mathematics

MDPI

Article

Clustering and Forecasting Urban Bus Passenger Demand with a Combination of Time Series Models

Irene Mariñas-Collado [1], Ana E. Sipols [2,*], M. Teresa Santos-Martín [3] and Elisa Frutos-Bernal [4]

[1] Department of Statistics and Operation Research and Mathematics Didactics, Universidad de Oviedo, 33007 Oviedo, Spain; marinasirene@uniovi.es
[2] Department of Applied Mathematics, Materials Science and Engineering and Electronic Technology, Rey Juan Carlos University, 28933 Madrid, Spain
[3] Department of Statistics, Institute of Fundamental Physics and Mathematics, Universidad de Salamanca, 37008 Salamanca, Spain; maysam@usal.es
[4] Department of Statistics, Universidad de Salamanca, 37007 Salamanca, Spain; efb@usal.es
* Correspondence: anaelizabeth.garcia@urjc.es

Abstract: The present paper focuses on the analysis of large data sets from public transport networks, more specifically, on how to predict urban bus passenger demand. A series of steps are proposed to ease the understanding of passenger demand. First, given the large number of stops in the bus network, these are divided into clusters and then different models are fitted for a representative of each of the clusters. The aim is to compare and combine the predictions associated with traditional methods, such as exponential smoothing or ARIMA, with machine learning methods, such as support vector machines or artificial neural networks. Moreover, support vector machine predictions are improved by incorporating explanatory variables with temporal structure and moving averages. Finally, through cointegration techniques, the results obtained for the representative of each group are extrapolated to the rest of the series within the same cluster. A case study in the city of Salamanca (Spain) is presented to illustrate the problem.

Keywords: forecasting; time series models; big data; clustering; cointegration; combination

MSC: 62R07; 62H12; 62H30

Citation: Mariñas-Collado, I.; Sipols, A.E.; Santos-Martín, M.T.; Frutos-Bernal, E. Clustering and Forecasting Urban Bus Passenger Demand with a Combination of Time Series Models. *Mathematics* **2022**, *10*, 2670. https://doi.org/10.3390/math10152670

Academic Editor: Christophe Chesneau

Received: 1 July 2022
Accepted: 26 July 2022
Published: 28 July 2022

1. Introduction

Advances in technology have allowed exponential growth in the volume of data that can be collected, especially in the field of transport. Public transport plays a key role in ensuring the movement of passengers within the city and between cities. Amongst them, the bus service is one of the most used means of transport due to its accessibility and low price. Forecasting methods used to make decisions need to be adjusted to the vast amount of information available nowadays.

This paper focuses on the modelling of transport data from the urban bus network in the city of Salamanca (Spain), in order to predict the behaviour of the users to help make decisions about the reform and management of said public service. First, the different bus stops are grouped into clusters; then, various prediction models are fitted, and their predictions are combined. Finally, cointegration techniques are used to study similar behaviour within each group.

Clustering is an essential tool for analysing big data. Shirkhorshidi et al. [1] reviewed the trend and progress of clustering algorithms to face the challenges of big data since the first proposed algorithms. Maharaj et al. [2] provides an overview of time series clustering and classification methods.

The combination of predictions assumes that the underlying process that explains a phenomenon cannot be identified by a single model. Each model may capture different aspects of the information, which lead to different predictions. Therefore, it may be desirable

to merge multiple forecasting methods to improve the precision of each prediction. There are different methods to combine predictions and the choice depends on the characteristics of the data and the degree of precision resulting from the adjustments [3].

The use of the public bus varies according to many variables of time and space, such as the day of the week, holidays, seasons, business centres, workplaces, residential areas and other factors such as weather. A number of methods have been developed in the literature for this type of analysis, most using clustering approaches [4]. There are two main approaches when analysing public transport passengers flow. On one hand, the stops can be grouped according to the temporal-spatial distribution characteristics of the passengers [5]. On the other hand, groups of passengers with similar boarding times along the week can be identified [6]. The k-means algorithm and hierarchical cluster analysis have been the most widely used methods. Wang et al. [7], Kim et al. [8] and Ding et al. [9] used gradient boosting decision trees. Hierarchical cluster analysis of passenger hourly entries is used in [10] to study the common characteristics of stations, whilst in [11] this was done using Tucker's decomposition.

Among the studies with bus transport data are [12], which uses Holt–Winters multiplicative models with data from Kerala (India) and [13], which proposed a hierarchical hybrid model based on different models of time series on the buses in Dalian, China. Comi and Polimeni [14] presented an approach to forecast travel time based on time series, using data from automated vehicle monitoring of bus lines sharing lanes with other vehicles in Rome (Italy) and Lviv (Ukraine). Ye et al. [15] proposed autoregressive models for forecasting data collected from bus cards.

In [16], ARIMA and artificial neural networks models were used for passenger flow of transit buses forecasting.

Cointegration techniques allow two series to be fitted at once using the same model if they share a common stochastic trend. Introduced in Engle and Granger [17], they are of great use in econometrics to measure relationships between economic variables. In the literature, works relating economic and environmental indices with the use of transport can be found, see, for example [18,19].

The main aim of this paper is to analyse the most commonly used time series models and improve their predictions when applied to transport data, more specifically to data from the Salamanca bus network and which can be extrapolated to any other city with similar characteristics, i.e. with no complementary metro network or other type of public transport network. Furthermore, given the temporal characteristics of the bus data, an improvement of the support vector machine is incorporated, using explanatory variables with temporal structure and moving averages to improve predictions.

The paper is organised as follows: Section 2 presents the data and the situation of the buses in Salamanca, in Section 3 the applied methodology is introduced: clustering, models and combinations of their predictions and cointegration techniques. In Section 4, the results from applying the steps proposed to the data from Salamanca are presented. Furthermore, in this section, it can be seen how the modification of the SVM method results in the predictions with the smallest errors. Using a representative and the cointegration techniques, instead of having to work with each series individually, saves computational time. Finally, Section 5 discusses the main conclusions and further lines of investigation.

2. Study Area

The city of Salamanca is located in western Spain and is the capital of the province of Salamanca in the autonomous community of Castile and León. It is close to the border with Portugal and just a couple of hours from the capital of Spain, Madrid. The province of Salamanca has 362 municipalities, 17 of which are less than 10 km from the city and about 30 are between 10 and 20 km, which make these municipalities dormitory populations for people who work in the city, who usually leave their vehicles on the outskirts and use urban transport to get around. The city has approximately 150,000 registered inhabitants and the main industries (apart from the university) are the service sector and agriculture.

Salamanca is known for being a university city. More than 30,000 people, which represents almost 20% of the population, are students and a large part of the inhabitants are directly or indirectly related to the university. Today, in addition to being a famous university city, Salamanca is a city that holds numerous international congresses and important cultural events. It is a UNESCO World Heritage City and, in 2002, it was named the European Capital of Culture. Furthermore, it is a popular destination for foreigners who want to learn Spanish.

In recent years, with the rapid development of intelligent transportation, the number of passengers taking a bus can be obtained through onboard instruments. This way, the number of boardings taken place at a bus stop at a certain hour can be obtained by adding up the entries at each bus. The data here studied consists of records for 272 bus stops where the hourly number of passengers is recorded. This paper focuses on the data from two consecutive weeks in May 2019, where the working hours of the different bus lines are from 7 a.m. to 11 p.m. Therefore, there are 17 daily entries for 14 days. A prior analysis of the data was performed, eliminating the stops whose average of passengers in the two weeks was less than 1, mainly corresponding to the last stops of the different lines in each direction.

Descriptive Data Analysis

The data provided by the transport company are the number of passengers at each of the different bus stops. When studying the daily total number of boardings, a great difference can be appreciated in those stops which are in the city centre (for example, Stop 4, where 6 out of the 13 bus lines pass through and can reach about 1600 passengers per day) and those in the areas surrounding the city, that are usually the last stops of the lines, and have barely any boardings.

Figure 1 shows the aggregated passenger count for each day of the week. A similar pattern can be observed in both weeks, as well as the (dis)similarities between and within weekdays and weekends. Figure 2 shows the boarding throughout the day for each day of the week for each week. It can be observed that both weeks have the same pattern, peak hours coincide every weekday (8 a.m. and 2 p.m.) and differ from Saturdays and Sundays, which are similar to each other.

