CIRRELT-2020-33

Forecasting of the Montreal Subway Smart Card Entry Logs with Event Data

Florian Toqué
Etienne Côme
Martin Trépanier
Latifa Oukhellou

September 2020

Bureau de Montréal

Université de Montréal
C.P. 6 128, succ. Centre-Ville
Montréal (Québec) H3C 3J7
Tél : 1-514-343-7575
Télécopie : 1-514-343-7121

Bureau de Québec

Université Laval,
2325, rue de la Terrasse
Pavillon Palasis-Prince, local 2415
Québec (Québec) G1V 0A6
Tél : 1-418-656-2073
Télécopie : 1-418-656-2624

Forecasting of the Montreal Subway Smart Card Entry Logs with Event Data

Florian Toqué[1,2,*], Etienne Côme[2], Martin Trépanier[1], Latifa Oukhellou[2]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Mathematics and Industrial Engineering, Polytechnique Montréal

[2] COSYS-GRETTIA, Université Gustave Eiffel, IFSTTAR, F-77447 Marne-la-Vallée, France

Abstract. One of the major goals of transport operators is to adapt the transport supply scheduling to the passenger demand for existing transport networks during each specific period. Another problem mentioned by operators is accurately estimating the demand for disposable ticket or pass to adapt ticket availability to passenger demand. In this context, we propose generic data shaping, allowing the use of well-known regression models (basic, statistical and machine learning models) for the long-term forecasting of passenger demand with fine-grained temporal resolution. Specifically, this paper investigates the forecasting until one year ahead of the number of passengers entering each station of a transport network with a quarter-hour aggregation by taking planned events into account (e.g., concerts, shows, and so forth). To compare the models and the quality of the prediction, we use a real smart card and event data set from the city of Montréal, Canada, that span a three-year period with two years for training and one year for testing.

Keywords: Forecasting, smart card data, Machine learning, events.

Acknowledgements. This research was partially supported by Thales and the Natural Science and Engineering Research Council of Canada (NSERC). The authors wish to thank the transport organisation authority of Montreal (Société de transport de Montréal) for providing the smart card and event data. They also acknowledge the computer infrastructure support from the IVADO and the Quebec Research Funds (FRQNT, FRQSC).

Results and views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect those of CIRRELT.

Les résultats et opinions contenus dans cette publication ne reflètent pas nécessairement la position du CIRRELT et n'engagent pas sa responsabilité.

* Corresponding author: florian.toque@gmail.com

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2020

1 Introduction

Public authorities currently play a significant role in encouraging sustainable development policies, giving impetus to sustainable urban mobility practices that aim to reduce the use of private cars and increase the use of sustainable transport modes, such as public transport. In urban and peri-urban areas, this transport strategy faces a number of challenges, including the regularity, quality of service and congestion of public transport. One of the major goals of stakeholders (operators and authorities) is to adapt as accurately as possible the schedules to the passenger demand during each specific period (e.g., normal period, period under events, disturbed period, special day, and so on). According to transport operators, another goal is to anticipate the demand for disposable ticket or pass (non-rechargeable smart cards) to match ticket availability to passenger demand during a specific period, particularly event periods (e.g., concerts, sports games, shows, exhibitions, and so forth). Furthermore, this information on the number of type of ticket or pass used per quarter hour can be used to provide mobility services adapted to the different types of passengers (regular/occasional). For example, a larger number of agents may be made available in the event of a high number of occasional passengers to help manage the extra passenger flow.

To address these issues, we propose a generic data shaping of contextual data, allowing the use of well-known regression models for long-term forecasting of passenger demand with fine-grained temporal resolution. In this study, we forecast the number of passengers entering each of the 68 metro stations in the city of Montréal, Canada, until one year ahead by taking calendar information and planned events into account. We also predict passenger demand per type of ticket or pass used to travel to address the problem of adapting ticket availability to passenger demand. The aggregation time window for the number of passengers has been chosen as 15 minutes, which permits the precise analysis of the impact of events on passenger demand and is relevant to adapting transport supply. We compare several well-known forecasting models, including basic, statistical and machine learning models. In this context, we analyse the use of contextual data such as information about the day and an event database provided by the public transportation authority of Montréal (Société de transport de Montréal, STM). This methodology aims to be reproducible to forecast the passenger demand for other transport networks around the world (depending on the availability of equivalent data sets in the other cities).

The main objective of this study is to determine whether it is possible to predict the number of passengers using the calendar and event information available in advance (in this case, available one year in advance), with the following innovative aspects:

- Predict the number of incoming passengers at each station of a transportation network (68 stations in the Montréal metro network in Canada)
- Propose a generic data shaping of contextual data
- Carry out the study over a long period of time (2 years for the learning set and 1 year for the test set)
- Predict the number of passengers aggregated with a fine temporal resolution (15 minutes)
- Perform a detailed analysis of the forecasting results during different periods (e.g. event periods and periods without events)
- Forecast the number of passengers and perform an analysis of this forecast based on the type of ticket or pass used to travel.
- Compare several forecasting methods, including basic methods, statistical methods and machine learning methods

First, early forecasting of the number of aggregated passengers per quarter hour at each station is useful to transport operators to help them improve the planning of the transport supply schedule (e.g., the number of subways per quarter hour, when planning to increase the supply of related transport systems such as buses) to match it as closely as possible to passenger demand. In addition, this demand forecast can be used to plan the presence of agents, secure stations in the case of excessive passenger traffic and allow passengers to avoid overcrowded situations.

Note that this approach can only be applied on fixed networks. To be effective, the approach requires a historical data set that includes the occurrence of events at a station; otherwise, the forecasting model will not be able to take into account the event information at a station that never hosted an event in the historical database.

The remainder of this paper is organised as follows. Section 2 details the related work. The case study is presented in Section 3. Section 4 details the forecasting methods and the data shaping that we have developed. Section 5.1 describes the forecasting results on the global aggregation of type of ticket or pass used to travel, while Section 5.2 provides an analysis of the forecasting performance per type of ticket or pass used to travel. Finally, some possibilities for future research and conclusions are outlined in Section 6.

2 Related Work

Since 2004, the use of smart card data to analyse mobility in public transportation has received substantial attention from researchers. More recently, studies on mobility analysis have revolved around passenger demand forecasting. A distinction can be made between research that relates to forecasting OD matrices and research that attempts to forecast passenger flows at a specific point. Knowledge about these two factors is indeed essential for planning, operation and management in any transportation network, but each of these areas uses different types of data.

The passenger demand goals differ depending on the forecasting time horizon. For long-term forecasting, the aim is to forecast demand with data available at the long-term period in advance (e.g., time features and planned events), which can be very useful for improving transport supply scheduling. In contrast, the forecasting process can also account for the last observations, in which case it is generally referred to as short-term forecasting.

Going forward, in the case of an atypical situation, the main goal for transport operators is to use the forecasted passenger demand to optimise transport system operation to match transport supply to the atypical demand or propose to the passenger an alternative way to reach their destination.

2.1 Short-term Forecasting of Passenger Demand in Public Transport

Short-term forecasting, which corresponds to a few time steps ahead forecasting, has been studied with different models. [1] used multiscale RBF networks to forecast the number of alighting passengers at different Beijing subway stations multiple time steps ahead (t+15 and t+30 minutes) by taking the number of boarding passengers at the other station of the subway network into account. In this study, the authors performed an in-depth analysis of the results obtained under special event scenarios. Other examples of subway passenger flow forecasting include the work of [2], where the authors predicted passenger flows of the next time step (t+2 minutes). The authors used a Bayesian network model and predicted multiple passenger flows (entry and exit) at all the stations of a subway line of the Paris network. In the study of [3], the authors created a fuzzy nonlinear autoregressive exogenous model to predict the number of passengers at the next time step (t+1 hour). In addition to forecasting, [4] conducted an in-depth analysis of the influence of subway predictor variables, such as bus transfer activities, and temporal features on the forecasting results and showed that the most important short-term forecasting features are the past observation of the metro ($\sim$ 82.0%), the past observation of the bus ($\sim$ 10.4%) and the prediction time step ($\sim$ 3.6%). This study predicted the next time step (t+15 minutes) of passenger flows at 3 stations of the Beijing subway network.

