Spark Big Data Platform to Manage City Traffic

Pilar Rey del Castillo
Instituto de Estudios Fiscales
Avda.Cardenal Herrera Oria 378
28035 Madrid
Spain
{mpilar.rey@ief.hacienda.gob.es}

ABSTRACT: *Electronic sensors are able to generate statistical data based huge amount of data files. This is possible when the traffic dataset are available in open data portal. Electronic data is produced from traffic sensors and serve as a rich source of information, that provides speed, vehicle count and so on. It is important that traffic data needs to be processed at the micro-level using complex workloads. Thus the normal data processing tasks require big data specific tools. Initially we have used a few stages in producing short-term indicators of the evolution of the traffic flow variable in a city using the Spark big data platform. With the help of the data on the sensors' geographical location, the indicators are then analyzed to assess the impact of some recent local government measures used to ease the pollution and traffic flow.*

Keywords: Big Data, Short-term Indicators, Spark Platform, Traffic Measures

Received: 13 November 2019, Revised 10 April 2020, Accepted 20 April 2020

DOI: 10.6025/jitr/2020/11/3/94-104

1. Introduction

The local government of Madrid City offers an open data portal designed for the users to explore and download their publicly accessible data. The datasets available include data from traffic sensors located at strategic points in the roads and streets of Madrid City. These traffic sensors are a rich source of information, providing data not only on the vehicle count, but also, e.g., on its speed and geographical location. There have been a number of studies on traffic sensors[6,5] reporting that they provide, in general, accurate traffic measures.

The volume of the downloaded information cannot be processed using conventional statistical software and requires procedures specifically developed for this purpose. Apache Spark [13], an open source analytics engine for Big Data processing has been used on a single node for the first steps of collecting and pre-processing data. The volume of the downloaded information cannot be processed using conventional statistical software and requires procedures specifically developed for this purpose. Apache Spark [13], an open source analytics engine for Big Data processing has been used on a single node for the first steps of collecting and pre-processing data. The first aim of the paper is to study the traffic in the city from 2016, constructing daily indicators of its evolution. Monitoring the real evolution is a task more difficult than it appears at first glance. In order to obtain good enough indicators and before the final calculations to compute the indexes, it requires various steps to detect and correct logical inconsistencies in the data, impute missing information, and summarize at different granularity levels.

Once the indicators are available, the traffic evolution can be analyzed to learn significant patterns of behavior. The information on the sensors geographical location may help at this stage to discover similarities and differences between zones in Madrid City. On the other hand, combining all these data will allow to evaluate the results of the recent traffic measures taken by the local government addressed to improve the levels of air pollution within the city and surrounding areas.

The remainder of this paper is organized as follows: the next section presents a summary of the steps taken to construct the indicators; section 3 analyses the high-frequency series obtained; section 4 performs the assessment of the traffic measures; and, finally, a number of remarks and conclusions are shown in section 5.

2. Construction of the Daily Indicators

The raw data to be used as source for computing the time series consist on the datasets made available in the portal after the end of each month, including the figures of the previous month, for each one of the more than 4000 sensors, of a number of variables measured in 15-minutes intervals. This makes around 150 million of data points for each year and each variable. Besides the previously cited Apache Spark, the Python software [10]has been used for all calculations and analysis once the indicators have been obtained.

Although the datasets provide information on more variables, this paper only studies a single variable, the intensity measured by the number of vehicles by time unit, as an example of the analysis that could be performed. A daily intensity indicator will be computed for the whole city, and also split into the urban area and the M30 ring road. For this purpose, the calculations are performed in some stages. Given that the names and/or categories of the intensity and sensor type variables have changed in the datasets through the times, the first step of preprocessing is done, treating the data to make them homogeneous. After this, as the daily level of time granularity has been chosen, the total number of vehicles per sensor and day is calculated.

As next stage, data editing must be performed to ensure completeness and validity because the transmission of information from some sensor nodes may sometimes fail. To detect these failures, data with more than a certain proportion of missing information in the readings are not validated. These data together with missing data are imputed by a procedure described later.

Since the intensity of the traffic in a road is defined by the number of vehicles passing the road in a period of time, the natural way to measure the intensity in an area would be by the average number of vehicles in all the roads and streets located in the area. As there are not sensors in all the roads and streets, it could be approximated by the average number of vehicles in all the sensors located in the area during the period. But the transmission of information from some nodes may sometimes fail due to environmental interference, physical damage or lack of power. Therefore, changes in the averages could be motivated by changes in the sensors location and/or activity and not necessarily by changes in the traffic intensity in the area.

