Forecasting using Dynamic Factor Models with Cluster Structure at Barcelona subway stations

I. Mariñas-Collado[a], A.E. Sipols[b], M.T. Santos-Martín[c] and E. Frutos-Bernal[d]

[a]Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo; [b]Department of Applied Mathematics, Materials Science and Engineering and Electronic Technology, Rey Juan Carlos University; [c]Institute of Fundamental Physics and Mathematics. Department of Statistics, University of Salamanca; [d]Department of Statistics, University of Salamanca.

ARTICLE HISTORY
Compiled October 7, 2022

ABSTRACT
Dynamic factor models are a powerful technique for analysing vast volumes of data, more precisely, time series. However, the large volumes of data that come from public transport networks tend to have heterogeneity and a cluster structure. In this paper, Dynamic Factor Models with Cluster Structure (DFMCS) are used to forecast hourly entrances in the different stations of the Barcelona subway network. The main and most novel contribution lies in the use of clustering techniques to make an initial grouping of the behaviour of the elements belonging to the time series, in order to subsequently be able to predict future patterns.

KEYWORDS
Time series; Forecasting; Dynamic Factor Models; Dependency measures; Public transport;

1. Introduction

Barcelona is regarded as a major success story in European urban development. Given that it is Spain's second-largest city, with the growth this implies, it has made a big effort to lead the way in smart cities (Bakıcı, Almirall, and Wareham 2013). Its urban area extends to numerous neighbouring municipalities, making it the fifth most populous urban area in the European Union. Barcelona is a major cultural, economic, and financial centre, with a rich cultural heritage, which makes it an important tourist destination. This is the reason it is such an interesting case to investigate. Figure 1 shows the Barcelona subway map.

The majority of public transportation systems employ automated fare collecting (AFC) systems, which are regarded as a secure way of user validation and fare payment and provide new potential for creative and flexible fare structuring (Dempsey 2007). Furthermore, AFC systems generate massive amounts of highly precise data about on-board transactions (Pelletier, Trépanier, and Morency 2011). The fare system of the Barcelona metro only records entry passes, so the end of the trip is unknown. Although it lies outside the aims of this paper, public transportation destination estimation is one of the primary challenges for smart card data implementation, and there are numerous

CONTACT I. Mariñas-Collado. Email: marinasirene@uniovi.es

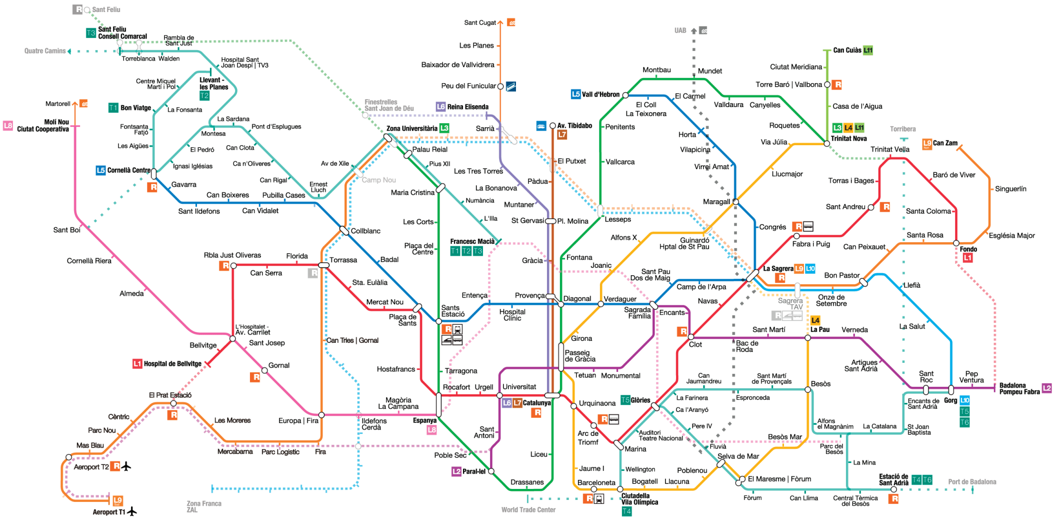


Figure 1.: Barcelona subway (https://www.metrobarcelona.es/mapas.html)

techniques (see, for example, Li et al. (2018b); Alsger et al. (2018); Alexander et al. (2015); Jun and Dongyuan (2013)).

The metro network's passenger volume might change based on the time of day and the location. The use of the subway varies between the different days of the week, on holiday periods, between residential areas and business centers or workplaces, and it's affected by other factors like weather, for example. Different methodologies are used in the literature for this sort of studies, the most commonly used involve clustering techniques (Briand et al. 2017). When performing a cluster analysis with smart card data, two main approaches have been proposed in the literature. The first consists in grouping the stations based on spatio-temporal passenger data (Chen, Chen, and Barry 2009), and the other approach is to directly cluster travellers with common behaviours (El Mahrsi et al. 2017). The k-means algorithm and hierarchical cluster analysis have been the most widely used methods. Wang, Lo, and Liu (2015), Kim et al. (2017), Ding, Cao, and Liu (2019) used gradient boosting decision trees. Montero, Vilar et al. (2014) developed an R-package, TSclust, for time series clustering. There are many alternative dissimilarity measures to compare time series and TSclust is the result of integrating these to conduct time series clustering. The package includes commonly used dissimilarity measures, such as complexity-based measures, model-based measures, feature-based measures and the prediction-based dissimilarity introduced by Vilar, Alonso, and Vilar (2010). Furthermore, in most articles, data related to one or two weeks are used to carry out the analyzes. Losing, in some cases, the information that large databases can provide.

To model a large number of time series, dynamic factor models (DFMs) are very effective (see, for example, Mestekemper, Kauermann, and Smith (2013); Luciani (2014)), but they need to be adapted if there is group structure within the set of series. Golay et al. (1998) and Chouakria and Nagabhushan (2007) used modifications of the instantaneous cross-correlation, and Ando and Bai (2017) studied dependency clustering considering the hypothesis that the time series vector has a Dynamic Factor Model structure where some factors influence different groups of series. Different approaches on forecasting with Dynamic Factor Models are gathered in Escribano, Peña, and Ruiz (2021).