2.2 Short-term Mobility Forecasting with Spatiotemporal Focus

A closer examination of the most recent studies about short-term forecasting in the transportation field reveals high spatial and temporal values in such prediction problems. For example, in ride-sharing demand forecasting, a research team from Uber ([5]) studied Uber ride-sharing demand data with a focus on the temporal values for extreme event forecasting. [6] focused on capturing knowledge from the spatiotemporal information of the ride network via a deep learning approach. Similar approaches have been performed by [7] to predict citywide crowd flows and by [8] to predict taxi demand. Studies that spotlight the spatiotemporal aspect of traffic forecasting have also been conducted by [9, 10] with a combination of convolutional and recurrent neural network models and by [11] with a graph convolutional neural network model.

2.3 Event Data Usage in Short-term Forecasting

Some studies have shown the importance of external data, especially event data, for improving the prediction accuracy of forecasting models. Events such as concerts, shows, and sports games are sources of disturbance regarding human mobility. [12] developed short-term prediction approaches to forecast subway passenger flows for the next 4 hours using social media data. The authors focused on predicting the total number of passengers (sum of entry and exit) of one subway station of the New York City network. They proposed a two-step methodology: hashtag-based event detection followed by the combined use of linear regression and a seasonal autoregressive moving average model. More recent studies conducted by [13, 14] involved automatic event data collection, where the authors worked on the short-term forecasting of taxi demand in two distinct locations in New York city by using deep learning methods. In these studies, the model comparison showed that event categorisation could significantly help forecasting models obtain better results.

As shown in Table 1, numerous studies consider short-term forecasting with various methods and forecasting horizons.

Table 1: Related work on short-term forecasting

Reference	Method	Mode	Aggregation	Horizon	Event
[1]	RBF	Subway	15 min	1,2	No
[2]	Bayesian	Subway	2 min	1	No
[3]	AR	Subway	1 h	1	No
[4]	MLP	Subway	15 min	1	No
[5]	LSTM	Taxi	1 day	1	No
[6]	CRNN	Taxi	1 h	1	No
[7]	CRNN	Taxi&Bike	1 h & 30 min	1&1	No
[8]	CRNN	Taxi	30 min	1	No
[9]	CRNN	Traffic	5 min	1	No
[10]	CRNN	Traffic	15 min	1,2,3,4	No
[11]	GCNN	Traffic	15 min	1,2,3	No
[13]	GP	Taxi	1 h	1	Yes
[14]	LSTM	Taxi	1 day	1	Yes

RBF represents radial basis function network. AR represents autoregressive method. MLP represents multilayer perceptron. LSTM represents long short-term memory introduced by [15]. CRNN represents a different architecture of neural network with convolution and recurrent neural network. GP represents Gaussian process. GCNN represents graph convolutional network.

2.4 Long-term Passenger Demand Forecasting

To the best of our knowledge, only a few resources related to long-term forecasting with fine-grained resolution are available in the literature, unlike short-term forecasting. The study most related to our work is the study of [16]. The authors worked on long-term forecasting approaches using event data extracted from the web as features to forecast the aggregated number of passengers per half hour of tap in/out of 3 subway and 11 bus stops assigned to 5 venues in the city of Singapore. Their study was performed on a data set with a total period of 16 days. They demonstrated that using event information (online information) combined with public transport data can improve the quality of transport prediction under special events.

In this study, we investigate the problem of long-term (one year ahead) passenger demand forecasting, represented as the number of tap ins aggregated by 15-minute intervals of all the subway stations (68 stations) in the city of Montréal, Canada, and the use of an event database given by the transport organisation of Montréal. The real data set spans a long period (3 years). We propose a data shaping method that allows the use of well-known regression models for long-term forecasting of passenger demand. Moreover, we study the forecasting of the passenger demand per type of transit fare to provide an in-depth analysis of the passenger demand forecasting, thus helping transit operators adapt the availability of specific pricing during special events.

3 Case Study

The forecasting of transport demand at each station of a public transport network is a challenging task, mainly due to the influence of several well-known factors introduced by [7] on crowd flows and on transport demand. These factors can be summarised as follows: temporal factors, including time interval and the type of day, i.e., Monday, Tuesday, ..., Sunday; public or school holidays; and extra day off. Spatial factors include the type of area where the station is located (e.g., residential, office, shopping, and areas of interest). Predictable factors include weather, events, transport operator strikes and renovations. Unpredictable factors include transport network disruption that could be induced by a technical problem (rail problem, fire accident), a passenger problem or another factor that could severely impact the transport supply.

In this study, we aim to perform one-year-ahead forecasting by taking the temporal, spatial and contextual factors into account. To this end, temporal and contextual data that are available one year ahead will be used as inputs of the forecasting models. In the following sections, we detail the smart card entry logs, the time features and the event database.

3.1 Smart Card Entry Logs

The real dataset used to evaluate the proposed methodology was provided by the transport organisation authority of Montréal, Canada (Société de transport de Montréal, STM). The ticketing logs used in our study are obtained thanks to the validation of passes and tickets on automated fare collection systems for each user's entry into the transport network. We address all 68 subway stations in the city. The data set consists of ticketing logs aggregated by 15-minute intervals during 2015, 2016 and 2017. The studied subway network handles more than 670k passengers every day. We also forecast the number of passengers by the type of ticket or pass used to travel. We have aggregated the passengers according to their type of ticket or pass: STM monthly pass, regional monthly pass, book tickets and occasional passes. Disposable tickets include tickets used occasionally (occasional passes), 1- or 2-way tickets, 1- or 3-day passes, weekend passes, special event tickets, etc.

From Figure 1, we can see the percentage of passengers entering the subway network according to their type of ticket or pass during the global period from 2015-2017 and during the event period (pairs of days and stations with events). During the global period, the most used pass is the STM monthly pass, with approximately 140M entries per year, which represents 58% of the passenger demand. On the other hand, during event periods, the percentage of occasional passes increases significantly to 29.2%, versus 15.7% during the global period.

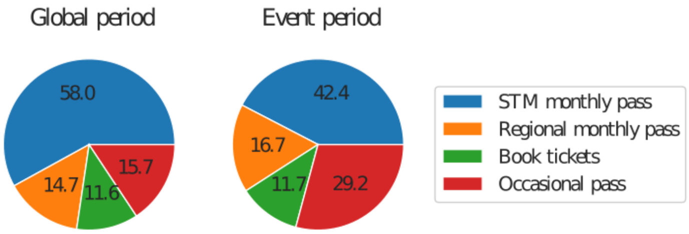

Figure 1: Use of the type of ticket or pass in percentage (2015-2017): global passenger demand on the left, passenger demand during event period on the right (pairs of days and stations with events)

3.2 Detailed Calendar Information

Passenger demand mainly depends on the day type; therefore, we created a list of nine day features, as follows:

- Name of the day of the week (e.g., Monday, Tuesday, and so on)
- Month (e.g., January, February, and so on)
- Holiday (e.g., Christmas day, New Year's day and so on.)
- 24th of December
- 31st of December
- Christmas holiday
- Summer university holiday part 1 (intensive session, Université de Montréal)
- Summer university holiday part 2 (regular session, Université de Montréal)
- Renovation period that took place at the Beaubien station over 4 months in 2015

This list of features is certainly specific for this transport network and this city, but it could easily be modified to suit another transport network and city.

3.3 Event Data

Passenger demand strongly depends on different contextual factors. Some factors cannot be planned far in advance (e.g., weather, transport network disruptions, and so forth), whereas others can be planned in advance, such as the presence of large events in a city (sports games, festivals, concerts, and so on). We could manually create or even automatically extract such event databases, as shown in previous studies conducted by [17, 13, 14]. In our case, a real data set of events was provided by the STM operator, who manually built a calendar of events in the city of Montréal occurring during the three years of the study. Each event is characterised by a location (the nearest subway station), the event start and end times (format is "Y-m-d H:M:S", approximately 80% of the event end times are available) and a manually built short text description of the event (description does not follow the same construction pattern, e.g., the same event could have different descriptions). We manually defined 10 topics as event categories to have an event categorisation able to be taken into account by the forecasting models. Taking these type of data into account is a challenge because it involves a large and sparse encoding of the data that is difficult to treat with regression models. Moreover, the end time of the event is not available for each event, which makes the interpretation of the event difficult.

Figure 2 shows the number of events per station and per category. We consider an event by the presence of a start time in the database (e.g., if the same event occurs on 4 consecutive days and is represented by 4 start times in the database, it will be counted as 4 events). As shown, most of the events occur near three stations: Lucien-L'Allier, Jean-Drapeau and Bonaventure.