Being flow data, a simple aggregative index [9] could be used to compute the evolution of the intensity. Instead, to solve the previous problem in measuring the evolution, the indicators are computed as change estimators or chain-linked index

$$I_t = I_{t-1}.\frac{\sum_k x_{kt}}{\sum_k x_{kt-1}} \tag{1}$$

where the sum is extended to the k sensors having data validated for both periods t and $t-1$. The indexes I_0 for the first period, the first of January 2016, are calculated as the average by sensors in the area of the total number of vehicles this day. Once the indexes of a day are computed and before calculating the indexes of the following day, the sensors having missing data on this day are imputed as

$$\widehat{x}_{it} = x_{it-1} \cdot \frac{\sum_k x_{kt}}{\sum_k x_{kt-1}}$$

where the sum is also extended to all k sensors having data validated for both periods t and $t - 1$. Then the imputed values are validated and the indexes are re-calculated, obtaining the same previous values. In this way, the imputed data are available for the calculation of the following day indexes. It can be shown that using this simple method of imputation, the indexes are always computed using all the information available, and they are not deteriorated by a repeated lack of information on some sensors.

After the imputations are computed in this way, there is a remaining problem: there are days for which there are no data for any sensor and indexes cannot be calculated. The daily changes series are then considered to complete the missing days using time series predictions. The first attempt for forecasting was made using LSTM (Long Short-Term Memory) Deep Neural Networks [8], a class of artificial neural networks that allows exhibiting temporal dynamic behavior. These networks have proven to be able to outperform state-of-the-art univariate time series forecasting methods. However, in our case, having less than 4 years of data, forecasts from ARIMA models, following the Box-Jenkins methodology [1], have obtained better results in terms of minimum mean square error of forecast.

As a final stage, once microdata have been imputed and missing daily changes have been predicted, intensity indicators are computed for the whole city, the M30 ring road and the urban area.

3. High-frequency Series Analysis

Fig. 1 and Fig. 2 show the three daily indicators obtained following the described steps for the period between January 2016 and August 2019. This section presents some features of their behavior from the time series analysis perspective. It can be seen that, for all series, the day-to-day movement has a lot of noise, with a large number of rises and falls, and there is also a clear common pattern of seasonal decreasing in August. Although having a similar evolution, the intensity level is much higher (around 8 times) in the M30 ring road than in the urban area.

In order to extract some meaning from the indicators through the seasonalpatterns, their periodogram spectrum estimates using Welch's method [12] are shown in Fig. 3. The peaks in the spectrum indicate the frequencies of cyclical movements. Being daily indicators, they might potentially have up to 4 periodic components: a weekly cycle (7 days), a monthly cycle

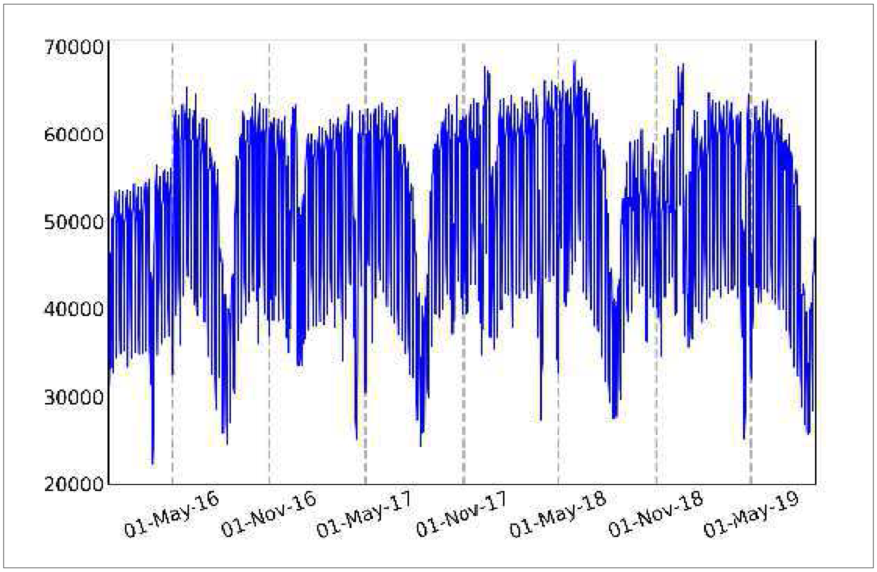


Figure 1. Global intensity indicator

(average length of 30.4369 days), a quarterly cycle (average length of 91.3106 days) and an annual cycle (average length of 365.2425 days). Vertical lines have been added at the frequencies corresponding to annual, monthly and weekly periods (frequencies = 1/number of days per cycle and its harmonics).

