 0.75 and $R_2 = 3$ minutes with a probability of 0.25 and therefore $\mathbb{E}(R_2) = 0.75$ minutes. This example confirms that the correct way to compute the expected secondary delay takes into account the probability distributions of the TT. Moreover, and at a more fundamental level, since the planned duration of a trip $i$ is usually set to a value close to $\mathbb{E}(T_i)$, by computing the expected secondary delays using the expected TTs, any potential delay propagation is ignored. To take into consideration the fact that some trips are more uncertain than others, we must compute the expected secondary delays based on the complete probability distributions of the TT. In reality, these distributions are much more complex than the one presented in the above example, justifying the need to explore models for the prediction of the probability distributions of the TT.

There are two types of TT prediction: short-term and long-term. Both types can predict either a segment or a complete trip TT. The former is usually performed less than one hour before a trip and uses online information as well as external factors (e.g., weather). This type of prediction can be integrated to the operator's operations control system and provides online information to the users about the estimated arrival time of a bus. On the other hand, the long-term TT prediction can be performed a few days before the trip and helps for transit planning. In this work, we are interested in computing long-term TT predictions.

We frame the long-term prediction of the density of the TT (PDTT) as a supervised learning problem which aims at predicting, for each trip $i$ in a set of unseen trips (test set), an estimate of the complete conditional probability density function (PDF) of its TT, $\hat{p}(T_i|x_i)$, given $x_i$ the set of characteristics of trip $i$. We assume that the conditional PDF of the TT does not depend on scheduling decisions, i.e., the TT uncertainty is exogenous to the resource allocation. The end-goal of this problem is to accurately estimate the expected secondary delays of trips in a test set. To this end, we perform simulations using the predicted probability distributions of the TT to approximate the true expected secondary delays. The model that generates the most accurate approximations of the expected secondary delays is selected. This information can then be used by the operator's schedulers to evaluate and compare vehicle schedules in terms of their reliability or by a computer to optimize over a large number of possible vehicle schedules. Probabilistic models are compared to a Random Forests (RF) model, which provided the most promising results among the three regression models studied in the work of Moreira, Jorge, Sousa and Soares (2012). We introduce two types of probabilistic models, namely the similarity-based density estimation models and the smoothed logistic regression model for probabilistic classification, and present experimental results on a large-scale dataset of more than 41,000 trips and 50 bus routes. Our contribution is threefold:

- The state-of-the-art for long-term prediction of public bus TT is almost nonexistent (Moreira-Matias, Mendes-Moreira, de Sousa and Gama, 2015). This work tries to fill this gap and, in addition, it is to our knowledge the first work to propose probabilistic models for the long-term prediction of public bus TT.
- We propose a novel method to approximate the expected secondary delays based on the probability distributions of the TT.
- To the best of our knowledge, it is the first study in the field of public transport that empirically studies such a large number of bus route's TTs simultaneously. We hope this can make our results relevant to other bus networks.

The remainder of this paper is organized as follows. In Section 2, we review the literature on TT analysis. The dataset used for the PDTT is presented in Section 3. We overview the main bus route characteristics and portray a preliminary analysis of the features. Section 4 describes the methodology that can be applied for the PDTT. A Monte Carlo simulation to compute the expected secondary delays based on the results of the models for the PDTT is introduced in Section 5. In Section 6, data preparation as well as features and parameters selection are presented, before the evaluation metrics are specified and the performance of all models for the PDTT is compared. Thereafter, in Section 7, a preview of an optimization model that uses the approximations of the expected secondary delays in an attempt to improve the reliability of bus schedules is presented. Section 8 summarizes our findings.

## 2. Related works

The introduction of automatic vehicle location (AVL) data has given rise to a flourishing number of studies in the field of public transport on speed, arrival time and TT analysis. Because TT and arrival time measures are closely related, studies on both measures are treated without distinction. Indeed, the arrival time $A_i$ of a trip $i$ is given by

$$A_i = D_i + T_i. \tag{4}$$

In this section, three topics are covered: long-term TT prediction, TT variability analysis and TT distribution modeling. The latter fits the TT distribution of trips that occurred during a given period in order to analyze the shape and nature of the PDF of the TT, without trying to predict future events, as in the PDTT. We extend the field of the first topic to all road-based transport, but the subsequent topics are restricted to the public transport field. Approaches proposed for the PDTT are inspired by lessons learned through the review of the literature on these topics.

Compared with the literature on short-term TT prediction, studies on long-term TT prediction are rare and to the best of our knowledge, only the works of Chen, Liang and Chu (2020), Moreira et al. (2012) and Klunder, Baas and Op de Beek (2007) proposed or reviewed long-term TT prediction methods. In a survey on improving the planning of public transit using AVL data, Moreira-Matias et al. (2015) suggested that the long-term TT prediction should be valid for an horizon of at least the entire forecasting period. They divided models found in the literature for short-term TT prediction in four categories: (i) machine learning and regression, (ii) state-based and time-series, (iii) traffic theory-based and (iv) historical databased, and suggested that some regression algorithms applied for short-term TT prediction

could be adapted for long-term TT prediction. Gradient boosting is an example of a model that was successfully applied to the short-term TT prediction and then adapted by Chen et al. (2020) to the long-term TT prediction of trips on a freeway segment in Taiwan. The features were ranked in order of relative importance: time of the day, day of the week, national holiday, day of long / consecutive holiday, big event / activity, electronic tool collection fee promotion and narrowing of roadway. Three regression models, namely a Projection Pursuit Regression, a Support Vector Machine and a Random Forest, were compared by Moreira et al. (2012) for the long-term TT prediction of trips of one public bus route in Porto. With a basic pre-processing work, the Random Forest had better results, but was slightly outperformed by a Projection Pursuit Regression when the authors added an instance selection step. Klunder et al. (2007) trained a $k$ Nearest Neighbours algorithm ($k$NN) with only time-based variables for the long-term TT prediction on a motorway network in the Netherlands.

The reasons for TT variability can be external or internal (Yetiskul and Senbil, 2012) and related to demand or capacity (Mazloumi, Currie and Rose, 2010) (see Figure 1). In an early study, Abkowitz and Engelstein (1983) suggested that shorter routes may have reduced TT variablity. Also, they reported that a running time deviation at the beginning of a route tends to propagate downstream. Hence, control actions to correct early deviations on a route could reduce TT variability. Strathman and Hopper (1993) reported that the afternoon peak period has higher TT variability, in particular because of the higher passenger demand. In a study on TT variability in the city of Ankara, Yetiskul and Senbil (2012) found major differences in regional TT variability and suggested that bus-stop spacing should depend on the neighborhood density. Comi et al. (2017) performed a time series decomposition of the TT and compared it to the temporal traffic patterns. The two are reported to have similarities. Also, the seasonality of the time series decomposition was most significant for the hour of the day.

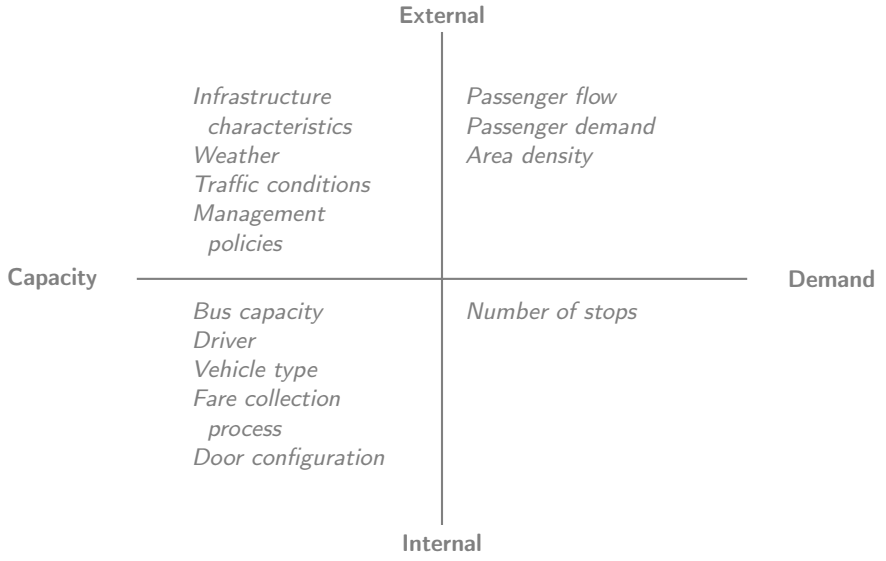



**Figure 1:** Reasons of TT variability

TT distribution modeling has been studied mostly with the objective of quantifying the reliability of a transit service. Most works on TT distribution modeling in public transit occurred after the introduction of AVL systems. We focus our review on the work of Mazloumi et al. (2010), Ma, Ferreira, Mesbah and Zhu (2016) and Büchel and Corman (2018), as they are, in our view, the most comprehensive studies, from which useful lessons can be learned for the PDTT. Mazloumi et al. (2010) assessed the shape and nature of the TT distribution over the course of the day and for different levels of temporal aggregation on segments of a bus route. The authors measured the level of temporal aggregation by the length of the departure time windows (DTW), which are time slots for which trips departing during the slot are aggregated for subsequent analyzes (e.g., 15 minutes, 30 minutes or 1 hour). The study concluded that for shorter DTWs, the TT distribution follows a Normal distribution. For longer DTWs, this result holds for peak periods, but not for off-peak periods. For the latter, the Log-Normal distribution fits better. Also, the contribution of a set of features to the TT was assessed through a linear regression analysis. The land use (industrial vs. residential)