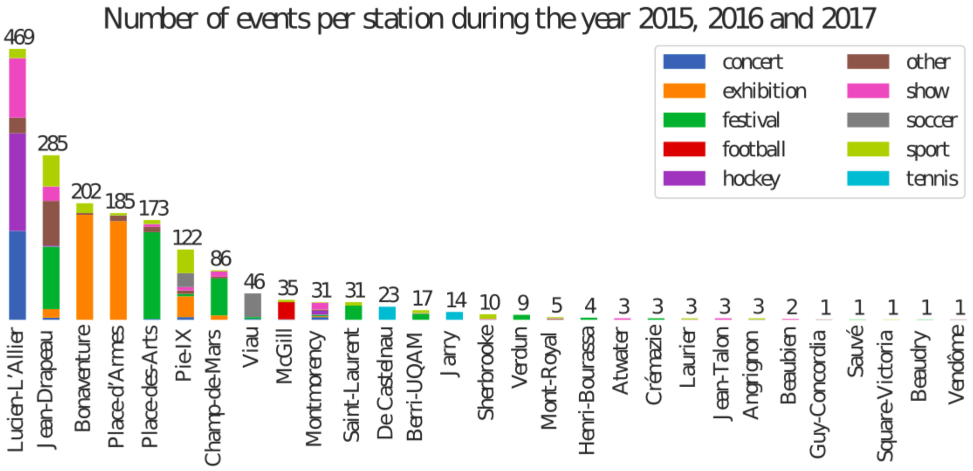

Figure 2: Number of events per metro station that host the event and per category.

To provide an overview of the smart card data set and the differences in passenger demand that could occur between the same type of day with or without the presence of an event, the numbers of passengers on three different Mondays of the same month (April 2017) at the station named "Lucien-L'Allier" are depicted in Figure 3. Monday, April 3, 2017, is depicted by the green line and could be considered a normal Monday. We can observe the typical morning and evening peaks of passenger demand. Monday, April 10, 2017, is coloured in orange, and this day is considered special because an event (Def Leppard concert that finished at 11:00 p.m.) occurred on this day near this station. Finally, Monday, April 17, 2017, which is a holiday (Easter Monday), is depicted by the blue line. We can observe a decrease in passenger demand throughout all Easter Monday (blue) compared to the normal Monday (green). However, we can see a highly concentrated increase in passenger demand due to the end of the concert during the Monday with the event (orange).

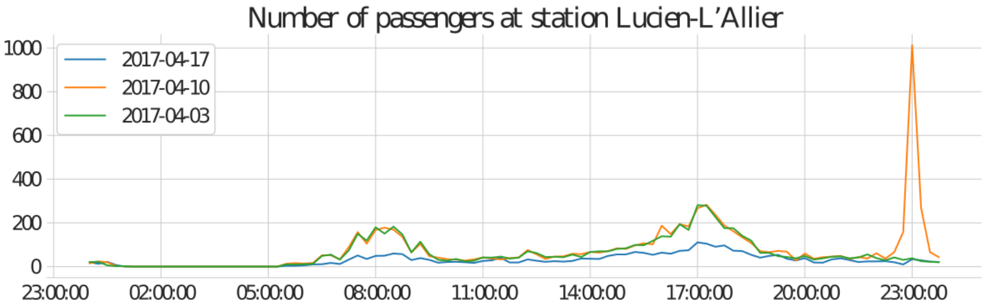

Figure 3: Number of passengers on three different Mondays. April 3, 2017, corresponds to a normal Monday; April 10, 2017, is a Monday with an event (Def Leppard concert that finished at 11:00 p.m.); and April 17, 2017, is a holiday (Easter Monday).

4 Forecasting Workflow

We aim to forecast the number of passengers entering each station of a transport network at each time step of a day until one year ahead. Here, we forecast the passenger demand of the 68 metro stations in the city of Montréal, Canada, at each quarter hour of a day (96 time steps per day) by taking planned events into account. We have compared the use of different sets of features as inputs of the forecasting models and different types of forecasting models. Section 4.1 details the data shaping and the compared set of features. The general description of the compared models is depicted in Section 4.2 and the evaluation metrics are described in Section 4.3.

4.1 Data Configuration

To evaluate the importance of each contextual data set, we trained the forecasting models with four input data sets (D1, D2, D3 and D4). Each of these input data sets corresponds to a specific concatenation of the following 4 sets of features:

- A: Month and name of the day of the week, encoded as one-hot vectors.
- B: Holiday, 24 of December and 31 of December, Christmas school holiday, university holidays part 1 and part 2, and Beaubien station renovation period. These features are encoded as one-hot vectors.
- C: Start event, end event and period event at each station that hosts an event. For each station that hosts an event (29 stations), at each time step of the day (vector size 96), we counted the number of time steps related to event schedule information (3 features), namely, the start time, end time and event period. For example, if we encode this information for one station that hosts an event on the day from 00:00 a.m to 00:45 a.m. We will obtain the following three vectors of size 96: (i) start time $[1, 0, ..., 0]$, (ii) end time $[0, 0, 1, 0, ..., 0]$ and (iii) event period $[1, 1, 1, 0, ..., 0]$.
- D: Category of the event (10 event categories). This has the same encoding as C, but the difference is that we counted the number of event per category. For each station that hosts an event (29 stations), at each time step of the day (vector size 96), for each category of event (10 categories), we counted the number of time steps related to event schedule information (3 features), namely, the start time, end time and event period.

The input data sets D1, D2, D3 and D4 are depicted in Table 2.

Table 2: Input data sets D1, D2, D3 and D4.

Data	A	B	C	D	Size
D1	✓				$11 + 6 = 17$
D2	✓	✓			$17 + 7 = 24$
D3	✓	✓	✓		$24 + 96 \times 3 \times 29 = 8376$
D4	✓	✓	✓	✓	$8376 + 96 \times 3 \times 29 \times 10 = 91896$

The set of features: A corresponds to the month and name of the day of the week, B corresponds to the detailed day features, C corresponds to the event features, and D corresponds to the category of the event features.

We have trained a specific forecasting model per station (total of 68 models) with daily multi-time-step output forecasting. This means that for each day, we perform a unique prediction that corresponds to a vector containing the forecasting of the number of passengers per quarter-hour intervals (output vector size is equal to 96, the number of 15-minute time steps in 24 hours). All the forecasting models (one model per station) have the same inputs and outputs, which are depicted in Figure 4. This figure depicts one input sample ($x_i \in X$) composed of features {A,B} and {C,D} corresponding to the features available until one year in advance of the forecasted day_i. Features A and B are detailed in Section 3.2 (e.g., day of the week, holiday, school holiday, and so forth). Meanwhile, features C and D are encoded per time step (96 quarters of hour per day) and correspond to the event features.

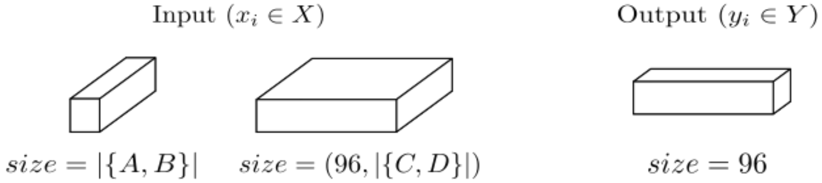

Figure 4: Data shaping of one input sample ($x_i \in X$) composed of features {A,B} (day features) and {C,D} (event features) and of one output sample ($y_i \in Y$) that corresponds to the forecasting of the number of passengers entering a station at each of the 96 time steps.

4.2 General description of the compared models

We aim to forecast the passenger demand until one year ahead with a fine-grained temporal resolution (quarter-hour aggregation). In this context, it is not possible to use and optimise the parameters of time series forecasting as autoregressive models (ARIMA, SARIMAX, and so forth) because of the too large training data set and the multi-time-step ahead prediction. Therefore, we compared different well-known models that could be used for regression problems. We computed a baseline model based on a historical average, a linear regression model, machine learning models such as random forest, gradient boosting decision trees, artificial neural network and kernel-based models including a support vector regressor and Gaussian process.

4.2.1 Historical Average

The historical average model is a baseline model that aims to predict passenger flows based on the average value of historical observations by type of day in the corresponding time step. The type of day could be defined by a set of features, such as those depicted in Section 3.2. For example, one may take the most basic feature, which is the name of the day of the week. In this case, the prediction for the time step of 8:00 a.m.-8:15 a.m., a Monday, corresponds to the average of all the historical values for Monday at 8:00 a.m.-8:15 a.m. The model computes the average number of passengers from the available historical dataset, representing two years in our case. This approach is the baseline of the long-term forecasting models.