Similar behavior can be seen for the three indicators: the highest frequencies correspond to weekly periods, there are small frequencies for annual periods, and the frequencies are only just different from zero for monthly periods. It could be interpreted that the most important cyclical oscillations correspond to weekly periods although these oscillations can hardly be seen in Fig. 1 and Fig. 2 due to the big number of data. Annual oscillations must be taken with caution because there are less than four years of data and they may also be hidden by the 7-day periodic component.

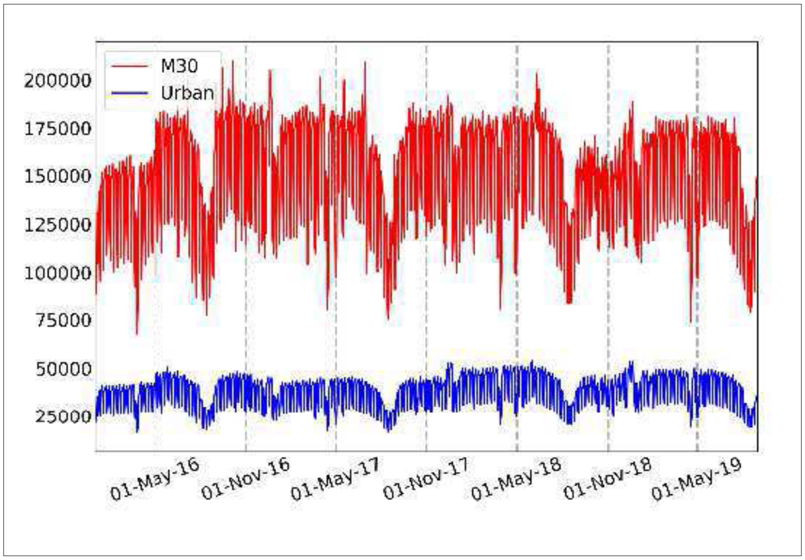


Figure 2. Indicators at M30 and Urban areas

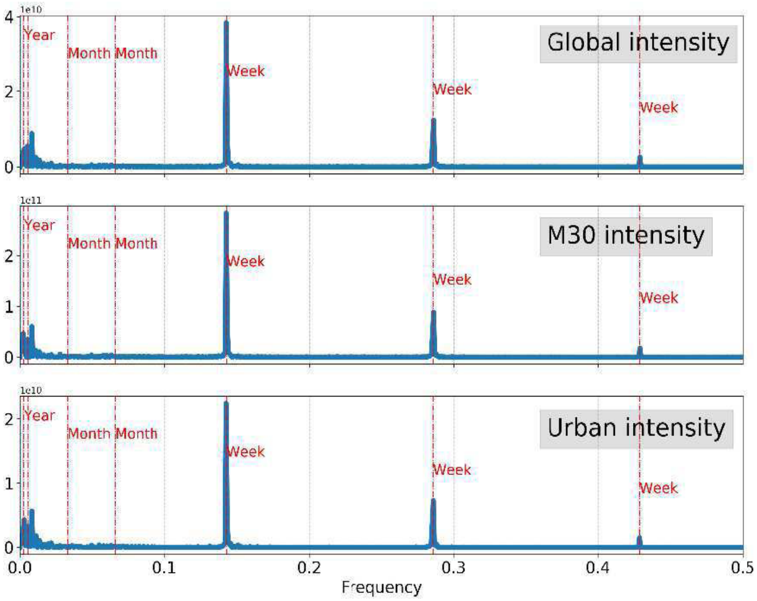


Figure 3. Periodogram spectrum estimates

Even though the temporal granularity chosen is of 1-day intervals, another aspect to consider is the distribution of the vehicles flows within the day. The traffic intensity for the combination day-of-the-week and hour may show interesting patterns. For this purpose, it has been calculated for each sensor the average of the traffic intensity per day of the week and hour, and later these averages have been divided by the maximum found traffic intensity at this sensor in an hour. The method provides

