Emerging Cutting-Edge Developments in Intelligent Traffic and Transportation Systems
M. Shafik (Ed.)

doi:10.3233/ATDE240016

Analysis and Characterization of Vehicular Traffic in a Low Emission Zone

Laura SALDARRIAGA[a,1] and Gustavo PATIÑO[b]
[a, b] *Faculty of Engineering, Universidad de Antioquia*

Abstract. Traffic congestion is one of the biggest problems for transportation, since it carries high costs for cities and health risks due to its impact on air pollution. Therefore, several strategies have been proposed to improve mobility considering pollution reduction as a critical factor. To accomplish that, methods like forecasting and traffic light control systems are used for traffic management and for them to be more efficient, prior analysis of traffic data is needed. This paper is focused on that analysis for a Low Emission Zone located in Medellin, Colombia. The most relevant characteristics related to the temporal component and variables of the dataset are visualized and analyzed through statistical methods such as correlation and visualization approaches such as PCA and bar plots, also used for feature selection.

Keywords. Vehicular Traffic, Low Emission Zone, Time Series, Principal Component Analysis (PCA), Correlation

1. Introduction

Traffic congestion refers to the excess of vehicles in a portion of the road, caused by various factors, including the increase in the number of vehicles, traffic accidents, and weather conditions [1], resulting in slower traffic speeds and leading to situations of stopped traffic, extending travel times, and affecting the quality of citizens' life.

This issue not only affects mobility but impacts air quality and health by emitting polluting particles during engine combustion processes, a phenomenon that worsens in situations of high traffic congestion [2]. Given this relationship between the transport and environmental pollution, large cities such as Rome, Milan, and London began to engage in more conscious mobility planning through strategies such as higher parking fees to incentivize the use of public transportation, standardizing speeds for specific urban zones, definition of restricted traffic zones during specific hours [3], and tagging Low Emission Zones (LEZ) [4, 5], defined as geographic delimitations of areas on which different actions are planned and executed to improve air quality and protect the health of citizens by reducing pollutants produced by the vehicle fleet [6].

To accomplish an adequate implementation of the above, a prior analysis of traffic data with variables such as traffic flow, speed, travel time, density, and the temporal component of data (periodicity, time variation, trends, etc.) is needed. This can be performed with different methods, e.g. by using statistic information like distribution, average, Principal Component Analysis [7], and correlation [8], commonly supported by

[1] Laura Saldarriaga, Calle 67 No. 53-108, Medellín, Colombia; E-mail: laura.saldarriagah@udea.edu.co.

different kind of graphs. The obtained results are used to plan and establish more efficient strategies with a bigger impact on air quality.

On account of the importance of data analysis of traffic data for a subsequent approach to traffic management strategies and considering that the city of Medellín has established a LEZ, this paper describes a traffic characterization for said zone, following the next outline: Section II presents some previous work related with Low Emission Zones. Further on, Section III describes the case study LEZ, the dataset and the treatment given to data, considering preprocessing and cleaning procedures performed prior analyses and visualizations of Section IV which mainly focus on the temporal component of data, considering aspects such as periodicity, monthly behavior, and changes in traffic during holidays. Then, in Section V statistical and visualization methods are used to determine the main features of the dataset. Finally, conclusions are presented on the analysis carried out regarding the temporal component and the selected features.

2. Related work

Protected air zones in urban areas have mainly implemented in European cities to reduce emissions of pollutants such as PMx, CO2, and NOx [9]. To achieve this effectively, traffic and road infrastructure have to be characterized and two important components for this task are monitoring and measuring traffic. For this, strategies such as vehicle counting, identification of the type of vehicles, and variables such as circulation speed [10], lane occupancy, and traffic volume, combined also with temporal information, such as the day of the week [11], have been measured.