4.2.2 Linear Regression

Linear regression is a statistical model that assumes a linear relationship between the dependent variable and one or more explanatory variables (or independent variables). We use as input more than one explanatory variable, and we predict more than one dependent variable (total of 96 dependent variables). In this context, we choose a multivariate linear regression. To prevent the collinearity phenomenon with categorical features, such as the name of the day of the week, we formatted the data by deleting one of the categories. To avoid overfitting, we computed the linear regression with the elastic net regularisation introduced by [18] that linearly combines the L1 and L2 penalties of the lasso and ridge methods. We optimised the hyperparameters alpha and l1_ratio, where alpha is a constant that multiplies the penalty terms L1 and L2, and l1_ratio corresponds to the penalty term associated with the L1 method and (1-l1_ratio) to the L2 method.

4.2.3 Random Forest (RF)

Random forest is a well-known machine learning model whose effectiveness for performing regression or classification problems has been widely proven for many real-world applications. The model introduced by [19] is an ensemble learning algorithm based on the average prediction of different decision trees (forest). Each tree is fit on different parts of the data, which were created by applying two sampling methods: random sampling with replacement, which is also known as the bootstrap aggregation or bagging method, and random selection of features, which is called feature bagging. The bagging methods and the averaging of the results obtained by the different trees make the RF more robust and accurate than a simple decision tree. We optimised the hyperparameters n_estimators, which corresponds to the number of trees used by the model; min_samples_split, which corresponds to the minimum number of splits required to split an internal node; min_samples_leaf, which is the minimum number of leaves required to be at a leaf node; and max_features, which is the number of features to consider when searching for the best split.

4.2.4 Gradient Boosting Decision Tree (GBDT)

Gradient boosting, introduced by [20], is a machine learning model for regression or classification tasks that uses an ensemble of weak prediction models, such as decision trees in our case, to create a prediction model. Similar to most of the other boosting methods, GBDT builds weak learners (decision trees) one at a time, where each new tree helps to correct the errors made by previously trained trees. After a tree is added, the data weights are readjusted. Correctly classified input data lose weight, and misclassified examples receive a higher weight. This technique helps future trees focus more on input data that were misclassified by previous trees. We optimised the same hyperparameters as the random forest model detailed in Section 4.2.3.

4.2.5 Artificial Neural Network (ANN)

An artificial neural network, also known as a neural network, is a computational model based on the structure and functions of biological neural networks. Each neuron receives inputs and biases, multiplies them by their weights, sums them and combines that sum with their internal state (activation function) to produce an output. In our case, we used the rectified linear unit (relu) function as the activation function of the hidden layer neurons, and the identity function for the neurons of the last layer (used as default by the scikit-learn library for regression problems). We optimised the number of layers and the number of neurons per layer, and we used the early stopping technique in order to stop the training of the model automatically.

4.2.6 Gaussian Process Regressor (GP)

[21] developed the Gaussian process, which is a generic supervised learning method; more specifically, it is a kernel method designed to solve regression problems. The prediction is probabilistic (Gaussian) and interpolates the observations. One of the advantages of this model is that it is able to compute confidence intervals in addition to the prediction. The main disadvantage of Gaussian processes is that they lose efficiency in high-dimensional space and

that they use the entire sample of feature information to perform the prediction, which could lead to overfitting. We optimised the hyperparameter alpha, which specifies the noise level in the targets.

4.2.7 Support Vector Regressor (SVR)

The support vector regressor is a supervised machine learning model and is based on the kernel method introduced by [22]. This model can efficiently perform a nonlinear regression using the kernel trick, implicitly transforming the data into a high-dimensional feature space to make it possible to perform the linear regression. The implementation is based on support vector machine, which is effective in high-dimensional spaces. We optimised the hyperparameters kernel; gamma, which corresponds to the kernel coefficient; and C, which is the penalty parameter of the error term.

4.2.8 Trend Factor

The main disadvantage of the forecasting method described in this study is that the models do not take into account the global trend of the number of passengers from year to year. The heatmap in Figure 5 shows the percentage increase between the years 2015 and 2016 and between 2016 and 2017 of the average number of passengers per time step and per station (we do not take into account the Beaubien and Rosemont stations, which were severely impacted by renovations in 2016). As shown, for 60% of the stations, the increase is of the same sign (positive or negative) between 2015-2016 and 2016-2017.

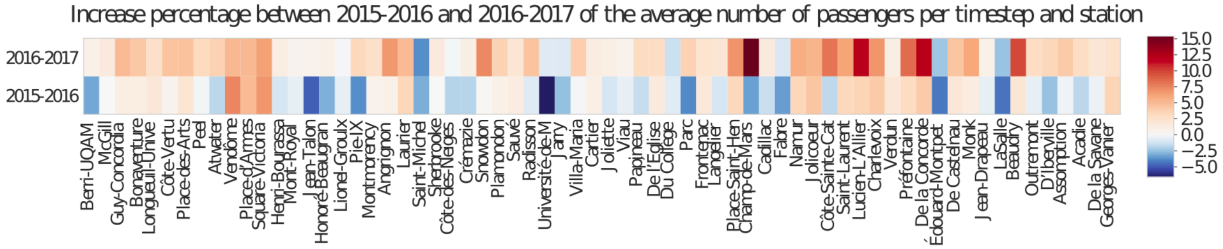

Figure 5: Trend factors between 2015-2016 and 2016-2017 per station. Note that 60% of the stations have the same sign of trend factor between 2015-2016 and 2016-2017.

To take this trend into account in the forecast, we multiplied the forecasted passenger demand at each time step by the trend factor depicted in Equation 1, obtained between 2015 and 2016 (training set). We set the trend factor of the Beaubien and Rosemont stations to 1 to not take the trend factor of these stations into account. To adjust the forecast of the number of passengers per type of ticket or pass used to travel, we calculated a specific trend factor between 2015 and 2016 for each type of ticket or pass used to travel.

$$trend\_factor_{2015-2016}(s) = \frac{\frac{1}{T_{2016}} * \sum_{t_1=0}^{T_{2016}} x_{2016}^{t_1}(s)}{\frac{1}{T_{2015}} * \sum_{t_2=0}^{T_{2015}} x_{2015}^{t_2}(s)} \tag{1}$$

where

x = Number of passengers
T_y = Number of time step of year y, with $t \in T$
s = Station s

This first attempt to introduce a trend factor in the forecasting model is basic. Further investigations are needed to improve the forecasting capability of the models.

4.3 Evaluation Methods

To evaluate the models, we split the entire dataset into two different parts: (i) a training dataset used to fit the models with data spanning the years 2015 and 2016 and (ii) a testing dataset used to compare the performance of the models with data of the year 2017. We evaluate the results obtained by the different forecasting models with several well-known metrics. To obtain a better understanding of the errors, three measures of prediction accuracy were used, namely, the root mean square error (RMSE), the median absolute error (MAE) and the mean average percentage error at v (MAPE@v).

These errors can be expressed as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{s=1}^{S}\sum_{t=1}^{T}(\hat{y}_s(t) - y_s(t))^2}{T \times S}} \tag{2}$$

$$\text{MAE} = \frac{1}{T \times S}\sum_{s=1}^{S}\sum_{t=1}^{T}|\hat{y}_s(t) - y_s(t)| \tag{3}$$

$$\forall \mathrm{y_s(t)} > \mathrm{v},\ \mathrm{MAPE@v} = \frac{100}{T \times S} \times \sum_{s=1}^{S}\sum_{t=1}^{T}\left|\frac{y_s(t) - \hat{y}_s(t)}{y_s(t)}\right| \tag{4}$$

where $\hat{y}_s(t)$ is the forecast value of station s at time step t, $y_s(t)$ is the actual value, and S is the station number.

4.4 Implementation and Optimisation of the Models

In this section, we detail the setups of the different forecasting models. We discuss the optimisation of the hyperparameters and the library and resources used to build the models.

4.4.1 Implementation

We used Scikit-Learn developed by [23], a famous Python library, to compute the following models: elastic net regression, Gaussian process regressor, random forest, gradient boosting decision tree, artificial neural network and support vector regressor. We used the MultiOuputRegressor class of Scikit-Learn to perform multi-output forecasting with SVR and GBDT models that are single-output regressors.

4.4.2 Optimisation

We performed a grid search with 5 random fold cross-validation to optimise all the statistical and machine learning models. We fixed the computation time for the optimisation of each model to a maximum of 2 days. We used the Scikit-learn default hyperparameters for the model GBDT with input data sets D3 and D4 because of the computation time being too long. The tested hyperparameters are presented in Table 3. The experiments were conducted in parallel on 20 cores.