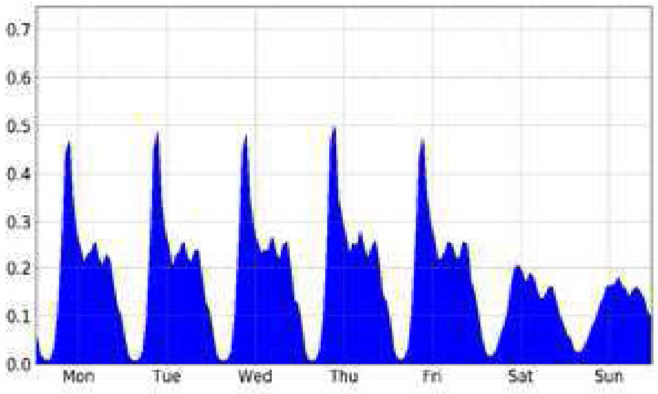


Figure 4. Example of the weekly profile for a particular sensor

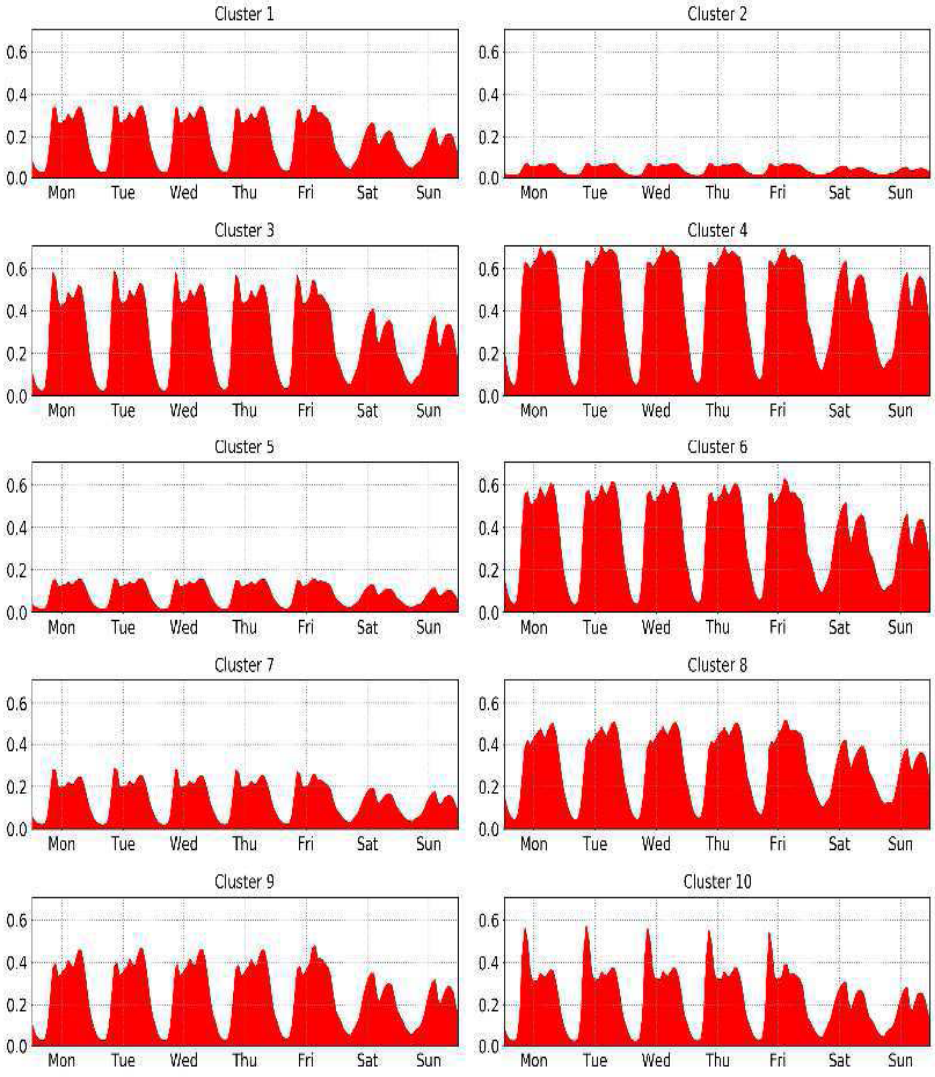


Figure 5. Cluster centers of the traffic intensity weekly profiles

an approximate idea of the average level of occupancy during the week of the road or street on which the sensor is located.

Fig. 4 shows an example of the profile for a particular sensor (tick marks indicate noon for each day) where it can be seen the decay on weekend and a peak around 9 a. m. each weekday. These profiles form 168-dimensional points. Clusters of these points using the K-means algorithm and the Euclidean distance [4] have been built to explore and summarize the results. Fig. 5 shows the centers of the clusters for k = 10 clusters.

Although the elbow method [11] to determine the optimal number of clusters is not totally conclusive, this number has not a big impact on the results: similar graphs and conclusions could be obtained with another number of clusters. Asgeneral patterns for all roads or streets, besides a decay on weekends, it is found that the traffic intensity decreases during night hours (from 1 to 5 a. m.), especially on weekdays, and that there are generally decaying around noon and 3 p. m. Besides these general features, there are big differences between the levels of occupancy, extending from light in clusters 2 and 5 to heavy in clusters 4 and 6. It can also be seen that sensors in clusters 3 and 10 have maximum traffic on weekdays at morning commuting hours, while sensors in 4, 8 and 9 have the top at afternoon hours. Therefore, there are two aspects that may characterize the sensors weekly behavior and may be of interest to explore and describe: the global level of occupancy, and the time of the day at which the intensity on weekdays is the highest.

