Machine Learning Methods to Forecast Public Transport Demand Based on Smart Card Validations

Brunella Caroleo[1(✉)], Silvia Chiusano[2], Elena Daraio[2], Andrea Avignone[2], Eleonora Gastaldi[2], Mauro Paoletti[3], and Maurizio Arnone[1]

[1] LINKS Foundation, Via P.C. Boggio 61, 10135 Turin, Italy
{brunella.caroleo,maurizio.arnone}@linksfoundation.com
[2] Politecnico di Torino, c.so Duca degli Abruzzi 24, 10129 Turin, Italy
{silvia.chiusano,elena.daraio,andrea.avignone}@polito.it,
eleonora.gastaldi@studenti.polito.it
[3] Granda Bus, Via Circonvallazione 19, 12037 Saluzzo, Italy
mauro.paoletti@grandabus.it

Abstract. This paper explores the forecasting of public transport demand using mobility data obtained from electronic tickets and smart cards. The research aims to estimate the demand for a selected route at a specific bus stop on a given day and time slot. The study utilizes a large dataset of historical demand data, including approximately 10 million validations collected in 2019 by the Piedmont transport operator Granda Bus, and combines it with additional information such as weather conditions, anonymized user data, and temporal segmentation of the yearly calendar. To identify the peculiarities in demand forecasting for each bus route and stop, a clustering analysis is performed, resulting in the identification of six cohesive and homogeneous clusters. Various machine learning models are tested and compared to determine the most suitable model for forecasting public transport demand at each stop within one-hour time slots. The results demonstrate that machine learning algorithms consistently outperform average-based techniques: the machine learning algorithms exhibit a significant improvement (up to 50% compared to the baseline) when demand uncertainty is greater. The proposed methodology framework is replicable and transferable to other areas, providing a valuable tool for optimizing resource allocation and network planning, while enhancing user satisfaction by accurately forecasting passenger demand at each stop and desired time slot.

Keywords: public transport demand · machine learning · clustering · forecasting

1 Introduction

Estimating the public transport demand has become a great concern for the public transport agencies: it would allow to improve the service offered to the customer and to optimize the physical resources and the operating costs to the service provider.

A. L. Martins et al. (Eds.): INTSYS 2023, LNICST 540, pp. 194–209, 2024.
https://doi.org/10.1007/978-3-031-49379-9_11

This objective has many challenges to overcome: the number of passengers which need to travel at specific place and time may depend on several factors, thus subject to a great variability. The potential directions are toward a reliable forecasting method, able to consider the specific features of the territory, of the service and of the demand, or towards the design of an on-demand transport service.

To estimate the demand for a certain trip, travel documents analysis can provide useful information to create suitable models [1]. Among the ticketing typologies, smart cards generate large amounts of data revealing more insights in passengers' travel behavior [2], allowing the analysis of the current public transport usage and the prediction of the future one.

This paper addresses the problem of forecasting public transport demand at a certain bus stop for each route (denoted as *bus stop-route couple*) of a geographical zone at each hour of the day of the following week.

Data examined in this paper come from the public transport operator Granda Bus of the Piedmont Region (Italy). Smart cards of this operator allow to charge both subscriptions and transport credit documents. The first case (subscriptions) refers to a fare which gives the right to travel within a certain area and for a certain time of validity, which can start with the purchase or with the first validation. In the latter case (transport credit), a certain amount of money is charged on the card and, at each travel, the corresponding cost is subtracted according to the departure and the arrival location. Smart cards are usually validated only when on-boarding (tap-in/check-in), especially in the case of subscriptions. If check-out is not validated, the final destination of the journey can be inferred as stated in the reference literature [3, 4]. If properly exploited, information coming from the smart cards can be extremely precious.

To forecast public transport demand at a certain *bus stop-route couple* of a geographical zone at a given 1h-timeslot, this paper proposes a novel approach that combines already known techniques but in an innovative and customized way according to the context of analysis. The proposed approach is based on the main concept of the bus stop-route couple, meaning that each bus stop of the public transport infrastructure is analyzed separately for each route that passes by it (details in Sect. 3). Starting from this concept and to pursue the forecasting goal, the proposed approach is structured as a two-level methodology designed in the following way. The first level is based on a clustering analysis, whose objective is to identify similarities among the bus stop-route couples. The second level is based on a regression analysis contextualized with respect to each cluster, previously identified through the first level of the methodology, whose objective is to evaluate the best suitable model and its configuration to forecast the demand. Within the proposed methodology, the data about service demand and supply are enriched with additional information about the meteorological conditions and other temporal information (details in Subsect. 3.1).

This paper is organized as follows: after a review of the literature in Sect. 2, Sect. 3 contains a description of the proposed approach, while Sect. 4 reports the main obtained results and their analysis. Finally, conclusions and recommendations for future research are provided in Sect. 5.

2 Related Works

The literature presents different issues that can be dealt using smart card data: segmenting customers according to their personal data or to their mobility patterns [1, 2], forecasting the most likely destination given the boarding stop [4–6], forecasting individual mobility and trip chain [7, 8], estimating the time at which a vehicle will arrive a certain stop [9, 10], predicting travel and dwell time [11, 12], predicting bunching and preventing it [13], reorganizing the routes of the public transport basing on travel/dwell times and on demand at each stop [14]. Similarly, clustering techniques are employed to reveal insights about popular stations and group of passengers [15], as well as for the mining of travels [16] and for characterizing the structure of cities [17].

