The Role of Local Urban Traffic and Meteorological Conditions in Air Pollution: A Data-based Case Study in Madrid, Spain

Ibai Laña[a], Javier Del Ser[a,b,*], Ales Padró[a],
Manuel Vélez[b], Carlos Casanova-Mateo[c]

[a] *TECNALIA. P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain*
[b] *Dept. of Communications Engineering. University of the Basque Country UPV/EHU. Alameda Urquijo S/N, 48013 Bilbao, Spain*
[c] *Dept. of Civil Engineering: Construction, Infrastructure and Transport, Universidad Politécnica de Madrid, Spain*

Abstract

Urban air pollution is a matter of growing concern for both public administrations and citizens. Road traffic is one of the main sources of air pollutants, though topography characteristics and meteorological conditions can make pollution levels increase or diminish dramatically. In this context an upsurge of research has been conducted towards functionally linking variables of such domains to measured pollution data, with studies dealing with up to one-hour resolution meteorological data. However, the majority of such reported contributions do not deal with traffic data or, at most, simulate traffic conditions jointly with the consideration of different topographical features. The aim of this study is to further explore this relationship by using high-resolution real traffic data. This paper describes a methodology based on the construction of regression models to predict levels of different pollutants (i.e. CO, NO, NO_2, O_3 and PM_{10}) based on traffic data and meteorological conditions, from which an estimation of the predictive relevance (*importance*) of each utilized feature can be estimated by virtue of their particular training procedure. The study was made with one hour resolution meteorological, traffic and pollution historic data in roadside and background locations of

*Corresponding author: javier.delser@tecnalia.com (Prof. Dr. Javier Del Ser). OPTIMA Unit. TECNALIA. P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain. Tl: +34 946 430 50. Fax: +34 901 760 009. E-mail: javier.delser@tecnalia.com.

the city of Madrid (Spain) captured over 2015. The obtained results reveal that the impact of vehicular emissions on the pollution levels is overshadowed by the effects of stable meteorological conditions of this city.

Keywords: Urban air pollution; Traffic Flow; Meteorological Conditions; Supervised Learning; Random Forests

1. Introduction and Related Work

Madrid is the capital city of Spain, with 3.1 million inhabitants and a densely populated urban area (5225 inh/km^2) situated at an elevation of 667 meters over the sea level. As shown in Figure 1, the star-shaped design of the Spanish road network makes Madrid the central transport hub of the entire country. This fact, combined with the 4.2 million registered vehicles in the region, yields a heavy traffic supporting metropolis undergoing severe congestion issues through its road network. As a consequence of this, road traffic is widely acknowledged as the main source of air pollutants in Madrid [1]. In quantitative terms, NO_x and CO emissions are related to traffic in more than 80% in the city [2], 48% of PM_{10} mass was proven to be contributed by vehicle emissions [3], and 65% of tropospheric O_3 formation is on account of traffic-related precursors [4]. This close relationship between traffic and pollution comes along with severe health implications: indeed, worldwide epidemiological and toxicological studies have linked these traffic related pollutants to respiratory issues [5, 6], cardiovascular health effects [7] and lung cancer risk [8]. In 2013, the specialized cancer agency of the World Health Organization – the International Agency for Research on Cancer (IARC) – announced that outdoor air pollution has been officially classified as an carcinogenic agent for humans (Group 1) [9].

Even though the number or vehicles has increased significantly over the last two decades [10], levels of NO, NO_2, CO and PM_{10} have featured a decreasing trend in Madrid [11] as a result of the pollution abatement policies promoted by the European Parliament (Directive 98/69/EC [12]). The implementation of such regulatory laws and other subsequent sets of measures involves not only administrations, which are compelled to materialize control and management over traffic, home and industry pollutants, but also vehicle manufacturers, with more severe regulations for the exhaust emissions. Another relevant factor for this decreasing trend is the economic recession, which started in Spain in 2008 and has led to lower levels of fuel consump-

tion [11]. On the other side, and aligned with this NO_x reduction, an upward trend is found in tropospheric O_3 concentration in the last decade [4]. O_3 is formed within a complex photochemical process that requires, among others, anthropogenic and natural sources of NO_x, and solar radiation. When radiation conditions are met, O_3 is created and NO_x is consumed, bringing along the aforementioned trends.