Alonso and Peña (2019) proposed an algorithm to find groups in large time series vectors using the Generalised Cross-Correlation (GCC), a metric of how similar two time series are. This measure compares the determinant of the correlation matrix of the bi-variate vector with those of the two uni-variate time series. The resulting dissimilarity matrix can then be used in any cluster procedure which requires this kind of input. The proposed procedure can be used in exploratory analysis of a large set of time series and, also, can be very useful to build models with grouped factor structure. In this paper, Dynamic Factor Models with Cluster Structure (DFMCS) are used, with some factors that are group-specific and others global.

In the literature, DFMCS have been studied with different assumptions. Wang (2008) presented a theory for analysing large dimensional factor models with a multi-level factor structure and derived conditions for identification of these models. In their multi-factor model, in each group, the series are influenced by global and specific factors. Hallin and Liška (2011) studied the structures of dynamic interrelations within and between blocks of time series. They proposed a model with two clusters in which the factors define four orthogonal sub-spaces: first, the variables which are common to both groups and are, therefore, common factors; second, for the two groups, significantly idiosyncratic variables; finally, those which are considered a common factor for a cluster and idiosyncratic to the other. In these studies, the number of groups and the allotment of the series is assumed to be known. On the other hand, Ando and Bai (2017) introduced a more generic model, assuming unknown membership and Blasques et al. (2021) used an observation-based strategy, which assumed the factors as dynamic processes formulated as functions of previous data.

To model and predict passenger flow, both classical methods and machine learning techniques are commonly used. The latter are increasingly popular due to computational developments. Tang et al. (2018), designed a non-parametric nonlinear regression model to capture passenger flow fluctuation characteristics. Li et al. (2018a) proposed a hybrid model that combines a symbolic regression model and auto-regressive moving average (ARIMA) model. Ye, Liu, and Xue (2021) proposed three kinds of time series models: AR, ARIMA and quadratic ARMA, to forecast passenger flow. Gensuo, Liqin, and Miao (2015) and Sun, Leng, and Guan (2015) forecast transfer passenger flow for rail transit using a support vector machine (SVM) model. Habtemichael and Cetin (2016) proposed a non-parametric and data-driven methodology for short-term traffic forecasting using an enhanced KNN algorithm. Chang et al. (2012) constructed a KNN-NPR model to predict dynamic multi-period traffic volume.

The use of Neural Networks has also increased in recent years. Long Short-Term Memory (LSTM) artificial recurrent neural networks and Convolutional Neural Networks (CNN) are used to explore spatial and temporal relations. Chen et al. (2021) used a convolutional long short-term memory (Conv-LSTM) network to extract spatial and temporal characteristics to solve the short-term prediction problem of the subway congestion delay in the network structure. Xiong et al. (2019) used two deep learning neural networks to predict an urban rail transit passenger flow time series and spatio-temporal series, respectively. Zhang et al. (2020) proposed a deep learning architecture combining the residual network, graph CNN and LSTM network, to forecast short-term passenger flow in urban rail transit on a network scale. Liu, Liu, and Jia (2019) proposed an end-to-end deep learning architecture, to forecast the metro inbound/outbound passenger flow. Nagaraj et al. (2022) proposed a graph learning-based spatial-temporal graph convolutional neural network for traffic forecasting which is founded on graph learning (Hu et al. 2021).

This work focuses on the robust procedure to build DFMCS proposed by Alonso, Galeano, and Peña (2020), which is applied to predict the hourly entrances in the Barcelona subway stations. Passenger flow varies according to different station characteristics, such as the location, the population and the district where the metro station is located, as well as the day of the week or hour of the day. The methodology allows building generalised dynamic factor models to model large time series matrices and perform predictions. The approach includes defining the factors that influence all series at a global level, as well as those factors that are specific to each of the clusters, i.e. they affect only the series included in some clusters, but have no effect on the series belonging to other clusters. Both global and group-specific factors are modelled using ARMA models to fit and pre-specify the data. A comparison of the predictions obtained using DFMCS and a classical approach (ARIMA model) and a machine learning approach (KNN) is presented.

The database consists of 132 stations with the hourly entrances of each day of 2018. To perform the analysis, those times at which most of the stations are closed are omitted, resulting in 21 hours analysed daily for every station. The proposed approach can be applied on its own when working with large sets of time series or, alternatively, it may be considered a starting point for estimation techniques for these types of models. Moreover, this procedure is part of Big Data and Data Science technologies, which are not often used in the area of public transport studies but can produce accurate results.

The rest of the paper is organised as follows: Section 2 introduces DFMCS. Section 3 applies the methodology to fit the DFMCS to the data set from the Barcelona metro, performs forecasting for a week and presents the results. Finally, Section 4 highlights the main conclusions of the paper.

2. Dynamic factor models with cluster structure

Suppose a vector of zero-mean stationary time series, $\mathbf{x}_t = (x_{1t}, ..., x_{kt})'$. Each element of the observed series vector is assumed to be a linear combination of global and specific components in k clusters or groups (and some noise). Let these factors be represented:

- The global factors by the r_0-dimensional vector $\mathbf{f}_{0t} = (f_{01t}, ..., f_{0r_0t})'$.
- The global factor loading matrix by $\mathbf{P}_0 = [\mathbf{P}'_{0,1} \mid \cdots \mid \mathbf{P}'_{0,s}]'$, with dimensions $k \times r_0$ and where $\mathbf{P}_{0,i}$, for $i = 1, ..., s$, is the $k_i \times r_0$ loading matrix for the k_i series of the i-th group.
- The specific factors for the i-the cluster as the r_i-dimensional vector $\mathbf{f}_{it} = (f_{i1t}, ... f_{ir_it})'$.
- The matrix of specific factors loadings, that only affect the k_i time series in the i-th group, by $\mathbf{P}_i = [\mathbf{0}'_{i,1} \mid ... \mid \mathbf{P}'_{i,i} \mid ... \mid \mathbf{0}'_{i,s}]$, with dimensions $k \times r_i$.

Then, the Dynamic Factor Model with Cluster Structure (DFMCS) can be written as:

$$\mathbf{x}_t = \mathbf{P}_0\mathbf{f}_{0t} + \sum_{i=1}^{k} \mathbf{P}_i\mathbf{f}_{it} + \boldsymbol{\eta}_t. \tag{1}$$

The series are considered to be in order, and therefore, the first k_i series belong to the first group and so on. Then, $\sum_{i=1}^{k} k_i = k$. This does not incur loss of generality. The idiosyncratic term or noise, $\boldsymbol{\eta}_t = (\eta_{1t}, ..., \eta_{mt})'$, is a general sequence of stationary time series with mean $\mathbf{0}_m$, and weak dependency. For more details, see Alonso, Galeano, and Peña (2020).