The focus of this paper is the problem of forecasting mobility demand of public transport at a certain stop to predict how many people will need a particular mobility service, at a certain place, within a specific time slot. There are different approaches in the literature, mainly differing in terms of type of model used for the prediction, input variables, and time horizon of the prediction. Some previous studies focused on predicting bus passenger demand with deep learning and machine learning techniques. For example, [18] proposes the use of a SAE-DNN model (a hybrid deep network of unsupervised SAE and supervised DNN) to predict the hourly passenger flow using a three-stage deep learning architecture. [19] used Gradient Boosting Decision Trees to forecast the number of alighting passengers up to 15–30 min, using also the demand data of the adjacent bus stops. The authors of [20] used a LSTM (Long short-term memory) recurrent neural network (RNN) architecture to forecast the demand using also weather features, with a 10-min time horizon. LSTM technique has also been used in [21] in comparison with SVR (Support Vector Regression). Other approaches are presented in [22] (ARIMA/SARIMA) and [23] (Random Forest).

However, none of these works has analyzed the entire supply network of a public transport operator, crossing with the historical demand data, and detecting the peculiarities in terms of prediction of each route/stop with respect to the other ones.

After a preliminary exploration conducted in [24, 25], this paper proposes a novel data analysis approach to cover this gap. Specifically, it aims at identifying the most proper model to predict public transport demand at each stop of the entire transit network covered by a transport operator for each 1-h timeslot. The features characterizing each bus stop are: time series of the demand at that bus stop, working/school holiday days, weather conditions, type of users (students/retirees/others), segmentation of the yearly calendar depending on the transport supply, the cluster to which the bus stop belongs.

3 Methodology

The proposed methodology is based on data coming from a public transport operator, that consists in the public transport supply and demand of the whole year 2019. The dataset is provided by Granda Bus, a consortium of 16 transport agencies founded in 2004, mostly operating in the area of Cuneo, in North-Western Italy. Data provided by Granda Bus consortium are totally anonymous and comply with the specifications of the privacy authority and the GDPR.

From the supply side, the involved public transport operator collects and stores data in the General Transit Feed Specification (GTFS) format. This is a standard representation of the scheduled public transport services, including also geographical information. In the selected year, the service includes 237 routes, 6,069 trips and 7,371 stops.

From the demand side, for the whole 2019, Granda Bus provided data of users' validations, the typology of the tickets and the category of users. In more detail, from each validation there are: (i) *temporal information*, i.e. date and time at which the validation occurs, (ii) *spatial information*, i.e. the stop at which the validation occurs and the corresponding bus route, (iii) *ticketing information*, i.e. its typology (e.g., single ticket, carnet, weekly, annual subscriptions) and its type (if the validation refers to a check-in or a check-out), and (iv) *anonymized personal information*, as age, sex, birthplace and category (students/retirees/others) in case of smart cards (i.e. subscriptions).

Data retrieved from the public transport operator have been enriched in this study with external data sources, as detailed in Subsect. 3.1. The proposed framework is made of three main building blocks represented in Fig. 1, namely the *Data acquisition, collection and enrichment*, the *Clustering analysis*, the *Forecasting model: analysis and assessment*. Each block will be detailed in the next subsections.

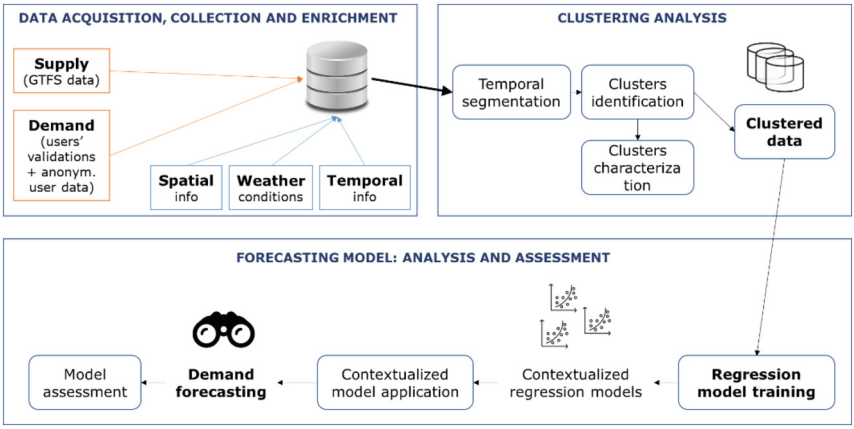

Fig. 1. The proposed framework.

The current methodology is based on the core concept that each bus stop can be better modelled if considered with respect to the route that goes by it. As an example, the route that connects the two municipalities of Saluzzo and Cuneo is identified by route_id = B91 (route_long_name = SALUZZO-CUNEO). This bus route has a stop at Saluzzo railway station (stop_id = 1). The same bus stop also serves other bus routes in addition to B91 (e.g., B95TO, B104, B105). Since each bus stop is characterised by different validation trends according to the considered route, each possible bus stop-route couple has been considered separately. For this reason, from now on we will refer to this concept as the bus stop-route couple, e.g., the couple (1, B91) refers to the *stop_id* = 1 (Saluzzo railway station) of the route_id = B91 (SALUZZO-CUNEO).