5 Forecasting Results and Discussion

First, the results of the passenger demand forecasting with an overall aggregation are presented in Section 5.1. We compare the forecasting results and present some forecast and observation examples during two different periods: a normal period and an event period. Regarding the methodological context, we show which model performs the best in terms of forecast accuracy and the importance of each feature in the forecasting. In a more focused system transport context, we show the forecasting results per station. Then, we detail the results obtained for each type of ticket or pass used to travel in Section 5.2.

All the results correspond to the forecast obtained by the different models in addition to the trend factor method explained in Section 4.2.8. The trend factor method improves the results by approximately 0.80%.

5.1 Forecasting Result Aggregation of All Types of Ticket or Pass

5.1.1 Global Forecast Analysis

To obtain a global comprehension of the results obtained by the models using different sets of features as model inputs, we studied the aggregated errors mentioned in Section 4.3 of all the stations during the entire training and testing. Table 4 depicts the RMSE, MAE and MAPE@150 errors on the training and test sets obtained by the forecasting models described in Section 4. The models are used to forecast the number of ticketing logs aggregated per 15 minutes. Each model has been computed with different input data sets (D1, D2, D3 and D4) detailed in Section 4.1. We can observe that the best results are obtained with models using the combination of all the input data sets (D4), except for the ANN model, which is not able to capture the event information because of the excessively low number of training samples due to the special formatting of the data, and the Gaussian process, which overfits because of the excessively

Table 3: Grid search hyperparameters of the forecasting models.

Model	Hyperparameter	Tested values
LR	alpha	0.1, 1, 10
	l1_ratio	0.25, 0.5, 0.75, 1
	normalise	True, False
GP	alpha	0.1, 0.5, 1
	normalise_y	False, True
RF	n_estimators	100, 150, 200
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 5, 10
	max_features	’auto’
GBDT	n_estimators	100, 150, 200
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 5, 10
	max_features	’auto’
SVR	kernel	’rbf’, ’linear’
	gamma	1, 0.1, 0.01, 0.001
	C	0.001, 0.01, 0.1, 1.0, 10
ANN	solver	’adam’
	batch_size	16
	max_iteration	5000
	early_stopping	True
	hidden_layer_sizes	(10), (100), (300), (10, 10), (100, 10), (100, 100), (300, 100)

The value ’auto’ of the hyperparameter max_features of the RF and GBDT models corresponds to the total number of features. The kernel ’rbf’ of the SVR model corresponds to the radial basis function kernel. These two values corresponds to the default values in the scikit-learn library. Concerning the hyperparameter hidden_layer_sizes of the ANN model, the ith element represents the number of neurons in the ith hidden layer.

large and sparse information caused by the additional event and category of the event features. The best prediction model is obtained with the RF model, with 38.53 and 13.13% for the RMSE and MAPE@150 error, respectively. The historical average (HA) model is a basic method in terms of its implementation (it calculates the average number of passengers based on the day type). Unlike the SVR and LR models, where the number of parameters corresponds to the number of features, the number of parameters of the HA model corresponds to all possible combinations of features. For example, for the data set D1, the HA model contains 84 parameters (7 days * 12 months). This explains why this model succeeds in obtaining better results than the LR, and SVR models. On the other hand, it becomes more difficult to predict with this model when the number of features increases (data set D2) and even impossible to predict if the number of features is too large (data sets D3 and D4).

We have seen that the random forest model obtains the best forecast results using the D4 data set over the global period. We will see in Table 5 that despite the difference in the number of features between D3 and D4, it is preferable to predict the number of passengers with the D4 data set, since the difference in performance may increase depending on the forecast period.

As shown in Figure 6, the MAPE@v error highly depends on the threshold value. For example, with the best input data (D4), the RF model has a MAPE@5 (MAPE considering all of the observation passenger numbers greater than 5) of approximately 20% and a MAPE@150 of 13%. We choose the threshold value of MAPE as 150 to obtain a better estimation of the performance of passenger number forecasting when there is a considerable amount of demand that could impact ticketing demand and transport supply.

The observation and forecasting of the random forest model with all the sets of input data (D1, D2, D3 and D4) at Guy-Concordia station are depicted in Figure 7. The passenger demand for this station is largely related to the activity of students at Guy-Concordia University. Indeed, we can see that on Monday, September 18, 2017, the passenger demand appears to follow a regular pattern with activity peaks corresponding to the end of the courses at the university. In this case, the model with input data set D4 succeeds in accurately predicting passenger demand and is slightly better than the models with other input data.

Because events could impact the forecasting results, we analysed the forecasting results of the best forecasting model (RF model) considering periods with and without events (see Table 5). Seventeen stations with events in 2017 (test set period) were extracted. The filtering of the event period is performed by selecting the day/station pair with events. The period without an event represents the remaining data in the considered period. We can observe that the choice of input data significantly impacts the RMSE error during the event period. Indeed, RMSE is slightly improved during the

Table 4: Errors on the training and test sets of the different forecasting models with different input data sets (D1, D2, D3 and D4).

		Train (2015-2016)			Test (2017)		
Data	Model	RMSE	MAE	MAPE	RMSE	MAE	MAPE
D1	HA	45.41	18.07	12.69	50.36	21.47	15.28
	LR	49.71	20.51	13.87	52.04	22.46	15.51
	RF	46.97	18.76	13.01	50.39	21.32	15.17
	GP	46.09	18.62	12.85	**50.32**	21.56	15.17
	SVR	55.98	23.72	14.12	57.54	25.40	15.58
	GBDT	48.21	19.44	13.35	51.19	21.77	15.82
	ANN	49.93	20.57	14.51	53.13	22.76	16.18
D2	HA	32.24	13.04	9.73	44.31	19.16	13.84
	LR	41.15	17.99	12.44	44.96	20.13	13.99
	RF	35.12	14.66	10.65	**41.35**	18.19	13.20
	GP	33.68	14.04	10.30	41.42	18.54	13.42
	SVR	45.67	20.42	12.59	49.15	22.47	14.04
	GBDT	37.62	16.18	11.56	42.33	18.84	13.91
	ANN	40.2	16.6	12.08	43.83	18.75	13.63
D3	LR	34.56	16.59	11.57	43.74	20.37	14.21
	RF	26.79	12.67	9.29	**39.66**	17.99	13.16
	GP	17.13	7.00	4.91	79.71	36.39	22.66
	SVR	36.83	18.93	12.32	51.11	24.98	16.18
	GBDT	26.38	13.39	9.82	42.75	18.90	14.04
	ANN	40.01	18.52	13.62	55.2	25.43	18.61
D4	LR	33.79	16.62	11.65	42.62	20.27	14.18
	RF	26.60	12.63	9.29	**38.53**	17.88	13.13
	GP	16.90	6.96	4.85	80.71	36.98	23.06
	SVR	37.04	19.20	12.36	51.14	25.35	16.37
	GBDT	26.10	13.33	9.79	40.79	18.77	14.01
	ANN	41.44	19.80	14.52	63.57	29.62	21.55

The data are represented by different sets of features (D1, D2, D3 and D4) described in Section 4.1. The different models are described in Section 4. The evaluation metrics RMSE, MAE and MAPE@150 are defined in Section 4.3.

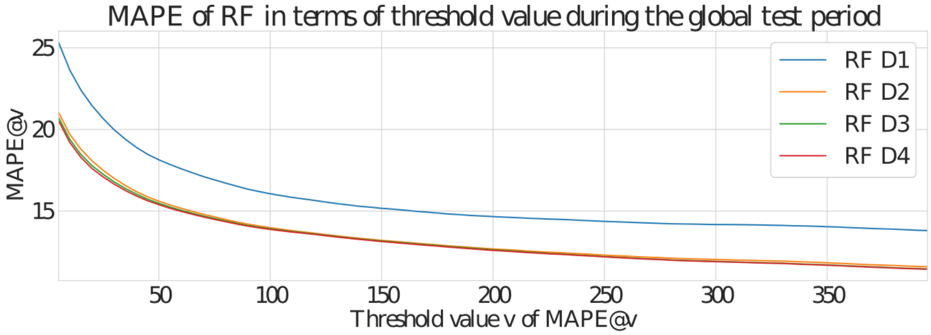

Figure 6: MAPE of random forest models with input data sets (D1, D2, D3, D4) in terms of the MAPE threshold value.

period without an event depending on the use of the input data set D2 or D4 (50.71 against 48.83 of RMSE), whereas this error is largely improved during the event period when D4 is used (153.34 against 124.72 of RMSE). The model with input D1 is too basic to be relevantly compared during periods without events with the model with input D4.