and the length of the segment were those affecting the most the TT variance. Ma et al. (2016) studied intensively the influence of temporal and spatial aggregation on the TT distribution, with the objective of providing common grounds for modeling and evaluating the performance. To this end, several settings of temporal and spatial aggregations were assessed and an evaluation approach based on a statistical hypothesis test was proposed. The TT of a trip starting at stop $i$ and ending at stop $j$ was decomposed in dwelling times $DT_k$ at each stop $k$ of the trip and running times $RT_{(k,k+1)}$ between each pair of stops

$$T_{(i,j)} = \sum_{k=i}^{j-1} DT_k + RT_{(k,k+1)}. \tag{5}$$

Results concerning the normality of the TT distribution were in line with the ones of Mazloumi et al. (2010). Also, the analysis suggested that spatial aggregation tends to decrease the multimodality of TT distribution. A multimodal distribution is defined as a probability distribution with several modes. A Gaussian mixture model (GMM) was proposed to address the multimodality of the link level TT distribution. Büchel and Corman (2018) found that the Log-Normal distribution was, out of four unimodal statistical distributions, the best fit for the TT distribution modeling.

## 3. Data

Before introducing the data, it is essential to distinguish terms that are used in the following and that should not be confused, namely bus routes, bus lines, timetabled trips and trips. First, we define a bus route as an ordered sequence of road segments and bus stops, where the first and the last stops are called terminals. Second, a bus line usually has two associated routes, each one going in opposite directions (e.g., North-South or East-West axis). Third, timetabled trips are generated during service planning, which is performed for typical days in the planning horizon. For example, service planning for the next two months can be reduced to planning for a typical weekday, Saturday and Sunday. A timetabled trip is associated with a given route, time and typical day and is valid for the planning horizon. It is therefore not associated with a given date. Fourth, trips are a unique event associated with a timetabled trip and a given date. Buses record trip data as they travel, so each data point in the dataset is associated with a trip.

The dataset used for this study was collected during a 2-month period from 08/28/2017 to 10/29/2017 by in-car Advanced Public Transport Systems (APTS) installed in buses running in the city of Montréal, Canada. Those systems collect automatically at every stop of a trip the corresponding trip identifier, route identifier, direction identifier, stop identifier, date, scheduled departure time, scheduled arrival time, actual departure time, actual arrival time and number of passengers loading or unloading, among other things. The scheduled departure, scheduled arrival, actual departure and actual arrival times are stored in milliseconds. The actual TT of a trip is the difference between its actual arrival time and its actual departure time at the terminals, whereas its primary and secondary delays are the differences between its actual TT and its scheduled TT and between its actual departure time and its scheduled departure time, respectively. Hence, for every trip, only the first and last stops (i.e., terminals) data is kept. Since the APTS were embedded in approximately 20% to 30% of the vehicles at that time, weekends and holidays had an insufficient number of trips recorded. Indeed, during weekends and holidays, the service is reduced and thus the number of trips recorded during those days is too small to conduct relevant data-driven analysis. For that reason, weekends and holidays are not studied and are removed from the dataset. After removing weekends and holidays, the dataset has more than 116,000 trips. Of the 408 routes in the dataset, only the 50 most frequent are kept for the remainder of the study, resulting in a dataset of over 41,000 trips. The 50 selected routes run between 4:00AM to 1:59AM (+1 day) during weekdays. To facilitate the notation, we add two extra hours to the usual 24-hour daily period. Thus, we say that the selected routes run between 4:00AM to 25:59PM.

Figure 2 shows the average secondary delay per scheduled departure hour and the average secondary delay plus the standard deviation ($\sigma$) of all 41,000 trips. Note that the secondary delay cannot be negative (see equations (1) - (3)) and thus the average secondary delay lies between 0 to 1.5 minutes all day long. The variability of the secondary delays is high in the late afternoon (approximately from 15:00PM to 18:59PM) likely because delays accumulate during the day and because this is a period of high mobility.

### 3.1. Route's characteristics

The distribution of the average TT per scheduled departure hour is presented in Figure 3, where each piecewise linear curve represents the evolution of the average TT of a route. It is possible to distinguish the morning peak

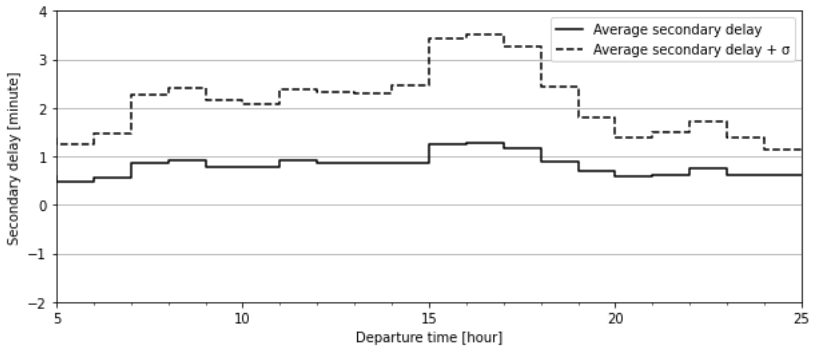



**Figure 2:** Average secondary delay and variability per hour

approximately from 6:00AM to 8:59AM, and the afternoon peak, approximately from 14:00PM to 17:59PM for all routes. The afternoon peak usually has an average TT higher than the morning peak and the average TT is generally decreasing after 17:00PM.

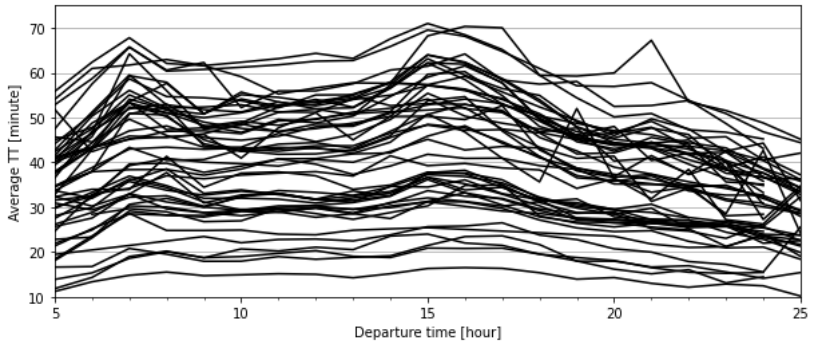



**Figure 3:** Average TT per scheduled departure hour

The main characteristics of the 50 selected routes, namely the number of stops, the distance traveled and the type of operational region, are presented in Table 1. Each route has a unique combination of line identifier and direction, such that *A-East* and *A-West* are two different routes of the same bus line, but in opposite directions. We categorized the type of operational region in 6 categories: residential areas, crossing city center (CC), from city center (to a residential area), to city center (from a residential area), from an industrial (indust.) area (to a residential area), to an industrial area (from a residential area). Bus line B is the only one crossing the city center; it starts in a residential neighborhood, crosses the city center and ends in another residential neighborhood. Industrial areas are characterized by a high density of factories. The city center and industrial areas are usually regions where a large number of people commute to work every day. The majority of bus routes operate in residential areas (32 out of 50). Those bus routes may, for example, connect two residential areas or a residential area to a subway or train station. The number of stops per bus route ranges from 17 to 74 stops, while the distance traveled ranges from 3.0 km to 15.3 km. In general, as the number of stops goes up, the distance traveled goes up as well. Line P is a counterexample, because it has a large number of stops close to each other. Note that the number of stops and the distance traveled of two routes of the same line are generally not equal, as the path in one direction is usually not symmetric to the path in the other direction (e.g., because some streets are one-way).