For example, in the city of Brussels, the estimated pollutant emissions, the number of vehicles and the average speed of the vehicle fleet were collected from a remote sensing system [5] to identify the most recurring vehicle types, as well as the estimation of the pollutants emitted. This provided important information for the first implementation of the LEZ, so that after a period of operation, its impact on the air quality of the region was evaluated.

On the other hand, the city of Lisbon considered the characterization of the vehicle fleet based on vehicle counting, age of the vehicles, and interviews with the drivers to estimate the effects of the introduction of the LEZ. With this information they evaluated daily traffic in three areas of the city and concluded that the LEZ could be more efficient in reducing PM10 than NOx and that emission reduction does not only depend on the number of vehicles but also on the type of vehicle, speed, and distance traveled [4].

In this article, we describe a traffic characterization for the city of Medellín, carried out based on available data, and the identification of relevant traffic, temporal variables, and their influence on the LEZ of the city [12], as well as the most important features that could be used for future technology-based strategies to improve mobility in this Zone.

3. Low Emission Zone

In 2018, the Aburrá Valley Metropolitan Area (*Área Metropolitana del Valle de Aburrá*, AMVA) [13] defined a LEZ based on the traffic monitoring station called *Tráfico Centro* [12], which collected data between 2014 and 2017. This LEZ is delimited from south to north between streets *San Juan* and *Calle Echeverri* and from west to east between *Avenida del Ferrocarril* and *Carrera Girardot*, covering the area of downtown Medellín.

As part of the initiatives planned for this area, various entities in the environmental and mobility areas are looking for strategies to reduce emissions from mobile sources, supported by different organizations, like CITRA, which stands for *Centro Integrado de Tráfico y Transporte* (Integrated Traffic and Transportation Center). It is an agency that processes traffic data and information to take data-driven decisions to manage mobility through technological infrastructure, and information systems [14]. This organization provided a dataset, used to characterize vehicular traffic in Medellín's LEZ. These data are described in the next section.

3.1. Description of the dataset

A time series is a set of data collected and stored over time, as a product of monitoring processes for specific events [15]. This is the case of the available traffic dataset, which contains 5,290,071 observations of different variables resulting from the capture of information through cameras located in nine roads of the LEZ between 2020 and 2023. The dataset contains the following fields:

Table 1. Fields of the traffic dataset shared by CITRA

Variable	Description	Variable	Description
Class_1	No. of vehicles with a length between 0 and 3 meters	**Direction of travel**	Direction in which data is captured (NS, SN, WE, EW)
Class_2	No. of vehicles with a length between 3 and 6 meters	**Date_time**	Timestamp in YYYY-MM-DD format
Class_3	No. of vehicles with a length higher than 6 meters	**Road**	Name of street or avenue
Class_4	No. of motorcycles	**Records**	No. of captured records
Speed	*Speed* in km/h	**Occupancy**	Lane occupancy percentage
Location	Latitude and longitude	**Intensity**	Total number of vehicles

3.2. Data processing

Real-world data usually contains missing fields, irrelevant data, and errors. Therefore, it has to be properly preprocessed [16] to obtain reliable results and take better decisions.

After an initial analysis, observations from the year 2020 were removed due to their atypicality attributed to the mobility restrictions imposed during Covid-19 quarantines. Afterwards, observations with null data were removed and outliers were filtered. Also, some columns were added to the dataset (hour, day, month, year, holiday indicator) for subsequent visualization and analysis.

Additionally, considering that the received information is a product of grouped data (records), the average was calculated for variables *Class_X* and *Intensity* by dividing the number of vehicles per observation by the number of records.

4. Data visualization and analysis

Analyses presented below will mainly focus on the features related to lane occupancy, speed, and vehicular intensity to describe the behavior of traffic, as they have a direct relationship with this phenomenon [11] and are also used in tasks like vehicular traffic forecasting, among others [17].