Although vehicle emissions, industry and heating produce most of the atmospheric pollutants, the climatological characteristics of the region play an important role in how the pollutants are dispersed or conserved. The related literature (e.g. [13, 14, 15, 16] and references therein) has evinced that while precipitation and wind can help dissipating heavier and lighter pollutants respectively, their lack combined with high-pressure atmospheric conditions curb pollutant dispersion, and high UV radiation levels unchain the forenamed O_3 effects. A study made in Oslo (Norway) in 2004 [17] analyzed how distinct meteorological conditions impact on different pollutants, and elucidated that the number of vehicles is the most important factor in a city under such conditions. In a city like Madrid, with stable atmospheric conditions, the pollution should be strongly dominated by the prevailing meteorological conditions, setting traffic aside in a less relevant role, as pollutants accumulate until they are washed away by meteorological agents. Thence, two cities with similar anthropogenic emission levels may have acutely different pollution levels if they have antithetic meteorologic features. The dry and stable climate of Madrid, with less than 60 days of precipitation in 2015, results in a highly meteorology-dependent pollution. Furthermore, topography and urban street disposition and building types play also an important role in pollution concentration and dispersion. The term *street canyon* referred originally to narrow streets flanked by buildings, although this definition has been updated and characteristics like the height of buildings of each side, the length of the street, the number of crossing streets and the number of openings in the walls configure different types of street canyons [18]. Street canyons may produce diverse effects in pollutant concentrations depending on the direction and speed of wind, creating vortexes of pollution when wind is perpendicular to the street and in-flow channels when wind runs parallel to street.

In relation with the meteorological influence, 2015 has been the warmest year ever recorded in a global scale [19]. This fact, along with the incipient overcoming of the economic crisis, has implied high pollution issues over the city of Madrid during late 2015. Evidences abound: all over the year levels

of NO_2 exceeded the 200 $\mu g/m^3$ limit up to 95 times in some districts, and an average of 23 times a year for all air quality stations deployed in the city. Likewise, levels of tropospheric O_3 exceeded the 120 $\mu g/m^3$ limit an average of 10 times a year with a top of 68 excesses at one of the monitoring stations, while PM_{10} and CO remain below the recommended limits [20]. This series of data evidences has motivated authorities to undertake traffic containment measures such as speed and parking limitations or public transport reinforcement [21], to the point of foreseeing stringent traffic restrictions in the inner city should the previous measures not lower down pollution to admissible levels. However, such countermeasures are not new, as similar action plans have been put to practice for years in other cities [22, 23]. Effectiveness of the implementation of these policies is not strongly evidenced [24]; they do have an impact on pollutant levels in some areas [25], but it is slightly significant in other cities. Concurrently, researchers have modeled pollution from different approaches in order to predict and manage it efficiently. On one hand, research efforts have been invested on explaining the behavior of traffic emissions [26, 27] in order to understand how traffic pollutants are produced and how this knowledge should be exploited so as to diminish them. A report on this subject [28] showed that the factors influencing traffic emissions can range from drivers' aggressiveness to the number of stops they make if the traffic is congested. The first kind of factors are out of reach for traffic management, but the latter can be tuned so as to reduce emissions. However, this tuning might cause a negative impact in the level of service of the road network, and therefore should be implemented with caution.

Pollution models can help traffic managers to take decisions efficiently, by selecting the most adequate traffic management strategy [29]. In literature, meteorological data are the main input for the models [30, 31], while some researchers use only traffic data [32], and a slighter proportion of researchers build their models with both traffic and meteorological data as inputs [33, 34]. This manuscript will examine the relevance of road traffic variables and meteorological conditions in order to understand and predict the levels of pollutant agents in different kinds of locations of the city of Madrid, using historic traffic, pollution and meteorological data of 2015 as inputs. To this end, a methodology based on supervised machine learning will be followed; in this respect, most of the prediction models proposed in the literature hinge on artificial neural networks [35]. Variations in the neural network model and improvements in the pre-processing of input data are introduced by [36] to enhance the predictive capabilities. Other machine learning techniques such

as decision trees [37] or support vector machines [38] have also been used to predict pollution from meteorological data [39, 40]. Linear regression has also been used to model PM_{10} concentrations [34]. This paper joins these previous works from a new perspective: not only it explores the performance when predicting pollution using combinations of meteorological and traffic inputs and an ensemble supervised learning model, but also analyzes a quantitative measure of the importance of each variable as estimated during the training process of the model itself. Furthermore, the selected locations of air quality stations and traffic loops utilized in this study are characterized by different configurations in regards to their surrounding topography and urban street disposition. As discussed and concluded from the performed data analysis the impact of meteorological conditions do prevail on the pollution levels of this city, which might ultimately outgain any traffic-based countermeasure promoted by relevant authorities and stakeholders.

2. Materials and Methods

In this study open data provided by the Madrid City Council [41] and the Meteorological State Agency of Spain (AEMET) are used. Data correspond to the year 2015, from January to November, as December traffic data are undisclosed at the time of the development of this study. Meteorological conditions operate in large areas and traffic loops are available all over the city; therefore, the selection of the input data has been made by defining a set of targeted air quality stations and their closest traffic loops with enough diversity to represent different urban topographic characteristics and neighborhood types (e.g. downtown, residential).

2.1. Pollution Data

Madrid Air Quality System maintains 24 stations in the metropolitan area, which provide a variety of pollutant readings. Equipment to measure NO_x levels is available at all stations. There are two so-called "super-stations" which provide, besides NO_x, SO_2, CO, PM_{10}, $PM_{2.5}$, O_3, heavy metals and benzopyrenes readings, whereas for the rest the set of available measurements vary. According to the European and Spanish legislation there are three types of air quality stations: Urban Background (□), representative of urban population exposure to pollutants, in which the pollution levels should be distributed among different sources [42]; Roadside Traffic (☆), representing mainly emissions originated in close roads; and Suburban (◯),

in the outskirts of the city, which record the highest levels of ozone. Figure 2 depicts the location and type of the air quality stations deployed in Madrid.