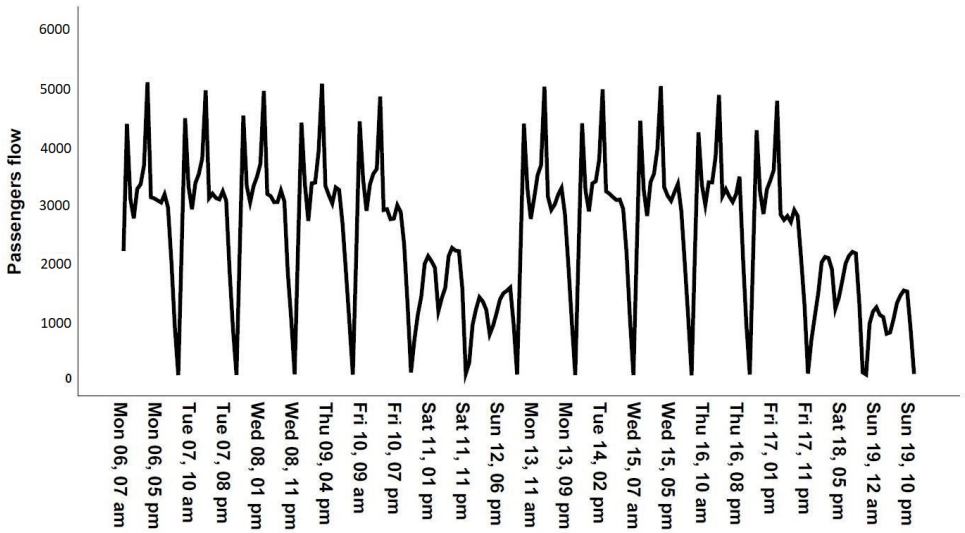

Figure 1. Series for the 14 days, 17 h by day.

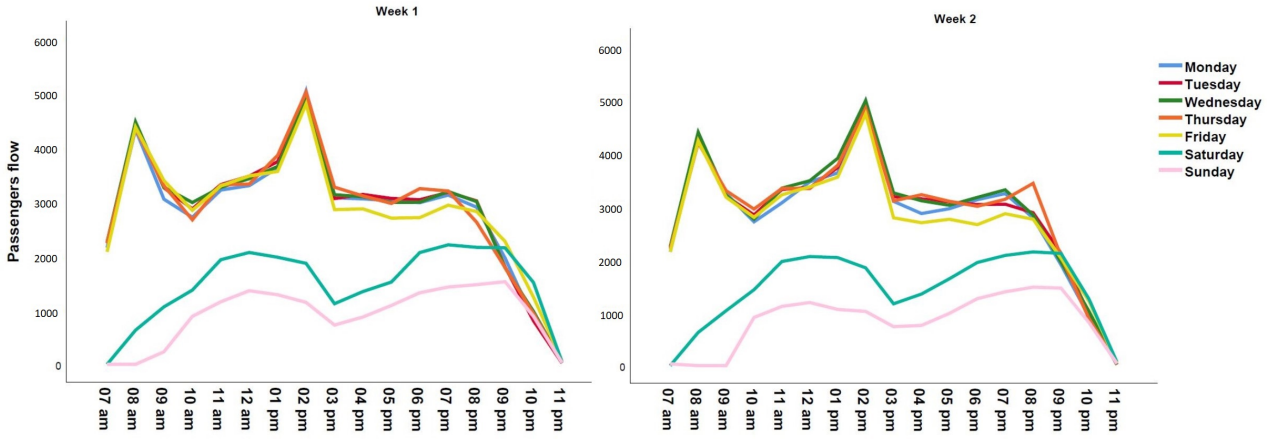

Figure 2. Time-varying diagram of passenger flow by weeks.

In Figure 3, the daily boxplots, where the differences in passengers between weekdays and weekends are appreciated, are shown. This decrease in the number of passengers is what causes the frequency of buses to be lower on weekends. Asymmetry can be observed on Thursdays and Fridays, which is caused by the university nature of the city of Salamanca, since many schools do not have classes on Fridays, causing a weekend eve effect on both days. In addition, many companies on Fridays work intensive hours only in the morning, which also affects the use of the bus.

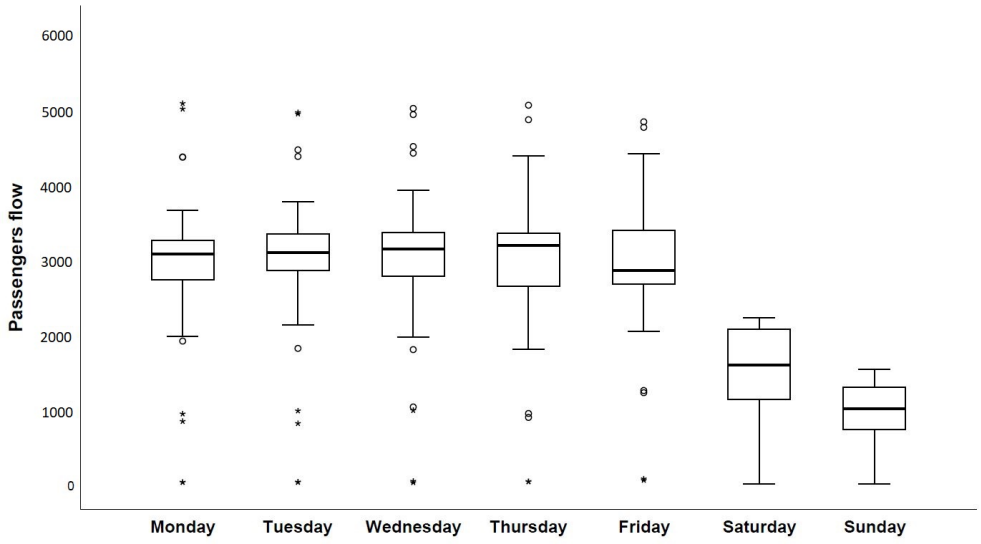

Figure 3. Passenger flow boxplots by day.

3. Methodology

Clustering is an unsupervised learning task that aims to divide a data set into homogeneous groups or clusters. The partition is done in such a way that the elements in the same group are more similar to each other than the elements in different groups according to some defined criterion, which marks the measure of similarity.

Clustering techniques are divided according to whether the number of partitions to be created is known in advance (partition clustering) or if the number of clusters is not known, but observations are grouped according to their similarity to a structure hierarchical (hierarchical clustering). Moreover, clustering methods require a metric that defines the distance, either similarity or dissimilarity, between the observations. Selecting an appropriate distance measure is a key aspect of the clustering process. In the specific context of time series data, the concept of dissimilarity is particularly complex due to the dynamic nature

of the series. Differences that are generally considered in the conventional grouping cannot work well with time-dependent data because they ignore the interdependence relationship between values.

The first important question is to decide whether grouping should be governed by a 'form-based' or a 'structure-based' concept of dissimilarity [20,21]. In the context of time series, establishing what makes two objects to be considered 'similar', i.e., that should belong in the same cluster, is particularly complex due to the dynamic character of the series. Dissimilarities usually considered in conventional clustering could not work adequately with time-dependent data because they ignore the interdependence relationship between values. Several authors have considered distance measures based on the estimated autocorrelation functions (see e.g., [22–24]).

Amongst the different clustering techniques, the hierarchical cluster is performed. To select the optimal number of clusters k, different methods are compared. A simple and popular solution consists of inspecting the dendrogram produced to see if it suggests a particular number of clusters. However, this approach is very subjective. Fortunately, there are several indices and methods that have been published for identifying the optimal number of clusters. This method is well summarised in Charrad et al. [25]. In this study, the elbow method, which looks at the total within-cluster sum of square (WSS) as a function of the number of clusters, is also looked at.