Instead of visually studying the graphs to assign a level of occupancy for each sensor, they are automatically classified into three levels, depending on the computed area under the normalized by the maximum weekly profile curve. Fig.6 shows the average level of occupancy obtained from the sensors in Madrid City Fig. 7. Weekday profiles of usage boroughs. It can be seen that most of the areas with Light traffic intensity are outside the central part of the city.

Similarly, the sensors can be automatically classified into three groups depending on the time of the day at which the intensity on weekdays is the highest (a sensor belongs to Morning commuting/Afternoon group when its average for weekdays exceed by more than 20% the Afternoon/Morning commuting average, respectively, being Morning commuting between 7 and 9 a.m. and Afternoon between 2 and 9 p.m.; otherwise belongs to All day group). Fig. 7 provides an idea of the typical weekday profile of usage of the roads and streets.

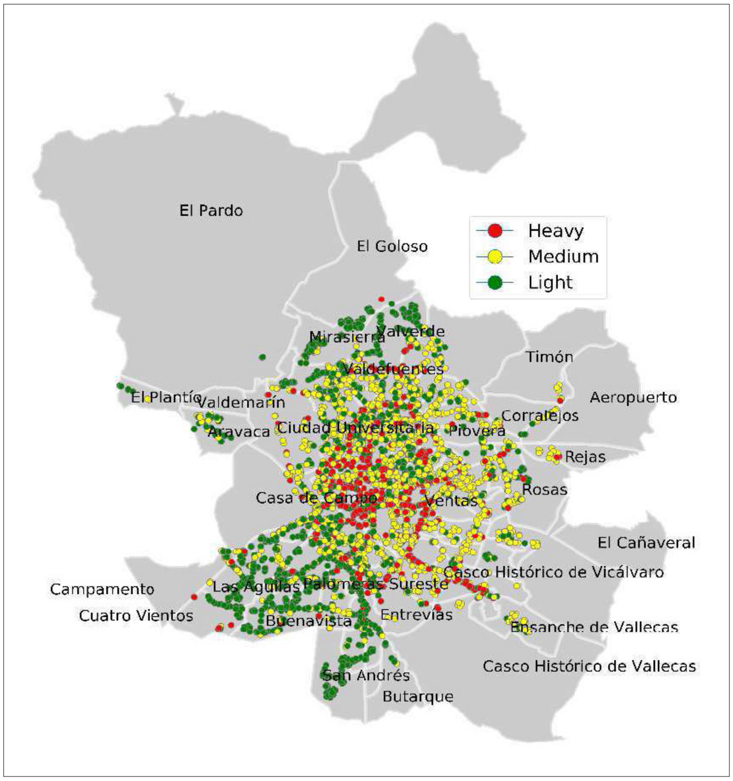


Figure 6. Average levels of occupancy

4. Assessment of the Impact of the Traffic Measures

The local government of Madrid City has taken in the last years, some measures addressed to reduce pollution. Although the current understanding of the air pollution impacts from traffic congestion on roads is limited [14] , it seems that vehicle emissions and traffic-related pollution are typically one of the largest contributors to air pollution in cities. This paper studies just one variable, the traffic intensity, and, consequently, the evaluation refers exclusively to the effects on traffic reduction, and not directly to the effects on air pollution. The most important traffic measures taken may be summarized in Table 1.

As the measures have been gradually taken, a first assessment of the impact on the whole city can be done from the annual average rates in Table 2. The global indicator reflects the behavior of the whole Madrid City area and the other indicators (M30 and Urban) extend also over all area. For this reason, it is not likely to find any effect of the traffic measures because they refer to only some zones and there may also exist opposed effects in other parts.

To check the hypothesis of a possible effect on any of the indicators, ARIMA models with intervention analysis [2,3] have been used. Thus, a basic multiplicative ARIMA model with weekly seasonality has been fitted to each series using the Scikit-learn software library [7] . There have also been included as regressors some additive outliers and a specific variable to measure the effect of Easter, a relevant moving holiday for daily data. Then, different intervention variables, trying to gather the effects of the traffic measures (with different structures and different dates) have been tested. But the value of the corresponding parameter estimates has never been significantly different from zero.

In any case, the assessment must be better referred to zones that can be affected by the measures. The information about the geographical location of the sensors, provided also in the open data portal of Madrid City, can be used. Two zones probably affected have been considered: Madrid Central, the area with borders defined by the local government and which some of the traffic measures refer to, and another area defined as a crown of 300 meters surrounding Madrid Central, which will be named Crown. The delimitation of the zones appears in Fig. 8.