The procedure to fit the DFMCS consists of the following steps: first, the observed times series are cleaned, including removal of additive outliers, level changes, and outlying time series. Then, the factors and factor loadings are estimated in the first place by minimising:

$$\mathrm{SE}_1 = \sum_{t=1}^{T} \|\mathbf{x}_t - \mathbf{P}\mathbf{f}_t\|^2 , \tag{2}$$

where $\|\cdot\|^2$ represents the Euclidean vector norm. The resulting estimates of the factor loading matrix, $\hat{\mathbf{P}}$, are the eigenvectors associated to the r_c largest eigenvalues of the time series' sample covariance matrix (the test proposed in Ahn and Horenstein (2013) is used to specify r_c, the number of factors). Then, $\hat{\mathbf{f}}_{i,t} = \hat{\mathbf{P}}'_{\mathbf{i}}\mathbf{x}_t$ estimates the factors and the common component is $\mathbf{c}_t = \hat{\mathbf{P}}'\mathbf{x}_t$. Then, the clustering algorithm proposed by Alonso and Peña (2019) to divide the time series in groups with similar linear dependence, is used to locate the groupings in these common components.

Once the series are divided into groups, those inside each one are used to estimate the new factors and their loadings. The method by Ahn and Horenstein (2013) is used again to find the number of factors for each group, r_i^s, $i = 1, \ldots, k$. Analogously to the global factors and their loadings, the specific loading matrices are estimated from the eigenvalues of the time series sample covariance matrices in each group. The specific loading matrices $\hat{\mathbf{P}}_{\mathbf{i}}$, of dimension $m \times r_i^s$ and columns $\hat{\mathbf{P}}_{\mathbf{i1}}, \ldots, \hat{\mathbf{P}}_{\mathbf{ir_i^s}}$ are built by adding a set of zero values for the observations which are not in the group, to the eigenvectors of the largest r_i^s eigenvalues in the ith group. $\hat{\mathbf{f}}^s_{ij,t} = \hat{\mathbf{P}}'_{\mathbf{ij}}\mathbf{x}_t$, with $j = 1, \ldots, r_i^s$, estimates the factors of each group. The primer set of r_c factors and the second set of $\sum_{i=1}^{k} r_i^s$ factors are divided into global or specific. To determine whether a factor belongs to a specific factor set, the empirical canonical correlation between the factor and those in the set is calculated. Then, the residuals $\mathbf{v}_t = \mathbf{x}_t - \hat{\mathbf{P}}_{\mathbf{0}}\hat{\mathbf{f}}_{0t}$ are computed, where the factors have been estimated using ARMA models. $\hat{\mathbf{f}}_{\mathbf{0t}}$ and $\hat{\mathbf{P}}_{\mathbf{0}}$ are the final vector of estimated global factor and its loading matrix respectively. The specific factors are re-estimated using the series v_{it} corresponding to each group. Then, we check whether all the groups have at least one specific factor. Finally, with the estimated factors, groups and loadings, the residuals, or idiosyncratic component, can be computed and the SE minimised.

In summary, the series are first cleaned up, outliers are eliminated, and the common elements of the series are computed. The series are then separated into clusters using generalised cross correlation and factors are determined for each cluster. Finally, in order to forecast the series, the factors are adjusted with seasonal ARIMA models.

3. Analysis and results of Barcelona data

The data provided by the Transport Metropolitan of Barcelona correspond to the hourly daily entrances of 2018 of the different subway stations. To apply the methodology, a total of 2772 series are used. In order to carry out the clustering taking into account the temporal characteristics of the data, each series is subdivided by the hours of the day. A new matrix is built, where in each row only the day of the year is taken into account and where the columns collect the number of passengers at each station in each hour, that is, the new matrix has as many rows as days of the year and 2772 columns that correspond to the 132 stations times 21 hours.

The procedure by Alonso, Galeano, and Peña (2020) starts with the selection and estimation of the common factors, calculated using all the time series. Therefore, the clustering is done in the common part so that the effect of the noise is "eliminated". Figure 2 shows the total variability explained by each global factor. The first three factors explain over 80% of the observed variability.

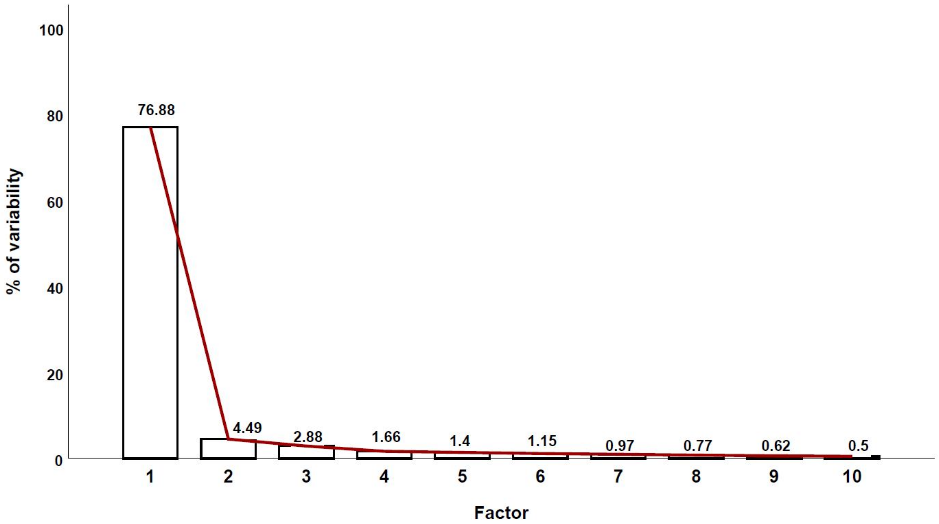


Figure 2.: Variability explained by each global factor.

The clusters are now calculated following Alonso and Peña (2019). Figure 3 shows the dendrogram obtained from applying hierarchical clustering algorithm with single linkage to the dissimilarity matrix.

