Bus line classification using neural networks

Felipe Jiménez *, Francisco Serradilla, Alfonso Román, José Eugenio Naranjo

Technical University of Madrid (UPM), University Institute for Automobile Research (INSIA), Campus Sur UPM, Carretera de Valencia, km. 7, 28031 Madrid, Spain

ARTICLE INFO

Keywords:
Cluster
Urban buses
Neural network
Bootstrap method

ABSTRACT

Grouping urban bus routes is necessary when there are evidences of significant differences among them. In Jiménez et al. (2013), a reduced sample of routes was grouped into clusters utilizing kinematic measured data. As a further step, in this paper, the remaining urban bus routes of a city, for which no kinematic measurements are available, are classified. For such purpose we use macroscopic geographical and functional variables to describe each route, while the clustering process is performed by means of a neural network. Limitations caused by reduced training samples are solved using the bootstrap method.

Introduction

Representing the complete vehicle operation of an urban bus transit system using a single drive cycle may be an excessively simplistic approach and many details concerning differences among lines may remain unobserved (Jiménez et al., 2013). Therefore, several methodologies to group routes into clusters of lines with homogeneous characteristics are proposed. This classification could provide relevant information to the bus fleet company in order to organize its vehicles in the most appropriate way, for example, to save fuel or reduce emissions considering each line's characteristics.

Clustering (Jain et al., 1999) or classification (Beniwal and Arora, 2012) algorithms differ because grouping into clusters is a non-supervised process; while a classification procedure requires that groups are predetermined and their properties must be previously identified.

These techniques are frequently used in diverse areas related with the transport industry. In addition to studies on peer transportation companies (Giuliano, 1981; Fielding et al., 1985), or road classifications according to traffic conditions, traffic characteristics and fleet compositions (Chen et al., 2008); clustering is also used for developing driving cycles or during the categorization of routes. For instance, Fotouhi and Montazeri (2013) use the centroid (k-means) method to group a sample of microcycles that represent traffic patterns in Tehran. André and Villanova (2004) study the bus network in Paris using two methods to evaluate and classify the lines from an environmental perspective: the first approach uses statistics related with route characteristics, travel time, commercial speed, annual statistical data, the irregularity of travel and information on the problems encountered; while the alternative procedure considers aspects related with the socioeconomic peculiarities of each route's setting. In André et al. (2005), the bus routes are characterized by linking analyses of bus operating conditions and urban characteristics collected and managed using a Geographic Information System (GIS).

Due to the difficulty, the high costs required, and the time consumed in the acquisition of detailed data (for example, speed profiles) for the complete operation, most studies perform the route classification process without pre-established groups and generally based on macroscopic variables. Subsequently, this classification is utilized as a guideline in order to select a limited sample of routes in which to measure the kinematic variables.

This article proposes a different approach by assigning a characteristic operational cycle to each line, considering that speed cycle information is a more reliable representation of the route's operation and, therefore, contains additional information that may be used during the clustering process. In this case, the preliminary cluster organization based on the kinematic data contained in a limited sample of routes has already been structured (Jiménez et al., 2013), and consequently, this paper describes the methodology utilized to allocate the remaining lines to these clusters using macroscopic variables that are available for the complete set of lines, since kinematic variables have only been gathered for the sampled routes. For this purpose, a neural network is used in order to process the macroscopic variables as input data and provide output values with the appropriate cluster recommendation for each line.

Methodology

Jiménez et al. (2013) includes a proposal of algorithms that may be utilized to classify a set of urban bus routes using the kinematic data related with the operation. However, collecting this information for a large amount of lines is usually a costly and time-consuming process, and therefore, a more practical approach consists in identifying macroscopic variables that influence the operation of the vehicles. Consequently, a method based on the analysis of macroscopic variables is required in order to assign lines to the set of clusters previously obtained from kinematic data.

Macroscopic variables

Table 1 shows the list of macroscopic variables used in this study. These variables include features related with the operation of the vehicles, the characteristics of the routes, the infrastructure on which the lines are set and the areas of the city being serviced. The metropolitan surface is segmented according to two different criteria: a broader perspective which entails zones that are delimited by the main beltways of the city, and a more detailed partition based on district and neighborhood limits. It should be noted that some of the macroscopic variables can be adapted depending on the city's geography (for instance, defining more areas in variables 15 and 16), but this circumstance does not influence the remaining application of the methodology.