## 3.2. Features analysis

The PDTT has to be based upon features (i.e., explanatory variables) that are available a few days or weeks in advance. For example, meteorological conditions are likely to influence the TT duration. However, since it is an information that is not available when solving service planning problems, it is not considered. Likewise, the TT of

**Table 1**
Characteristics of bus routes studied

| Line | Dir. | #stops | Dist. (km) | Type of Region | Line | Dir. | #stops | Dist. (km) | Type of Region |
|---|---|---|---|---|---|---|---|---|---|
| A | East | 52 | 13.9 | Residential | M | West | 18 | 4.0 | Residential |
| A | West | 49 | 14.5 | Residential | N | East | 28 | 5.3 | Residential |
| B | East | 46 | 13.2 | Cross CC | N | West | 29 | 5.3 | Residential |
| B | West | 46 | 12.0 | Cross CC | O | North | 35 | 3.0 | To indust. |
| C | North | 40 | 12.1 | Residential | O | South | 40 | 3.4 | From indust. |
| C | South | 45 | 12.3 | Residential | P | East | 74 | 9.0 | Residential |
| D | North | 33 | 10.3 | From indust. | P | West | 67 | 8.5 | Residential |
| D | South | 36 | 10.4 | To indust. | Q | East | 40 | 7.0 | Residential |
| E | East | 50 | 14.2 | Residential | Q | West | 38 | 7.0 | Residential |
| E | West | 52 | 13.3 | Residential | R | East | 37 | 5.3 | Residential |
| F | East | 34 | 7.8 | Residential | R | West | 35 | 5.3 | Residential |
| F | West | 36 | 7.7 | Residential | S | East | 47 | 11.8 | From indust. |
| G | North | 17 | 4.6 | Residential | S | West | 51 | 11.6 | To indust. |
| G | South | 19 | 4.3 | Residential | T | North | 34 | 8.5 | To indust. |
| H | North | 37 | 9.3 | Residential | T | South | 30 | 8.5 | From indust. |
| H | South | 40 | 10.8 | Residential | U | North | 46 | 11.1 | Residential |
| I | East | 71 | 15.3 | Residential | U | South | 42 | 10.7 | Residential |
| I | West | 68 | 15.3 | Residential | V | East | 46 | 9.5 | Residential |
| J | North | 28 | 7.1 | From CC. | V | West | 49 | 8.5 | Residential |
| J | South | 30 | 7.1 | To CC | W | East | 43 | 10.3 | Residential |
| K | East | 53 | 11.1 | To CC | W | West | 47 | 11.6 | Residential |
| K | West | 51 | 11.4 | From CC | X | North | 30 | 6.6 | From CC |
| L | East | 35 | 5.9 | Residential | X | South | 34 | 7.5 | To CC |
| L | West | 36 | 6.0 | Residential | Y | North | 30 | 8.0 | To CC |
| M | East | 18 | 4.4 | Residential | Y | South | 28 | 8.0 | From CC |

**Table 2**
Long-term features

| Feature | Type | Possible values |
|---|---|---|
| Day of the week | Categorical | {Monday, Tuesday, ..., Friday} |
| Region | Categorical | {residential, crossing CC, ..., to indust.} |
| Route identifier | Categorical | {A East, A West, ..., Y South} |
| Distance (km) | Non-categorical | [3, 15.3] |
| Number of stops | Non-categorical | {17, 18, . . . , 74} |
| Scheduled departure time | Non-categorical | [4:00AM, 25:59PM] |
| Week number | Non-categorical | {35, 36, ..., 44} |
| Year | Non-categorical | {2017} |

the previous trip is not considered. The list of possible features includes the day of the week, type of region, route identifier, distance, number of stops, scheduled departure time, week number and year. Possible values and types of features (categorical or non-categorical) are listed in Table 2.

The feature year is discarded because our dataset is spread over 2017 only. The statistical significance of the features scheduled departure time, day of the week and week number can be analyzed visually by looking at Figures 3, 4 and 5, which present the average TT per route depending on each of the feature respectively. Figure 3 suggests that the scheduled departure time has a high importance. The relationship between the TT and the scheduled departure time

is not linear. Generally, the average TT of a bus route increases during peak hours and is steady between the morning and the afternoon peaks. Second, Figure 4 suggests that the relationship between the TT and the day of the week is less important. This is hardly surprising given that Saturdays and Sundays are not considered. Interestingly, there is no common pattern between the routes; for example some routes have a slightly higher average TT on Tuesdays than on Mondays and Wednesdays, while some others have an inverse pattern (i.e., the average TT on Tuesdays is slightly lower than on Mondays and Wednesdays). Third, Figure 5 suggests that the relationship between the TT and the week number is significant only for a handful of bus routes. In Figures 3, 4 and 5, we can observe that the average TT differs greatly from one bus route to another, unsurprisingly as each route has its own characteristics (as discussed earlier).

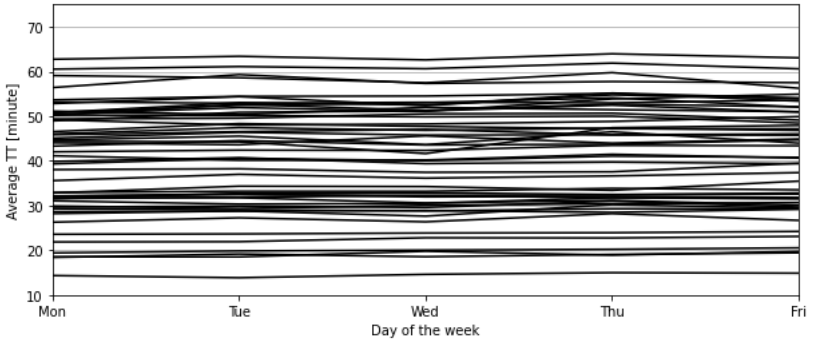



**Figure 4:** Average TT per day of the week

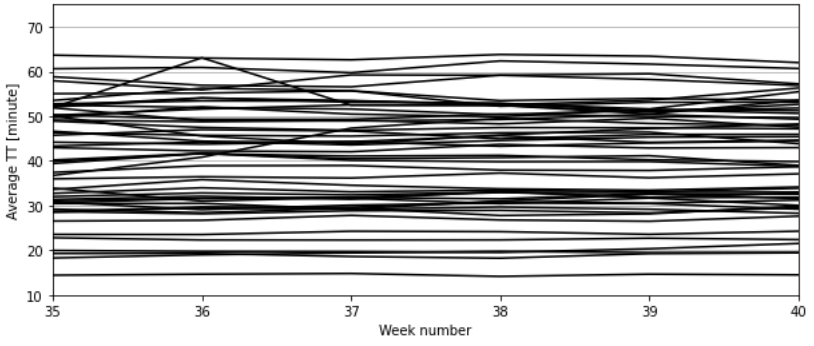



**Figure 5:** Average TT per week number

## 4. Methodology

Several approaches have been proposed to compute the conditional PDF of a random variable in a probabilistic fashion. Gaussian process based models are among the most popular. The work of Dutordoir, Salimbeni, Deisenroth and Hensman (2018) proposed a Gaussian process-based model to estimate a conditional PDF using latent variables in order to model non-Gaussian probability distributions. Also, Bishop (1994) developed Mixture Density Networks, which is a type of artificial neural network predicting multimodal conditional density distributions. The main drawback of Mixture Density Networks is that they perform poorly when the size of the dataset is not large enough. In the work of Yeo, Melnyk, Nguyen and Lee (2018), the prediction of a continuous PDF is converted into a classification task by using a discretization technique. This simplifies the learning task and traditional probabilistic classifiers can be used to predict the probability mass function, which can be smoothed later on into a PDF. The focus of this paper is on frequentist models (Koller and Friedman, 2009). In the remainder of this section two approaches for the PDTT are presented. The first one estimates the PDFs of the TT of a set of similar trips using parametric, semi-parametric or non-parametric density estimation models. The second approach, namely the smoothed Logistic Regression for

probabilistic classification, is similar to that of Yeo et al. (2018), but fits a Logistic Regression instead of a Recurrent Neural Network estimator.

## 4.1. Similarity-based density estimation

Similarity-based density estimation models are a two-step process: for each trip, (1) find the set of similar trips and (2) estimate the density of this particular set, by fitting a parametric, semi-parametric or non-parametric model. Next, we will define two similarity-based methods and introduce some density estimation models.