Once the different clusters are defined, a representative is chosen for each of the clusters randomly among those stops with the largest number of boardings (therefore, the most used stops) and different models are fitted:

- *Holt–Winters seasonal exponential smoothing*. Holt [26] and Winters [27] extended Holt's method to capture seasonality. The Holt–Winters seasonal method comprises the forecast equation and three smoothing equations and is used for forecasting time series data that exhibits both a trend and a seasonal variation. The unknown parameters are determined by minimising the squared prediction error. More details can be found, for example, in [28–30].
- *The Arima model or Box–Jenkins method*. Introduced by Box et al. [31], this method focuses on the autocorrelation between the observations, describing each value as a linear function of previous data and errors due to chance, being able to include a cyclical or seasonal component. The acronym ARIMA stands for auto-regressive integrated moving average and its a generalisation of an auto-regressive moving average (ARMA) model.
- *The K-nearest Neighbours (KNN) method*. KNN is a very popular algorithm used in classification and regression. This algorithm stores a collection of examples. Each example consists of a vector of features that describe the example and, in our case, its numeric value (for prediction). Given a new example, KNN finds its k most similar examples, called nearest neighbours, according to a distance metric and predicts its value as an aggregation of the target values associated with its nearest neighbours. The multiple input multiple output (MIMO) strategy to forecast multiple steps ahead, commonly applied with KNN, with $k = 2$, is used.
- *Autoregressive neural networks (ARNN)*. This method is based on a combination of the multilayer perceptron method with an autoregressive linear model. For time series data the lagged (autoregressive) values of the time series are used as inputs to a neural network. The objective is then to determine how many lags to include in the input layer and how many neurons to include in the hidden layer to produce a forecast that minimises the error. The ARNN is trained to make use of the R Package developed by Velásquez et al. [32].
- *Support vector machines (SVM)* are a type of neural network that can be used for prediction in time series. Parameter estimation is done by minimising a risk function where the empirical error between the model and the data and a regularisation component that depends only on the weights is measured. In this work, a modification of the SVM procedure is presented, in which explanatory variables are incorporated to contribute

to the accuracy of both the fit and the prediction. Without this modification, the SVM model does not capture the temporal dynamics of the data (hours, days, weeks, ...). First, variables to represent the hour and the day of the week are constructed by means of indicator variables (dummies). In addition, autoregressive variables and lags smoothed by means of a moving average are included to capture the dynamics of the series more accurately.

- *Exponential smoothing state space model with Box–Cox transformation, ARMA errors, trend and seasonal components (TBATS)*. TBATS is an acronym for key features of the model: *T*: trigonometric seasonality; *B*: Box–Cox transformation; *A*: ARIMA errors; *T*: trend; *S*: seasonal components. The main aim of this model is to forecast time series with complex seasonal patterns using exponential smoothing. The trigonometric seasonality expression can significantly reduce model parameters at high seasonality frequencies and at the same time offer the model plasticity to compromise with complex seasonality [33].

Once the models have been fitted by the different methods described above, to choose the most accurate one, the estimation errors (difference between the observed value y_i and the predicted value $\hat{y}_i$) are analysed through the following measurements of precision, mean squared error (MSE) and mean absolute error (MAE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

To assess the stability of the model over time and its forecast accuracy, a rolling-window analysis of the models is done. For this, first, a rolling window size, m, is chosen, i.e., the number of consecutive observations per rolling window. In this case, we are working with hourly boardings, with 17 h a day, for 2 weeks: 238 data points. A window of 5 days (85 h), is chosen. Then, the forecast horizon, h, is set to be 1 day (17 h). The number of increments between successive rolling windows is also chosen to be 1 day. Then, for each rolling window sub-sample, the model is fitted over the m historical data and the h-step-ahead forecast is done. Finally, the forecast errors, MSE and MAE, for all the predictions through the different moving windows are calculated. The MAEs and MSEs among the models are compared and the model with the lowest set of errors has the best predictive performance.

Once the best prediction models have been chosen, the combination of predictions will be used for the final model, combining the different forecasts obtained from each model into one, providing the information collected by each of the models individually to the combined model [34]. There are many different ways to perform the combination of models, such as the arithmetic mean of the predictions obtained by the individual methods, the weighted average based on variances where the weights are obtained based on the error variance of the predictions [35] or a weighted mean based on regression where the weights are obtained by a regression model, for which there is a method that was first proposed by Granger and Ramanathan [36], amongst others.

After the predictions for the representatives have been made, if the series within each cluster are cointegrated, the results obtained for the representatives can be used to adjust and predict the behaviour of the rest of the cluster stops by cointegration. Two series are said to be cointegrated if they move together in time and the differences between them are stable. The cointegration tests of Johansen [37] and Johansen et al. [38] allow to test the cointegration between series. In this work, the trace test will be used. To estimate the cointegration relationship, linear regression is adjusted for the cointegrated series, evaluating the stationarity of the residuals. In this way, the settings for all stops can be obtained using the information provided by the representative, without having to adjust the models to each of the series.

4. Results

The results shown below have been obtained using *R* [39], *EViews 10* [40] and *IBM SPSS 26* [41].

4.1. Clustering Analysis

First of all, the series are standardised. Centring is done by subtracting the series means and then scaling is done by dividing the (centred) series by their standard deviations.

To calculate the clusters, autocorrelation-based dissimilarity is used. This performs the weighted Euclidean distance between the simple autocorrelation coefficients. The total within-cluster sum of square as a function of the number of clusters is shown in Figure 4, pointing to 2 clusters as the optimal solution. The periodogram-based distance was also explored, pointing to the same results while being much more computationally expensive. The dendrogram in Figure 5 shows that two clustering solutions are possible. The four-cluster solution is chosen as it provides a more detailed segmentation of the stations. A representative of each of the clusters is chosen, based on those that present larger variability in the number of passengers. It should be noted that the stops are grouped according to time and location, with the different lines that operate through them not being particularly relevant. The different lines of the bus network start from the peripheral areas and cross the city through the centre. The most important aspect of the network, for this study, is the number of passengers per stop, so as to perform the appropriate modifications.

There are 27 stops in Cluster 1. Its representative is Stop 2, a stop located in a peripheral area, through which 3 lines pass. In this cluster, there are mainly stops in peripheral neighbourhoods, where the peaks of boardings correspond to the start times of school, first thing in the morning. Few lines pass through these stops (1 or 2 maximum). Compared with the series from the rest of the clusters, they are stops with fewer passengers. In cluster 2, there are 79 stops. Its representative, Stop 33, is a stop that borders the pedestrianised old town of the city and 6 lines pass through it. These are stops whose main use corresponds to leaving work and schools to return to the suburban residential neighbourhoods. For Cluster 3, with 44 stops, the representative is Stop 6. In this group, there are stops that are on large avenues. These are stops through which more than two lines operate, and they may serve as transfer stops on the way back home. Weekends, on the other hand, have little movement compared to weekdays. The 93 stops in Cluster 4 are represented by Stop 309, a stop located in a non-central area. These are stops farther from the historical centre, without becoming peripheral neighbourhoods. The largest number of passengers is concentrated mainly in the first hours of the day. More than two lines pass through most of the stops.

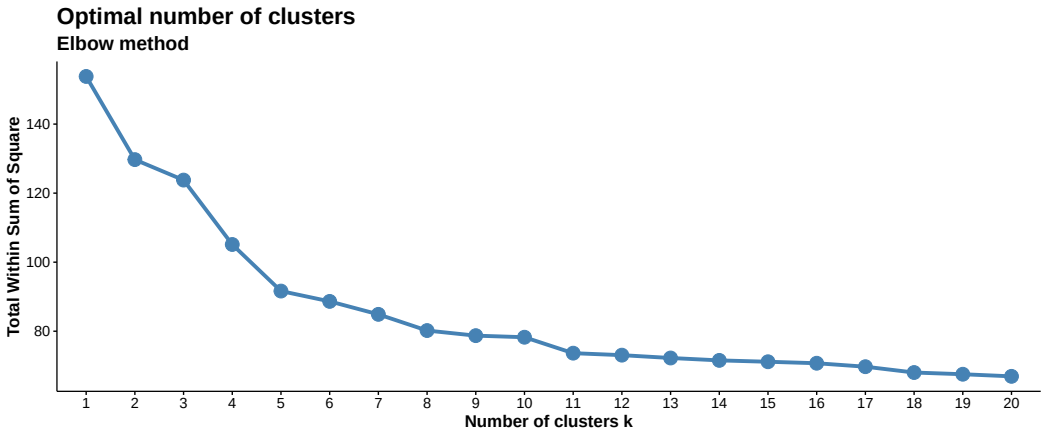

Figure 4. Graphical representation of elbow method to determine the optimal number of clusters, using the ACF distance.