4.1. Time series decomposition

In the case of the dataset described in Section III, a decomposition shown in Fig. 1 was made to observe how the measurements change over time for *Occupancy*. From top to bottom, the first graph illustrates the values taken by the variable over time, the second shows the trend for the variable, and the third shows the seasonal behavior of data.

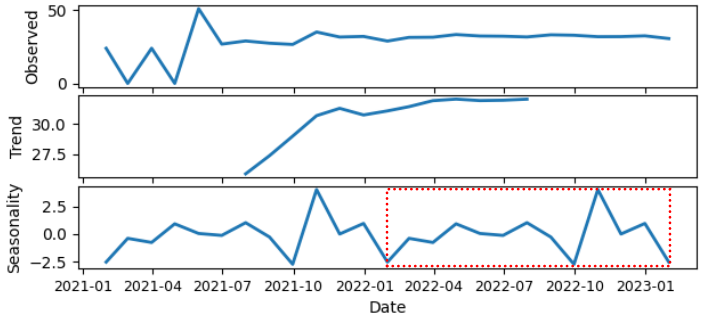


Figure 1. Temporal decomposition for the *Occupancy*

The same procedure was carried out for *Intensity*, which presented a similar trend, but less fluctuating than *Occupancy*; and for *Speed*, that reached its maximum values in April but decreased in May before increasing towards another peak between June and July, where they finally stabilized (Their graphs are not shown due to space constraints).

As for the seasonal component, the three variables presented a cyclical behavior i.e., the values they take form patterns that repeat at the same periods of the following year. The pattern of *Occupancy* is indicated by the red dashed rectangle in Fig. 1.

4.2. Analysis by month and day of the week

Since data in the set presented cyclic behavior over time and vehicular traffic has temporal variations, monthly averages were calculated for *Occupancy* and *Speed* for the years 2021, 2022, and January 2023 to identify the most critical months for mobility in the LEZ and to evaluate how traffic varies between weekdays.

In the case of Occupancy, there was a general increase between 2021 and 2022, except for the months of May and October, which reached higher values in 2021.

Considering 2021 and 2022, January was the month with the lowest average lane occupancy and October presented the highest. Similarly, when analyzing the monthly variation of the *Speed* variable for the same period, there is an irregular month-to-month variation, however, considering the monthly averages for the years 2021 and 2022, August was the month with the highest circulation speed, followed by July. *Intensity* reached its highest averages in December and May.

When observing the behavior of these variables on different days of the week, Sundays and Mondays presented the highest circulation speeds, while Fridays and Tuesdays were the days with the lowest. For *Occupancy* and *Intensity* variables, the days with the highest values were Fridays and Saturdays, while the lowest values were obtained for Sundays and Mondays.

Finally, to identify if there is a significant difference in mobility between regular days and holidays, monthly and daily averages were evaluated for *Occupancy*, *Intensity* and *Speed* variables, reflecting lower occupancy and intensity during holidays for all

months of the year. As for the average speed, higher values have been presented on holidays. In this case, no observations were obtained for February, April, and September.

4.3. Analysis by roads

Although all data capture points are in the LEZ, each road may exhibit different behavior, reflected in traffic descriptor variables (*Occupancy*, *Intensity*, and *Speed*), as well as in the type of vehicles that circulate.

When evaluating average values of *Occupancy* and *Speed* for LEZ roads, *Calle 57 - Avenida Oriental* and *Avenida Oriental - Calle 52* presented the highest occupancy values. On the other hand, *Avenida Oriental* recorded the highest speeds. In both cases, for the highest average values of *Occupancy*, very low speeds were identified, and vice versa. This relationship was not as clear when comparing *Intensity* and *Speed* variables.

Fig. 2 shows the average amount of records identified in each road, according to the category indicated by the label *Class_X*, showing that *Class_1* vehicles are the most circulating in all roads of the LEZ, notably surpassing the number of vehicles of the other categories. In addition, the amount of motorcycles (*Class_4*) is low compared to the other types of vehicles, as their presence is only evident on three roads: *Carrera 43 - Girardot*, *Avenida Oriental - Calle 52*, *and Carrera 57 - Avenida Oriental*. Larger vehicles (*Class_2* and *Class_3*), circulate mainly on *Avenida Ferrocarril - Calle 48*.