Figure 8 depicts the MAPE@v error according to the threshold value v during the event test period of the RF models. As shown, the best performance is not obtained by the same models with a threshold that is lower or greater than 120, which could be explained by the fact that the calculation of the MAPE@v error is clearly impacted by the passenger number observation. It disadvantages forecasting with a high value when the observation corresponds to a small value over forecasting with a low value when the observation relates to a high value. To improve transport supply and ticket availability in cases of high demand, the usage of a threshold that is greater than a certain value is more relevant than a

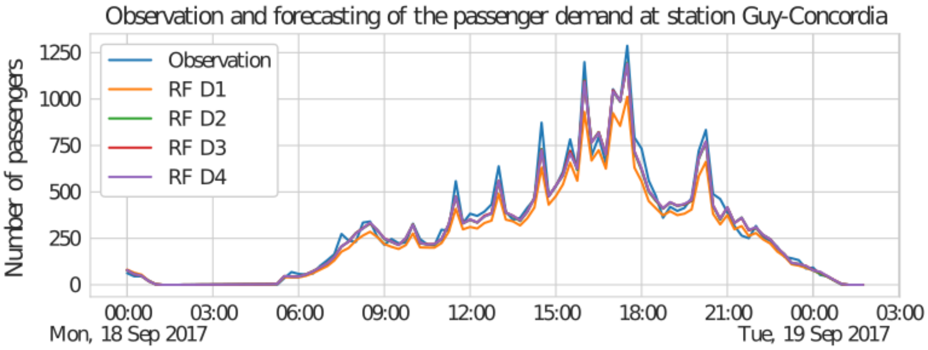

Figure 7: Observation and forecasting of the passenger demand at Guy-Concordia station, Monday, September 18, 2017.

Table 5: Errors of the random forest model applied to the test set period, 2017 (event period and the period without event), on the 17 stations that host events in 2017.

	Period without event			Event period		
Data	RMSE	MAE	MAPE	RMSE	MAE	MAPE
D1	61.54	28.48	14.95	159.13	46.96	23.59
D2	50.71	24.73	13.18	153.34	43.44	22.20
D3	49.13	24.10	13.19	137.69	43.51	21.37
D4	**48.83**	23.98	13.12	**124.72**	40.70	21.07

The data are represented by different sets of features (D1, D2, D3 and D4) described in Section 4.1. The different models are described in Section 4. The evaluation metrics RMSE, MAE and MAPE@150 are defined in Section 4.3.

lower threshold for a model comparison. According to this data set, the threshold of 150 seems to be a good compromise for the evaluation of the models.

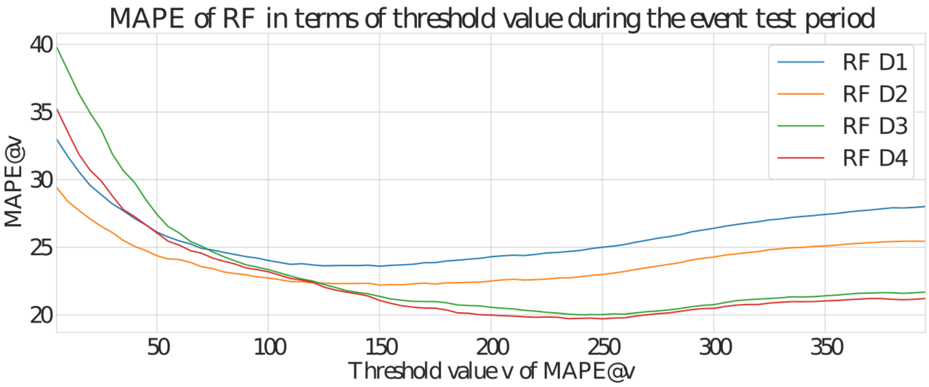

Figure 8: MAPE of random forest models with input data sets (D1, D2, D3, D4) in terms of threshold value of MAPE during the event test period.

Taking the presence of events into account may be essential for forecasting the number of passengers with precision. As shown in Figure 9, the random forest with input data set D1 or D2 (detailed information about the day) is not able to predict the high increase in the passenger demand due to the end of a hockey game at Lucien-L'Allier station. However, with the help of event and event category information (input data set D4), the random forest model accurately forecasts the passenger demand peak.

Figure 10 shows the passenger demand observation during the event named "Nuit Blanche", which induces a very specific pattern due to numerous events occurring during the night in the event area of Place-des-Arts station and the opening of the metro all night. We can observe a high increase in the number of passengers that has been successfully forecasted by the random forest model with input data set D4.

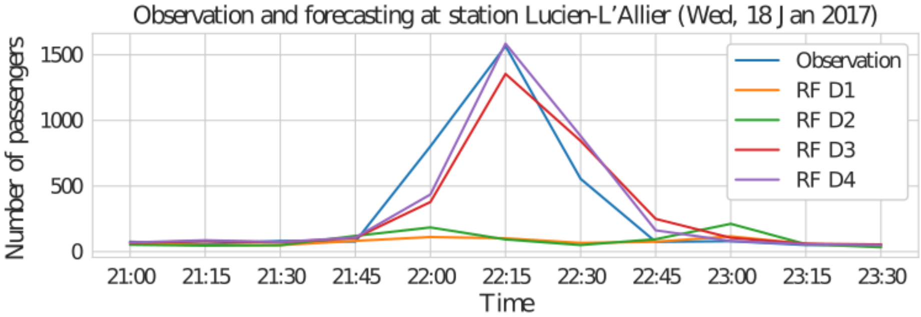

Figure 9: Observation and forecasting of the passenger demand at Lucien L'Allier station, Wednesday, January 18, 2017. The information about the event is the following: start time, 19:30; end time, 21:30; station, Lucien-L'Allier; and category, hockey.

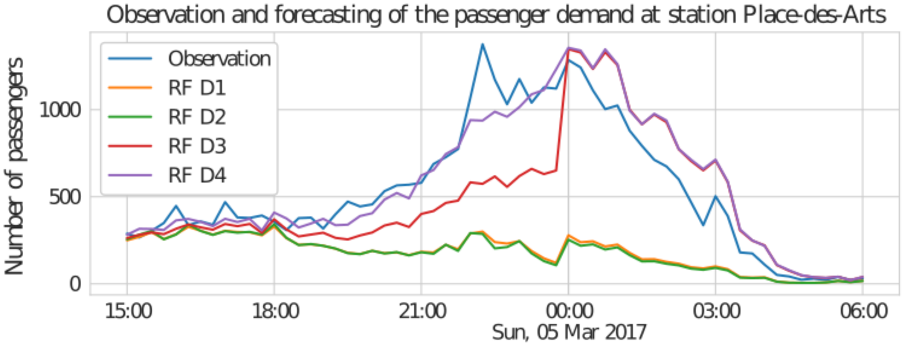

Figure 10: Observation and forecasting of the passenger demand at Place-des-Arts station, Sunday, March 5, 2017. The information about the event is as follows: start date and time, 2017-03-04 18:00:00; end date and time, 2017-03-05 05:00:00; station, Place-des-Arts; and category, other.

5.1.2 Feature Importance on Specific Station

Models such as random forest allow quantification of the feature importance of the input data. Because one model has been trained per station, we are able to investigate with precision the feature importance per station, which could be interesting for understanding how the models work. Figure 11 shows the feature importance of the random forest model with input data set D4 for 3 stations with particular locations (stations are depicted in Figure 12). The feature ranking denoted as f is computed with the "mean decrease impurity" used for regression trees introduced by [24]. The importance of feature i, denoted as f_i, is given by:

$$f_i = \frac{\sum_{j:\text{node j splits on feature i}} n_j}{\sum_{j\in\text{all nodes}} n_j} \tag{5}$$

With n_j the importance of node j,

$$n_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{6}$$

where w_j is the weighted number of samples in node j, C_j is impurity in this node that corresponds to the within node variance of the output value, and left(j) and right(j) are its respective child nodes. The feature importance is given in percentage and has been aggregated in the following categories: the information about the date detailed in Section 3.2, events (that corresponds to the sum of the feature importance of all the start, end and period event features of all the stations with events) and category (it corresponds to all the information about the event category available in all the station with events). The most important feature is the name of the day of the week, with 60.31%, 83.78% and 53.49% feature importance for the Place-des-Arts, Square-Victoria and Guy-Concordia stations, respectively. For these three stations we can see that the importance of the features December 24 and 31 are less than 1%. This is explained by the few days with these features in the training database (4 days). Nevertheless these features are still important to predict those special days that cannot be categorized with the other features. We can see that Place-des-Arts is a station largely impacted by the event and event category features (approximately 8% and 11% of feature importance), which is explained by the presence of many events located near this station. Square-Victoria-OACI is a station located in a business area; in contrast to the Place-des-Arts station, we find that the most important features are holiday and Christmas school holidays. Finally, the Guy-Concordia station is the station of the Guy-Concordia University, which explains the importance of the features Christmas school holidays and school holidays parts 1 and 2.