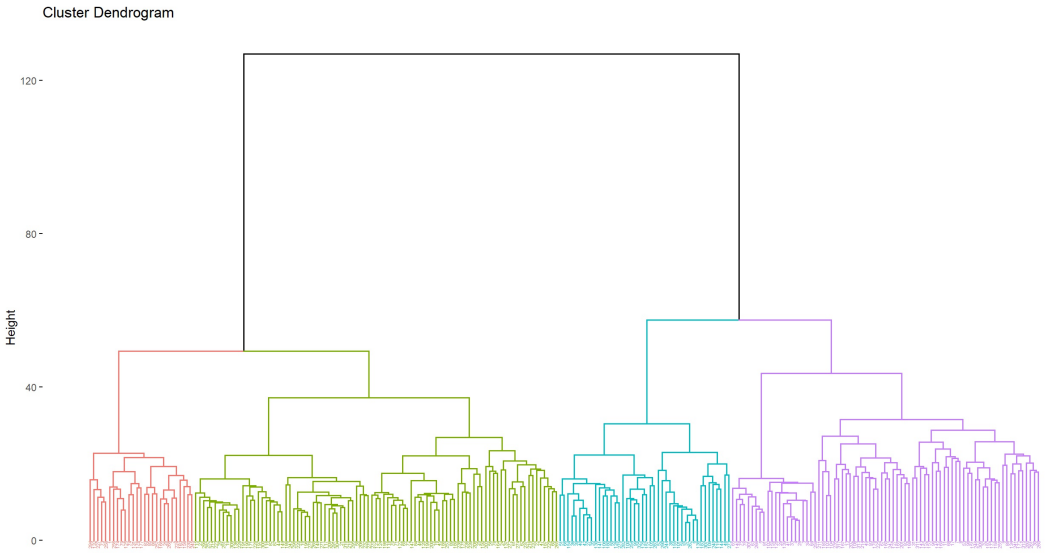

Figure 5. Hierarchical clustering with ACF distance dendrogram.

Figure 6 shows the four representatives chosen. Although some patterns may seem similar, the differences in the y axes must also be taken into account. Clusters 2 and 3 representatives, for example, have far more boardings than the other two.

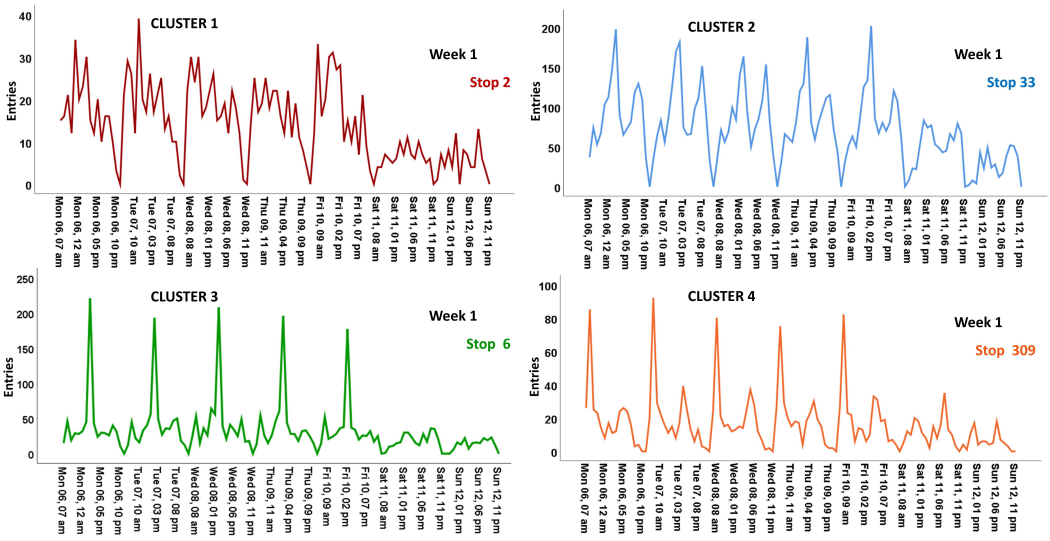

Figure 6. Representatives for each cluster.

4.2. Forecasting Ridership Patterns

Figure 7 shows the MAEs for the different models in each cluster's representative. The overall MAEs are shown in Table 1. SVM and TBATS are always best. In particular, SVM, which has the explanatory variables previously described, is always the one with the smallest MAE. The third best is between ARIMA and Holt–Winters (H-W), the more traditional methods.

Table 1. Overall MAE and MSE for each model in each cluster.

		ARIMA	H–W	KNN	ARNN	SVM	TBATS
CLUSTER 1	MAE	7.77	5.73	7.83	8.51	4.85	5.34
	MSE	98.90	62.18	100.30	117.22	42.68	49.23
CLUSTER 2	MAE	32.02	35.34	39.39	42.54	17.21	25.02
	MSE	2047.28	2325.61	3303.37	3365.08	475.58	1190.70
CLUSTER 3	MAE	14.46	15	18.68	24.18	9.72	13.58
	MSE	882.53	1019.44	1213.24	1749.24	167.32	598.37
CLUSTER 4	MAE	9.29	8.71	11.46	14.11	5.81	8.57
	MSE	267.34	240.89	364.13	424.30	50.99	182.10

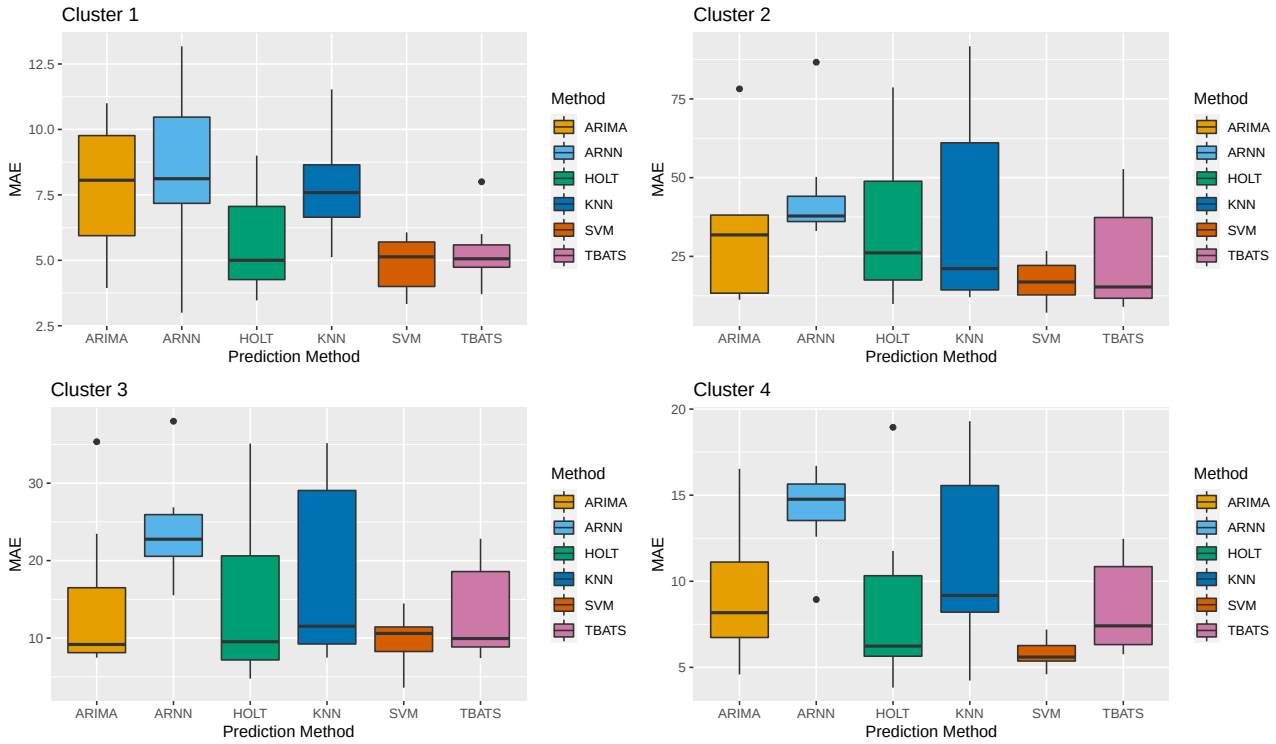

Figure 7. Boxplot MAE cluster.