Neural network

Artificial Neural Networks are recently developed computational models that are very useful in classification and clustering processes (Bishop, 1995; Hagan et al., 1996).

Table 1
Definition of the macroscopic variables.

N°	Variable	Units
Bus operation		
MV 1	Average speed	km/h
MV 2	Bus frequency (time between 2 consecutive bus arrivals)	minutes
MV 3	Number of passengers index	dimensionless
Route characteristics		
MV 4	Number of bus stops per km	dimensionless
MV 5	Number of streets per km	dimensionless
	Coincidences with other means of transport	
MV 6	Coincidences with Metro stations per km	dimensionless
MV 7	Coincidences with Metro lines per km	dimensionless
Bus route infrastructure		
Bus lane		
MV 8	% of the route distance driven along conventional bus lane	%
MV 9	% of the route distance driven along bus lane with barrier	%
MV 10	% of the route distance driven along independent bus lane	%
MV 11	% of the route distance driven along bus lane	%
City areas		
MV 12	Streets fiscal index (index between 1 – highest index and 9 – lowest index)	dimensionless
Distances to city center		
MV 13	Distance of the nearest route terminal to the city center	m
MV 14	Distance of the furthest route terminal to the city center	m
MV 15	Proximity index to the city center (index between 1 and 3 depending on the city area)	dimensionless
MV 16	Distance index to the city center (index between 1 and 3 depending on the city area)	dimensionless
Circulation along city areas limit streets		
MV 17	% of the route distance driven along district limit streets	%
MV 18	% of the route distance driven along neighborhood limit streets	%
MV 19	% of the route distance driven along district or neighborhood limit streets	%

Based on the aforementioned 19 variables, an expert system has been developed in order to classify each line within the most appropriate group included in the set of N predefined clusters which already includes the routes contained in the sample. For this purpose, a neural network (whose output variable is the assigned cluster) was trained. Different types of network architectures have been tested, including deep-learning networks (those with more than two hidden layers) based on autoencoders.

The training process is implemented using the SALMON (System for Automated Learning: Modelling of Operative Networks) application (Naranjo et al., 2012). This tool does not only provide the means to train a multi-layer perceptron, but also offers an estimation of the error that the network will generate in a production environment. Additionally, it provides advanced training methods such as bootstrapping or sensitivity analysis to identify the relative weight (i.e. importance) of the input variables of the network.

An important limitation of this classification procedure is originated by the undersized sample used to train the network, since the attainment of microscopic information on the operation of the routes is a demanding procedure. Predictions must be made for a considerable amount of lines based on limited training information, and accordingly, generalization becomes a difficult task. If the network does not generalize properly, it may occur that the level of errors is small during the training phase and, however, the number of mistakes increases significantly throughout the system's utilization phase, this being an undesirable effect. This problem is known as overfitting.

During the training stage, a procedure of cross-validation should be implemented in order to estimate the system's level of inaccuracy when processing unclassified data during the productive phase (Krogh and Vedelsby, 1995). This method consists in extracting a certain sample of random data from the training information and avoiding its use during the training process. Later on, this information may be used in order to calculate the percentage of success using such test data, which, in every way, is not known by the network. If the training set is extremely small, detaching a significant number of cases is undesirable, since the problem may not be well represented and the system could provide a very low performance. This inconvenience has been overcome using the bootstrap method to evaluate the level of error (Franke and Neumann, 2000; Bakker and Heskes, 2003; Allende et al., 2004). The aforementioned method provides a robust estimation of the error when only a small number of training examples are disregarded. In our case, only one random sample was extracted in each sequence. The network was then trained and its result evaluated in order to measure the degree of successful classification achieved for that particular example. This process was repeated 10,000 times and the percentage of successful events was calculated for the total amount of tests. This type of training simulates the variability of the information included in the sample and provides an enhanced adjustment of the neuronal network for the subsequent classification of the remaining routes.

Application

The previous methodology has been applied on the set of 160 urban bus lines in Madrid (Spain) and makes use of the cluster structures defined in Jiménez et al. (2013), in which a sample of 25 lines with known kinematic cycles was analyzed. These 160 lines correspond with the total amount of bus lines in Madrid, with the exception of the night and special services. The previously defined macroscopic variables were obtained from several sources:

- Representation of the bus line itineraries in a GIS using the mapping information from the Spanish National Geographic Institute.
- Line management data provided by the Municipal Bus Company of Madrid (EMT).
- Street information made available by the Madrid Council and the Statistics Institute of Madrid.