In this research, the focus is set on those pollutants most closely related to traffic [1, 43]. Several of those pollutant agents, such as sulfur dioxide (SO_2), are originated mainly in industrial processes. In contrast, other pollutants like nitrogen oxides (NO_x) and particulate matters ($PM_{2.5}$ and PM_{10}) are directly related to road traffic [1, 44]. Ozone (O_3) levels depend mainly on meteorological conditions, but changes in NO_x emissions strongly influence O_3 trends [45]. Therefore it embodies a good indicator of traffic-related pollution, specially in locations where cloudy weather conditions are infrequent as occurs in Madrid. Consequently, air quality stations have been selected for the study considering their capabilities to measure NO, NO_2, O_3, $PM_{2.5}$ and PM_{10}, their location over the city and their proximity to traffic measuring points; the latter condition leaves suburban stations out of the study. The selected stations are depicted in Figure 3 and described as follows:

- *Escuelas Aguirre* (RS-EA): Roadside "super-station" with equipment capable of measuring CO, NO, NO_2, SO_2, O_3, benzene, hydrocarbons, $PM_{2.5}$ and PM_{10}. This station is located in the junction of three main roads in the center of Madrid with four or more lanes. Although it supports large amounts of traffic, it is placed next to El Retiro, a 350 acre park, which may mitigate the effects of surrounding traffic. The station is placed in an open area, flanked by the park and six-story buildings, not forming a typical street canyon.

- *Barrio del Pilar* (RS-BP): Roadside station located in a residential area in the north of the city. The station is placed in a park, next to a junction of two four-lane streets. The station is surrounded by a wide open area, but closer streets form street canyons flanked by thirteen-story buildings. The M30 ring highway, one of the roads with heavier traffic in Madrid, is located 230 meters north. Interestingly for the purpose of this study, measurements recorded by this station violated the NO_2 limit levels 95 times in 2015. RS-BP provides readings of CO, NO, NO_2 and O_3.

- *Plaza de Fernández Ladreda* (RS-FL): Roadside station in a residential area, supplying measurements of CO, NO, NO_2, and O_3. It is located in a junction of 4 important streets and a highway, but surrounded by

trees of a small park nearby. In front of its location there is a greater park (*Parque Emperatriz María de Austria*).

- *Plaza del Carmen* (UB-PC): Urban background station located less than 2 km west of the RS-EA station and close to *Gran Via*, an important artery with heavy bus traffic. Nonetheless, the station is placed in a pedestrian public square, surrounded by buildings that isolate it from the direct impact of *Gran Via* traffic. CO, NO, NO_2, and O_3 levels are provided by this station.

- *Arturo Soria* (UB-AS): Urban background station in a location similar to that of RS-BP: a residential working-class district with a main road crossing 50 meters south. However, this station is not so close to highways, and the surrounding buildings are lower (i.e. three-story), hence avoiding street canyons with their implications in pollution. It also has trees in both sidewalks and in the median strip. The station provides CO, NO, NO_2, and O_3 measurements.

- *Farolillo* (UB-FA): Urban background station in a residential working-class district with no important roads in 1 km around. The station is placed in a small public square surrounded by short buildings that conform typical street canyons with low traffic impact. Besides CO, NO, NO_2, SO_2 and O_3, this station provides PM_{10} readings, which will be helpful in order to assess the impact of close traffic on this pollutant.

It should be noted that the above stations supply the most complete sets of pollutants, with NO, NO_2, and O_3 covered in all of them, and PM_{10} in two of them. The rest of stations provide only part of these pollutant agents, with several of them missing, not being useful for comparison purposes. Moreover other criteria have also been applied when selecting stations for this study: as to mention the station located in *Casa de Campo*, a suburban "super-station", has not been selected because the high distance from it to any road makes the effects of traffic too indirect and could bias and mislead the conclusions from this study. Readings of the selected pollutants are published by the stations in $\mu g/m^3$ with a hourly resolution.

2.2. Traffic Data

Generally speaking Automatic Traffic Recorders (ATR) are magnetic loops embedded underneath the road that count the vehicles passing over

them. This captured information makes it possible to compute road usage metrics of inherent utility for different traffic management purposes:

- Flow: Number of vehicles that would pass in an hour extrapolated from the current number of vehicles passing (measured in vehicles/hour).
- Occupancy: Percentage of vehicles in relation to the total vehicle capacity of the road segment where the loop is placed (measured in %).
- Load: an undisclosed value from 0 to 100 that depends on the flow, occupancy and the infrastructure features.
- Level of service: a simplified scale with disclosing purposes used to denote the state of the road in understandable terms. It takes value 0 if load $\leq$ 60; 1 if 60 $<$ load $\leq$ 70; 2 if 70 $<$ load $\leq$ 90; and 3 if 90 $<$ load $\leq$ 100.