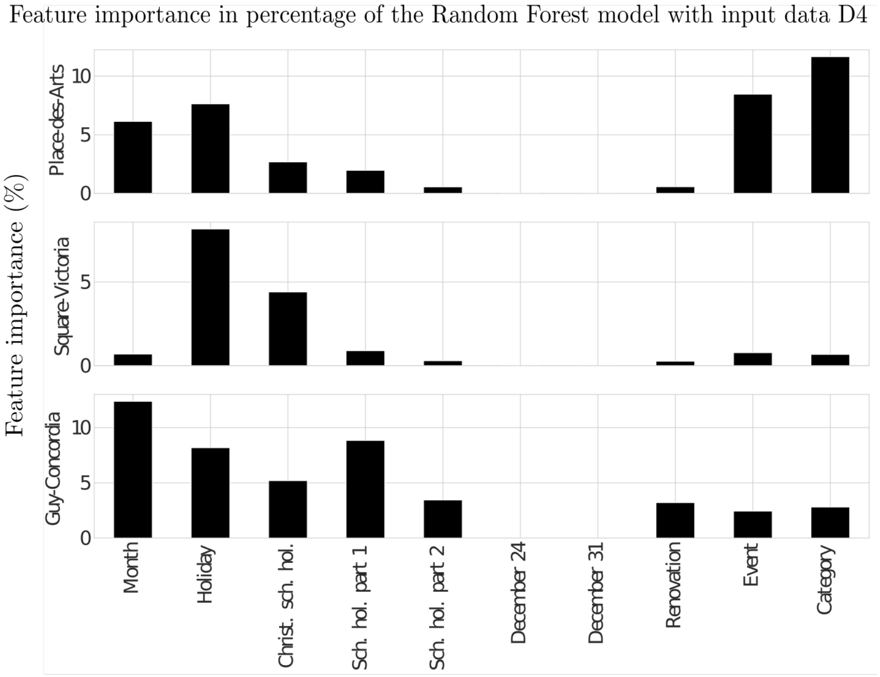

Figure 11: Aggregated feature importance of random forest model with input data set D4 in Place-des-Arts, Square-Victoria-OACI and Guy-Concordia stations

5.1.3 Forecast Analysis by Station

In addition to the global analysis detailed in Section 5.1.1, it is also important to analyse the results per station because each station has its own activity pattern. Figure 12 shows the MAPE@150 error of the best forecasting model, which is the random forest with the full set of features as input data (D4 corresponds to the information about the day, the event and the category of the event). As shown, the model obtains an error greater than or equal to 17% as MAPE@150 in some special stations. Université de Montréal and Edouard-Montpetit stations are located on the University of Montreal campus, which implies a passenger demand impacted by the university calendar (MAPE@150 equal to 17% and 20%). The Lucien-L'Allier station is difficult to predict (MAPE@150 equal to 20%) because this station is the one that hosts most of the events in the city. Finally, the several events that took place at the Jean-Drapeau station, located on an island without habitation, make this station the hardest to predict (50% of MAPE@150).

5.2 Forecasting Results per Type of Ticket or Pass

One of the goals of transport operators is to accurately estimate the demand for certain types of ticket or pass used to travel to adapt ticket availability to passenger demand. In this context, we compare the forecasting results of the random forest model with a focus on the forecasting of subsets of data corresponding to the number of passengers by type of ticket or pass used to travel. The type of ticket or pass are aggregated into the following categories: STM monthly pass (SMP), regional monthly pass (RMP), book tickets (BT) and occasional pass (OP).

5.2.1 Forecast Analysis per Type of Ticket or Pass During the Global Period

According to the results shown in Table 6 and as expected, the occasional transport demand (pass OP) is the most difficult to forecast. The MAPE@150 is 27.93% for this type of pass that represents 15.7% of the total passenger demand against a MAPE@150 of 12.23% for the type of pass SMP that represents 51% of the total passenger demand. The use of input data set D4 related to event information in addition to the information about the day is necessary to obtain the best results for the forecasting of occasional passenger demand BT and OP passes. This is due to the particularity of book tickets and the occasional pass that are mainly used during events.

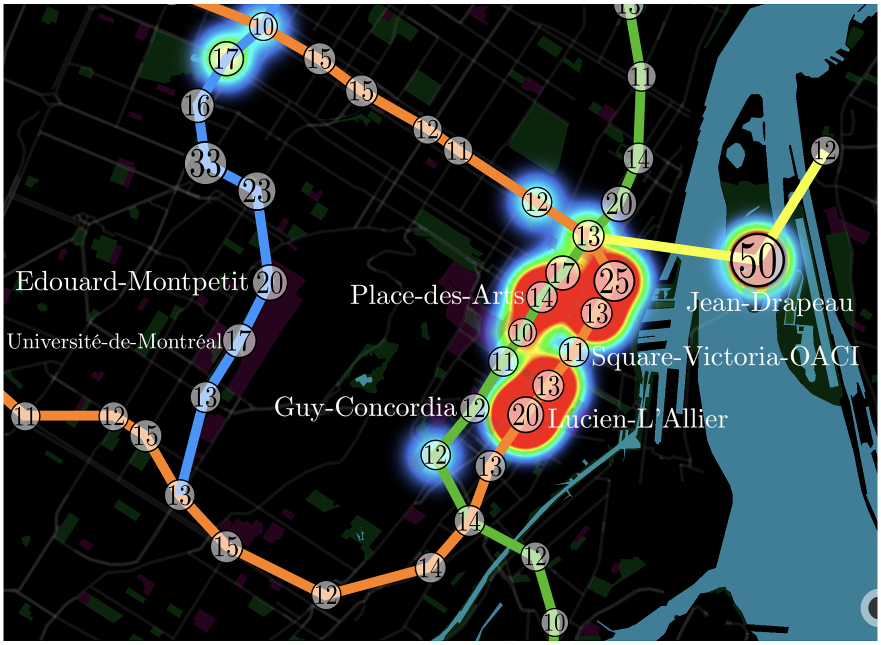

Figure 12: MAPE@150 error per station during the global test period (2017) of the random forest model with input D4. The metro lines of Montreal are depicted by the blue, orange, green and yellow lines. The heatmap (blue zone to red zone) depicts the event activity during 2017. The green background represents parks, and the pink background indicates school or university.

Table 6: Errors of the random forest model on the training period from 2015-2016 and the test set period, 2017, per type of ticket or pass used to travel.

		Train set (2015 and 2016)			Test set (2017)		
Pass	Data	RMSE	MAE	MAPE	RMSE	MAE	MAPE
SMP	**D2**	16.37	8.65	9.67	**20.04**	10.76	12.23
	D4	14.10	7.71	8.51	20.08	10.72	12.24
RMP	**D2**	8.30	3.08	9.28	**10.17**	3.76	11.97
	D4	7.46	2.84	8.28	10.28	3.75	12.12
BT	D2	5.58	3.06	16.23	6.73	3.63	20.48
	D4	5.06	2.89	13.17	**6.57**	3.58	19.19
OP	D2	19.49	4.88	28.28	21.41	5.66	30.42
	D4	11.22	4.07	18.93	**17.86**	5.40	27.93

The data are represented by different sets of features (D2 and D4) described in Section 4.1. The evaluation metrics RMSE, MAE and MAPE@150 are defined in Section 4.3. The aggregation of the types of passes is as follows: STM monthly pass (SMP), regional monthly pass (RMP), book tickets (BT) and occasional pass (OP).

5.2.2 Forecast Analysis per Type of Ticket or Pass Used to Travel During the Event Period

The STM monthly pass and regional monthly pass are slightly impacted by events. As shown in Table 7, the RMSE of the 17 stations with events increased from 24.99 to 28.01 during the event period for the STM monthly pass and from 12.48 to 13.27 during the event period for the regional monthly pass. Meanwhile, we can observe that book tickets and occasional passes are highly impacted by the presence of events. The random forest model obtains the best scores for these two types of passes with the input data set D4 in both periods: with and without event periods.