### *4.1.1. Similarity-based methods*

Consider a trip $i$ with a feature vector $(\boldsymbol{x}_1^{(i)}, \dots, \boldsymbol{x}_d^{(i)})$. We want to select, based upon one of the following similarity-based methods, the set of trips in the (reduced) training set that have similar attributes:

- Equivalent DTW (eDTW) : select all trips from the same route that have a scheduled departure time in the same departure time window (DTW) (Mazloumi et al., 2010; Büchel and Corman, 2018; Ma et al., 2016) as trip $i$.
- $k$ Nearest Neighbors ($k$NN): select the $k$ nearest neighbors of the trip $i$. The distance between trip $i$ and trip $j$ is the Euclidean distance between their feature vectors and is computed as

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{\ell=1}^{d} (x_\ell^{(i)} - x_\ell^{(j)})^2}. \tag{6}$$

### *4.1.2. Density estimation models*

The conditional probability $p(T_i = t_i \mid \boldsymbol{x}_i)$ is estimated by fitting a given density estimation model on points close to $i$ in the (reduced) training set, which are either trips in the same eDTW or close neighbors.

*Parametric models*

Parametric density estimation considers a restricted set of common probability distributions. Each of these distributions has a small number of parameters that have to be estimated from the data. We consider the Normal, Log-Normal, Logistic, Log-Logistic, Gamma, and Cauchy probability distributions. The Gamma distribution is a family of probability distributions containing the Exponential, Erlang and Chi-Squared distributions. In the work of Ma et al. (2016), the first four distributions were successful at modeling the TT distribution at a route level. For each trip, parameters of these probability distributions are found using the Maximum Likelihood Estimation (MLE) algorithm.

*Semi-parametric model: Gaussian Mixture Model*

GMMs are a sub-category of mixture models composed of $K$ normal components. GMMs are relevant when the population modeled is multimodal and has undefined subpopulations or states, such that each component represents a state. It is common in transport to use three components, one for each of the traffic states: free flow, recurrent and non-recurrent traffic (Ma et al., 2016). The PDF of a $K$-components GMM is given by

$$\hat{p}(T_i = t_i \mid \boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(t_i \mid \mu_{ik}, \sigma_{ik}^2). \tag{7}$$

The vector of positively defined coefficients $\pi_i = (\pi_{i1}, ..., \pi_{iK})$, such that $\sum_{k=1}^{K} \pi_{ik} = 1$, and the vector of the model's $k^{th}$ component's parameters $(\mu_{ik}, \sigma_{ik}^2)$, are found by applying the expectation maximization (EM) algorithm.

*Non-parametric model: Kernel Density Estimation (KDE)*

A KDE model infers the PDF of a random variable based on a sample of its population. To estimate the PDF of a trip $i$, the model uses $m$ points close to $i$ in the training set. It is a data smoothing problem that allows to find, in a non-parametric fashion, the curve of the PDF given a sample. The Gaussian kernel, $K(\cdot)$, is the most widely used, but

any function that integrates to unity ($\int K(t)dt = 1$) can replace it. The smoothness of the estimator is adjusted by the bandwidth parameter $h$ as

$$\hat{p}(T_i = t_i \mid \boldsymbol{x}_i) = \frac{1}{m}\sum_{j=1}^{m} K\left(\frac{t_i - t_j}{h}\right). \tag{8}$$

### 4.2. Smoothed Logistic Regression for probabilistic classification (LR-PC)

Probabilistic classifiers are a type of machine learning model that can predict the probability that a given input belongs to a set of classes, instead of only predicting the class with the highest probability. When a numerical discretization is applied to the random variable $T$, such that the TT is categorized in bins of 1 minute, the PDTT task can be translated into a probabilistic classification one: estimate the probability that $T$ takes a value that falls into class $c \in \{0, ..., C-1\}$. A model is fitted per bus route, because this setting yields better experimental results than learning a unique model for all bus routes (see Section 6.2).

A question arises: how to choose the number of classes for a bus route? Our approach was to use, for a given bus route, the difference between the trip in the training set with the shortest duration, $t_{min}$, and the trip with the longest duration, $t_{max}$, as $C$, the number of classes. Thus, $P(T_i = c)$ is the probability that $T_i$ takes a value in $[c + t_{min}, c + 1 + t_{min}[$. We disregarded the fact that trips in the test set can have shorter or longer duration than $t_{min}$ and $t_{max}$ respectively, as the smoothing discussed later implicitly solves this issue.

Multinominal Logistic Regression is naturally probabilistic and is commonly used for probabilistic classification tasks. Classes' probabilities of a multinomial Logistic Regression are defined as

$$\hat{P}(\lfloor T_i \rfloor = c + t_{min} \mid \boldsymbol{x}_i, \mathbf{w}_c) = \frac{exp(\mathbf{w}_c^T \boldsymbol{x}_i)}{\sum_{c'=0}^{C-1} exp(\mathbf{w}_{c'}^T \boldsymbol{x}_i)}, \tag{9}$$

where $\mathbf{w}_c$ is the vector of parameters of the class $c$, found using a stochastic average gradient descent solver.

Logistic Regression outputs a probability mass function that can be smoothed into a PDF subsequently. As proposed by Yeo et al. (2018), the output of the multinomial Logistic Regression, which takes the form of a p.m.f., is passed thought a one-dimensional convolution layer. This step has the effect of enforcing a spatial correlation in the output. The convolution layer is analogous to a KDE with a bandwidth $h$ and a kernel $K(\cdot)$, but uses the probability mass function, $\hat{P}(\lfloor T_i \rfloor \mid \boldsymbol{x}_i, \boldsymbol{w}_c)$, instead of a sample of trips:

$$\hat{p}(T_i = t_i \mid \boldsymbol{x}_i) = \sum_{c'=0}^{C-1}\left[K\left(\frac{t_i - (c' + t_{min})}{h}\right) \times \hat{P}(\lfloor T_i \rfloor = c' + t_{min} \mid \boldsymbol{x}_i, \mathbf{w}_{c'})\right]. \tag{10}$$

## 5. Simulation framework to measure the delay tolerance of a vehicle schedule

Consider again a vehicle schedule $s = \{1, 2, ..., m_s\}$ with $m_s$ trips. After the schedule has been performed, the secondary delay $R_i$ of a trip $i$ can be computed using equations (1)-(3), as the actual value of the TT ($T_{i-1}$) and the actual departure time ($D_{i-1}$) of the previous trip are then available. However, before the schedule is performed, $\mathbb{E}(R_1), \mathbb{E}(R_2) \ldots, \mathbb{E}(R_{m_s})$ must be computed using the PDFs of the TT and it is impossible to do this exactly. Instead, we propose a Monte Carlo simulation of $K$ iterations, where at each iteration a TT is randomly sampled, when it is possible, from $\hat{p}(T_i|x_i)$ for each $i = 1, \ldots, m_s$ and delays are propagated from the first to the last trip in order to compute the secondary delay of each trip. Before going further, we must state the situation in which it is not possible to sample the TT. This situation occurs because trips in $s$ may not be included in $\mathcal{B}$, the dataset of 41,000 trips and 50 bus routes presented in Section 3. A detailed explanation of the reasons for this is provided in Section 6.3 and for now just remember that this may be the case. If trip $i \notin \mathcal{B}$, then we have no information about the TT distribution of this trip and we have to use the scheduled duration of the trip directly. The scheduled duration of trip $i$ is the difference between its scheduled arrival time $a_i$ and its scheduled departure time $d_i$. After running the $K$ iterations, it is possible to approximate the expected secondary delays of all the trips in the schedule. This approximation is given by:

$$\mathbb{E}(R_i) \approx \bar{R}_i = \frac{\sum_{k=1}^{K} R_i^k}{K}, \quad \text{for } i = 1, \ldots, m_s, \tag{11}$$

with $R_i^k$ the secondary delay of trip $i$ computed at iteration $k$. The exact expected secondary delay of trip $i$ is approximated by $\bar{R}_i$, its average secondary delay over $K$ iterations. At each iteration, the randomly sampled TTs of trips $1, \dots, m_s$ must have a duration that lies between $MinTT_i$ and $MaxTT_i$, the smallest and the largest observed TTs of the trips on the same bus route as trip $i$. Otherwise, a new TT is sampled until that condition is fulfilled. In other words, we truncate the PDF of the TT below and above the times never recorded and therefore for which we have no information.

The pseudo-code in Algorithm 1 summarizes the Monte Carlo simulation used to compute the approximation of the expected secondary delays for all trips of a schedule $s$. In essence, at each iteration of the simulation the TT of each trip in the schedule $s$ is either sampled or set to the planned duration and the delays are propagated from the first trip to the last one. The simulation outputs the average secondary delay over $K$ iterations for each trip $i \in s$. Note that we assume that the first trip of a schedule always starts on time, i.e., its secondary delay is null (see equation (3)).