The Madrid City Council publishes open historic datasets of around 3800 traffic sensors with aggregated readings of the above variables in intervals of 15 minutes. In this study regression models were built using traffic flow data. To this end, a distinct subset of the ATRs deployed over the city is chosen for each air quality station based on their proximity to each other. Although pollutant agents can be dispersed in greater areas and affected by diverse factors, this study is focused in the direct impact of close traffic; consequently, a 100-meter radius has been set around each station[1] so as to discriminate the subset of ATRs that characterizes the traffic in the surroundings of the station at hand. Differences in traffic volume are expected to have a slighter impact in pollution than the meteorological conditions, which is close to be the same for all areas.

2.3. Meteorological Data

The present study requires meteorological data with both good quality and high temporal resolution. Consequently, although it is possible to access daily resumes of many of the meteorological observatories hosted by AEMET [46], we decided to use the meteorological observations provided by the Aerodrome Meteorological Office of the Adolfo Suárez Madrid-Barajas International Airport. Specifically, we have used the METAR and SPECI reports.

[1]Except for UB-FA, which is 520 meters from the closest ATR.

METAR is a coded report normally generated every half an hour throughout the worldwide network of operative airports. On the other hand, SPECI stands for the code of an aerodrome special meteorological report. SPECI briefings are generated when there is significant deterioration or improvement in airport meteorological conditions (for more information on METAR and SPECI reports see ICAO ANNEX 3 to the Convention on International Civil Aviation [47]). Both reports are freely available on the Internet and provide interesting meteorological data. For this study, we have selected the following data:

- Precipitation, extracted from the group that informs about present weather phenomena observed at or near the aerodrome.
- Temperature and Dew Point, obtained from the group that informs about both variables (measured in degrees Celsius).
- Wind intensity, obtained from the wind group. For this study, we have converted this variable from knots to kilometers/hour.
- Cloud Type and Cover, obtained from the group used to report sky condition for atmospheric layers aloft. This group informs only about some cloud types (mainly Cumulus Congestus and Cumulonimbus), and gives interesting information regarding cloud cover, which is encoded into five categories: Sky Clear, Few, Scattered, Broken and Overcast. Not having UV radiation data makes these variables relevant for the regression of ozone levels.

Finally, regarding the quality of the meteorological data used in this work, it is interesting to note that all the meteorological observing systems that the Aerodrome Meteorological Office Staff utilize for preparing METAR and SPECI reports (e.g. thermometers, anemometers, and ceilometers, among others) are managed under a Quality Management System certified to ISO 9001:2008. Besides the aforementioned fine-grained temporal data, the open data bank of the Madrid City Council provides monthly aggregated temperature in Celsius and precipitation in millimeters since 1988, as well as monthly aggregated maximum wind speeds in kilometers/hour and UV radiation levels in Joules/m2 since 2012, all captured from the Madrid-Retiro meteorological observatory (latitude: 40.41 N, longitude: 3.68 W, altitude: 667 m). These data are not used in the regression models, as they do not have

enough resolution; instead, they provide the means to compare the climate in Madrid during 2015 to the averaged sequence of previous years.

2.4. Regression Model and Feature Importances

As anticipated in the introduction we will resort to one of the most utilized ensemble methods for supervised learning problems, Random Forests (RF), which have gained momentum in the last decade by virtue of their ability to handle multidimensional classification and regression problem with excellent accuracy and small chance to overfit [48]. In their naive definition RF consists of an ensemble of weak tree learners, each trained on a sampled subset of the available data, from where the predicted output is taken by aggregating and averaging the individual predictions of all such compounding trees. This particular construction method, which blends together the concepts of bagging and random feature selection, have been demonstrated to improve performance over other machine learning algorithms and linear regression models [49], due to their high accuracy, low over-fitting, and reduced tuning requirements.

As a byproduct of this training procedure RF also provide an embedded method for quantifying the predictive importance that each of the predictor variables in the dataset possesses with regards to the target variable to be predicted. Specifically, the value of the feature importance reflects the mean decrease in accuracy when testing the model with out-of-bag (oob) observations [49], either for classification or regression problems. Specifically, the importance value of the j-th feature after training the RF model is computed by first permuting the values of this feature among the training data and next computing the average out-of-bag score difference between the original, unaltered dataset and that obtained after permuting the variable. Scores are normalized by the standard deviation of such differences in performance. This provides a numerical estimation $I_j \in [0, 1]$ that denotes the importance of such a variable during the training process of the model, i.e. $I_j \to 1$ if the k-th feature is predictively relevant for the variable to be predicted, and will approach 0 otherwise.