Cluster	No. of Series	Percentage
1	2098	75.69%
2	674	24.31%

Table 1.: Series in each cluster.

There is a strong dependency between the series, which is confirmed below by the existence of two clearly global factors. There is a large group of series with a very strong relationship and therefore, it is decided to consider one unique cluster for those series with a dependency level below 0.001. The other group consists of series that are dependent but with smaller dependency levels. Table 1 shows the resulting number of series in each cluster and the percentage they represent over the total number of series.

The factors in each group are then calculated. All specific factors are presumed to be part of the group factors, together with some (or all) global factors. After this, together with the initial estimations, there are a total of 3 factors that need to be classified as global or specific. This is done using the canonical correlations between

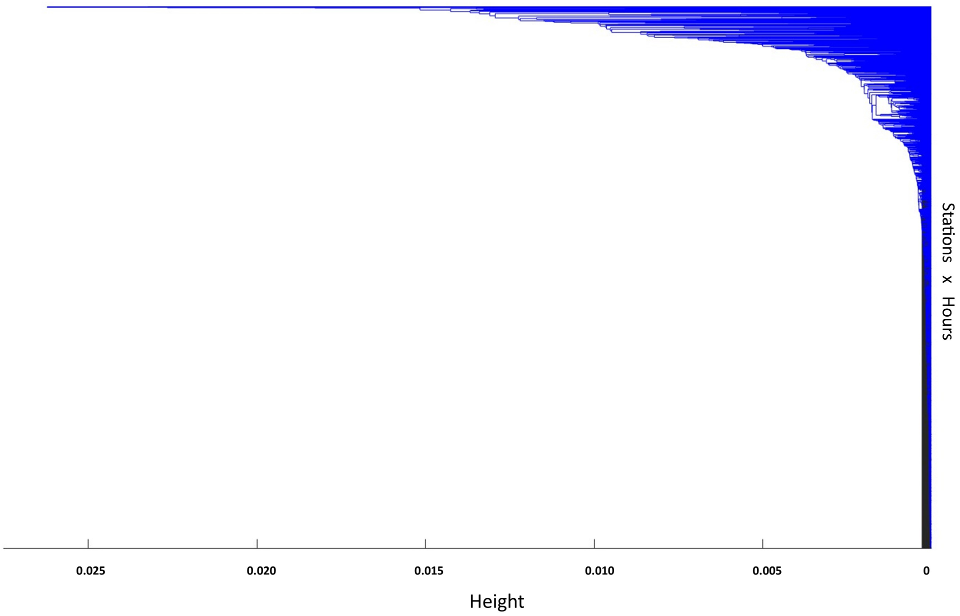


Figure 3.: Dendrogram of the series.

the factors and the clusters.

Table 2 shows the correlation between each factor and the clusters. It can be noted how the first factor (Factor 1) is highly correlated with the first cluster and somewhat less with the second. Factors 2 and 3 are correlated with both clusters similarly. For this reason, Factor 1 is classified as a specific factor of Cluster 1, while Factors 2 and 3 are categorised as global, given there is not a clear very strong correlation with just one of the two clusters. Moreover, it is found that the factor for Cluster 1 explains 88% of the variability in the group, after removing information from the global factors. When estimating specific factors, it also appeared that there may be two factors for Cluster 2, however, the variability explained by them is very small and therefore it is concluded that there are no specific factors for the second cluster. Figure 4 shows the variability explained by the specific factors of each cluster. The final model has, therefore, two global factors and one specific factor for Cluster 1.

	Cluster 1	Cluster 2
Factor 1	1.0000	0.6106
Factor 2	0.9154	0.9504
Factor 3	0.8926	0.7037

Table 2.: Correlation between factor and cluster.

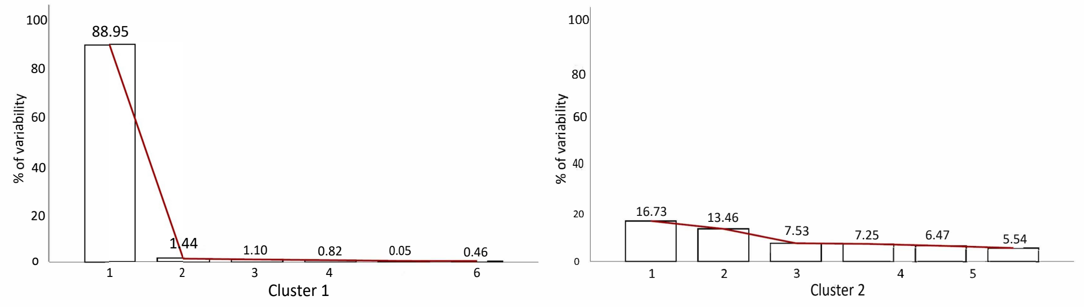


Figure 4.: Variability explained by each factor in Cluster 1 (left) and Cluster 2 (right).

Out of sample predictions are used to assess the performance of the models when the cluster effects are included or not. Thus, two models are considered: Model 1, a DFM with two global factors and Model 2, a DFMCS model with two global factors and one specific factor. In both models, the idiosyncratic term is supposed to be random and seasonal ARIMA models are fitted to the factors, following García-Martos and Conejo (2013) and Alonso, Bastos, and García-Martos (2016). Throughout the testing period, one-day ahead forecasting is done, using a rolling window approach.

Forecasting results comparison

The aim now is to evaluate whether the models fitted adjust well to the data. For this, the validation is done using dividing the data in training and test data-sets. An extended practice is to choose a model merely upon how well it fits the data. However, while a model may fit and reproduce the data very well, this does not necessarily imply that it can predict future data well (usually, the better the model fits the data, the worse it predicts future data, which is called over-training). To avoid this, the series are split into two groups, one to estimate the model (training data), and one to measure the quality of the model predictions (test data). Specifically, the method used in this paper is Leave-one-out cross-validation (LOOCV). This is a particular case of 'leave-p-out' cross-validation, making a more efficient use of the data, since omitting one or several observations does not eliminate all the information associated with them due to the inherent correlation between observations in a time series. Moreover, a rolling forecast origin approach was adopted. The testing period chosen is a 7 days forecast using the previous 90 days.