**Algorithm 1:** Monte Carlo simulation to approximate the expected secondary delays

$\text{SumR}_i \leftarrow 0, \forall i \in s$
**for** $k \leftarrow 1$ **to** $K$ **do**
  **for** trip $i \leftarrow 1$ **to** $m_s$ **do**
    **if** trip $i = 1$ **then**
      $D_i \leftarrow d_i$
    **else**
      $D_i \leftarrow \max\{D_{i-1} + T_{i-1} + l_{i-1,i}, d_i\}$
      $\text{SumR}_i \leftarrow \text{SumR}_i + (D_i - d_i)$
    **end**
    **if** trip $i \in \mathcal{B}$ **then**
      **repeat**
        $T_i \leftarrow$ sample from $\hat{p}(T_i|x_i)$
      **until** $\text{MinTT}_i \leq T_i \leq \text{MaxTT}_i$
    **else**
      $T_i \leftarrow a_i - d_i$
    **end**
  **end**
**end**
**for** each trip $i \in s$ **do**
  $\bar{R}_i \leftarrow \text{SumR}_i / K$
**end**

## 6. Experimental results

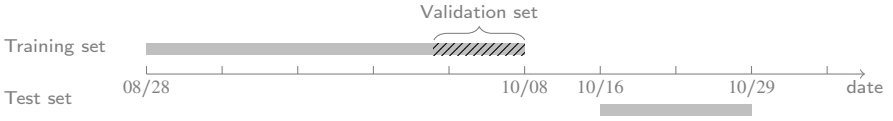
Using the dataset described in Section 3, we next explain how we fit probabilistic models for the PDTT, before comparing the performance of these models. First, we describe how the data is filtered and split for the PDTT. Second, we go through features and parameters selection for each type of model and detail the selection of the temporal aggregation level. Third, metrics to evaluate the performance of the models on the test set are presented. Finally, all probabilistic models are compared to a Random Forests model in terms of their performance on the test set.

### 6.1. Data preparation

The dataset is filtered in order to remove erroneous information and special situations that we do not want to cover in the PDTT. Incomplete trips, vias and trips with detours are discarded. A via is a trip that deviates from the main trip or an express trip. To remove erroneous trips which were not filtered by the previous step, the Median Absolute Deviation or MAD (Hellerstein, 2008) with a 6-delta criterion is used. A trip is discarded if it has a TT longer or shorter than the median TT value of the trips associated with the same route plus (minus) 6 times the corresponding standard deviation. This method also removes trips with extended TT due to exceptional scenarios (e.g., bus failure) that the PDTT problem should not cover, because when such exceptional scenarios occur, an additional bus is usually dispatched to recover the schedule and prevent severe delay propagation.

We use a hold-out method and split the dataset in two sets: a training and a test set. The validation set, a hold-out subset of the training set, is used for features and parameters selection and the test set is used for model evaluation. Figure 6 summarizes the dataset split. We split the dataset such that the training data starts on 08/28/2017 and ends on 10/08/2017. The set of test data is composed of the trips from 10/16/2017 to 10/29/2017. Hence, a complete week, from 10/09/2017 to 10/16/2017 is discarded to simulate real-life settings where the planning is done at least a few days ahead. The training set is split again in a validation set and a reduced training set by slicing the last trips recorded per route from the original training set. The reduced training set contains 80% of the trips in the original training set and the validation set contains the remaining 20%.

**Figure 6:** Dataset split




## 6.2. Models training

In the training process, the features and parameters of the estimator $\hat{p}(\cdot)$ are selected by fitting each model to the reduced training set and evaluating them on unseen data in the validation set. The performance of all the probabilistic models is evaluated by the negative log-likelihood (NLL) score over the validation set (containing $n_{val}$ points), computed as

$$NLL_{val} = -\sum_{i=1}^{n_{val}} \log(\hat{p}(t_i|\boldsymbol{x}_i)), \tag{12}$$

with $t_i$ the true TT of trip $i$. The performance of the Random Forests model, for its part, is evaluated by the mean squared error (MSE) of the output. The pair of features and parameters which obtains the best performance on the validation set is selected.

Similarity-based density estimation models use one of the similarity-based methods, namely the eDTW or the $k$NN method. While the eDTW method does not require feature selection, as the features used are always the route identifier and the scheduled departure time, the $k$NN method does require feature selection. Indeed, the distance between neighbors depends on the specified feature vector. Estimating the TT density of a trip $i$ does not require feature selection; it fits a probability density to a sample containing trips similar to trip $i$. Thus, here the features selection problem is reduced to finding a feature vector for the $k$NN method.The selection of features is carried out in parallel with the selection of parameters. The parameters of the similarity-based methods are the DTW duration for the eDTW method and the number $k$ of neighbors for the $k$NN method. For the KDE, the validation set is used to select the bandwidth $h$ and the kernel function.

We found that similarity-based density estimation models and the LR-PC model fit the data better when they are fitted per bus route. By doing so, the features describing the bus route characteristics, namely the number of stops, distance traveled, route identifier and type of region, become uninformative to the model. Indeed, all trips used to train the model of a given bus route have exactly the same values for these features. The remaining features to consider are the scheduled departure time, the week number and the day of the week.

To select the features of the similarity-based density estimation using $k$NN models as well as the LR-PC and the Random Forests models, we applied the permutation feature importance technique (Breiman, 2001). The latter reports the statistical significance of a set of possible features by measuring the increase of a predictor score when the values of a feature are permuted. The importance of a feature $\ell$ is the difference between a model's score over the original dataset and the average (over 10 shuffles) score over a corrupted dataset (with the values of the feature $\ell$ permuted). Results of this analysis are presented in Table 3. In line with the preliminary feature analysis presented in Section 3.2, the results indicate that the scheduled departure time has a higher statistical significance than the week number and the day of the week. For all models, features with a relative statistical significance of more than 1.00% are selected. Thus, for all similarity-based density estimation models except the Cauchy with $k$NN and the LR-PC model, the scheduled

**Table 3**
Relative statistical significance (%) of features

| | $k$NN* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Cauchy | Gamma | Normal | Log-Norm. | Logistic | Log-Log. | KDE | LR-PC | RF |
| Number of stops | - | - | - | - | - | - | - | - | 59.03 |
| Distance | - | - | - | - | - | - | - | - | 17.75 |
| Sched. departure time | 99.74 | 96.71 | 97.02 | 97.01 | 97.98 | 98.21 | 97.43 | 99.51 | 17.37 |
| Route identifier | - | - | - | - | - | - | - | - | 2.25 |
| Week number | 0.50 | 2.47 | 2.10 | 2.32 | 1.61 | 1.49 | 1.69 | -0.46 | 2.09 |
| Region | - | - | - | - | - | - | - | - | 1.20 |
| Day of the week | -0.24 | 0.81 | 0.88 | 0.67 | 0.41 | 0.30 | 0.87 | 0.95 | 0.30 |

**Table 4**
NLL (lower is better) of the similarity-based density estimation models using eDTW method with different levels of temporal aggregation on the validation set

| | 5 periods | | 60 minutes | | 30 minutes | |
|---|---|---|---|---|---|---|
| Model | Train | Validation | Train | Validation | Train | Validation |
| Cauchy | 3.24 | 3.30 | 2.72 | 2.89 | 2.60 | 2.96 |
| Gamma | 3.08 | **3.15** | 2.58 | 2.83 | 2.47 | 5.98 |
| Normal | 3.09 | 3.16 | 2.59 | 2.87 | 2.48 | 6.57 |
| Log-Normal | 3.09 | **3.15** | 2.57 | 2.81 | 2.46 | 5.72 |
| Logistic | 3.09 | 3.16 | 2.59 | 2.77 | 2.48 | 2.91 |
| Log-Logistic | 3.09 | **3.15** | 2.58 | **2.75** | 2.47 | **2.88** |
| GMM | 3.07 | 3.16 | 2.26 | 5.81 | 1.82 | 24.96 |
| KDE | 2.99 | **3.15** | 2.49 | 2.76 | 2.77 | 2.89 |

departure time and the week number are selected. Only the scheduled departure time is selected for the Cauchy with $k$NN and the LR-PC models. The number of stops, the distance, the scheduled departure time, the route identifier, the week number and the region are selected for the Random Forests model.