4.2.1. Predictions and Combinations by Cluster

Once the three best models are chosen, the combination of predictions is carried out using the arithmetic mean (AM), the Bates and Granger weighted mean (B&G) and the weighted mean based on constrained least squared (CLS) regression. For the latter, the variant of the method implemented adds the restriction that combination weights must be non-negative and is combined with the condition of forcing the weights to sum up to one. To illustrate the combinations, a week is chosen from Friday to Thursday and the following Friday is predicted.

Cluster 1

Figure 8 shows the forecasts from each model and the real values for Stop 2, the representative of Cluster 1. The predictions are shown together with the real number of boardings, as well as the last three previous days. It should be noted that in this cluster, the number of boardings is significantly smaller than in other clusters. While Stop 33, for example, reaches a maximum of almost 200 passengers, in Stop 2 the maximum does not reach 50. The MAEs and MSEs are shown in Table 2, together with the errors for the combination of the best three models. For Cluster 1, the best three models are SVM, TBATS

and Holt–Winters. The best combination, in this case, is the Bates and Granger weighted mean (B&G). Figure 9 shows the real data, the best model and the best combination.

Table 2. MAEs and MSEs for each model in Cluster 1 for one-week historical and one-day predictions and for the different combinations of the best three models.

Models	ARIMA	H-W	KNN	ARNN	SVM	TBATS
MAE	5.47	4.12	6.29	10.88	3.33	4.41
MSE	57.24	38.59	70.65	206.29	22.80	41.47
Best 3 Comb.	AM	B&G	CLS			
MAE	3.33	3.13	3.33			
MSE	30.40	28.20	22.80			

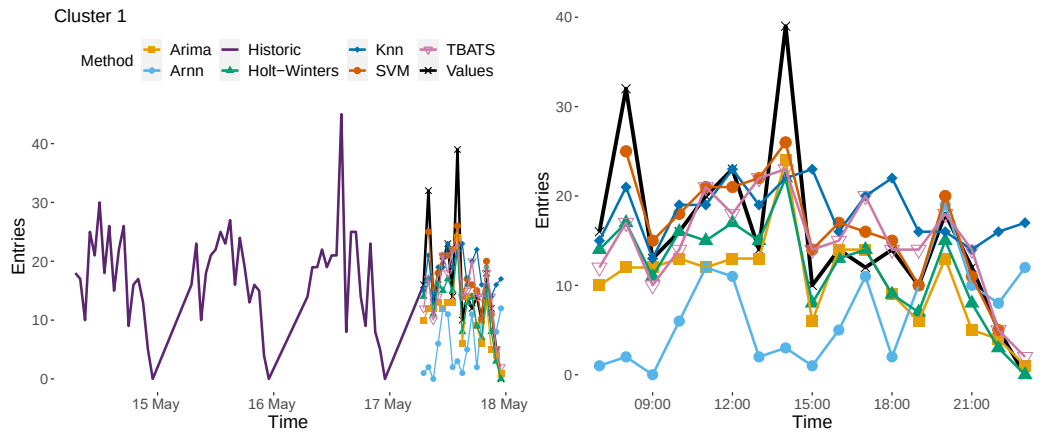

Figure 8. Cluster 1: two last days of historical data plus one-day predictions and real values (**left**). On the (**right**), a close-up of the predictions and real values is shown.

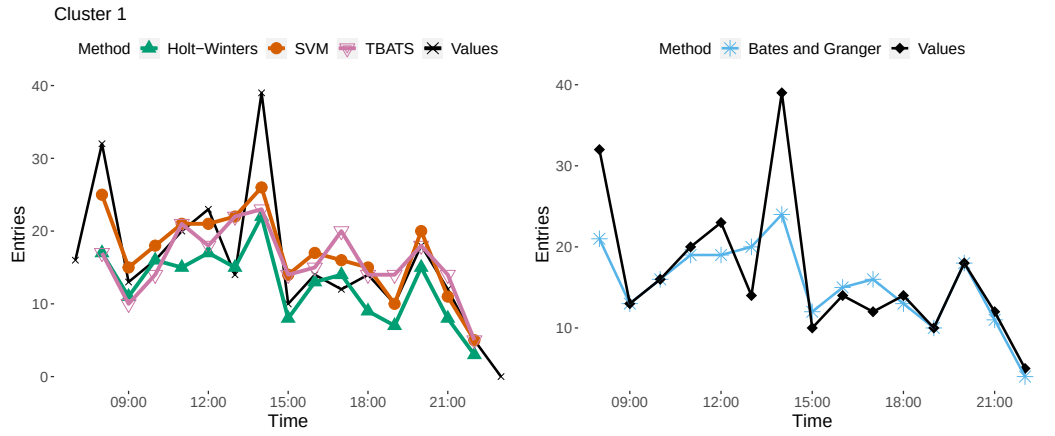

Figure 9. Cluster 1 best 3 models (**left**) and final combined predictions (**right**).

Cluster 2

For the representative in Cluster 2 (Stop 33), the forecasts from each model are shown in Figure 10. The predictions are shown together with the real number of boardings, as well as the last three previous days. The MAEs and MSEs are shown in Table 3, together with the errors for the combination of the best three models. In this case, it can be seen that the best models are SVM and TBATS, as they are in every cluster, and the third best model is ARIMA. Therefore, these are the three models combined.

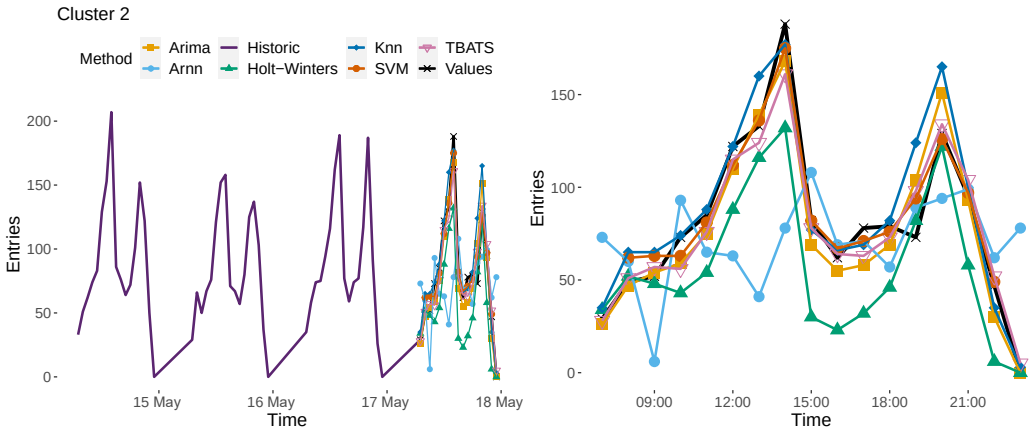

Figure 10. Cluster 2: three last days of historical data plus one-day predictions and real values (**left**). On the (**right**), a close-up of the predictions and real values is shown.

Table 3. MAEs and MSEs for each model in Cluster 2 for one-week historical and one-day predictions and for the different combinations of the best three models.

Models	ARIMA	H-W	KNN	ARNN	SVM	TBATS
MAE	11.24	26.12	12.06	30.47	7.13	9
MSE	197	1006	325.12	1694	81.67	136.53
Best 3 Comb.	AM	B&G	CLS			
MAE	8.53	7.86	7.13			
MSE	122.53	103.07	81.67			

The best combination is constraint least squares (CLS), which is actually setting all the weights to select the predictions from SVM, which was the best model. Figure 11 shows the real data, the three chosen models and the best predictions.