3.1 Data Acquisition, Collection and Enrichment

The first block of the framework refers to the acquisition and collection of public transport data, in terms of both supply and demand and including the personal users' data when available (in case of subscriptions). The following additional information has been added to enrich the information retrieved from the public transport operator:

- *Weather conditions*: meteorological information is retrieved from 3bMeteo [26] with daily or hourly temporal granularity (one station for each municipality). Temperature (minimum and maximum) and quantity of rain precipitation are hourly information, while the categorical description of the weather condition (i.e., sunny, partly cloudy, cloudy, variable, rainy, snowy, rainy-snowy, stormy, foggy) is daily;
- *Temporal information*: temporal information is computed from the validation timestamp to further characterise the data in terms of *weekday* (Monday - Sunday), *day type* (working day, holiday day or pre-holiday day) and *school holiday* (boolean equal to 1 for the days in which the schools are closed);
- *Spatial information*: type of zone in which the bus stop of the validation is located (residential, working, or mix).

For the cleaning process of validation data, the following set of filters was designed:

- *User's information incoherent filter*: it removes all validation related to users with incoherent registry information and ticket typologies (e.g., middle-aged user with student subscription);
- *Frequency validation filter*: it discards the validations related to the same user ID whose count in the same day is greater than 10 (threshold value defined together with the transport provider);
- *Missing stop ID filter*: it removes all the validations without any bus stop ID;
- *Average speed filter*: it removes the validations that implies a user average speed greater than the average speed for the relative trip;
- *Synchronisation filter*: it evaluates the feasibility of two validations associated to the same user/customer and at the same time instant. In particular, it checks if the customer could cover the distance between the reported stops in less than 1 min, which corresponds to the sensitivity of the recorded timestamp.

Due to the presence of categorical variables (i.e., the day of the week, the day type and the meteorological conditions), the one-hot-encoding has been selected to exploit the translation into numerical variables.

3.2 Clustering Analysis

The second building block of the proposed methodology exploits cluster analysis to partition the bus stop-route couples and find the optimal forecasting model for each stop. The model which best fits the data related to the most representative bus stop-route couple of the cluster (i.e., the centroid) is used for forecasting the demand at the other stops of the same cluster. It entails two main steps of the analysis: (i) the *temporal segmentation of analysed data*, to properly analyse the mobility data by a temporal point of view, and (ii) the first level of the proposed two-level methodology, which consists

of the *clustering method selection*, together with its tuning and the evaluation of the obtained clusters. Both these steps are here after described more in details.

Temporal Segmentation: according to the characteristics of the public transport service supply, it is possible to identify different time periods that follow external conditions variations, such as the school holiday period which highly affects the daily routines. Three temporal segments have been identified:

- *Working segment*: it refers to the period during which schools are open and days are referred to as working days (from the second week of January to the first week of June and from the second week of September until the third week of December);
- *Holiday segment*: it refers to the period during which schools are closed and days are referred to as holiday days (from the second week of June to the first week of September, plus the Christmas holidays);
- *Hybrid segment*: it refers to the weeks composed by both working and holiday days (e.g. Carnival's week, Easter's week and the 25th April's week).

Clusters Identification: the cluster identification step aims to select the most suitable clustering algorithm to identify clusters of similar bus stop-route couples.

To model the bus stop-route object, the current methodology is based on the following concepts: given a time bin *tb*, each bus stop-route couple is modelled in terms of demand *d* at *tb* (i.e. d_{tb}). For each bus stop-route couple, the corresponding demand d_{tb} is computed based on two features: (i) the *sum of the validations occurred in each time bi*n d_{tb}^{sum}, and (ii) *the variance of demand at each time bin* across the days in the dataset d_{tb}^{var}. According to this model, each bus stop-route is characterized by $time\_bins \cdot 2$ features. To evaluate the right cluster set and if it is possible to keep it as fixed across different time periods, the clusters' identification needs to be performed on temporal segments while comparing the obtained results.

The proposed methodology explored the suitability of the algorithms available in literature by evaluating different kinds of approaches: the k-means algorithm, the hierarchical agglomerative algorithm and the DBSCAN density-based algorithm. Each of them requires the fine-tuning of parameters configuration through a grid search.

To evaluate the quality of the identified cluster set, the following indices are chosen:

- the *Sum of Squared Errors* (SSE): it is defined as the sum of the squared distances between the centroid and each member of the cluster. It evaluates the cluster compactness and the best number of clusters for k-means using the Elbow method;
- the *Davies-Bouldin index* (DB index): it is defined for each cluster as the maximal ratio between the sum of the spatial dimensions among itself and another cluster and the distance between the two clusters. Then, the values are averaged. It is a measure of compactness and separation from other clusters.
- the *Silhouette Score* (SS): it is defined as the ratio between *b-a* and *max(a, b)*, where *a* is the mean distance from the other elements of the same cluster, while *b* is the mean distance from the elements of the nearest cluster. It ranges between -1 (worst case) and 1 (best case). It is also a measure of compactness and separation from other clusters.

3.3 Forecasting Model: Analysis and Assessment

The third building block entails the second level of the proposed two-level methodology, which consists of the regression model selection and fine tuning. Within this research activity, the authors selected some of the most known algorithms, such as Random Forest (RF) and its version with most important features only (RF-MF), Gradient Boosted Decision Tree (GBDT), Support Vector Regressor (SVR) and Seasonal Auto Regressive Integrated Moving Average (SARIMA).