The evaluation of the quality indicators used during the clustering process suggested that the sample of 25 lines could be arranged in 8 clusters (Jiménez et al., 2013), so this classification is used as a basis in order to allocate the remaining 135 lines. Nevertheless, the results indicate that the neural network provides excessively poor results, with approximately 19% of successful classifications compared to the 12.5% that would be obtained using a random classification. This circumstance is caused by the limited information contained in the macroscopic variables, as it does not sustain such an accurate characterization of the lines, with an added difficulty, since some clusters contain only 1 or 2 lines, thus complicating the application of the method and the learning process of the neural network. Furthermore, the effect of the reduced number of lines in the sample which is used during network training increases when so many categories are feasible for the output variable. Consequently, the number of clusters is reduced using the hierarchical method with the selection criterion that maximizes the least unfavorable combination of route groupings as shown in Jiménez et al. (2013). After performing a diversity of tests, the 3 cluster option is preferred because this grouping integrates within the same cluster both traveling directions (considered as different lines) of the single loop route in the city.

Initially a perceptron network without hidden layer was trained in order to evaluate to which extent the problem is linear, and the result provided 55% of successful assignments. The introduction of a hidden layer improved the proportion to values above 60%. The success ratios using the bootstrap method are summarized in Table 2.

These results reveal that the problem is not linear; therefore, a multi-layer perceptron is an appropriate tool for the classification of this information. Several architectures were tested and the structure with the best success ratio with the

bootstrap method was selected (architecture 19-9-3). Its 60.9% success ratio, although still moderate due to the limited information contained in the macroscopic variables, improves considerably the percentage of success that is expected when classifying randomly in 3 groups, which is 33.33%. Furthermore, the Central Limit Theorem substantiates that mistaken assignments conform a normal distribution with a standard deviation of 0.0049, thus deducing that the percentage of success is included in the interval [59.95, 61.85] with a confidence level of 95%.

Nevertheless, it is relevant to analyze the confusion matrix (Table 3). The data indicates that the network commits assignment mistakes between clusters 1 and 2 (around 50% classification hits), while assignments to cluster 3 seem to be suitably reliable (above 90% classification hits). This outcome is reasonable because clusters 1 and 2 contain primarily urban lines, whereas cluster 3 is composed of lines with a significant presence of periurban or higher speed roads in their itinerary. Accordingly, the network is capable of differentiating with minor errors those lines with periurban journeys, but the information contained in the macroscopic variables is insufficient in order to discriminate both groups of urban lines adequately.

As a result, the perceptron network was trained with the data gathered from the 25 lines contained in the sample. A sensitivity analysis is implemented in order to determine the relative significance of each input variable. The calculation of the sensitivity for the input variables is carried out as follows. Firstly, every single input is fixed to its average value (computed from all of its possible values in the training set). Then, for each input variable, the output is computed by setting the maximum and minimum values in the input variable involved, maintaining the others at the average value. Both outputs represent the change that caused a single variable in the classification. At this point we have obtained the variation produced for each input variable on the output of the network when varying each input variable between its maximum and minimum values. Finally, these values are normalized by dividing by the maximum range obtained, which is the variable that produces higher output variation. Thus, we can compare the variation produced by each of the input variables. Table 4 shows the relative importance of each input variable over the output, 1 representing maximum influence and 0 no influence.

There are 6 variables that exercise a significantly higher influence than the rest. These correspond to:

- Average speed, which is higher in peripheral lines.
- The percentage of the route's distance driven along district or neighborhood limit streets, since these are main roads and usually have a higher number of lanes than other less significant streets.
- Distance index, which discriminates lines that travel far-off from the city center from others that remain in the proximity.
- Number of bus stop coincidences with Metro stations and lines per km, considering that concurrence occurs in areas with high passenger transit.

However, circulation in bus lanes, regardless of which type, is not significant. Neither is the number of bus stops or the number of different streets along a route, which is a variable that is associated with the number of street crossings. That is, other variables related with these have an improved route classification capacity.

The trained network has the capability to classify the remaining 135 lines, for which no kinematic data is available, into 3 final clusters, as shown in Fig. 1.