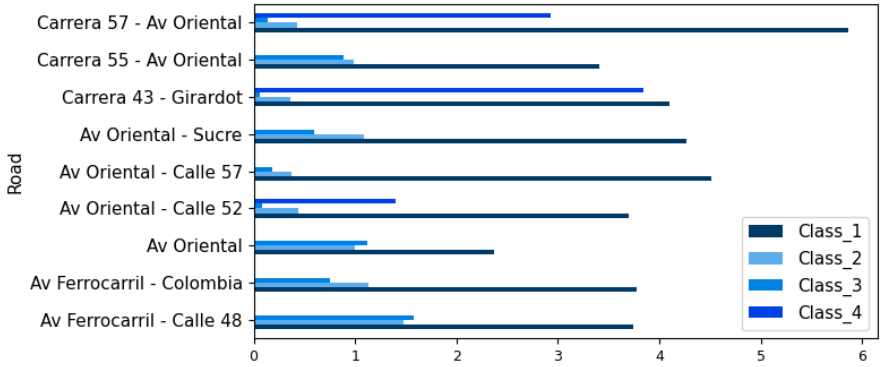


Figure 2. Average vehicle count circulating in LEZ according to *Class_X* variables

5. Feature selection

Feature selection is a process that involves reducing the number of variables, so that the most consistent and relevant ones are identified. This avoids redundancy of features and reduces the computational cost during the execution of processes and algorithms. For example, if predictive models were implemented, irrelevant features for the variable to be predicted would not be included [18], improving their performance and accuracy. For the LEZ data, this selection was carried out through different methods, described next.

5.1. Correlation

Correlation is a statistical measure that indicates the relationship between two variables through a coefficient that ranges from -1 to 1, where a magnitude close to 1 indicates that

the variables are highly related in a linear manner, either directly (if positive) or indirectly (if negative) [8]. For the case study dataset, correlation coefficients obtained for the quantitative variables are shown in Fig. 3 and the pairs that present significant coefficients are highlighted with red boxes.

The results indicate that *Class_1* vehicles (length between 0 and 3 meters) and the *Intensity* have a strong correlation (0.83). This may be an indication that this category has the highest circulation in the area, followed by vehicles of *Class_2* and motorcycles (*Class_4*). In addition, *Occupancy* and *Intensity* variables present a significant correlation coefficient (0.49) that accounts for direct relationship between them.

On the other hand, despite not having a high degree of correlation, *Occupancy* and *Speed* show an inverse relationship, indicating that as lane occupancy increases, the speed of circulation tends to decrease.

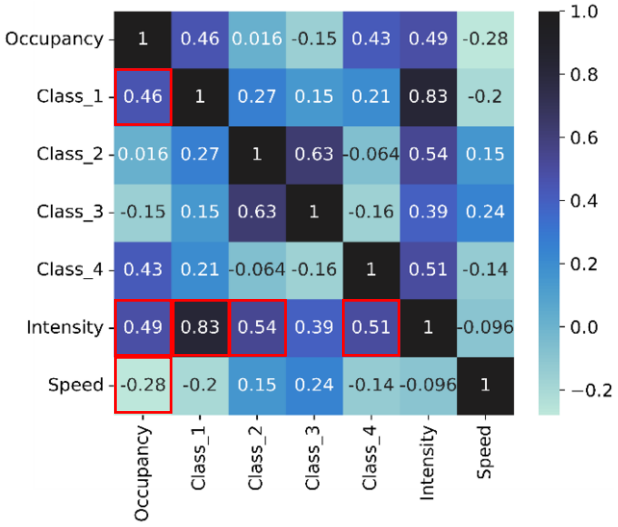


Figure 3. Correlation heatmap

5.2. Principal Component Analysis

Considering there are significant correlations between some of the variables, it is feasible to apply PCA. This is a technique used for dimensionality reduction in large datasets through the creation of new variables or principal components (PCs) that are linear functions of the original variables, preserving as much information as possible. Generally, the components used are those that can explain 80% of the total variance [7].