The performances of the abovementioned algorithms have been compared with two wise-baselines: the *Average Response (AR)*, which computes the average validation among the corresponding hours of the same type of day (working/holiday), and the *Median Response (MR)*, which is the same as the AR but it computes the median value. These wise-baselines are simple and they do not require parameters setting, so they are useful to evaluate the benefits of the adoption of machine learning techniques.

For each predictive model, a grid search has been conducted to determine the optimal configuration setting, both in terms of hyperparameters and in terms of training set length. In particular, the proposed methodology evaluates the model goodness when trained on multiples of a week, which means 7 days, 14 days, and so on.

Forecasting and temporal segmentation: the forecasting has been performed by taking into consideration the temporal segmentation previously described in Subsect. 3.2. To forecast the demand of the working segment, the model has been trained only on weeks of the working segment and analogously for the holiday segment. The approach for the hybrid segment is slightly different: these weeks have been evaluated combining the usage of working and holiday models, according with the type of each day of the week under analysis, as represented in Fig. 2.

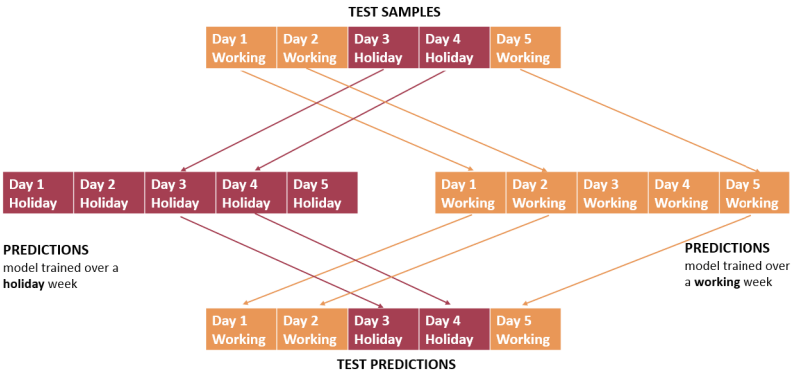

Fig. 2. Forecasting approach in the hybrid segment.

Model assessment: the last part of the proposed methodology provides evidences of the performance in the demand forecasting, obtained through the regression models introduced above. In particular, the quality of forecasting is quantified through a set of common quantitative metrics. Thus, as a result of this block, some useful guidelines to support the service provider in the decision-making process could be provided.

The contextualised models are evaluated in terms of: (i) *Mean Absolute Error* (*MAE*), measuring the difference between the predicted value and the real one; (ii) *Mean Absolute Scaled Error* (*MASE*), a scale-free error metric that never deals with undefined or infinite values, representing a good choice for intermittent-demand series, which can occur in our analysis context due to the service interruption during some timeslots (like during night); (iii) the coefficient of determination *R squared* (R^2), evaluating the ratio between the variance of the error and the variance of the measured data.

To evaluate the performance on each temporal segment, the authors propose to train as many models as possible that can be then tested on the week right after the training period, averaging the performance metrics computed for each model.

4 Results

The computing environment for the experimental evaluation was mainly based on Python, using the most popular libraries (e.g., pandas, numpy, sklearn, geopy).

4.1 Data Acquisition, Collection and Enrichment

The dataset refers to all the bus tickets validations of the Granda Bus consortium for the whole year 2019. The map in Fig. 3 shows an example of geographical distribution of the bus stops in one month of the year under examination (October 2019, retrieved from the GTFS provided by the transport operator), where the colour of the stop points ranges from white to blue depending on the number of incoming validations (at least one), while stops with no validations in October are in pink.

Data retrieved from each validation is: timestamp (date and time), stop_id (the bus stop where the validation occurred), trip_id and route_id (the trip and the route corresponding to the validation), ticket typology (single ticket/carnet/subscription/students' subscriptions/over65 subscriptions), and if it is a check-in or a check-out. In case of smart card, information of each user has been included. The recorded information includes about 1,000 travel documents and 100,000 users. The dataset has been enriched with weather, temporal and spatial information introduced in Subsect. 3.1.

Since the objective of the study is to forecast the demand at each hour of the day for the following week, data were resampled in hourly time slot. The study takes as unit of measure one week (24*7 = 168 records): this means that the size of the training dataset is a multiple of one week, and the size of the test dataset is fixed to one week, so that each week can be separately forecasted.

As a result of the data cleaning described in Sect. 3.1, 89% of raw data for the whole year have been considered for further analysis: this percentage can be interpreted as a good quality of the dataset provided by the transport operator.

4.2 Clustering Analysis

A clustering of all the bus stop-route couples has been performed according to the methodology described in Sect. 3.2, using validation data of one representative month (October) in 6,714 stop-route couples.

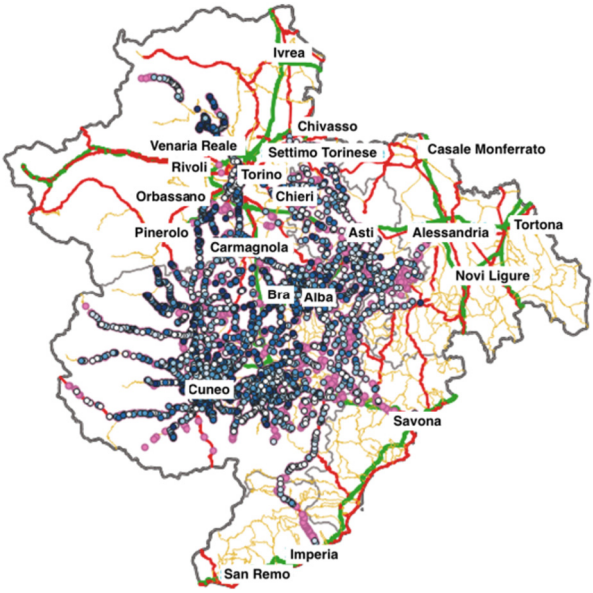

Fig. 3. Geographical distribution of stop points: example on the validations of October 2019. (Color figure online)