Temporal aggregation is a fundamental aspect of the PDTT since it has been shown to affect the shape and nature of the TT probability distribution (Mazloumi et al., 2010; Ma et al., 2016). Thus, the parameter associated with it, namely the DTW duration, is studied carefully. We select the DTW duration by analyzing how the density estimation models perform on the validation set for different levels of temporal aggregation. DTWs considered are, going from the most aggregated to the least aggregated, 5 periods per day (before morning peak, morning peak, in-between morning and afternoon peaks, afternoon peak and after afternoon peak), 60 minutes and 30 minutes. Table 4 compares the performance of models using the eDTW method, with the values in bold indicating the best NNL of the validation set for each level of temporal aggregation. For all models except the GMM, the NLL score over the validation set is better at DTWs of 60 minutes. The Log-Logistic model obtains the best results at all aggregation levels, matched by the Gamma, Log-Normal and KDE models at an aggregation level of 5 periods per day. The GMM has a similar performance to the parametric models for the most aggregated level, but it also has a poor performance for lower levels of temporal aggregation, both 60 and 30 minutes. Since the performance on the reduced training set is good, it indicates that the GMM overfits the training data. For the DTWs considered, the conditional PDF of the TT is most likely not multimodal. Thus, this model is discarded for the rest of the study. Interestingly, the training NLL decreases when the temporal aggregation level increases for all models except the KDE, while the NLL over the validation set increases from DTWs of 60 minutes to DTWs of 30 minutes for all models. This suggests that models are overfitting more at DTWs of 30 minutes than at DTWs of 60 minutes.

We can conclude that, between the three levels of temporal aggregation compared, the best one is the one with DTWs of 60 minutes. We denote the eDTW method with DTWs of 60 minutes as eDTW*. For the second similarity-

based method, namely the $k$NN, the value of $k$ can be chosen similarly to the duration of DTWs, by assessing the performance of the similarity-based density estimation models on the validation set for different values of $k$. Figure 7 shows that the NLL over the validation set decreases significantly for all models when the value of $k$ increases up to approximately $k = 13$. After that, the NLL stays almost constant. Thus, we set the number $k$ of neighbors to 13 and denote the $k$NN method with $k = 13$ as $k$NN*.

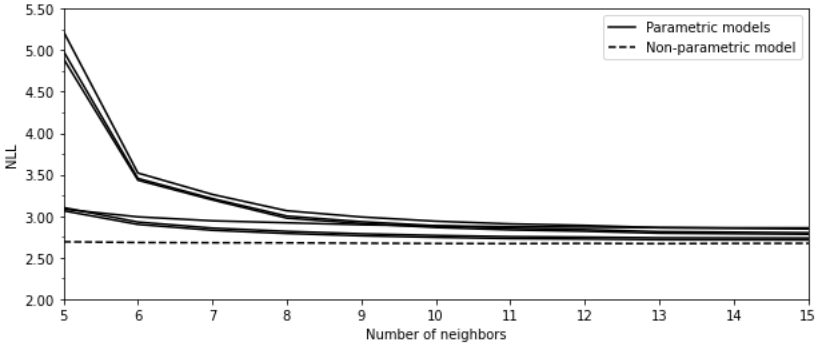



**Figure 7:** NLL (lower is better) of the similarity-based density estimation models using $k$NN method with different numbers of neighbors

The LR-PC model yields better performance when it considers transformations of the scheduled departure time to capture a non-linear relationship with the TT. First, the scheduled departure time is categorized in bins of 1 hour and 30 minutes using a one-hot encoding. Second, the sine and cosine of the scheduled departure time are computed. The total dimension of the feature vector is 69 (22 for the one-hot encoding of bins of 1 hour, 44 for the one-hot encoding of bins of 30 minutes, 2 for the sine and cosine of the scheduled departure time and 1 for the original scheduled departure time). The bandwidth and the kernel function of the LR-PC model are selected based on the performance on the validation set, along with the regularization strength of the Logistic Regression.

The number of trees in the Random Forest model, the maximum number of features considered when branching, the maximum depth of each tree and the minimum number of samples required to split an internal node are selected based on the performance on the validation set.

## 6.3. Models evaluation

After models training, the reduced training set is combined with the validation set and each model is trained on the complete training set using their selected features and parameters. The performance of the models for the PDTT is evaluated by the NLL score and the MSE of the expected secondary delay over the test set. The NLL score over the test set is analogue to the NLL over the validation set ($NLL_{val}$) and is computed as

$$NLL_{test} = -\sum_{i=1}^{n_{test}} \log(\hat{p}(t_i|\boldsymbol{x}_i)). \tag{13}$$

It quantifies the likelihood of the PDFs of the TT predicted by the models, with respect to the test points. The second metric, the MSE of the expected secondary delay over the test set, measures the accuracy of the approximation of $\mathbb{E}(R_i)$ for $i = 1, \dots, n_{test}$. It is given by

$$MSE_R = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(r_i - \bar{R}_i\right)^2, \tag{14}$$

with $r_i$ the true secondary delay of trip $i$ and $\bar{R}_i$ the average predicted secondary delay of trip $i$, for $i = 1, \dots, n_{test}$. The latter values are obtained by running the simulation presented in Section 5. This simulation takes as input a complete vehicle schedule $s = \{1, \dots, m_s\}$, the scheduled departure and arrival times of trips $i = 1, \dots, m_s$ and the probability distributions of the TT of trips $i \in s$ for which the information is available. Indeed, as mentioned in Section 3, the

**Table 5**
NLL (lower is better) and MSE (lower is better) of the expected secondary delay of similarity-based density estimation, LR-PC and Random Forests models on the test set.

| Model | Similarity method | $NLL_{test}$ | $MSE_R$ |
|---|---|---|---|
| Cauchy | eDTW* | 2.87 | 4.44 |
| | *k*NN* | 2.84 | 4.33 |
| Gamma | eDTW* | 2.79 | 4.45 |
| | *k*NN* | 2.80 | 4.33 |
| Normal | eDTW* | 2.82 | 4.47 |
| | *k*NN* | 2.84 | 4.37 |
| Log-Normal | eDTW* | 2.77 | 4.43 |
| | *k*NN* | 2.78 | 4.33 |
| Logistic | eDTW* | 2.74 | 4.42 |
| | *k*NN* | 2.73 | 4.34 |
| Log-Logistic | eDTW* | 2.72 | 4.39 |
| | *k*NN* | **2.70** | **4.31** |
| KDE | eDTW* | 2.75 | 4.41 |
| | *k*NN* | 2.73 | 4.43 |
| LR-PC | - | 2.71 | 4.72 |
| Random Forests (*point prediction*) | - | - | 4.75 |
| Random Forests (probabilistic interpretation) | - | 2.83 | 4.59 |

*DTWs = 60 minutes or $k = 13$

original dataset of 166,000 trips and 408 bus routes is reduced to a dataset of 41,000 trips and the 50 most frequent bus routes. Thus, many, if not nearly all, of the trips in the test set are part of a vehicle schedule that contains some trips that are discarded and for which we therefore have no information about the uncertainty of their TTs. In order to be able to propagate delays from the first trip to the last trip of $s$ using the recursive equations (1)-(3), the TTs of the trips for which no information is available are considered deterministic and equal to the planned duration.

## 6.4. Models comparison

Table 5 presents the NLL and the MSE of the expected secondary delay of all models over the test set. On the one hand, it is interesting to see that models using a non-Gaussian probability distribution, either a Gamma, Log-Normal, Logistic or Log-Logistic distribution, yield a lower NLL over the test set than those using a Normal distribution, which calls the normality of the conditional PDF of the TT into question. For the parametric models, the Cauchy, Logistic and Log-Logistic distributions have better NLL score over the test set when using the *k*NN* method, while the Gamma, Normal and Log-Normal distributions have rather the opposite results. The KDE models also have a better NLL over the test set when using the *k*NN* method than when using the eDTW* method. Overall, the Log-Logistic with *k*NN* model is the one which yields the best test NLL.

On the other hand, all parametric models predict more accurate expected secondary delays when using the *k*NN* method. On the contrary, the KDE models generate more accurate expected secondary delays when using the eDTW* method. The KDE using *k*NN* method and the LR-PC models surprisingly generate poorly accurate approximations of expected secondary delays, even though they achieved good NLL scores. It appears that smooth distributions, like parametric distributions, produce better approximations of the expected secondary delays. From the perspective that the end-goal of the PDTT is to approximate the expected secondary delays, the Log-Logistic with *k*NN* model should be selected because it yields the lowest MSE of the expected secondary delay, with a MSE of 4.31. Furthermore, the good test NLL of this model confirms this choice.