To compare the forecasting results of Model 1 and Model 2 (without and with cluster structure respectively) two error measures are used: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), calculated using the target values, y_i, and the predictions $\hat{y}_i$:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

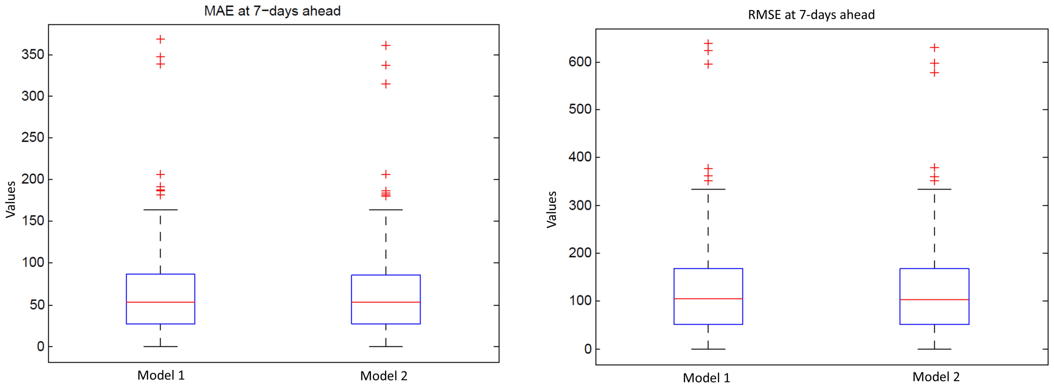
$$\mathrm{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)^{1/2} \tag{4}$$

As previously mentioned, the data from the Barcelona subway consists on 21 hours registered daily in 2018 in the 132 stations. Then, to make predictions for, say, a week after the day 90, then the errors would be calculated as:

$$\text{MAE} = \frac{1}{268 \times 132 \times 21} \sum_{d=90}^{365-7} \sum_{s=1}^{132} \sum_{h=1}^{21} |y_{d,s,h} - \hat{y}_{d,s,h}| \qquad (5)$$

$$\text{RMSE} = \left(\frac{1}{268 \times 132 \times 21} \sum_{d=90}^{365-7} \sum_{s=1}^{132} \sum_{h=1}^{21} (y_{d,s,h} - \hat{y}_{d,s,h})^2 \right)^{1/2} \qquad (6)$$

Prediction is done from 1 to 7 days. Table 3 shows the MAE and RMSE for Model 1 (without cluster structure) and Model 2 (with cluster structure) and the percentage of improvement achieved using the second model with respect to the first. The improvement percentage is always positive, which indicates that model 2 is superior to model 1 in all cases. Given that Cluster 2 has no specific factors, the same results are obtained using both models. The one-day forecast results are similar to the rest, with an improvement of almost 3% in the MAE using the model with cluster structure. This is even more evident in Cluster 1, where the percentages of improvement are generally larger.



(a) MAE

(b) RMSE

Figure 5.: MAE and RMSE for Model 1 (left) and Model 2 (right).

Figures 5a and 5b show boxplots for the MAE and RMSE for a 7-days forecast for both models. A comparison between the true values and the predicted supports the adequate predictive capacity of these models. Model 2 improves the predictions.

To further inspect the accuracy of predictions, two stations are studied. “Station 1”, which corresponds to the Airport and “Station 132”, in the university area, each belonging to a different cluster. Station 132 belongs to Cluster 1, i.e. there can be differences between the predictions using model 1 or 2, even if small. Moreover, predictions with a traditional method (ARIMA) and machine learning techniques (KNN) are compared. For the KNN model, the MIMO (Multiple Input Multiple Output) strategy is used. This strategy is commonly applied with KNN and it is characterised by the use of a vector of target values. The length of this vector is equal to the number of periods to forecast (Mariñas-Collado et al. 2022).

Table 4 shows MAES and RMSEs for predictions 1-7 days ahead, using the previous 90 days. A Rolling Window approach is used in which the 90 days window moves 30

	Global		Cluster 1		Cluster 2	
1 day	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	72.68	192.10	79.32	199.92	51.99	165.42
Model 2	70.57	189.29	76.54	196.35	51.99	165.42
% imp.	2.90	1.46	3.51	1.78	0	0
	Global		Cluster 1		Cluster 2	
2 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	68.99	182.20	74.66	187.41	51.33	164.94
Model 2	67.63	179.41	72.87	183.81	51.33	164.94
% imp.	1.96	1.54	2.39	1.92	0	0
	Global		Cluster 1		Cluster 2	
3 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	67.61	170.63	72.91	172.19	51.09	165.69
Model 2	66.20	167.68	71.06	168.31	51.09	165.69
% imp.	2.08	1.73	2.55	2.25	0	0
	Global		Cluster 1		Cluster 2	
4 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	69.98	177.78	76.13	183.39	50.83	159.06
Model 2	68.76	174.92	74.52	179.72	50.83	159.06
% imp.	1.74	1.61	2.11	2.00	0	0
	Global		Cluster 1		Cluster 2	
5 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	62.94	161.54	66.92	162.61	50.56	158.16
Model 2	61.09	158.44	64.48	158.53	50.56	158.16
% imp.	2.94	1.92	3.65	2.51	0	0
	Global		Cluster 1		Cluster 2	
6 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	65.29	175.91	70.21	182.67	49.96	152.98
Model 2	63.43	173.18	67.76	179.19	49.96	152.98
% imp.	2.84	1.55	3.48	1.90	0	0
	Global		Cluster 1		Cluster 2	
7 days	MAE	RMSE	MAE	RMSE	MAE	RMSE
Model 1	68.82	175.41	74.54	183.12	51.02	148.89
Model 2	67.69	172.75	73.05	179.74	51.02	148.89
% imp.	1.63	1.52	1.99	1.84	0	0

Table 3.: MAE and RMSE for Model 1 (without cluster structure) and Model 2 (with cluster structure) for all the series and both clusters.

days ahead each time, so that the last observed value of each window corresponds to days: 90, 120, 150 and 180. It can be noted that the errors are smaller for the predictions obtained with DFM (with or without cluster structure). This is done first independently for the two selected stations, and then the mean of the errors for all the stations together are shown.