Three clustering techniques have been compared: density-based DBSCAN, Ward hierarchical agglomerative, and k-means. Density-based clustering provided the worst results, while k-means resulted to be preferable with respect to Ward agglomerative clustering, due to a slightly higher silhouette and lower DB index with the same number of clusters. As regards the choice of the number of clusters, a slowdown in the decreasing of SSE can be observed at 6 and 7 clusters. Since the second value showed a local maximum of DB index and a local minimum of silhouette, the first option has been preferred. Thus, the algorithm chosen for the creation of the clusters is the k-means, with $k = 6$. The algorithm receives as input the data collected in October 2019 and provides in output six clusters of bus stop-route couples. Such partitioning has been obtained by the assessment of the number of validations and the relevance of each stop-route couple in terms of supply (number of trips, terminal stop, number of interchanges within the route and frequency provided) and the volume of the demand.

The distribution of the most significant variables characterizing each cluster is shown in Fig. 4 and in Fig. 5, separately for each partition (from #1 to #6) and related to working days. In Fig. 4, variables are related to the supply (*dens_pop* represents the density of population in the census zone of each stop_id, *num_trips* is the total number of trips passing by each stop, *terminal* is a binary variable denoting if the stop is the first or the last one of the trip, averaged over days), while Fig. 5 refers to the demand (*stud* is referred to validations coming from students' subscriptions, *ret* refers to validations coming from over65 subscriptions, while *other* to other ticket typologies).

Cluster #1 is the one with higher cardinality (6,303 samples), characterized by bus stop-route couples located in isolated places, with very few validations from all ticket

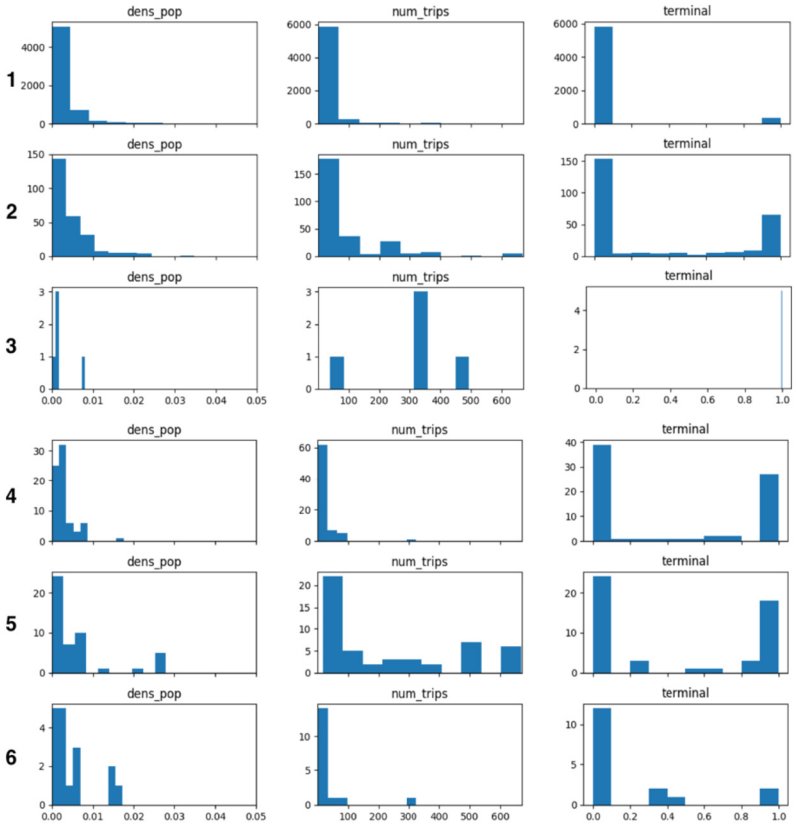

Fig. 4. Distribution of the most significant variables related to the supply across the clusters, numbered from 1 to 6.

typologies, a very low number of trips' interchanges; in addition, these stops are rarely terminals. The opposite holds for Cluster #3 (5 samples): it contains few stop-route couples, but corresponding to a very high demand, especially from students and others, which are the most likely to commute at train stations. The train stations of Saluzzo, Alba and another in the centre of Cuneo belong to this cluster. Also in Cluster #5, formed by 50 elements, there are stop points very relevant from the supply side (in terms of *num_trips* and *terminal*), but with a lower demand if compared with cluster #3. Stop-route couples belonging to Cluster #4 are located in high-density census zones, i.e., mainly in Cuneo and Turin. Cluster #2 (including 264 elements) is also characterized by high relevance in terms of supply and less in terms of demand, but stops belonging to this cluster are located in sprawled areas. Finally, the demand and the supply relevance of the stops are reduced for Cluster #6 (17 elements) and even more for Cluster #4 (75 elements, mainly in rural areas).