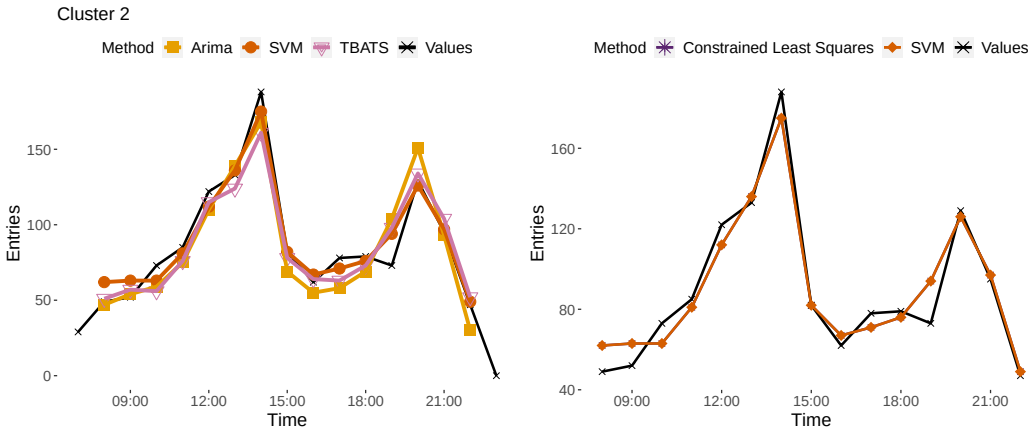

Figure 11. Cluster 2 best 3 models (**left**) and final combined predictions (**right**).

Cluster 3

Stop 6 is the representative for Cluster 3. In this cluster, the maximum number of boarding is over 200, as can be seen in Figure 12. Table 4 shows the MAEs and MSEs for this cluster. The improved SVM is again the best, followed by Holt–Winters and TBATS. The best combination is the same as in Cluster 1, the Bates and Granger weighted mean, shown in Figure 13. In this case, the combination is almost the same as the SVM predictions, just slightly improved.

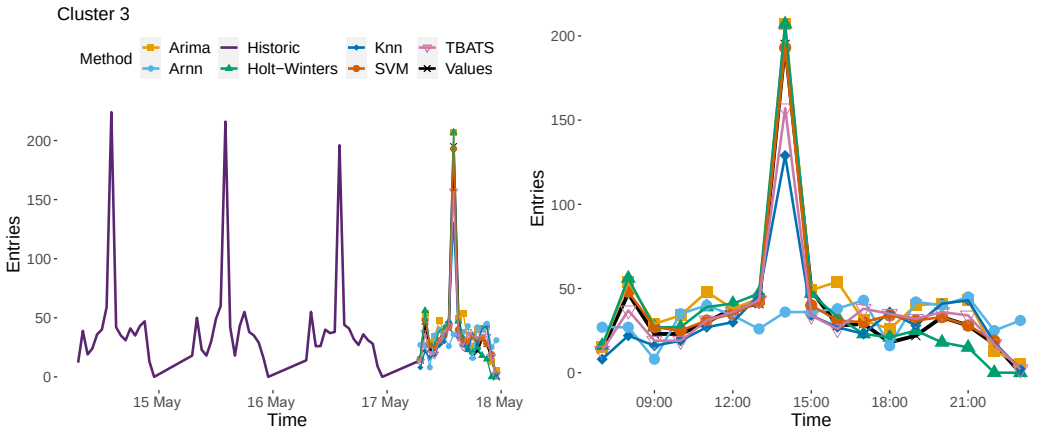

Figure 12. Cluster 3: two last days of historical data plus one-day predictions and real values (**left**). On the (**right**), a close-up of the predictions and real values is shown.

Table 4. MAEs and MSEs for each model in Cluster 3 for one-week historical and one-day predictions, and for the different combinations of the best models.

Models	ARIMA	H-W	KNN	ARNN	SVM	TBATS
MAE	8.47	6.65	11.54	21.76	3.60	7.41
MSE	122.47	67.83	358.59	1693.18	31.20	137.29
Best 3 Comb.	AM	B&G	CLS			
MAE	4.07	3.60	3.87			
MSE	28.87	24.87	23.47			

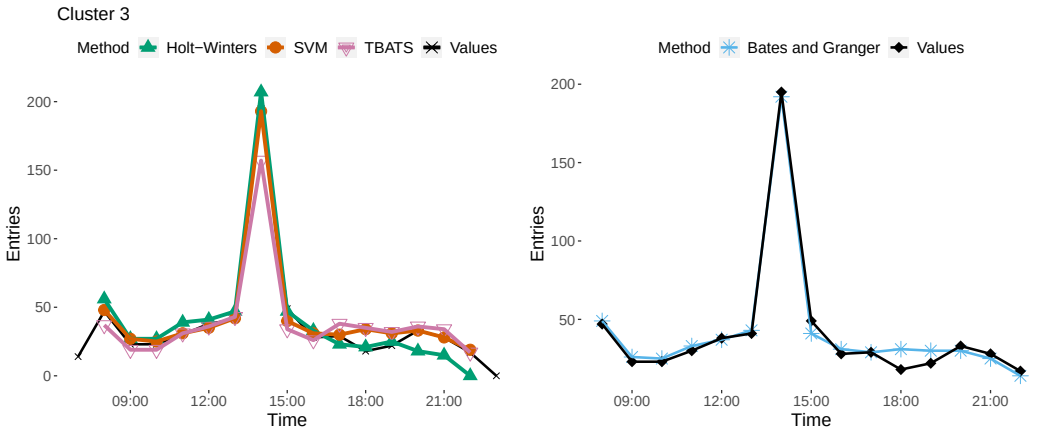

Figure 13. Cluster 3 best 3 models (**left**) and final combined predictions (**right**).

Cluster 4

The representative in Cluster 4 is Stop 309, which has a smaller number of boardings than those in clusters 2 and 3 but reaches almost twice as many passengers as Cluster 1. The forecasts from each model are shown in Figure 14, where, as above, the predictions are shown together with the real number of boardings, as well as the last two previous days of the historical data.

The MAEs and MSEs are shown in Table 5, together with the errors for the combination of the best three models, which are SVM, TBATS and Holt–Winters again. The best combination is, again, the weighted average of Bates and Granger. The best three models and final best predictions are shown in Figure 15.

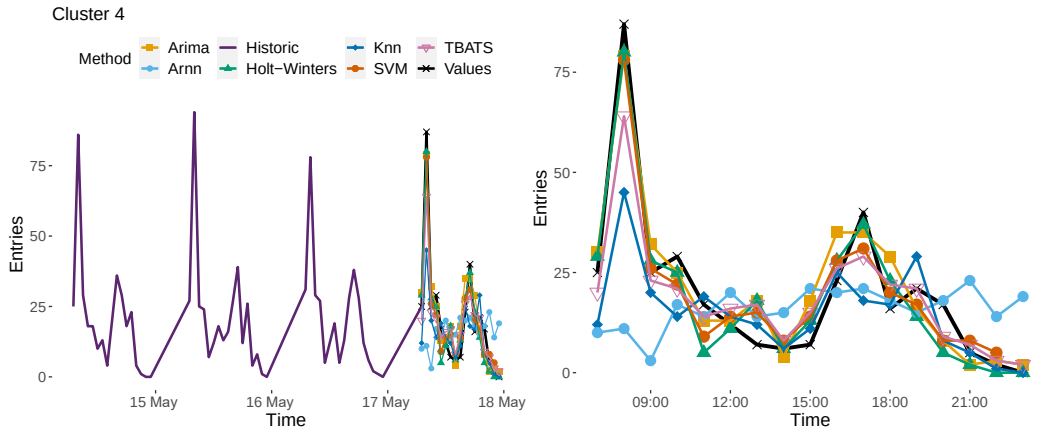

Figure 14. Cluster 4: three last days of historical data plus one-day predictions and real values (**left**). On the (**right**), a close-up of the predictions and real values is shown.

Table 5. MAEs and MSEs for each model in Cluster 4 for one-week historical and one-day predictions, and for the different combinations of the best three models.