Since this study deals with regression results discussed in what follows will be interpreted by jointly analyzing cross-validated predictive performance scores – in terms of the so-called coefficient of determination R^2 – and variable importances obtained for each dataset. This score measures how fit the model is to predict future data instances, with its best score being 1.0 for error-free prediction, and its worst value equal to −1 corresponding to a model that

performs worse than a constant prediction equal to the expected value of the target variable. Mathematically speaking the R^2 score computed over N predicted samples $\widehat{\mathbf{y}} = \{\widehat{y}_n\}_{n=1}^N$ corresponding to the true instances $\mathbf{y} = \{y_n\}_{n=1}^N$ will be given by

$$R^2(\mathbf{y}, \widehat{\mathbf{y}}) \triangleq 1 - \frac{\sum_{n=1}^N (y_n - \widehat{y}_n)^2}{\sum_{n=1}^N (y_n - \bar{y})^2}, \tag{1}$$

where $\bar{y} \triangleq \sum_{n=1}^N y_n / N$. An statistically meaningful estimation of the first-order statistics (mean, standard deviation) of the above score will be computed using K-fold shuffled cross-validation over the dataset in question.

2.5. Preprocessing of the Datasets

The data sources chosen in this manuscript deliver different temporal resolution data: hourly for pollution and meteorological data, and 15-minutely for traffic data. The first preprocessing step involves bringing resolution uniformity to the data, thence, 15-minute traffic slots were aggregated into one-hour slots, decreasing the noise produced by outliers, and maintaining the characteristics of original distribution. Each instance of a dataset is created with these 1-hour uniform data. Instances contain the hour of the day, the traffic flow value of each loop in the dataset for that hour, meteorological parameters during the same period and pollutant levels read by the monitoring station at hand. Besides, three additional columns accounting for the type of day, public holidays and month were added to the dataset, leaving the instances as shown in Figure 4.

A dataset is created for each monitoring station, each then split into four sub-datasets which are combinations of the three types of available input variables: all features (labeled with ① in the remainder of the paper); only traffic and meteorology features (marked with ②); only traffic and temporal features (correspondingly, ③); and finally, only meteorology and temporal features (④). This split into subsets aims at assessing the impact of the lack of each of the variables in the prediction of 4 pollutants: CO, NO, NO_2 and O_3. This process results in 16 datasets per location. Also, 2 additional datasets are created to build models for predicting PM_{10}. Datasets are initially created with 8760 instances (one per hour through all the year). There are, though, instances with incomplete data, such as all the ones corresponding to the month of December or the first part of January, for which there are no traffic data or lines with faulty meteorological readings. Such incomplete instances are deleted from the finally processed datasets.

3. Results and Discussion

As emphasized in the introduction, the experiments and results next discussed are oriented towards quantitatively assessing the influence of meteorological conditions and traffic variables on the local pollution levels in different parts of Madrid in 2015. For this purpose this section gravitates on the analysis of the interactions among such variables, which is done by both visual inspection and supervised learning.

3.1. Traffic Characteristics in Selected Zones

Besides the already highlighted criteria for selecting among station locations, the disparate traffic levels recorded thereat constitute another reason why they were chosen. These traffic behavioral differences are observable in Figure 5, which result after aggregating the data captured by the ATRs of each considered zone. Columns in each subplot represent the daily averaged traffic flow, i.e. traffic is averaged first over the available ATR readings for each hour, then the average over 24 hours delivers the value represented in this plot.

As expected, Figure 5 shows that the traffic flow in the most central zone (namely, RS-EA) is the highest one, whereas UB-FA supports the lowest traffic flow levels. It can be also observed that the area around RS-BP bears less traffic than those of UB-PC and UB-AS, being the first a roadside location and the latter urban background locations. RS-BP and UB-AS share characteristics: working class districts, close to main roads and more than 5 km far from the city center. Nevertheless, UB-AS almost doubles the average traffic in the nearby ATRs, and is considered a urban background measuring site, while RS-BP is contemplated as a kerbside location.

When it comes to pollution, an opposite trend is noted: according to [20], RS-BP has exceeded the limit levels of NO_2 up to 95 times during 2015, while UB-AS, holding more traffic, only exceeded the level 18 times (below the 20-times alert level) at similar O_3 levels. Although both locations share similar demographic features, the higher presence of trees on both sides and the center of the road, and the lower amount of crossings, which induce to stop and start the engines of the vehicles, are the main elements present in UB-AS that explain the opposite trends in traffic and pollution [50]. As aforementioned, UB-FA location is in a working class district with no important roads around, and this situation is observable in the traffic readings. Traffic levels are similar in UB-PC and UB-AS; both are urban

background locations, but the first is a small public square in the center, with ATRs in a one-way road (the rest of surrounding roads are pedestrian streets), while the latter is in a residential area and the ATRs are placed along two- or four-lane roads. Similar traffic flows in different type of roads imply higher occupation of the road in the smaller ones.

Aside from the discrepancies in the 6 zones, there are concurrences relative to the temporal dimension of the readings. In all 6 locations it is possible to observe a variably acute decrease in the beginning of the second half of the year. This decrease corresponds to August and late July, and it is more abrupt in center zones (RS-EA, UB-PC, where the working population is stable until the end of July), than in residential ones (RS-BP, UB-FA). There is also a noticeable decrease in the end of the first third of the year, corresponding to Easter holidays. This drop presents mostly the same length in any area, as these holidays have always the same duration and the date is predefined. Additionally, a pattern is observable in each week: traffic levels increase gradually on weekdays, and drop in weekends.