Clustering was assessed in an example month (October) characterised by the absence of special holidays and a pattern that is the prevailing one during the calendar year (working days). The analysis was then replicated in months characterised by: (i) public holidays

(isolated holidays, such as 25 April and 1 May), (ii) holiday periods (Christmas, Carnival, Easter), and (iii) periods characterised by a different supply, as summer holidays. The comparison reveals that: (i) the partitioning of stop-route couples in the different clusters is stable, and that (ii) the cohesion within the clusters varies by a percentage deemed negligible. Therefore, the clustering of stop-route couples in the sample month of October was identified as the reference one for the subsequent analyses.

Apart from the characterization of each cluster, as reported in Subsect. 3.2, the analysts' choice was to detect the most representative bus stop-route couple of each cluster (i.e. centroid), find the forecasting model that best fits data on this stop, and then use the same model for the other stops belonging to the same cluster. This hypothesis has been tested and validated through a quantitative evaluation of the clusters' cohesion. Thus, stop-route couples belonging to different clusters could be modelled differently to best capture the mobility patterns and the features characterizing each stop. This part of the analysis will be deepened in the next subsections.

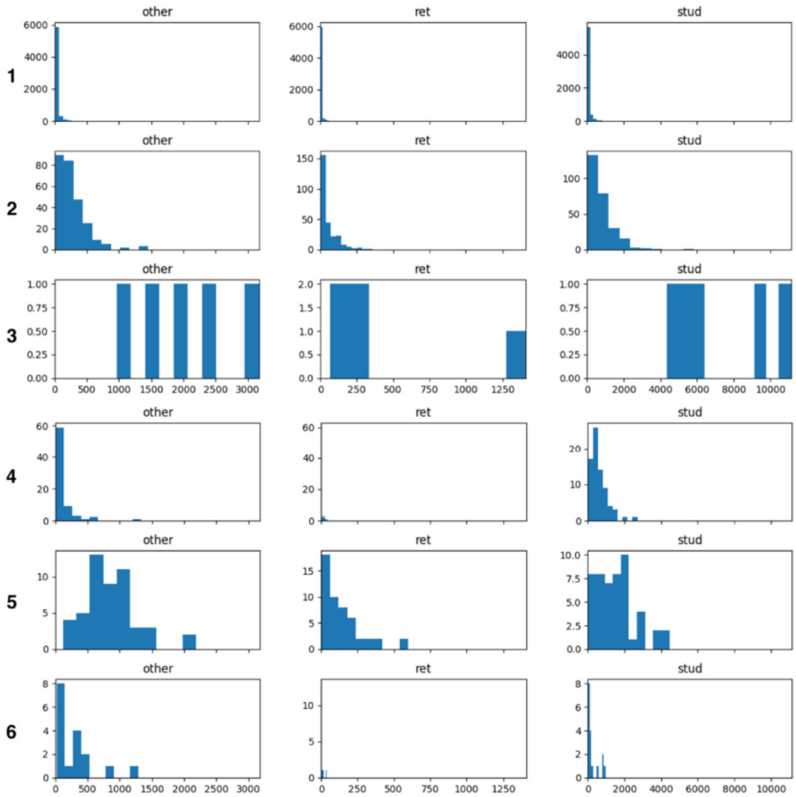

Fig. 5. Distribution of the most significant variables related to the demand across the clusters, numbered from 1 to 6.

4.3 Forecasting Model: Analysis and Assessment

Regression analysis was performed for each centroid of the clusters identified in the previous section. Table 1 reports the centroid information for each cluster, since it is the representative bus stop-route couple element of the specific cluster. The centroid is described by its stop-route name.

Table 1. Centroids of each cluster.

Cluster #	Centroid - Stop name (City)	Centroid - Route name
1	Corso Unione Sovietica (Torino)	Torino (TO) – Saluzzo (CN)
2	Bus station (Saluzzo)	Cuneo (CN) – Saluzzo (CN)
3	Piazza Caio Mario (Torino)	Torino (TO) – Saluzzo (CN)
4	Corso Giolitti (Cuneo)	Cuneo (CN) – Saluzzo (CN)
5	Bus station (Alba)	Bivio Cast. (CN) – Alba (CN)
6	Railway station (Mondovì)	Mondovì (CN) – Cuneo (CN)

As introduced in Subsect. 3.3, several Machine Learning (ML) techniques have been tested for each centroid, comparing the performance with two wise-baselines: the Average Response (AR baseline) and the Median Response (MR baseline).

By comparing the two wise-baselines, it has been observed that the AR baseline performs better with respect to the MR one in the working and in the hybrid temporal segments, while it is slightly worse in the holiday segment. The difference in this last temporal segment is negligible, so the authors selected the AR baseline as reference to evaluate the performance of the machine learning algorithms. With no need of parameters setting, the baseline is useful to evaluate the benefits of the adoption of ML with respect to simple average-based techniques that the Public Transport Operator can already perform without any deep knowledge of ML. For each abovementioned predictive model, a grid search has been conducted to determine the optimal size of the training dataset window (N, number of training days) and the hyperparameters.

The demand forecasting has been performed taking into consideration the temporal segmentation previously introduced (Fig. 2): working, holiday and hybrid segment. For each ML technique investigated, Table 2 shows the best size of the training window (N) obtained from the grid search in each temporal segmentation. The proper N size has been obtained after a deep joint analysis of MAE, MASE and R^2 (see Subsect. 3.3) for each possible value of N (multiples of 7 days, to obtain a number of training weeks required for forecasting).