Station 1				
Model	ARIMA	KNN	Model 1	Model 2
MAE	47.03	52.70	37.95	37.95
RMSE	62.77	68.21	50.07	50.07
Station 132				
Model	ARIMA	KNN	Model 1	Model 2
MAE	185.97	162.87	91.88	94.00
RMSE	286.9	233.78	146.91	149.26
All series				
Model	ARIMA	KNN	Model 1	Model 2
MAE	178.31	155.16	68.04	66.48
RMSE	346.16	294.95	176.51	173.67

Table 4.: MAE and RMSE for ARIMA, KNN, Model 1 (without cluster structure) and Model 2 (with cluster structure) for two stations separately and overall the stations.

The results for each station are shown in Figures 6 and 7 respectively. In each figure, only the previous 7 days are shown (although data from the previous 90 days has been used to make the predictions). The true series are shown in black and Model 1 and 2 in blue and green, respectively. Recall that in Station 1, belonging to Cluster 2, Models 1 and 2 are the same, since Cluster 2 has no specific factors.

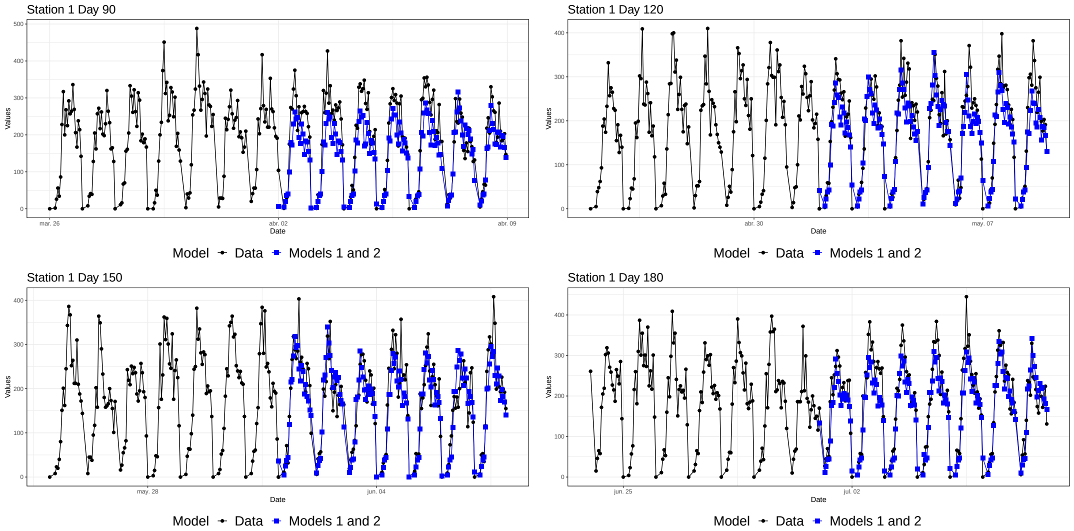


Figure 6.: Forecasting after day 90, 120, 150 and 180 in Station 1.

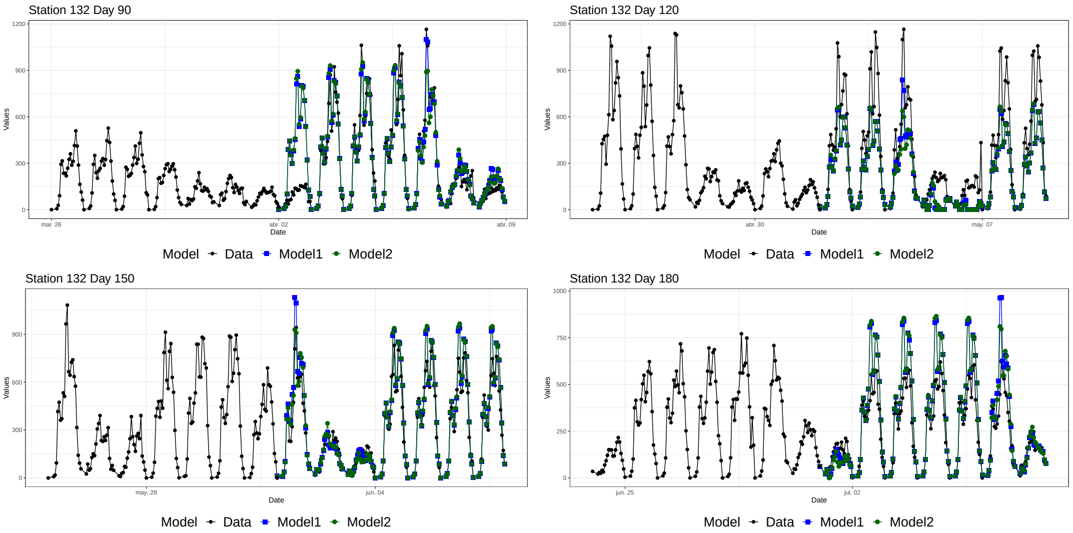


Figure 7.: Forecasting after day 90, 120, 150 and 180 in Station 132.

To illustrate the predictions with the different models, Figure 8 shows three days of predictions after one window in Station 132 (where Model 1 and 2 also differ). It can be noted that ARIMA repeats the same pattern, which the DFMCS are able to capture. The cluster structure in Model 2 allows for a better prediction of the patterns.

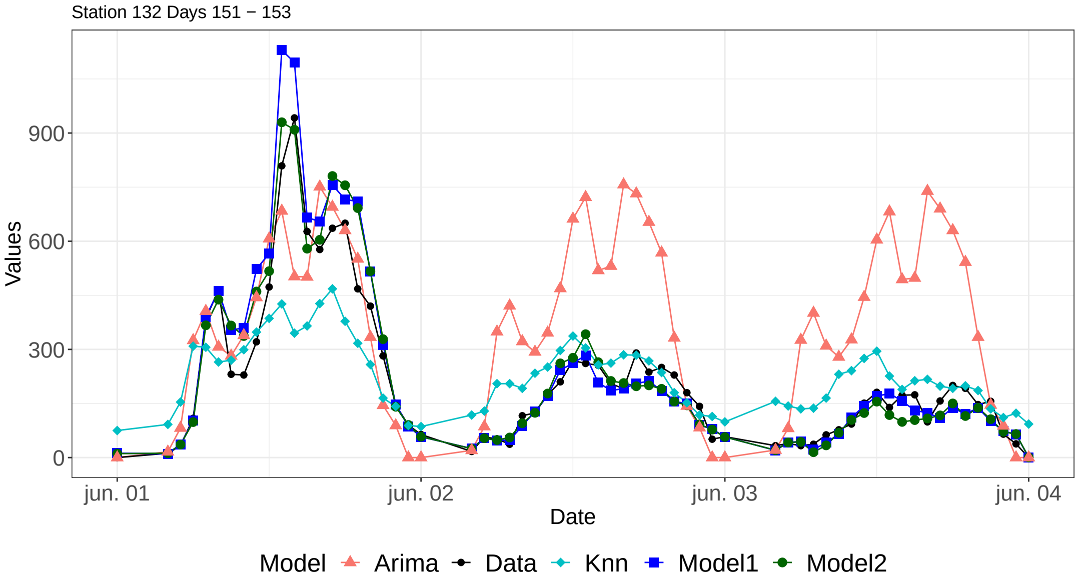


Figure 8.: Forecasting in Station 132 after day 150.