PCA components were obtained for the LEZ dataset using Python's Scikit-learn [19] for *Occupancy*, *Class_X* variables, *Intensity*, and *Speed* and the most significant component was PC1 with 86% of accumulated variance.

PCA can also be used to identify the importance of variables in a dataset in terms of their contribution to the principal components. This contribution can be evaluated with the variable weights: weights close to 1 or -1 indicate that the variable significantly influences a component. Fig. 4 shows the contribution of each variable to PC1, where the one that represents the highest contribution for the case study data is *Occupancy*.

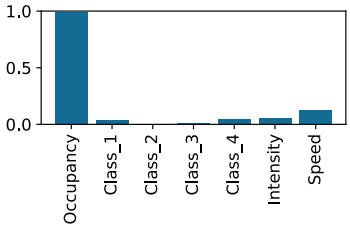
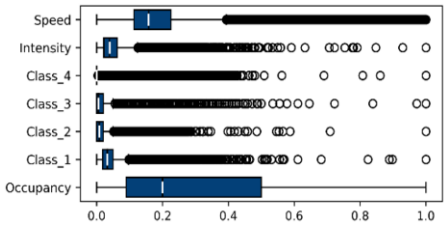


Figure 4. Contribution of the variables to the first component (PC1)

5.3. Box and whisker plot

Box and whisker plots are graphs used for the identification of distribution, minimum and maximum values, outliers, and bias of data [20]. They are used to compare variables, using the medians of their boxes as a reference: if the median line of one box protrudes beyond the limits of another box, there may be differences between the two variables.

In view of the above, box and whisker charts were plotted (Fig. 5(a)) using Python's Matplotlib [21] and the results are summarized in Fig. 5(b), where the Xs represent that the two variables being compared may have notable differences between them.



(a)

	Occupancy	Intensity	Speed	Class_1	Class_2	Class_3
Occupancy	-	-	-	-	-	-
Intensity	X	-	-	-	-	-
Speed	*	X	-	-	-	-
Class_1	X	*	X	-	-	-
Class_2	X	X	X	X	-	-
Class_3	X	X	X	X	*	-
Class_4	X	X	X	X	*	*

(b)

Figure 5. (a) Box and whisker plot **(b)** Variable comparison for box and whisker plot results

6. Conclusion

Considering the results from the analyses of the LEZ traffic data, monthly and daily variations result in changing traffic behaviors, so it is clear that the temporal component is highly relevant. Also holidays present significant differences when compared to regular days, thus the holiday column will be kept as part of the dataset.

On the other hand, by analyzing the relationships and contribution of variables with correlation and PCA analysis, *Occupancy* was identified as the most relevant feature, followed by *Speed* and *Intensity*. Also, correlation results between *Class_X* variables and *Intensity* could be considered to estimate the proportion of each class of vehicles, since future vehicular traffic management strategies may use information about the vehicles categories that circulate in the LEZ.

Regarding variable distributions, box and whiskers plot shown heavy tails for *Class_X*, *Intensity* and *Speed* due to measurements with very high values. Despite this, they were not removed since they were validated with CITRA as values that are still within the range considered as normal at certain times of the day.

As future work, different mechanisms for vehicular traffic management like traffic light control [22] and traffic prediction [20] will be evaluated for a possible application

for the city's LEZ, considering the features and temporal information found through the previous characterization. For example, if traffic prediction were implemented, *Class_X* and *Intensity* could be used as inputs, since they allow to avoid redundancies due to the differences between target variable *Occupancy* (best descriptor variable for traffic) and inputs when attempting to predict traffic behavior.