Finally, each model -fitted for the centroid of each detected cluster (Subsect. 4.2)- has been applied to the given dataset. Table 3 reports, for each cluster and for each temporal segment: (i) the best predictive model identified; (ii) the corresponding size of training window (N); (iii) the corresponding MASE; (iv) the *gain*, i.e., the MASE error gain of using the best identified forecasting model over using the AR baseline.

Thus, the gain provides a quantitative and comparable advantage of using forecasting models with respect to average-based techniques.

Table 2. Grid search results: size of the training window for each ML technique investigated.

ML technique	Temporal segmentation	N	Hyperparameters tuned
RF	Working Holiday Hybrid	21 21 14	Number of estimators (from 10 to 200, step 10) Depth of Trees ([3, 5, 7, None])
RF-MF	Working Holiday Hybrid	21 21 14	Number of estimators (from 10 to 200, step 10) Depth of Trees ([3, 5, 7, None])
GBDT	Working Holiday Hybrid	28 14 28	Number of estimators ([10, 20, 50, 100, 200, 500]) Maximum depth ([2, 3, 4, 5, 6, 7]) Learning rate ([0.0001, 0.001, 0.1, 1.0])
SVR	Working Holiday Hybrid	21 21 21	Kernel ([linear, polynomial, radial basis function]) C ([1, 10, 100, 1000, 10000]) gamma ([0.001, 0.01, 0.2, 0.5, 0.6, 0.9])
SARIMA	Working Holiday Hybrid	7 7 7	Autoregressive order, p (*auto_arima* function) Moving average order, q (*auto_arima* function) Differencing order, d (*auto_arima* function)

By comparing the wise-baseline with the ML algorithms results, the ML algorithms always outperform the results obtained through the AR baseline. As a matter of fact, the gain is always positive, and varies in the range [+2%, + 50%] according to the temporal segment. The best ML algorithm in most cases is the Support Vector Regression: across all temporal segments, the gain of SVR approach over the baseline improves performance by 14% (on average).

If the model that best fits each cluster is chosen within the working segment, the average gain of the MASE error is 10%, within the holiday segment it is 12% and within the hybrid it raises up to 26%. The gain from using ML algorithms is especially positive in the holiday and hybrid segments, where the centroid model is more effective than the AR baseline. This is due to the fact that these are segments with little data (because they are shorter periods during the year with respect to the working segment): in this case, therefore, the use of ML techniques is particularly effective.

Table 3. Best predictive techniques for each cluster in the three different temporal segments: gain of each model with respect to the AR baseline.

Cluster#	Centroid	Temporal segment	Best forecasting model	N	MASE	Gain wrt baseline
1	Corso Unione Sovietica (Torino)	Working Holiday Hybrid	SVR SARIMA SVR	21 7 21	0.57 0.97 0.85	**+38%** +28% +7%
2	Bus station (Saluzzo)	Working Holiday Hybrid	SVR RF SARIMA	21 21 7	0.23 0.77 0.37	+8% +2% **+50%**
3	Piazza Caio Mario (Torino)	Working Holiday Hybrid	RF SVR SARIMA	28 35 7	0.56 0.96 0.76	+3% **+30%** +23%
4	Corso Giolitti (Cuneo)	Working Holiday Hybrid	RF SVR RF	28 35 28	0.47 0.95 0.68	+2% +2% +7%
5	Bus station (Alba)	Working Holiday Hybrid	SVR SVR RF	28 21 28	0.71 0.95 0.73	+7% +6% **+31%**
6	Railway station (Mondovì)	Working Holiday Hybrid	RF SVR SVR	14 28 28	0.43 0.97 0.67	+4% +5% **+38%**

5 Discussion and Conclusions

This study is aimed at predicting public transport demand at each bus stop of the entire transit network covered by a transport operator. This objective has been achieved by the cross analysis of the supply network and the historical demand data (about 10 million validations collected in 2019), and by the enrichment of such data with other sources: weather conditions, users' type, temporal segmentation of the yearly calendar depending on the transport supply. In order to detect the peculiarities in terms of prediction of each route/stop with respect to the other ones, a clustering of all the bus stop-route couples of the transport network has been performed: six clusters were identified, being cohesive and homogeneous in terms of demand prediction. Different machine learning models have been tested and compared to identify the most proper model to forecast public transport demand for each 1-h timeslot at each stop of the transport network.

The first result of this study is the importance of the segmentation of stops resulting from the clustering: it allows to group together bus stop-route couples with similar features in terms of supply and demand so that all the elements of each cluster can be analysed using the same model within the same temporal segment, thus improving the performance of the subsequent forecasting.

As regards the forecasting, the results of the analysis show not only that machine learning algorithms lead always to better results with respect to average-based techniques, but also a quantitative assessment of such gain is provided. In more detail, the advantage in using ML models is more evident within the hybrid temporal segment (up to + 50% of gain with respect to the Average Response baseline), when the trend of validations is more variable over the weeks. Also, regarding holiday periods, ML algorithms lead to significant improvement in terms of forecasting performance (up to + 30% of gain) with respect to the baseline. So, the gain highly depends on the characteristics of the temporal segment: the contribution of ML algorithms is higher when the demand is more irregular (hybrid segment) and the wise-baseline is not able to catch its behaviour, while estimating the demand is still valuable for transport operators. When the demand is more regular (working period), a simpler model (e.g., average-based) is an acceptable alternative to more complex ML models. This is reasonable, since -in the working temporal segment- students and commuters are used to follow their own daily routine, thus the demand variability is lower. However, these cases represent times with more crowding, so even a slight increase in accuracy can lead to benefits in terms of operational efficiency and more comfortable travel experience.