A closer look at this pattern is provided in Figure 6 for the traffic captured in the RS-EA location. This traffic characterization is usual in the literature related to traffic modeling [51, 52] and temporal features were found to have a decisive relevance in long-term traffic forecasting [53]. Based on this rationale, temporal features are subsequently incorporated to the dataset in order to improve the performance of the regression techniques, as previously explained in Section 2.5.

3.2. Climate in Madrid during 2015

Climate in Madrid is typically dry, with maximum temperatures rounding the 35 °C during summertime, and negative values in January and February. Figure 7 displays the daily average, maximum and minimum temperatures recorded in the utilized aerodrome observatory during 2015. As described in Section 2.3, METAR and SPECI reports provide qualitative information about the precipitation episodes, but not quantitative. For this reason, and for visualization purposes, the number of hours in which precipitation has taken place each day are also included in the plot. Remarkably, only on 6 days of the entire year it rained for more than 8 hours, all of them aligned with temperature declines.

The monthly average historic data provided by the Madrid City Council show that temperatures are in line with the typical temperatures of this region through the year (Figure 8A), with some values above the average in

July, November and December. Late autumn and winter have been specially dry, though, with 29.1 mm and 4.2 mm precipitated in November and December, as opposed to 51.4 mm and 40.3 mm averages for the same months (Figure 8B).

This lack of precipitations hinders dispersion and reduction of some pollutants such as PM_{10} and $PM_{2.5}$ [13, 14, 15, 16], and are related to the pollution peaks recorded in the last part of the year in Madrid, which ultimately lead to traffic restrictions. The other two relevant factors analyzed in this study, wind speed (able to disperse pollutant particles, or bring them from somewhere else) and UV radiation (instigator of chemical reactions that transform some pollutants into others), have maintained values close to the historic records, which are constrained to 4 years, in the available public data (Figures 8C and 8D). The maximum wind speed attained in December is the most different recorded data (46km/h vs average 64km/h). This along with the previously exposed data buttress the relevance of the change in typical winter meteorological conditions in Madrid that may be behind the utmost pollution levels recorded in the last months of 2015 in this region of Spain.

3.3. Pollution Characteristics in Selected Zones

Air quality monitoring stations have been selected considering the pollutant agents they are able to measure, their distance to direct sources of traffic pollution, and the type of station, defined by their location. As described in Section 2.1, selected stations are urban background and roadside, and they are placed in locations with dissimilar characteristics; different levels of pollution and traffic are expected. Figure 9 shows the distribution of CO, NO, NO_2, and O_3 through the year.

General seasonal trends are visible in the figure. O_3 reaches its maximum in summer months (35-120 $\mu g/m^3$) when meteorological conditions facilitate its formation, whereas minimums (5-25 $\mu g/m^3$) are found in winter months (less solar radiation, shorter days), coherently with other studies carried out in southern Europe [45]. NO_2 levels remain stable through the year, closer to the 50 $\mu g/m^3$ line in roadside traffic stations and to 40 $\mu g/m^3$ in the urban background stations. Peaks attained by the CO and NO pollutants coincide when heating systems are active and ozone plunges.

PM_{10} levels are only available in 2 stations, and represented in Figure 10. RS-EA and UB-FA locations are the most dissimilar in the entire sample, as detailed in Section 2.1, and they are 5.2 km away. Yet, their PM_{10} measurements are resembling. Top and bottom peaks are produced in almost

the same parts of the year, and the values follow roughly coincidental lines. Aside from natural sources and Saharan dust being significant contributors of PM_{10}, traffic is also a proven relevant source of this pollutant [3]; thereby, vehicle emissions impact in PM_{10} are mainly contributed by background contamination, more than by local sources, as presented in [11]. On the basis of this evidence, and not having PM_{10} measurements for all the selected areas, this paper will focus in the analysis of CO, NO, NO_2, and O_3.

3.4. Relations among Pollution, Traffic and Meteorological Conditions in Selected Zones

Air pollution in big cities is produced in a significant level by road vehicle emissions, with the modifying influence of meteorological agents. When cross-matching the previously presented data it is possible to discern similar effects for the city of Madrid. In order to assess the impact of meteorological conditions and local traffic in local pollution levels, the concentration of CO, NO, NO_2 and O_3 data of each site were analyzed with different temporal scales and overlaid with traffic levels. Annual results for the RS-EA location are shown in Figure 11. The plot depicts pollutant levels running seasonally, with increased O_3 during the summer months and the consequent increment of NO_2 and decrease of NO. Winter months undergo peaks of NO, coinciding with less ozone presence and heating systems being active.