Models	ARIMA	H-W	KNN	ARNN	SVM	TBATS
MAE	6.06	5.12	7.71	14.47	5.40	5.76
MSE	51.47	40.06	168.65	486.35	36.87	62
Best 3 Comb.	AM	B&G	CLS			
MAE	5.46	5.26	5.40			
MSE	43.87	40.06	38.70			

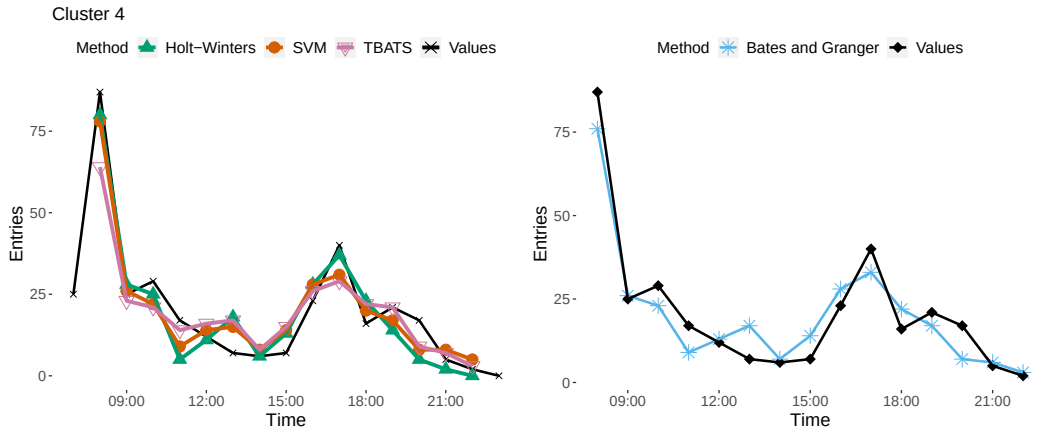

Figure 15. Cluster 3 best 4 models (**left**) and final combined predictions (**right**).

4.3. Cointegration Study

The Johansen trace test [37], with a 5% level, reflects the existence of cointegration relationships between all the data from the stops belonging to the same cluster with its representative, denoted $R1$, $R2$, $R3$ and $R4$, respectively. Therefore, it is not necessary to repeat the analysis shown in Section 4.2 for every series in each cluster, since the predictions of each one of them can be made using the cointegration equations. Table 6 shows a summary of the two stops of each cluster with the highest determination coefficient R^2. It should be noted that the minimum R^2 is still greater than 70% in all cases. Furthermore, the regression residuals are stationary, indicating the goodness of the fits. Figure 16 shows the fitted values of the stop with the highest R^2 in each cluster with respect to its representative and the residuals.

Table 6. Cointegration equations for two stops in each cluster and their representative.

	Cointegration Equation	R^2
CLUSTER 1	Stop(73) = 13.48 + 1.54$R1$	0.8354
	Stop(268) = 6.76 + 1.55$R1$	0.8619
CLUSTER 2	Stop(128) = 4 + 0.94$R2$	0.7434
	Stop(91) = 1.46 + 0.36$R2$	0.8821
CLUSTER 3	Stop(101) = -0.64 + 0.05$R3$	0.8854
	Stop(138) = -1.84 + 0.19$R3$	0.9123
CLUSTER 4	Stop(116) = 2.47 + 0.46$R4$	0.7000
	Stop(136) = 1.11 + 0.39$R4$	0.7718

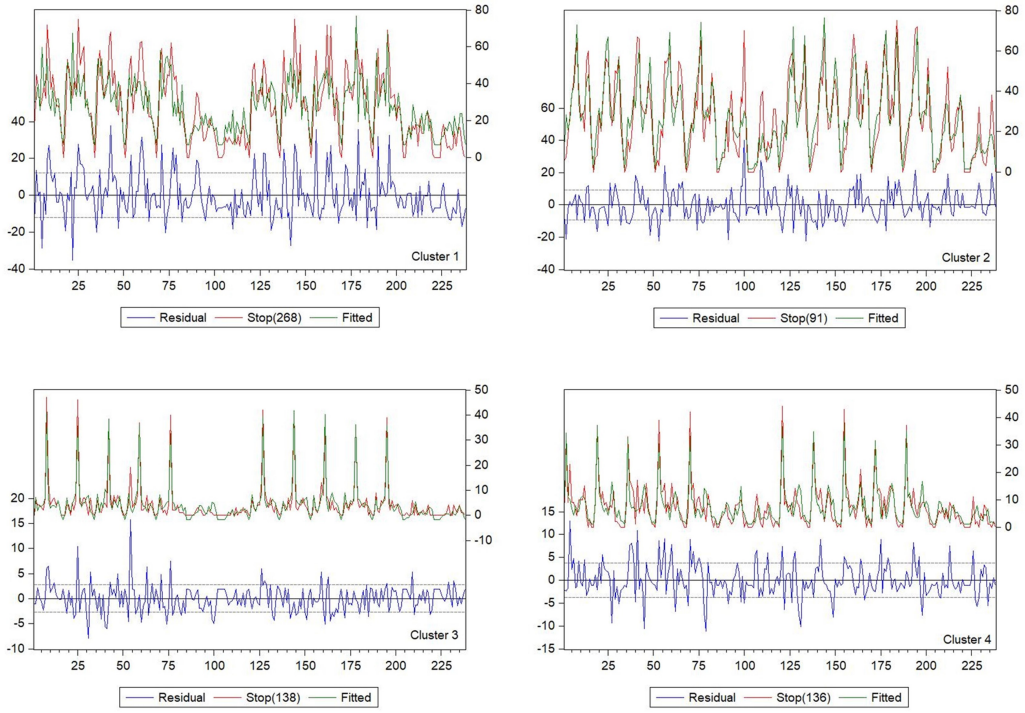

Figure 16. Fitted values from cointegration equations for one stop in each cluster, together with the real values and the residuals of the model.

5. Conclusions

The bus stops from the city of Salamanca (Spain) have been grouped attending to passenger demand and location. The clustering analysis results in four large groups with 27, 79, 44 and 93 stops each, respectively. The stops in each cluster have their own characteristics, as can be seen in Figure 6. The flow of passengers is determined not only by the location of the stop but also by the time slot, which is a true reflection of the daily activity of the city. Different models and methods have been applied to study the hourly passenger demand. The models used allow for robust predictions of passenger data on the bus network. Moreover, the combination of forecasts from conceptually different models (machine learning and traditional methods) effectively reduces prediction errors and, therefore, provides an improvement in accuracy. Finally, for the rest of the stops in each cluster, instead of repeating the whole process of forecasting, the cointegration equations calculated can be used.

The modification performed to the SVM method, with the incorporation of time dummies combined with autoregressive and moving averages, shows that SVM provides the best fitting model, independently of the slightly different pattern that each cluster may have, followed by more traditional methods such as Holt–Winters exponential smoothing.

When it comes to the combination of different predictions, the weighted means, specifically Bates and Granger, have been shown to reduce the errors better than the simpler arithmetic mean, although the differences are not very large. In this case, it is also clear that the weights are all in favour of the values predicted by SVM, which are already accurate before the combination. Future research may include the exploration of other different combinations.

The methodology used, and the results obtained provide valuable information regarding the restructuring of the transport network in the city, which is immersed in a process of change with the opening of the new hospital and the expansion of the peripheral neighbourhoods. The approach proposed not only categorises the bus network's stops but also enhances hourly predictions of the number of passengers. With this data, the frequency of buses may be increased at times when there is a high influx of users, routes can be modified, extended, etc. Knowing the behaviour of the passengers helps make decisions such as the modification of current stops or the suspension of those with low user counts.

The production of large volumes of massive data, big data, opens interesting possibilities to understand the mobility flows of our cities. The proposed steps (clustering, choosing a representative, combinations of predictions, and cointegration techniques) ease the understanding of passenger demand in bus networks and can be extrapolated to other cities where the bus network is the only public transport route too. Future lines of research include completing the analysis by taking into account the different bus lines that pass through each stop, which would increase the complexity since it would multiply the number of series. Moreover, it could be interesting to compare these results (pre-pandemic) and those after the social distancing measures have been relaxed, to evaluate whether the use of public transportation is back to normal after the pandemic.