The results shown were obtained modelling the procedure in Matlab (MATLAB 2021), using the TS function: TRAMO-SEATS (Román 2022), which allows to automatically determine the optimal parameters of ARMA models in time series. ARIMA and KNN models haven been implemented in *R* (R Core Team 2021).

4. Conclusions

The prediction of passenger hourly entries at each Barcelona metro station is a complex task to perform as there is a large volume of data from all stations on the network. Making use of the procedure by Alonso, Galeano, and Peña (2020) to estimate a Dynamic Factor Model with Cluster Structure, has made it possible to characterise the stations based on the hourly distribution. These forecasts would serve to obtain accessibility improvements throughout the Barcelona Metro network.

Some subway stations have a highly polarised temporary demand, so their efficiency decreases. Achieving a more balanced demand over time is an essential objective to ensure better use of metro infrastructure and services. Since the temporal distribution of the travellers who access the stations depends, to a large extent, on the characteristics of the station environment, land use policies could improve the efficiency of the metro system by promoting the mix of uses and activities.

The novel methodology used allows to obtain good estimations in passenger prediction in Barcelona subway stations using the information provided by the time series, which implies the analysis of large amount of data. It has been proved to be a very robust procedure for prediction in Big Data. The procedure rearranges the series to better capture the dynamics of the temporal frequency, it classifies them into different clusters and then extracts global and common factors, which are predicted with ARIMA models to compute forecasts. Moreover, since the ARIMA models to predict the factors allow the incorporation of regressor variables, these can be used to deal with special events and extended the applications of the method. It is important to highlight that this methodology has not been previously used to predict passenger flow, although it is able to capture the temporal structure of the data, as well as the information from the different clusters of series. The fundamental contribution lies in its ability to group and predict jointly, using the information of the factors that are calculated to make the groupings for the adjustment and prediction. The approach commonly used in these cases consists in defining the clusters first and then taking a representative of each cluster, without incorporating information related to the cluster to which it belongs. This results in a more precise fitting of the data that improves forecasting, as can be seen in the comparison with other used methodologies such as ARIMA applied directly to the series and the KNN algorithm.

The results that can be obtained with this methodology can help make decisions to determine public transport strategies. Controlling the number of passengers at rush hour, increasing the number of wagons and the frequency of the subway, orienting passengers to enter or leave stations that are not as busy at certain times... are some of the measures that can be taken knowing the hourly forecast of passengers in the Barcelona subway stations.

Acknowledgements

This work has been supported by the Spanish MINECO project PGC2018-098623-B-I00, the Agencia Estatal de Investigación (PID2019-108311GB-I00/AEI/10.13039/501100011033) and by the Castilla y León Government project SA105P20. The authors extend their gratitude to the Transport Metropolitans of Barcelona.

References

Ahn, Seung C, and Alex R Horenstein. 2013. "Eigenvalue ratio test for the number of factors." *Econometrica* 81 (3): 1203–1227.

Alexander, Lauren, Shan Jiang, Mikel Murga, and Marta C González. 2015. "Origin–destination trips by purpose and time of day inferred from mobile phone data." *Transportation Research Part C: Emerging Technologies* 58: 240–250.

Alonso, Andrés M, Guadalupe Bastos, and Carolina García-Martos. 2016. "Electricity price forecasting by averaging dynamic factor models." *Energies* 9 (8): 600.

Alonso, Andrés M, Pedro Galeano, and Daniel Peña. 2020. "A robust procedure to build dynamic factor models with cluster structure." *Journal of Econometrics* 216 (1): 35–52.

Alonso, Andrés M, and Daniel Peña. 2019. "Clustering time series by linear dependency." *Statistics and Computing* 29 (4): 655–676.

Alsger, Azalden, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. 2018. "Public transport trip purpose inference using smart card fare data." *Transportation Research Part C: Emerging Technologies* 87: 123–137.

Ando, Tomohiro, and Jushan Bai. 2017. "Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures." *Journal of the American Statistical Association* 112 (519): 1182–1198.

Bakıcı, Tuba, Esteve Almirall, and Jonathan Wareham. 2013. "A smart city initiative: the case of Barcelona." *Journal of the Knowledge Economy* 4 (2): 135–148.

Blasques, Francisco, Meindert Heres Hoogerkamp, Siem Jan Koopman, and Ilka van de Werve. 2021. "Dynamic factor models with clustered loadings: Forecasting education flows using unemployment data." *International Journal of Forecasting* .

Briand, Anne-Sarah, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. 2017. "Analyzing year-to-year changes in public transport passenger behaviour using smart card data." *Transportation Research Part C: Emerging Technologies* 79: 274–289.

Chang, H, Youngjoo Lee, B Yoon, and Sanghoon Baek. 2012. "Dynamic near-term traffic flow prediction: system-oriented approach based on past experiences." *IET intelligent transport systems* 6 (3): 292–305.

Chen, Cynthia, Jason Chen, and James Barry. 2009. "Diurnal pattern of transit ridership: a case study of the New York City subway system." *Journal of Transport Geography* 17 (3): 176–186.

Chen, Wei, Zongping Li, Can Liu, and Yi Ai. 2021. "A Deep Learning Model with Conv-LSTM Networks for Subway Passenger Congestion Delay Prediction." *Journal of Advanced Transportation* 2021.

Chouakria, Ahlame Douzal, and Panduranga Naidu Nagabhushan. 2007. "Adaptive dissimilarity index for measuring time series proximity." *Advances in Data Analysis and Classification* 1 (1): 5–21.