Date	Traffic measure
December 2016	Sporadic restrictions to private vehicles in some parts of the city center
December 1, 2017 to January 8	Restrictions to private vehicles in Gran Via and, sporadically, in other central areas
April 2018 to November 2018	Works for the reduction of the number of lanes in some of the main tracks
November 30, 2018	Starting of Madrid Central, a new big restricted area with cameras monitoring license plates of vehicles entering (without penalties)
March 16, 2019	Starting of Madrid Central (with penalties)

Table 1. Traffic measures taken in the last years

Year	Global	M30	Urban
2017	4.2	4.0	-0.4
2018	1.8	-1.9	10.1
Jan-Aug 2019/ Jan-Aug 2018	-3.6	-3.2	-3.8

Table 2. Average annual increase rates

What can be done now is to compute new indicators, following the rules in section 2, for the two zones, including in each one the data of the sensors within the corresponding area. Thus, intensity indicators for Madrid Central and Crown zones appear in Fig. 9 and Fig. 10.

For a first assessment, Table 3 shows the annual average rates where now possible effects appear. There is a gradual reduction in Madrid Central, probably reflecting the cumulative effect of the different measures. The Crown area, on its side, shows a clear increase in 2018, result of a plausible substitution or border effect. Nevertheless, this may revert as a result of the last traffic measures in 2019.

Fig. 11 and Fig. 12 present the corresponding monthly average and monthly average annual rates, respectively, of the traffic intensity at Madrid Central and Crown zones.

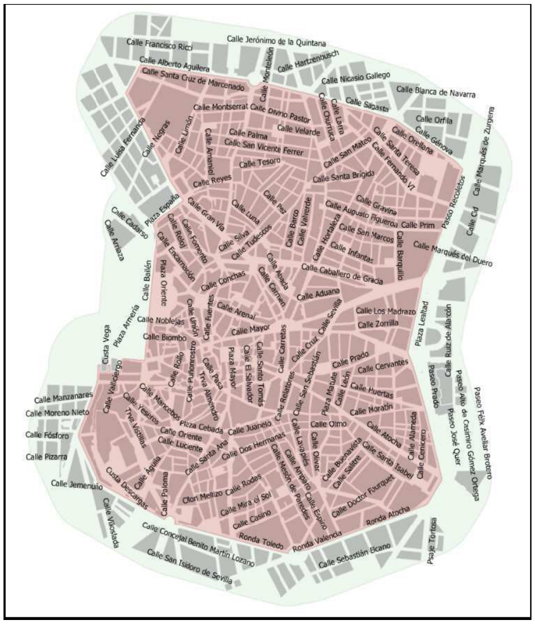
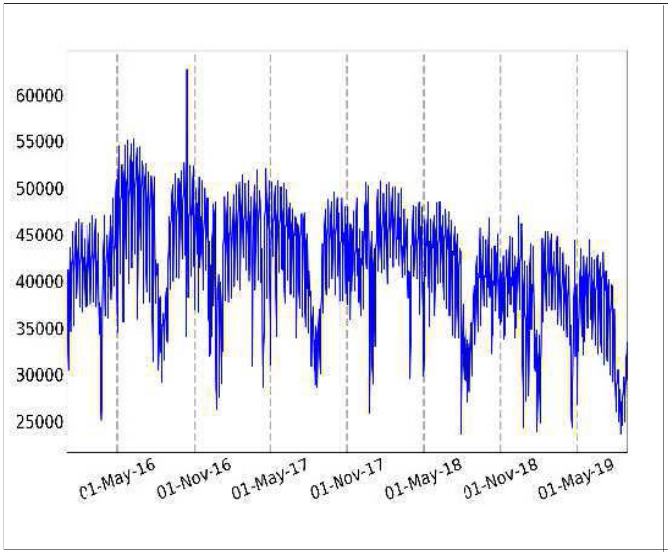


Figure 8. Madrid Central area (red) and Crown (green)