Acknowledgements

We thank Universidad de Antioquia for supporting the development of our project through the Research Development Committee (CODI), and CITRA, especially engineer Christian W. Quintero for his collaboration with traffic data and advice.

References

[1] FHWA. Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation; 2020 [Internet]. Available from: https://ops.fhwa.dot.gov/congestion_report/chapter2.htm .
[2] K. Zhang, S. Batterman. Air pollution and health risks due to vehicle traffic; 2014. [Internet]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243514/pdf/nihms641926.pdf .
[3] I. Allegrini, F. Costabile. An ITS based on traffic air pollution control. Air Pollution XII; 2004.
[4] F. Ferreira, P. Gomes, et al. Evaluation of the Implem. of a LEZ in Lisbon. Journal of Env. Prot.; 2012.
[5] Y. Bernard, T. Dallman, K. Lee, I. Rintanen, U. Tietge. Evaluation of real-world vehicle emissions in Brussels, Brussels; 2021.
[6] Energy Saving Trust. Guide to low emission zones [Internet]. Available from: https://energysavingtrust.org.uk/advice/guide-to-low-emission-zones
[7] J. C. Ian T. Jollife. PCA: a review and recent developments. Philosophical Transactions A 2016.
[8] JMP. Correlation [Internet]. Available from: https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-correlation.html.
[9] D. Ku, M. Bencekri, et al. Rev. of European LEZ Policy. Chem. Eng. Trans., vol. 78; 2020.
[10] D. Allende, F. Castro, S. Puliafito. Air Pollution Characterization and Modeling of an Industrial Intermediate City. International Journal of Applied Environmental Sciences, vol. 5; 2010.
[11] J. Ye, S. Xue, A. Jiang. Attention-based spatio-temporal graph convolutional network considering external factors for multi-step traffic flow prediction. Digital Communications and Network, vol. 8; 2022.
[12] Secretaría de Movilidad de Medellín. Zona Urbana de Aire Protegido, Alcaldía de Medellín [Internet]. Available from: https://www.medellin.gov.co/movilidad/gerencia-de-movilidad-humana/zona-urbana-de-aire-protegido-medellin.
[13] Área Metropolitana del Valle de Aburrá. ¿Quiénes somos? [Internet]. Available from: https://www.metropol.gov.co/area/Paginas/somos/quienes-somos.aspx .
[14] Secretaría de Movilidad. Centro Integrado de Tráfico y Transporte [Internet]. Available from : https://www.medellin.gov.co/es/secretaria-de-movilidad/centro-integrado-de-trafico-y-transporte/
[15] Universidad de Sonora. Series de Tiempo [Internet]. Available from: http://www.estadistica.mat.uson.mx/Material/seriesdetiempo.pdf .
[16] Power Data. Calidad de datos en minería; 2016 [Internet]. Available from: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/calidad-de-datos-en-mineria-de-datos-a-traves-del-preprocesamiento .
[17] Shabarek A, Chien S, Hadri S. Deep Learning Framework for Freeway *Speed* Prediction in Adverse Weather. Transp Res Rec 2020;2674(10):28-41.
[18] J. Brownlee. How to Choose a Feature Selection Method. Machine Learning Mastery; 2020 [Internet]. Available from: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ .
[19] Scikit-learn. PCA [Internet]. Available from: https://scikit-learn.org/stable/modules/decomposition.html .
[20] Open Learning. Interpreting data: boxplots and tables [Internet]. Available from: https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/interpreting-data-boxplots-and-tables.
[21] Matplotlib, Boxplots, [Internet]. Available from: https://matplotlib.org/stable/gallery/statistics/
[22] L. Qi, M. Zhou, W. Luan. Emergency traffic-light control system design for intersections. IEEE Transactions on Intelligent Transportation Systems. , vol. 17; 2015.