Similarity-based density estimation models and the LR-PC model are compared with two different interpretations of the Random Forests model. The first interpretation, which is our benchmark, considers the Random Forests model as non-probabilistic, i.e., used exclusively for *point prediction* (Dutordoir et al., 2018). The emphasis is on modeling the mapping between an input $\boldsymbol{x}$ to its output $y$ rather than on predicting the conditional PDF $p(y \mid \boldsymbol{x})$ (Dutordoir et al.,

2018). For this interpretation, the NLL over the test set is not computed because the prediction is not a PDF, but rather a point. Moreover, the expected secondary delays are computed slightly differently than for the other models because the TTs are considered deterministic in this interpretation. Thus, the expected secondary delays can be computed directly using equations (1) - (3) and without using the simulation presented in Section 5. As shown in Table 5 (see row "Random Forests (*point prediction*)"), this interpretation of the Random Forests is outperformed by all the models in terms of the MSE of the expected secondary delay. Thus, the experimental results show that there is an added value in modeling the conditional PDF of the TT using probabilistic models. A Random Forests model optimized with a sum-of-squares function can also be interpreted as fitting a Gaussian on the conditional distribution of the TT, where the mean of the Gaussian is given by the output of the Random Forests and the variance is constant for all the inputs and equal to the MSE of the output. The MSE of the output is not to be confused with the MSE of the expected secondary delay that we use to compare models performance. The results of this second interpretation can be found in row "Random Forests (probabilistic interpretation)" in Table 5. Note that the MSE of the expected secondary delay over the test set is better for this second interpretation than for the *point prediction* one, supporting our earlier assertion about the added value of probabilistic models. However, this interpretation is still outperformed by all models except the LR-PC model, showing the poor adequacy of the Random Forests model with a probabilistic interpretation to actual TTs and delays. As outlined by the NLL over the test set of parametric models before, the conditional PDF of the TT is most likely not Gaussian. Thus, it comes as no surprise that the Random Forests model, which assumes Gaussian noise, performs poorly both in terms of the NLL and the MSE of the expected secondary delay over the test set. Another key factor explaining the dominance of the Random Forests model by truly probabilistic models could be that the variance of probabilistic models is not constant for every trip, whereas it is the case for the Random Forests model as we interpreted it. In sum, because both interpretations of the Random Forests model, the non-probabilistic and the probabilistic ones, generate less accurate expected secondary delays than truly probabilistic models, the latter prevail for the prevision of the expected secondary delays of bus trips.

## 7. Preview of an integration in an optimization problem

The conditional PDF of the TT can be integrated in many service planning problems in an attempt to improve the delay tolerance of the service. For example, we are currently working on a variant of the vehicle scheduling problem with stochastic TTs that aims at computing vehicle schedules based on the expected secondary delay of their timetabled trips. The complete methodology and results of this work will be presented in a subsequent work. Nevertheless, we provide below a preview of the formulation of this optimization problem for interested readers.

The vehicle scheduling problem has been widely studied over the last half-century (Bunte and Kliewer, 2010) and consists of assigning vehicles to cover a set of timetabled trips, in such a way that every timetabled trip is covered exactly once and at minimal costs. When the operator's fleet is spread in two or more depots, it is referred to as the Multiple Depot Vehicle Scheduling Problem (MDVSP). We introduce an extension of the MDVSP, namely the reliable MDVSP with stochastic TTs, that exploits the long-term prediction of the PDFs of the TT studied in this work. This model takes the set $\mathcal{V}$ of $n$ timetabled trips and the long-term prediction of the PDF of the TT of each of these timetabled trips in input in order to output cost-efficient and delay tolerant vehicle schedules. Let $D$ be the set of depots, $S$ the set of all feasible vehicle schedules, and $S^d$ the subset of schedules starting and ending at depot $d$. The problem is to find a subset of vehicle schedules in $S$ that covers exactly once each timetabled trip while respecting the number of available buses $b_d$ at each depot $d \in D$ and minimizing a weighted sum of the total planned vehicle operating cost and the total expected secondary delay. To formulate this problem, we define for each timetabled trip $i \in V$ and schedule $s \in S$ a binary parameter $a_{is}$ which is equal to 1 if schedule $s$ covers timetabled trip $i$ and 0 otherwise, and denote by $c_s$ the cost of schedule $s$ (including delay penalties). Furthermore, we introduce for each schedule $s \in S$, a binary variable $y_s$ that takes value 1 if schedule $s$ is selected in the solution and 0 otherwise. The reliable MDVSP with stochastic TTs can then be expressed as the following integer linear program:

$$\min \quad \sum_{s \in S} c_s y_s \tag{15}$$

$$\text{s.t.} \quad \sum_{s \in S} a_{is} y_s = 1, \quad \forall i \in \mathcal{V} \tag{16}$$

$$\sum_{s \in S^d} y_s \leq b_d, \quad \forall d \in \mathcal{D} \tag{17}$$

$$y_s \in \{0,1\}, \quad \forall s \in \mathcal{S}. \tag{18}$$

Constraints (16) ensure that each timetabled trip is covered by a selected vehicle schedule, whereas constraints (17) impose vehicle availability at each depot.

The objective function (15) minimizes the total cost of the selected schedules which combines planned operational costs and delay penalties. The planned costs usually include a fix cost per vehicle used and a variable cost that depends on the traveled distance and the waiting time attended by a bus driver. Consider a vehicle schedule $s = \{1, 2, ..., m_s\} \in \mathcal{S}$ of $m_s$ timetabled trips. The total cost $c_s$ of the vehicle schedule $s$ is a weighted sum of the planned costs $q_s$ and the sum of the expected secondary delay $\mathbb{E}(R_i)$ of all timetabled trips $i$ covered by the schedule $s$, weighted by a factor $\beta$:

$$c_s = q_s + (\beta \sum_{i=1}^{m_s} \mathbb{E}(R_i)). \tag{19}$$

Note that the PDTT is trained and tested using trips data whereas the reliable MDVSP with stochastic TTs deals with timetabled trips. Fortunately, the selected model, namely the Log-Logistic with $k$NN*, can easily compute the PDF of the TT of timetabled trips, used to approximate the expected secondary delays. To that end, the model training is done as presented in Section 3.2 using data on past trips. The selected model for the PDTT is based on three features, the route identifier, the scheduled departure time and the week number. Thus, the prediction of the PDF of the TT of a timetabled trip $i \in \mathcal{V}$ depends on its feature vector $x_i = (x_1^{(i)}, x_2^{(i)}, x_3^{(i)})$, with $x_1^{(i)}$, $x_2^{(i)}$ and $x_3^{(i)}$ the route identifier, the scheduled departure time and the week number of timetabled trip $i$, respectively. Since timetabled trips are not associated with a given date, it is not clear how to define $x_3^{(i)}$, the week number. However, it is straightforward to see that, for the model selected, the PDF of the TT of a timetabled trip $i$ is the same regardless of the value of $x_3^{(i)}$, as long as it is a week number in the future planning horizon. Thus, the PDF of the TT of a timetabled trip can be computed by setting its week number to any week number in the future planning horizon. Then, an approximation of the expected secondary delay $\mathbb{E}(R_i)$ of all timetabled trips $i$ covered by a given schedule $s$ can be computed based on the PDFs of the TT by running the Monte Carlo simulation detailed in Section 5.

Since the number of feasible schedules is typically huge, it is impossible to enumerate them all. Instead, schedules are generated using column generation. The costs $c_s$ are computed during the solution process by taking into account the PDFs of the TT.

When solving the reliable MDVSP with stochastic TTs, the scheduled departure and arrival times are fixed for every timetabled trip. Thus, the choice of timetabled trip connections is the only lever to tackle reliability. The sequence of timetabled trips in each vehicle schedule must take into account the uncertainty of the TT of these trips, i.e., the connection between an uncertain timetabled trip and the next should allow enough idle time to avoid delay propagation. The $\beta$ factor can be modulated according to the operator's level of aversion to delay propagation. Of course, the higher the $\beta$ factor, the greater the tolerance to delay, but the higher the planned costs.

## 8. Conclusions

In public transport, reliability has become a key challenge for operators wishing to attract new users. In this work, we proposed a method to measure, in order to eventually improve, the reliability of bus schedules. To that end, we presented a simulation model to approximate the delay tolerance of a vehicle schedule based on the long-term conditional PDF of the TT. We framed the prediction of this conditional probability distribution, that we referred to as the PDTT, as a supervised learning problem. We verified if probabilistic models could predict more accurately the complete conditional PDF of the TT and generate more accurate approximations of the expected secondary delays than a Random Forests model. In fact, the latter is not inherently probabilistic and is typically used for *point prediction*. Also, we compared the performance of several probabilistic models.

To train and test the PDTT models, we used a 2-month dataset collected by buses equipped with APTS in the city of Montréal, involving 50 bus routes and a total of over 41,000 trips. The bus routes studied have various attributes (e.g., number of stops, frequency, traveled distance, etc.) and constitute a diverse sample from which we hope to obtain results relevant to other bus networks. Based on previous works on TT variability analysis, we determined a set of features, the number of stops, distance, scheduled departure time, route identifier, week number, type of region and day of the week, which we ranked in order of statistical significance for each model.