In addition, the line classification results provided by the preferred 19-9-3 network were compared with those obtained with other network structures, like 19-16-3 and 19-19-3, which also performed adequately during the training process. The number of the line assignment coincidences with the selected network was of 110 (81.5%) and 118 (87.4%) respectively, signifying a high grade of stability within the solution.

Table 2
Percentage of success achieved using the bootstrap method with different neural network architectures.

Network architecture (inputs-hidden layers-outputs)	Bootstrap classification hits (%)
19-3 (no hidden units)	55.0
19-4-3	58.9
19-6-3	58.9
19-9-3	60.9
19-12-3	59.4
19-16-3	60.3
19-19-3	60.2
19-25-3	59.5

Table 3
Confusion matrix (%).

		Are classified by the network in			
		C1 (%)	C2 (%)	C3 (%)	(%)
Lines belonging to	C1	48.1	51.9	0.0	100
	C2	35.6	53.4	11.0	100
	C3	1.3	7.9	90.8	100

Table 4
Sensitivity study on the influence of each macroscopic variable.

	MV	Variable	Range	Mean	Std. dev.	Sensibility indicator
1	MV 19	% of the route distance driven along district or neighborhood limit streets	3.8–94.6%	38.0%	19.5%	1.00
2	MV 1	Average speed	9.4–33.7	13.8	3.5	0.96
3	MV 4	Coincidences with Metro stations per km	0.1–1.8	0.7	0.4	0.92
4	MV 5	Coincidences with Metro lines per km	0.09–3.21	1.10	0.7	0.87
5	MV 16	Distance index to the city center	1–3	2.7	0.6	0.84
6	MV 17	% of the route distance driven along district limit streets	0–94.6%	13.6%	16.6%	0.80
7	MV 9	% of the route distance driven along bus lane with barrier	0–35.8%	4.4%	7.2%	0.56
8	MV 14	Distance of the furthest route terminal to the city center	2300–12565	6919.5	2303.8	0.53
9	MV 10	% of the route distance driven along independent bus lane	0–29.2%	1.3%	4.6%	0.51
10	MV 3	Number of passengers index	6.96–85.0	42.3	16.5	0.51
11	MV 5	Number of streets per km	0.4–4.3	2.1	0.6	0.35
12	MV 4	Number of bus stops per km	0.4–4.3	3.3	0.6	0.29
13	MV 2	Bus frequency	3–3	9.3	4.1	0.24
14	MV 13	Distance of the nearest route terminal to the city center	47–8748	2984.8	2099.5	0.22
15	MV 11	% of the route distance driven along bus lane	0–77.4%	13.3%	15.3%	0.17
16	MV 8	% of the route distance driven along conventional bus lane	0–41.7%	7.6%	9.4%	0.11
17	MV 12	Streets fiscal index	1.6–7.8	4.39	1.5	0.10
18	MV 18	% of the route distance driven along neighborhood limit streets	0–84%	24.3%	17.8%	0.08
19	MV 15	Proximity index to the city center	1–3	1.6	0.8	0.03

	Clustering of 25 routes	Clustering of 160 routes
Cluster 1 (red)	40.0%	36.9%
Cluster 2 (green)	36.0%	13.7%
Cluster 3 (blue)	24.0%	49.4%

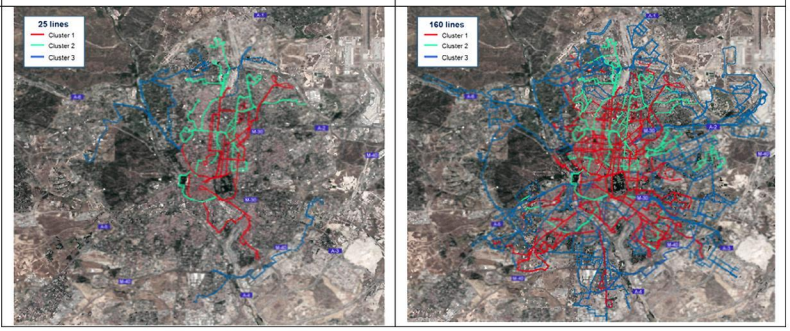

Fig. 1. Mapping of the lines assigned to each cluster.