Author Contributions: The authors confirm their contribution to the paper as follows: conceptualization, M.T.S.-M., A.E.S., I.M.-C. and E.F.-B.; methodology, M.T.S.-M., A.E.S. and I.M.-C.; formal analysis, I.M.-C.; draft manuscript preparation: I.M.-C., M.T.S.-M., A.E.S. and E.F.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Spanish MINECO project PGC2018-098623-B-I00, the Castilla y León Government project SA105P20 and by the Agencia Estatal de Investigación (PID2019-108311GB-I00/AEI/10.13039/501100011033).

Acknowledgments: The authors extend their gratitude to Salamanca de Transportes, S.A, and also to the Area of Development, Urban Planning, Citizen Protection, Traffic and Transport of the Salamanca Town Hall, represented by J.F. Carabias Acosta.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shirkhorshidi, A.S.; Aghabozorgi, S.; Wah, T.Y.; Herawan, T. Big Data Clustering: A Review. In Proceedings of the 14th International Conference on Computational Science and Its Applications—ICCSA 2014, Guimarães, Portugal, 30 June–3 July 2014; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8583.
2. Maharaj, E.A.; D'Urso, P.; Caiado, J. *Time Series Clustering and Classification*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019.
3. De Menezes, L.M.; Bunn, D.W.; Taylor, J.W. Review of guidelines for the use of combined forecasts. *Eur. J. Oper. Res.* **2000**, *120*, 190–204. [CrossRef]
4. 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.* **2017**, *79*, 274–289. [CrossRef]
5. Chen, C.; Chen, J.; Barry, J. Diurnal pattern of transit ridership: A case study of the New York City subway system. *J. Transp. Geogr.* **2009**, *17*, 176–186. [CrossRef]
6. El Mahrsi, M.K.; Come, E.; Oukhellou, L.; Verleysen, M. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 712–728. [CrossRef]
7. Wang, W.; Lo, S.; Liu, S. Aggregated metro trip patterns in urban areas of Hong Kong: Evidence from automatic fare collection records. *J. Urban Plan. Dev.* **2015**, *141*, 05014018. [CrossRef]
8. Kim, M.K.; Kim, S.P.; Heo, J.; Sohn, H.G. Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area. *KSCE J. Civ. Eng.* **2017**, *21*, 964–975. [CrossRef]

9. Ding, C.; Cao, X.; Liu, C. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. *J. Transp. Geogr.* **2019**, *77*, 70–78. [CrossRef]
10. Mariñas-Collado, I.; Frutos-Bernal, E.; Santos-Martin, M.T.; del Rey, A.M.; Casado-Vara, R.; Gil-González, A.B. A Mathematical Study of Barcelona Metro Network. *Electronics* **2021**, *10*, 557. [CrossRef]
11. Frutos-Bernal, E.; Martín del Rey, Á.; Mariñas-Collado, I.; Santos-Martín, M.T. An Analysis of Travel Patterns in Barcelona Metro Using Tucker3 Decomposition. *Mathematics* **2022**, *10*, 1122. [CrossRef]
12. Cyril, A.; Mulangi, R.H.; George, V. Modelling and forecasting bus passenger demand using time series method. In Proceedings of the 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 29–31 August 2018; IEEE: New York, NY, USA, 2018; pp. 460–466.
13. Zhai, H.; Tian, R.; Cui, L.; Xu, X.; Zhang, W. A novel hierarchical hybrid model for short-term bus passenger flow forecasting. *J. Adv. Transp.* **2020**, *2020*, 7917353. [CrossRef]
14. Comi, A.; Polimeni, A. Bus Travel Time: Experimental Evidence and Forecasting. *Forecasting* **2020**, *2*, 309–322. [CrossRef]
15. Ye, Y.; Liu, R.; Xue, F. Application of time series method to the passenger flow prediction in the intelligent bus transportation system with big data. In *Sensor Networks and Signal Processing*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 497–520.
16. Gummadi, R.; Edara, S.R. Prediction of passenger flow of transit buses over a period of time using artificial neural network. In Proceedings of the Third International Congress on Information and Communication Technology, London, UK, 15–16 November 2018; Springer: Singapore, 2019; pp. 963–971.
17. Engle, R.F.; Granger, C.W. Cointegration and error correction: Representation, estimation, and testing. *Econom. J. Econom. Soc.* **1987**, *55*, 251–276.
18. Abdallah, K.B.; Belloumi, M.; De Wolf, D. Indicators for sustainable energy development: A multivariate cointegration and causality analysis from Tunisian road transport sector. *Renew. Sustain. Energy Rev.* **2013**, *25*, 34–43. [CrossRef]
19. Wen, X.; Yang, T.; Guo, X.; Hu, Y. An Analysis of Cointegration Relationship between Public Transportation and Air Quality of Healthy Cities. In Proceedings of the 20th COTA International Conference of Transportation Professionals (CICTP 2020), Xi'an, China, 14–16 August 2020; pp. 2892–2903.
20. Lin, J.; Li, Y. Finding structural similarity in time series data using bag-of-patterns representation. In Proceedings of the International Conference on Scientific and Statistical Database Management, New Orleans, LA, USA, 2–4 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 461–477.
21. Corduas, M. Mining time series data: A selective survey. In *Data Analysis and Classification*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 355–362.
22. Peña, D.; Galeano, P. Multivariate analysis in vector time series. In *DES—Working Papers. Statistics and Econometrics. WS*; Universidad Carlos III de Madrid: Getafe, Spain, 2001.
23. Caiado, J.; Crato, N.; Peña, D. A periodogram-based metric for time series classification. *Comput. Stat. Data Anal.* **2006**, *50*, 2668–2684. [CrossRef]
24. D'Urso, P.; Maharaj, E.A. Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets Syst.* **2009**, *160*, 3565–3589. [CrossRef]
25. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. Determining the Best Number of Clusters in a Data Set. 2015. Available online: https://cran.rproject.org/web/packages/NbClust/NbClust.pdf (accessed on 1 November 2021).
26. Holt, C.C. Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Memo.* **1957**, *52*, 5–10.
27. Winters, P.R. Forecasting sales by exponentially weighted moving averages. *Manag. Sci.* **1960**, *6*, 324–342. [CrossRef]
28. Holt, C.C. Forecasting seasonals and trends by exponentially weighted moving averages. *Int. J. Forecast.* **2004**, *20*, 5–10. [CrossRef]
29. Gardner, E.S., Jr. Exponential smoothing: The state of the art—Part II. *Int. J. Forecast.* **2006**, *22*, 637–666. [CrossRef]
30. Hyndman, R.J.; Athanasopoulos, G. Forecasting: Principles and Practice. 2013. Available online: https://www.otexts.org/fpp (accessed on 15 February 2018).
31. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
32. Velásquez, J.D.; Zambrano, C.; Vélez, L. ARNN: Un paquete para la predicción de series de tiempo usando redes neuronales autorregresivas. *Rev. Av. Sist. Inf.* **2011**, *8*, 177–181.
33. Karabiber, O.A.; Xydis, G. Electricity price forecasting in the Danish day-ahead market using the TBATS, ANN and ARIMA methods. *Energies* **2019**, *12*, 928. [CrossRef]
34. Timmermann, A. Forecast combinations. *Handb. Econ. Forecast.* **2006**, *1*, 135–196.
35. Bates, J.; Granger, C. The combination of forecasts. *J. Oper. Res. Soc.* **1969**, *20*, 451–468. [CrossRef]
36. Granger, C.W.; Ramanathan, R. Improved methods of combining forecasts. *J. Forecast.* **1984**, *3*, 197–204. [CrossRef]
37. Johansen, S. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econom. J. Econom. Soc.* **1991**, *59*, 1551–1580. [CrossRef]
38. Johansen, S. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*; Oxford University Press on Demand: New York, NY, USA, 1995.
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
40. IHS Global Inc. *EViews 10 for Windows*; IHS Global Inc.: Englewood, CO, USA, 2017.
41. IBM Corp. *IBM SPSS Statistics for Windows*; IBM Corp.: Armonk, NY, USA, 2019.