Although these trends are expectable, there is no apparent relation with traffic. When traffic plummets in August, NO and CO levels are maintained. NO even peaks over 30 $\mu g/m^3$ on August 26^{th}, a day with 9 daylight hours with *overcast* cloud coverage, which might be behind the low levels of O_3 (under 50 $\mu g/m^3$). In Figure 7 it is possible to observe a week in the late March when it rained for several days, and temperatures decreased in 6-7 °C relative to the previous trend. This corresponds in Figure 11 to the NO peak and O_3 valley after day 50 (which aligns with middle March due to the absence of the first days of January). Traffic was about the same as the previous week, but pollution increased. Examining the same data for UB-FA (the most different location), a similar detachment is found as unveiled in Figure 12. Traffic is flatter through the year, and although pollution levels are lower, they follow a similar trend.

This detachment is not found if the time scale is varied. In Figure 13, traffic levels are plotted by day, divided by working and non-working days, and summer (April to October) and winter (November to March) months. Working and non-working days separation affects the traffic levels, and season sep-

aration impacts on both traffic and pollution levels. In this daily inspection, a closer relation between traffic and pollutants is discovered: when traffic starts in the first hours of the morning, specially in working days, formation of NO and CO is triggered. After 10 AM, and particularly in summer, ozone formation increases, combining itself with NO, and reducing its levels. The decay of both starts at the same time as the night (less traffic producing NO and less sunlight inducing O_3). Lighter traffic in weekends provokes higher levels of O_3, which is known as the *weekend effect* [45]. Differences between these two scopes have been already been noted by [44, 42, 14, 45] for diverse pollutants and over different cities.

These variables are used to build the datasets defined in Section 2.5. Random Forest regression models are built for each of the 96 datasets (Figure 4) and evaluated applying shuffled cross-validation with $K = 10$ folds. For instance, the first dataset is created with temporal, traffic and meteorological variables as features, and CO measurement as the response variable in the RS-EA location. For every hour of the year, one instance is created resulting in 8760 samples that are later cleansed by removing the ones with missing entries. For the cross validation, resulting dataset is split into 4 sub-datasets, each containing a shuffled random fourth portion of the original instances. The model is then trained with 3 of the 4 sub-datasets and tested with the remaining one, rendering performance metrics that are stored for subsequent processing. This process is repeated 10 times for each model, with different compositions of each sub-dataset, and the overall performance is averaged among the results of the 10 executions. Coefficients of determination are extracted to observe how the response variables are fitted by the model. Table 1 compiles the averaged R^2 values of the 96 models.

These performance values were achieved after several iterations in which the Random Forest model was refined, setting its parameters by way of a grid search procedure. Even after optimizing the estimator, the obtained R^2 scores are in general low (under 0.7), thus the developed model would not be useful for predicting pollutant levels except for O_3. If that would be the case, input features to the model would need to be predicted as well. For instance, to predict NO at a certain moment in the future, traffic and meteorological conditions at that moment would be required; being a future instant would make necessary to predict the traffic and meteorological conditions at that point. This would probably lead to even worse results. However, these results provide a comparative insight on the predictability of pollution based on traffic and meteorological conditions. In the first place, the worst results

are obtained when no temporal information is provided to the model (Table 1, figures under the ② label). Regardless of the particular location or predicted pollutant, the better scores are achieved for every model in any other combination of features, which is a revealing indicator of the seasonality of the data along time. Scores labeled under ③ in Table 1 correspond to the results obtained without meteorological data, and even if slightly better than those of ②, they are still far from the best obtained. Required temporal variables combined with traffic levels seem to perform poorly without meteorological information. Among these results locations with greater traffic flow levels achieve the best scores in traffic-emitted pollutants (NO and CO): RS-EA and RS-FL are close to important junctions, while UB-PC and UB-FA are far or blocked from main arteries. O_3 is not linked to traffic as directly as CO and NO, hence its scores teeter among locations not connected with each traffic density; this effect is transferred to NO_2 scores. PM_{10} is also a pollutant directly related to traffic, but as seen in Section 3.3, it is affected mainly by general contributions, not local. Obtained results buttress this previous hypothesis: Table 1 (① and ④) contain the best scores, being very similar. Although ① comprehends most of the best scores obtained (in bold type), the differences between ① and ④ are in the 10^{-2} or even 10^{-3} order of magnitude. Once again, the higher improvement when incorporating traffic is produced in RS-EA.

A non-parametric Wilcoxon hypothesis test has been performed in order to confirm this conjecture and to shed light on the statistical significance of such performance gaps. Table 2 shows Wilcoxon test results (p-values); in general, statistically significant discrepancies (p-values close to 0) are few. This means that for most cases, there is no evidence that the medians of the score sets being compared differ. UB-AS location presents, though, the opposite outcome: R^2 values have the larger differences between traffic and no-traffic datasets, and Wilcoxon p value rejects the hypothesis that the difference is due to chance. Figures 9 and 5 bolster this idea: UB-AS and UB-PC share very similar traffic levels, but NO and CO concentrations read in UB-AS are lower, with NO ranking from 1 to 15 $\mu g/m^3$ in UB-AS and from 10 to 30 $\mu g/m^3$ intervals in UB-PC. These differences might be caused by a variety of factors: natural ventilation of the area, presence of vegetation or others. Regardless of the reason, UB-AS – a background pollution measuring station – is more affected by local traffic than the other stations, and a deeper analysis, out of the reach of this study, ought to be made to determine its causes, possibly by resorting to topographical urban models. In

those locations where O_3 concentrations are best estimated with traffic data (RS-EA, UB-PC, UB-AS, UB-FA), Wilcoxon test outcome shows a relevant difference not due to chance. Road traffic emissions are linked to O_3 as they modify its concentrations when combined.