Conclusions

This article proposes a method to classify urban bus lines into clusters considering only macroscopic variables in a process that is independent from the procedure used in the generation of the initial groups. These clusters were created in order to group routes with similar kinematic behavior in a limited sample obtained from the totality of the lines. Nevertheless, in case it is unviable to gather detailed kinematic information for the complete set of lines of the city, a neural network is developed with the purpose of processing input data consisting of macroscopic information related with the operation, the infrastructure and the city's zonal distribution. This approach improves other proposals that generate clusters based on macroscopic information only and do not assess kinematic data during the process. The method has been applied using real data obtained from the urban bus operation in Madrid (Spain), so microscopic data for the first classification described in Jiménez et al. (2013) were obtained using onboard recording equipment and macroscopic data for the second line-clustering step were retrieved from the Spanish National Geographic Institute, Municipal Bus Company, the Madrid Council and the National Statistics Institute databases. The proposed method provides satisfactory results even if the training set is limited, although the reliability of the outcome is highly dependable on the quality of the information contained in the selected macroscopic variables. Since classification using macroscopic data mislays many details of each line's operation, the microscopic

classification that is previously carried out should provide clusters with very different characteristics, because the network cannot distinguish slight differences considering the features of the information being managed. Furthermore, selecting significant macroscopic data is essential for obtaining good results.

The results of this classification can be used to assign different bus types to each route considering its characteristics because each type of vehicle could have a different performance depending on the route's properties. One of the final intentions of such classification could be the analysis of fuel consumption and exhaust emissions emitted by an urban bus fleet when detailed emission models that consider driving cycles are used. The classification could also be utilized to undertake a reorganization of the vehicle fleet servicing the routes, considering that some bus types are more convenient for specific routes due to the peculiarities identified when evaluating each line's representative kinematic cycle, so routes should first be classified in homogeneous groups. This reorganization would derive in fuel consumption savings and exhaust emission reductions since, depending on the characteristics of the vehicles, their use in some itineraries may be more adequate than in others.

Acknowledgments

The authors thank the Madrid EMT for providing the necessary data for this study.

References

Allende, H., Ñanculef, R., Salas, R. 2004. Robust bootstrapping neural networks. In: Lecture Notes in Computer Science. In: MICAI 2004: Advances in Artificial Intelligence, vol. 2972, pp. 813–822.

André, M., Villanova, A., 2004. Characterization of an urban bus network for environment purposes. Sci. Total Environ. 334–335, 85–99.

André, M., Garrot, B., Roynard, Y., Vidon, R., Tassel, P., Perret, P., 2005. Operating conditions of buses in use in the Ile-de-France region of France for the evaluation of pollutant emissions. Atmos. Environ. 39, 2411–2420.

Bakker, B., Heskes, T., 2003. Clustering ensembles of neural network models. Neural Netw. 16 (2), 261–269.

Beniwal, S., Arora, J., 2012. Classification and feature selection techniques in data mining. Int. J. Eng. Res. Technol. 1 (6).

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK.

Chen, H., Namdeo, A., Bell, M., 2008. Classification of road traffic and roadside pollution concentrations for assessment of personal exposure. Environ. Modell. Softw. 23, 282–287.

Fielding, G.J., Brenner, M.E., Faust, K., 1985. Typology for bus transit. Transport. Res. Part A: Gen. 19 (3), 269–278.

Fotouhi, A., Montazeri, M., 2013. Tehran driving cycle development using the k-means clustering method. Int. J. Sci. Iran., Part A. 20 (2), 286–293.

Franke, J., Neumann, M.H., 2000. Bootstrapping neural networks. Neural Comput. 12 (8), 1929–1949.

Giuliano, G., 1981. The effects of environmental factors on the efficiency of public transit service. In: 60th Transportation Research Board Annual Meeting, January 1981, Washington, DC, USA.

Hagan, M.T., Demuth, H.B., Beale, M., 1996. Neural Network Design. PWS Publishing Company, Massachusetts, USA.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31 (3), 264–323.

Jiménez, F., Román, A., López, J.M., 2013. Methodology for kinematic cycle characterization of vehicles with fixed routes in urban areas. Transport. Res. Part D: Transport Environ. 22, 14–22.

Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation and active learning. In: Touretzky, D.S., Tesauro, G., Leen, T.K. (Eds.), Advances in Neural Information Processing Systems, 7. MIT Press, Cambridge, MA, pp. 231–238.

Naranjo, J.E., Jiménez, F., Serradilla, F.J., Zato, J.G., 2012. Floating car data augmentation based on infrastructure sensors and neural networks. IEEE Trans. Intell. Transp. Syst. 13 (1), 107–114.