We can observe the impact of a hockey game on the passenger demand for each type of ticket or pass in Figure 13. This event is described as beginning at 07:30 p.m.; however, the ending time is not defined. We can see that every random forest with the input data set D4 (day information, event and category information) is able to forecast with a good accuracy the increase in the number of passengers between 10:00 p.m. and 11:00 p.m. The type of ticket or pass used to travel that is the most impacted by the event is the occasional pass with an increase of 1000 passengers during the passenger demand peak at 10:15 p.m.

Table 7: Errors of the random forest model on the test event period and test set period without events, 2017, per type of ticket or pass over the 17 stations with events during the year 2017.

		Test period without event			Test set period with event		
Pass	Data	RMSE	MAE	MAPE	RMSE	MAE	MAPE
SMP	D2	**24.99**	13.12	11.88	30.48	13.68	18.14
	D4	25.21	13.07	11.99	**28.01**	13.27	16.49
RMP	D2	**12.48**	5.27	11.20	13.47	5.67	11.18
	D4	12.67	5.28	11.41	**13.27**	5.57	11.23
BT	D2	8.95	4.74	17.92	16.16	6.54	51.70
	D4	**8.80**	4.65	17.44	**14.63**	6.36	41.03
OP	D2	21.85	8.57	30.06	114.60	23.72	55.70
	D4	**19.17**	7.79	29.33	**91.03**	21.95	43.98

The data are represented by different sets of features (D2 and D4) described in Section 4.1. The evaluation metrics RMSE, MAE and MAPE@150 are defined in Section 4.3. The aggregation of the types of passes is as follows: STM monthly pass (SMP), regional monthly pass (RMP), book tickets (BT) and occasional pass (OP).

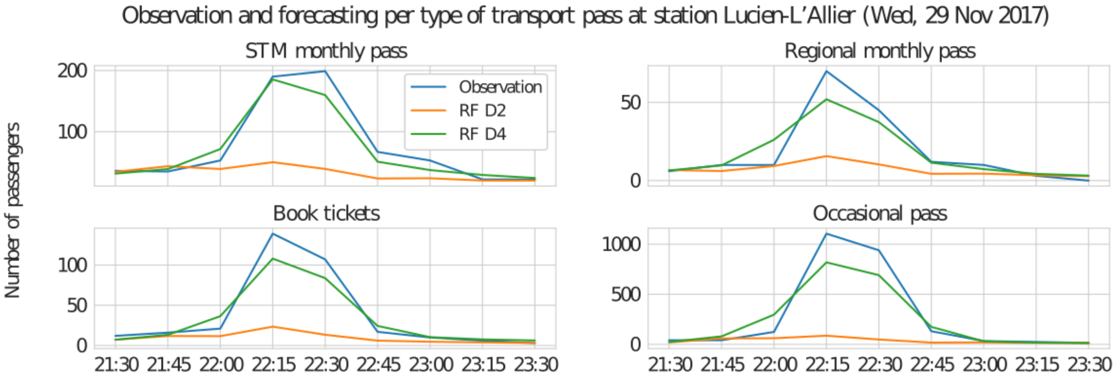

Figure 13: Observation and forecasting of the passenger demand per type of ticket or pass at Lucien-L'Allier station, on Wednesday, November 29, 2017. The information about the event is the following: start date and time, 2017-11-29, 19:30:00, end date and time Nan; station, Lucien-L'Allier; and category, hockey.

6 Conclusion

This paper has investigated the use of smart card, calendar and event data to forecast metro passenger demand per station with a long-term forecasting time horizon (until one year ahead) with fine-grained temporal resolution (15 minutes aggregation). We performed the forecasting task on real data (Montréal subway, Canada) by taking into account events in the city, such as concerts, hockey games, festivals, and so forth. The operational objectives were twofold: long-term forecasting can be useful for transport operators to adapt the transport supply and to adjust ticket availability to passenger demand. In this context, we have investigated the forecasting of the number of passengers per type of ticket or pass in addition to the forecasting of the global passenger demand.

We have proposed generic data shaping, allowing the use of contextual data (smart card, calendar and event data) as input for well-known regression models: basic, statistical and machine learning models. Global forecast analysis has proven that it is possible to obtain good long-term forecasting accuracy with fine-grained resolution even in the presence of events. The random forest model achieved the best forecasting results with the calendar information and event data as input. The forecasting results highlighted the importance of taking event data into account during the forecasting of passenger demand, particularly during an event period. This study has also illustrated the value for transport operators to use one regression model per station to understand which features mostly impact the passenger demand per station. We have studied transport-related results to better understand which station is difficult to predict. In the same line of work, we have shown that, as expected, passenger demand depending on certain types of ticket or pass used to travel (book tickets and non-rechargeable smart cards) is more impacted by events and requires event data to be accurately forecasted.

We have also proposed a basic method to reproduce the impact of the global year-to-year trend on the forecasting results. These results have demonstrated the effectiveness of the trend method in addition to the data shaping and machine learning method for such forecasting tasks. Nevertheless, further work is required to investigate in detail the trend problem in the long-term prediction task. Future work could investigate a medium-term forecast that could be placed between a long-term forecast that requires only the use of data available well in advance and a short-term forecast that requires recent observations of passenger numbers (collection and analysis of near real-time data). For this purpose, the medium-term forecast model could take as inputs, in addition to long-term data (calendar and event information), medium-term features such as the trend of the number of passengers observed recently (e.g., in previous days, weeks or months).

If we take the case of a transport network with a constrained spatial grid that can be defined by stations (e.g., metro, bus, train), it will be possible to use the same forecast method as well as the same data formatting method. These forecasting methods are applicable provided that the same types of data (calendar and event data) are used. On the other hand, in the case of an unmeshed network as in [7] (e.g., road traffic, free-floating bicycles, free-floating scooters), it will be necessary to spatially mesh the network in order to group the events into a number of fixed points of interest, as well as for the counting of transport flows, which will also have to be grouped into a fixed number of points. Because it is desired to be applicable in other public transport systems of the world, the forecasting methodology presented in this study could definitely help to create high added value mobility services for citizens.

Acknowledgement

This research was partially supported by Thales and the Natural Science and Engineering Research Council of Canada (NSERC). The authors wish to thank the transport organisation authority of Montreal (Société de transport de Montréal) for providing the smart card and event data. They also acknowledge the computer infrastructure support from the IVADO and the Quebec Research Funds (FRQNT, FRQSC).

References

[1] Yang Li, Xudong Wang, Shuo Sun, Xiaolei Ma, and Guangquan Lu. Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. *Transportation Research Part C: Emerging Technologies*, 77:306–328, 2017.

[2] Jérémy Roos, Stephane Bonnevay, and Gérald Gavin. Short-term urban rail passenger flow forecasting: A dynamic bayesian network approach. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 1034–1039. IEEE, 2016.

[3] Chunsheng Cui, Hongfei Jia, Liping Huang, and Xiaopeng Zhang. Fuzzy multivariate narx model for passenger entrance flow prediction in the shanghai subway system. *Journal of Intelligent & Fuzzy Systems*, 31(6):3047–3054, 2016.

[4] Chuan Ding, Donggen Wang, Xiaolei Ma, and Haiying Li. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability*, 8(11):1100, 2016.

[5] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. Time-series extreme event forecasting with neural networks at uber. In *International Conference on Machine Learning*, 2017.

[6] Jintao Ke, Hongyu Zheng, Hai Yang, and Xiqun Michael Chen. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies*, 85:591–608, 2017.

[7] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, pages 1655–1661, 2017.

[8] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, and Jieping Ye. Deep multi-view spatial-temporal network for taxi demand prediction. *arXiv preprint arXiv:1802.08714*, 2018.

[9] Yuankai Wu and Huachun Tan. Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework. *arXiv preprint arXiv:1612.01022*, 2016.

[10] Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. *arXiv preprint arXiv:1709.09585*, 2017.

[11] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional neural network: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[12] Ming Ni, Qing He, and Jing Gao. Forecasting the subway passenger flow under event occurrences with social media. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1623–1632, 2017.

[13] Ioulia Markou, Filipe Rodrigues, and Francisco C Pereira. Real-time taxi demand prediction using data from the web. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1664–1671. IEEE, 2018.

[14] Filipe Rodrigues, Ioulia Markou, and Francisco C Pereira. Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach. *Information Fusion*, 49:120–129, 2019.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, 19(3):273–288, 2015.

[17] Luís Moreira-Matias, João Gama, Michel Ferreira, João Mendes-Moreira, and Luis Damas. Time-evolving od matrix estimation using high-speed gps data streams. *Expert systems with Applications*, 44:275–288, 2016.

[18] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[19] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[20] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[21] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[22] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[24] Leo Breiman. *Classification and regression trees*. Routledge, 2017.