Dempsey, Paul Stephen. 2007. "Privacy issues with the use of smart cards." *Available at SSRN 3295908* .

Ding, Chuan, Xinyu Cao, and Chao Liu. 2019. "How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds." *Journal of Transport Geography* 77 (C): 70–78.

El Mahrsi, Mohamed K, Etienne Come, Latifa Oukhellou, and Michel Verleysen. 2017. "Clustering Smart Card Data for Urban Mobility Analysis." *IEEE Transactions on Intelligent Transportation Systems* 18 (3): 712–728.

Escribano, Alvaro, Daniel Peña, and Esther Ruiz. 2021. "30 years of cointegration and dynamic factor models forecasting and its future with big data." .

García-Martos, Carolina, and Antonio J Conejo. 2013. "Price forecasting techniques in power systems." *Wiley encyclopedia of electrical and electronics engineering* .

Gensuo, M, ZHAO Liqin, and L Miao. 2015. "Subway station passenger flow forecast based on mixed kernel support vector machine optimized by golden section chaotic particle swarm optimization." *Computer Engineering and Applications* 2015: 44.

Golay, Xavier, Spyros Kollias, Gautier Stoll, Dieter Meier, Anton Valavanis, and Peter Boesiger. 1998. “A new correlation-based fuzzy logic clustering algorithm for FMRI.” *Magnetic Resonance in Medicine* 40 (2): 249–260.

Habtemichael, Filmon G, and Mecit Cetin. 2016. “Short-term traffic flow rate forecasting based on identifying similar traffic patterns.” *Transportation research Part C: emerging technologies* 66: 61–78.

Hallin, Marc, and Roman Liška. 2011. “Dynamic factors in the presence of blocks.” *Journal of Econometrics* 163 (1): 29–41.

Hu, Na, Dafang Zhang, Kun Xie, Wei Liang, and Meng-Yen Hsieh. 2021. “Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting.” *Connection Science* 1–20.

Jun, Chen, and Yang Dongyuan. 2013. “Estimating smart card commuters origin-destination distribution based on APTS data.” *Journal of Transportation Systems Engineering and Information Technology* 13 (4): 47–53.

Kim, Mi-Kyeong, Sang-Pil Kim, Joon Heo, and Hong-Gyoo Sohn. 2017. “Ridership patterns at subway stations of Seoul capital area and characteristics of station influence area.” *KSCE Journal of Civil Engineering* 21 (3): 964–975.

Li, Linchao, Yonggang Wang, Gang Zhong, Jian Zhang, and Bin Ran. 2018a. “Short-to-medium term passenger flow forecasting for metro stations using a hybrid model.” *KSCE Journal of Civil Engineering* 22 (5): 1937–1945.

Li, Tian, Dazhi Sun, Peng Jing, and Kaixi Yang. 2018b. “Smart card data mining of public transport destination: A literature review.” *Information* 9 (1): 18.

Liu, Yang, Zhiyuan Liu, and Ruo Jia. 2019. “DeepPF: A deep learning based architecture for metro passenger flow prediction.” *Transportation Research Part C: Emerging Technologies* 101: 18–34.

Luciani, Matteo. 2014. “Forecasting with approximate dynamic factor models: the role of non-pervasive shocks.” *International Journal of Forecasting* 30 (1): 20–29.

Mariñas-Collado, Irene, Ana E Sipols, M Teresa Santos-Martín, and Elisa Frutos-Bernal. 2022. “Clustering and Forecasting Urban Bus Passenger Demand with a Combination of Time Series Models.” *Mathematics* 10 (15): 2670.

MATLAB. 2021. *version 9.10 (R2021a)*. Natick, Massachusetts: The MathWorks Inc.

Mestekemper, Thomas, Göran Kauermann, and Michael S Smith. 2013. “A comparison of periodic autoregressive and dynamic factor models in intraday energy demand forecasting.” *International Journal of Forecasting* 29 (1): 1–12.

Montero, Pablo, José A Vilar, et al. 2014. “TSclust: An R package for time series clustering.” *Journal of Statistical Software* 62 (1): 1–43.

Nagaraj, Nandini, Harinahalli Lokesh Gururaj, Beekanahalli Harish Swathi, and Yu-Chen Hu. 2022. “Passenger flow prediction in bus transportation system using deep learning.” *Multimedia tools and applications* 1–24.

Pelletier, Marie-Pier, Martin Trépanier, and Catherine Morency. 2011. “Smart card data use in public transit: A literature review.” *Transportation Research Part C: Emerging Technologies* 19 (4): 557–568.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Román, Juan Bógalo. 2022. *TS function: TRAMO-SEATS under Matlab*. MATLAB Central File Exchange.

Sun, Yuxing, Biao Leng, and Wei Guan. 2015. “A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system.” *Neurocomputing* 166: 109–121.

Tang, Liyang, Yang Zhao, Javier Cabrera, Jian Ma, and Kwok Leung Tsui. 2018. “Forecasting short-term passenger flow: An empirical study on shenzhen metro.” *IEEE Transactions on Intelligent Transportation Systems* 20 (10): 3613–3622.

Vilar, José Antonio, Andrés M Alonso, and Juan Manuel Vilar. 2010. “Non-linear time series clustering based on non-parametric forecast densities.” *Computational Statistics & Data Analysis* 54 (11): 2850–2865.

Wang, Peng. 2008. “Large dimensional factor models with a multi-level factor structure: identification, estimation and inference.” *Unpublished manuscript, New York University* .

Wang, WL, SM Lo, and SB Liu. 2015. “Aggregated metro trip patterns in urban areas of Hong Kong: Evidence from automatic fare collection records.” *Journal of Urban Planning and Development* 141 (3): 05014018.

Xiong, Zhi, Jianchun Zheng, Dunjiang Song, Shaobo Zhong, and Quanyi Huang. 2019. “Passenger flow prediction of urban rail transit based on deep learning methods.” *Smart Cities* 2 (3): 371–387.

Ye, Yinna, Ruoxi Liu, and Feng Xue. 2021. “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*, 497–520. Springer.

Zhang, Jinlei, Feng Chen, Zhiyong Cui, Yinan Guo, and Yadi Zhu. 2020. “Deep learning architecture for short-term passenger flow forecasting in urban rail transit.” *IEEE Transactions on Intelligent Transportation Systems* 22 (11): 7004–7014.