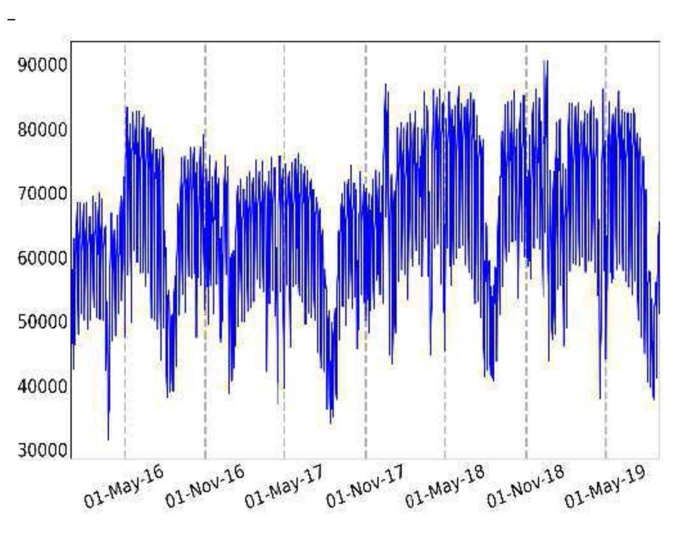


Figure 9. Madrid Central intensity indicator

Figure 10. Crown intensity indicator

With the aim to provide more detailed explanations, both series have been treated in a similar way to the previous for finding possible effects of the traffic measures. That is, basic multiplicative ARIMA models [2,3] with weekly seasonality, Easter variable, and additive outliers have been fitted, and later different intervention variables have been tested using the Scikit-learn [7] software library. Although at first glance, from Fig. 12, one of the most important measures, the starting of Madrid Central in March 2019, seems to be having some effects (both indicators show annual decreases from April 2019), no significant effects have been found. Nor have any other significant interventions related to the traffic measures been found, probably because of their gradual implementation that may be described by the ARIMA model.

Another interesting analysis to perform is to see whether there has been any effect on the weekly patterns of behavior for the roads and streets located in both areas. To simplify the study, the period since the complete implementation of all measures, (starting in March 16, 2019) is compared to an equivalent period in 2017 (March 16, 2017, to August 30, 2017), when hardly any traffic measure had begun to work.

As a summary result, Fig. 13 classifies the sensors on whether they have experienced an improvement or a worsening on the level of occupancy, computed as described in section 3, in these 2-years.

In general terms, after March 15, 2019 the level of occupancy has improved in the area of Madrid Central, with some exceptions. The border effect is concentrated in specific zones of the Crown area, while there are also in this area other parts that have experienced improvements in the level of traffic intensity.

Finally, in Fig. 14 are shown exclusively the sensors changing their profile of usage, calculated and defined as in section 3, between the same periods in 2017 and 2019. It must be noted that the sensors within Madrid Central have not changed to “All day” profile of usage, supporting that now the zone is not occupied through all hours. On the contrary, some sensors in the Crown area have worsened its level of occupancy and, at the same time, have now an “All day” profile of usage.

Year	**Madrid Central**	**Crown**
2017	-1.6	-1.3
2018	-4.4	13.1
Jan-Aug 2019/ Jan-Aug 2018	-11.5	-2.3