The proposed methodology, that is replicable and transferable in other zones, allows to forecast the passengers' demand at each stop in each desired timeslot. This is fundamental to optimize the allocation of resources (personnel and vehicles), the network planning, and therefore to reduce operating costs and increase users' satisfaction.

References

1. Costa, V., Fontes, T., Costa, P.M., Dias, T.G.: Prediction of journey destination in urban public transport. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015. LNCS (LNAI), vol. 9273, pp. 169–180. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23485-4_18
2. Briand, A.S., Côme, E., Trépanier, M., Oukhellou, L.: Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transp. Res. Part C: Emerg. Technol. **79**, 274–289 (2017)
3. Arnone, M., Delmastro, T., Giacosa, G., Paoletti, M., Villata, P.: The Potential of e-ticketing for public transport planning: the piedmont region case study. Transp. Res. Procedia **10**, 3–10 (2016)
4. Arnone, M., Delmastro, T., Negrino., Arneodo, F., Botta, C., Friuli, G.: Estimation of public transport user behaviour and trip chains through the piedmont region e-ticketing system. In: Proceedings of 14th ITS European Congress, Lisbon, Portugal, ITS-SP 2273 (2020)
5. Trépanier, M., Tranchant, N., Chapleau, R.: Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst.: Technol. Plan. Oper. **11**, 1–14 (2007)
6. He, L., Trépanier, M.: Estimating the destination of unlinked trips in transit smart card fare data. Transp. Res. Rec.: J. Transp. Res. Board **2535**, 97–104 (2015)
7. Toqué, F., Côme, E., El Mahrsi M.K., Oukhellou, L.: Forecasting dynamic public transport origin-destination matrices with long-short term memory recurrent neural networks. In: IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, pp. 1071–1076 (2016)
8. Zhao, Z., Koutsopoulos, H., Zhao, J.: Individual mobility prediction using transit smart card data. Transp. Res. Part C Emerg. Technol. **10**, 19–34 (2018)

9. Yu, H., Wu, Z., Chen, D., Ma, X.: Probabilistic prediction of bus headway using relevance vector machine regression. IEEE Trans. Intell. Transp. Syst. **18**, 1–10 (2016)
10. Sun, F., Pan, Y., White, J., Dubey, A.: Real-time and predictive analytics for smart public transportation decision support system (2016)
11. Othman, M.S., Tan, G.: Predictive simulation of public transportation using deep learning. In: Proceedings of 18th Asia Simulation Conference, Kyoto, Japan, 27–29 October, pp. 96–106 (2018)
12. Cristóbal, T., Padrón, G., Quesada-Arencibia, A., Hernández, F., de Blasio, G., García, C.: Using data mining to analyze dwell time and nonstop running time in road-based mass transit systems. Proceedings **2**(19), 1217 (2018)
13. Yu, H., Chen, D., Wu, Z., Ma, X., Wang, Y.: Headway-based bus bunching prediction using transit smart card data. Transp. Res. Part C: Emerg. Technol. **72**, 45–59 (2016)
14. Asmael, N., Waheed, M.: Demand estimation of bus as a public transport based on gravity model. MATEC Web Conf. **162**, 01038 (2018)
15. El Mahrsi, M.K., Côme, E., Oukhellou, L., Verleysen, M.: Clustering smart card data for urban mobility analysis. IEEE Trans. Intell. Transp. Syst. **18**(3), 712–728 (2017)
16. Kieu, L.M., Bhaskar, A., Chung, E.: Passenger segmentation using smart card data. IEEE Trans. Intell. Transp. Syst. **16**(3), 1537–1548 (2015)
17. Kim, K.: Identifying the structure of cities by clustering using a new similarity measure based on smart card data. IEEE Trans. Intell. Transp. Syst. **21**(5), 2002–2011 (2020)
18. Liu, L., Chen, R.C.: A novel passenger flow prediction model using deep learning methods. Transp. Res. Part C: Emerg. Technol. **84**, 74–91 (2017)
19. Ding, C., Wang, D., Ma, X., Li, H.: Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. Sustainability **8**, 1100 (2016)
20. Liu, Y., Liu, Z., Jia, R.: DeepPF: A deep learning based architecture for metro passenger flow prediction. Transp. Res. Part C Emerg. Technol. **101**, 18–34 (2019)
21. Guo, J., Xie, Z., Qin, Y., Jia, L., Wang, Y.: Short-term abnormal passenger flow prediction based on the fusion of SVR and LSTM. IEEE Access **7**, 42946–42955 (2019)
22. Milenkovic, M., Svadlenka, L., Melichar, V., Bojovic, N., Avramovic, Z.: SARIMA modelling approach for railway passenger flow forecasting. Transport, 1–8 (2015)
23. Toqué, F., Khouadjia, M., Côme, E., Trépanier, M., Oukhellou, L.: Short and long term forecasting of multimodal transport passenger flows with machine learning methods, pp. 560–566 (2017)
24. Gastaldi, E.: Forecasting public transport demand using smart cards data. https://webthesis.biblio.polito.it/20414/. Accessed 24 July 2023
25. Attili, A.: The demand for public transport: analysis of mobility patterns and bus stops.https://webthesis.biblio.polito.it/17338/. Accessed 24 July 2023
26. 3bmeteo. https://www.3bmeteo.com/. Accessed 30 May 2023