We proposed two types of probabilistic models for the PDTT, namely similarity-based density estimation models and the LR-PC model. The former is a two-step process that firstly find, for each trip, the set of similar trips and then estimate the density of this set using parametric, semi-parametric or non-parametric density estimation models. We proposed two types of similarity-based methods, namely the eDTW and the $k$NN, for which the temporal aggregation level and the number of neighbors had to be set, respectively. The LR-PC model applies a numerical discretization to the TT before fitting a Logistic Regression classifier per bus route. The output of the Logistic Regression is then smoothed into a PDF using a convolution layer analogous to a KDE.

Previous works on TT distribution modeling indicated that the level of temporal aggregation greatly affects the shape and nature of the TT distribution. Thus, we carefully selected the DTW duration based on the performance on the validation set. The GMM model had a poor performance for DTWs of 60 minutes and 30 minutes and thus we concluded that the conditional PDF of the TT is most likely not multimodal. This result is aligned with the one of (Ma et al., 2016) which observed that the multimodality of the TT decreases with spatial aggregation.

Models were compared in terms of both of their NLL and their MSE of the expected secondary delay over a test set. The first metric measures the likelihood of the probability distribution of the TT predicted while the second metric measures the accuracy of the approximations of the expected secondary delays outputted by the simulation model using the predicted probability distributions of the TT. From all the models tested, the density-based estimation model using $k$NN method and a Log-Logistic distribution yielded the best NLL and MSE of the expected secondary delay over the test set. Precisely, it produced approximations of the expected secondary delays that are about 9% more accurate than the benchmark model, the Random Forests. This result indicates that there is an added value in modeling the conditional PDF of the TT using probabilistic models. In particular, probabilistic models account for the variability of the TT whereas the Random Forests model does not intrinsically. Also, the Random Forests model as we interpreted it assumes, as many other *point prediction* models, that the noise of the TT is Gaussian. However, the normality of the TT was questioned because several similarity-based density estimation models using parametric distributions had better NLL over the test set than the models using the Normal distribution.

The Log-Logistic with $k$NN model generated accurate approximations of the expected secondary delays that schedulers can use to compare few bus schedule alternatives in terms of their reliability or to recommend changes to customers in the service planning parameters (e.g., minimum idle time between timetabled trips). Moreover, the expected secondary delays can be used to solve a reliable version of the MDVSP. We introduced this problem and proposed an integer programming model for it. In a forthcoming paper, we will propose a column generation algorithm for solving it.

*Declaration of interest:* None for all authors.

## Acknowledgements

The authors would like to thank Charles Fleurent and his team at GIRO Inc. for their valuable help and contributions and the Société de transport de Montréal (STM) for sharing the dataset. The advice of Didier Chételat and Maxime Gasse are greatly appreciated. We gratefully acknowledge the financial support provided by GIRO Inc. and the Natural Sciences and Engineering Research Council of Canada under the grant CRDPJ 520349-17 as well as the financial support provided by the Institute for Data Valorization (IVADO).

## References

Abkowitz, M.D., Engelstein, I., 1983. Factors affecting running time on transit routes. Transportation Research Part A: General 17A, 107 – 113. doi:https://doi.org/10.1016/0191-2607(83)90064-X.

Amberg, B., Amberg, B., Kliewer, N., 2019. Robust efficiency in urban public transportation: Minimizing delay propagation in cost-efficient bus and driver schedules. Transportation Science 53, 89–112. doi:https://doi.org/10.1287/trsc.2017.0757.

Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E: Logistics and Transportation Review 37, 191 – 229. doi:https://doi.org/10.1016/S1366-5545(00)00011-9.

Bishop, C.M., 1994. Mixture Density Networks. Technical Report NCRG/94/004. Department of Computer Science and Applied Mathematics, Aston University, UK.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Büchel, B., Corman, F., 2018. Modelling probability distributions of public transport travel time components, in: 18th Swiss Transport Research Conference. doi:https://doi.org/10.3929/ETHZ-B-000263929.

Bunte, S., Kliewer, N., 2010. An overview on vehicle scheduling models. Public Transport 1, 299–317. doi:10.1007/s12469-010-0018-5.

Chen, C.M., Liang, C.C., Chu, C.P., 2020. Long-term travel time prediction using gradient boosting. Journal of Intelligent Transportation Systems 24, 109–124. doi:https://doi.org/10.1080/15472450.2018.1542304.

Comi, A., Nuzzolo, A., Brinchi, S., Verghini, R., 2017. Bus travel time variability: some experimental evidences. Transportation Research Procedia 27, 101 – 108. doi:https://doi.org/10.1016/j.trpro.2017.12.072.

Desaulniers, G., Hickman, M.D., 2007. Chapter 2 public transit, in: Barnhart, C., Laporte, G. (Eds.), Transportation. Elsevier. volume 14 of *Handbooks in Operations Research and Management Science*, pp. 69–127. doi:https://doi.org/10.1016/S0927-0507(06)14002-5.

Dutordoir, V., Salimbeni, H., Deisenroth, M.P., Hensman, J., 2018. Gaussian process conditional density estimation, in: 32nd Conference on Neural Information Processing Systems.

Hellerstein, J.M., 2008. Quantitative Data Cleaning for Large Databases. Technical Report. UNECE.

Kieu, L.M., Bhaskar, A., Chung, E., 2015. Public transport travel-time variability definitions and monitoring. Journal of Transportation Engineering 141. doi:https://doi.org/10.1061/(ASCE)TE.1943-5436.0000724.

Klunder, G., Baas, P., Op de Beek, F., 2007. A long-term travel time prediction algorithm using historical data. Technical Report. TNO.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. The MIT press.

van Kooten Niekerk, M.E., 2018. Optimizing for Reliable and Sustainable Public Transport. Ph.D. thesis.

Kramkowski, S., Kliewer, N., Meier, C., 2009. Heuristic methods for increasing delay-tolerance of vehicle schedules in public bus transport, in: VIII Metaheuristics International Conference.

Kumar, B.A., Vanajakshi, L., Subramanian, S.C., 2014. Pattern-based bus travel time prediction under heterogeneous traffic conditions, in: Colloquium on Transportation Systems Engineering and Management. doi:https://doi.org/10.13140/RG.2.1.2338.5448.

Ma, Z., Ferreira, L., Mesbah, M., 2014. Measuring service reliability using automatic vehicle location data. Mathematical Problems in Engineering 2014, 1–12. doi:https://doi.org/10.1155/2014/468563.

Ma, Z., Ferreira, L., Mesbah, M., Zhu, S., 2016. Modeling distributions of travel time variability for bus operations. Journal of Advanced Transportation 50, 6–24. doi:https://doi.org/10.1002/atr.1314.

Mazloumi, E., Currie, G., Rose, G., 2010. Using gps data to gain insight into public transport travel time variability. Journal of Transportation Engineering-asce - J TRANSP ENG-ASCE 136, 623–631. doi:https://doi.org/10.1061/(ASCE)TE.1943-5436.0000126.

Moreira, J., Jorge, A., Sousa, J., Soares, C., 2012. Comparing state-of-the-art regression methods for long term travel time prediction. Intelligent Data Analysis 16, 427 – 449. doi:https://doi.org/10.3233/IDA-2012-0532.

Moreira-Matias, L., Mendes-Moreira, J., de Sousa, J.F., Gama, J., 2015. Improving mass transit operations by using avl-based systems: A survey. IEEE Transactions on Intelligent Transportation Systems 16, 1636–1653. doi:https://doi.org/10.1109/TITS.2014.2376772.

van Oort, N., 2011. Service Reliability and Urban Public Transport Design. Ph.D. thesis.

Strathman, J.G., Hopper, J.R., 1993. Empirical analysis of bus transit on-time performance. Transportation Research Part A: Policy and Practice 27, 93 – 100. doi:https://doi.org/10.1016/0965-8564(93)90065-S.

Yeo, K., Melnyk, I., Nguyen, N., Lee, E.K., 2018. De-rnn: Forecasting the probability density function of nonlinear time series, in: 2018 IEEE International Conference on Data Mining (ICDM), pp. 697–706. doi:https://doi.org/10.1109/ICDM.2018.00085.

Yetiskul, E., Senbil, M., 2012. Public bus transit travel-time variability in ankara (turkey). Transport Policy 23, 50 – 59. doi:https://doi.org/10.1016/j.tranpol.2012.05.008.