Table 3. Average annual increase rates

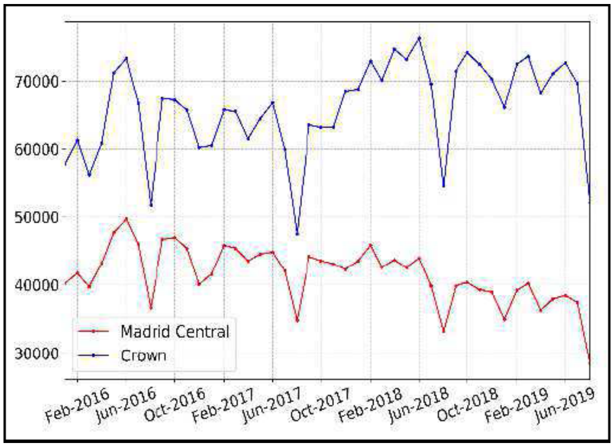


Figure 11. Monthly average

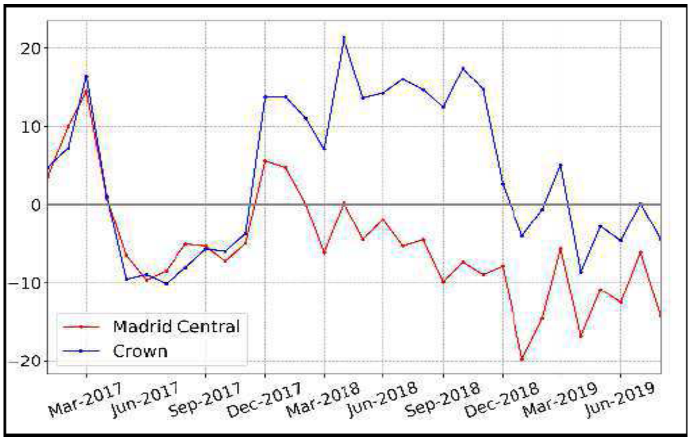


Figure 12. Monthly average annual rates

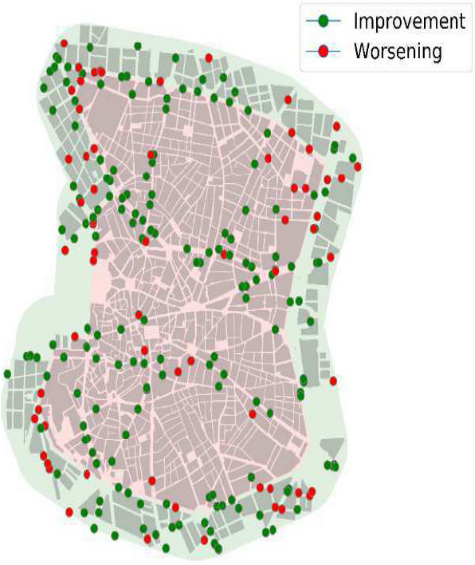


Figure 13. Changes in the occupancy levels after traffic measures

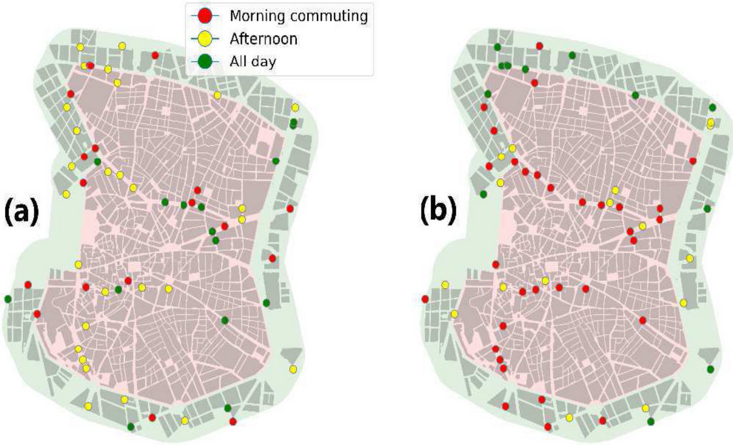


Figure 14. Weekday profiles of usage: (a) Before (b) After traffic measures

5. Final Remarks

This paper uses data about traffic sensors from the Madrid City open data portal to evaluate the impact of the traffic measures taken in the last years in Madrid. Being the first aim to study the behavior of the traffic intensity over time, it must be stressed the difficulties and complexities in measuring its evolution, requiring specific procedures.

The results obtained are very preliminary, first because only one of the variables available has been considered, and second because more periods would be needed to accurately measure the possible impacts.

Although the main objective of the traffic measures taken is to reduce air pol lution, what has been assessed here is the impact on the traffic volume, because it is considered one of the largest contributors to air pollution in cities. What has been found is that the actions implemented from 2017 seem to have reduced traffic congestion in Madrid Central and other areas especially from 2019. At the same time, in 2018 a first collateral border effect of increasing traffic intensity in the surrounding zones may exist, although this effect may revert in the next months as a consequence of the last actions undertaken.

Taking advantage of the spatial aspects of the information available, the methods proposed can be used to assess the effects of other traffic actions at the same or at more detailed geographical level, when data from more periods are available. The scope of the analysis can be widened when data from more periods are available and also by extending the procedures to other variables existing at the open data portal.

References

[1] Box, G.E., Jenkins, G. M., Reinsel, G. (1970). Time series analysis: forecasting and control holden-day san francisco. BoxTime Series Analysis: *Forecasting and Control Holden Day*, 1970.

[2] Box, G.E., Tiao, G. C. Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, 70 (349), 70–79.

[3] Chen, C., Liu, L.M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88 (421), 284–297.

[4] MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. *In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. vol. 1, p 281–297. Oakland, CA, USA (1967)

[5] Medina, J. C., Benekohal, R.F., Ramezani, H. (2012). Field evaluation of smart sensor vehicle detectors at intersections–volume 1: *Normal weather conditions. Tech. rep*.

[6] Mimbela, L.E.Y., Klein, L.A. (2000). Summary of vehicle detection and surveillance technologies used in intelligent transportation systems.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research,* 12 (Oct), 2825–2830.

[8] Schmidhuber, J., Hochreiter, S. (1997). Long short-term memory. *Neural Comput,* 9(8), 1735–1780.

[09] Stone, R., Prais, S. (1952). Systems of aggregative index numbers and their compatibility. *The Economic Journal* , 62 (247), 565–583.

[10] Team, P. C. (2017). Python: A dynamic, open source programming language, python software foundation.

[11] Thorndike, R.L. (1953). Who belongs in the family?, *Psychometrika,* 18(4), 267–276.

[12] Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics,* 15(2), 70–73.

[13] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., et al. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM,* 59(11), 56–65.

[14] Zhang, K., Batterman, S. (2013). Air pollution and health risks due to vehicle traffic. *Science of the total Environment* 450, 307–316.