The previous set of analyses is completed by Figure 14, where the feature importances of the different predictors – as provided by the RF regressor – is plotted as a heat map in order to analyze in detail the coupling of features and response variable. A noteworthy outcome is the low relevance of precipitation feature, which is in fact the less relevant for all datasets. This could be attributed to the lack of precipitations in 2015 in Madrid. Precipitations are indeed an important pollution-modifying factor, but when held in so infrequently it does not make a good general predictive feature. Also, the use of a discrete scale of precipitation levels, in the dearth of real millimeter readings, could reduce the relevance of this feature. On the other hand, cloud types are provided to the model in a similar discrete scale, and they are a generally relevant feature, and particularly the most relevant feature to predict O_3. Despite temporal features have been found to be the most determining ones, two of them – public holidays and day types – are scarcely relevant. These variables were useful in [53] to predict traffic, but in the long cycles of pollutants they have no effective value. On the other hand, months seem to be the most relevant (darker shade) for predicting CO and NO in almost every location, whereas hour of day influences specially in NO_2. CO and NO are produced by different sources and maintained in the air in cycles that depend on meteorological and season factors, and NO_2 levels are highly driven by the hour of day, as its production is linked to O_3 and the latter exhibit day and night cycles. Wind speed is a good NO_2 predictor, with importance values around 15% and peaking in 20% at the UB-FA location (the one with lower buildings, and in a flatter area more exposed to wind effects). ATR readings importance is in general low, confirming the R^2 scores and analysis, but also finding that in all cases traffic levels relevance for predicting CO, NO or NO_2 double their relevance for predicting O_3.

Although in Table 1 O_3 predictions always render better performance scores, the analysis of the feature importances shows that this performance is in any case linked to other variables, specially the cloud type. Other conspicuous outcome is the relatively high importance of traffic features in the UB-PC dataset; this is related to the distribution of the feature importance, which sums 1 for each dataset. Distributing the importance among less features adds relative weight to them. For the same reason, cloud type

feature predicting O_3 is more relevant as the dataset is smaller (0.35-0.4 for RS-EA and UB-FA, and 0.55 for UB-PC). Nonetheless, aggregated ATR measurements importance for each dataset gives an average 0.44 importance for all the ATRs in RS-EA and 0.27 for UB-PC, which confirm that traffic is more relevant in RS-EA than in UB-PC, as results in Table 1 clearly show. The aggregated importance is useful for comparison purposes, but not for feature importance analysis, because of the way Random Forest sub-samples the dataset. In each tree a random sub-set of features is used to build the model, and only a small portion of them will have all ATR reading features concurring in the same subset.

The fact that incorporating traffic levels to the predictive model could worsen the scores was unexpected, and makes local traffic readings become noise in some cases. This does not mean that traffic is irrelevant for pollution – it is indeed one of the main contributors to pollution in Madrid [1] – , but rather that its local effects on particular locations are limited and stringently linked to localities. Other research contributions such as [44, 54, 42] have addressed similar subjects and found significant disparities among urban background and roadside stations, being the latter highly influenced by vehicular emissions. In Madrid only measuring sites with heavier traffic and/or higher vegetation density have been shown to be influenced by traffic slightly over the direct effects of meteorology and seasonality.

4. Concluding Remarks

Many are the sources of air pollutant agents: industry, agriculture and livestock farming, road and air traffic, forest fires, natural sources like volcanoes or particles drawn by wind, among others. A wide variety of elements modify the concentration of different pollutants, mainly meteorological agents, but also topography, tree and shrub presence, building distribution or water streams like rivers. This research examined the effects of local road traffic, meteorological conditions and temporal variables on air pollution in Madrid. Data collected from 6 air monitoring stations, 33 ATRs and data from a meteorological observatory were used to build supervised learning models and analyze the relationships among these variables. Results showed that pollutant agent levels in the 6 evaluated locations were weakly linked to local vehicular emissions.

The outcomes provided by Random Forest regression and the analysis of the importance of the features during the training process of the model

suggest that seasonal features are the most relevant when predicting CO, NO and NO_2, with daily seasonality affecting both residential and downtown areas. PM_{10} concentrations have been studied for two of the locations, unveiling a slight impact of local traffic in these particles. Within the Madrid urban area, CO, NO, NO_2 and O_3 concentrations are strongly influenced by meteorological factors, specifically wind speed and cloud type, with less temperature influence. Precipitations, a usually influential actor in pollution alteration, have been proven to have a minor effect in pollutant concentrations over Madrid during 2015. The fact that this year has been specially dry, and the measurement scale used might be the most likely contributors to the poor predictive performance of this variable. Meteorological agents like precipitation in millimeters, wind direction, humidity or pressure have been left out of the study in the lack of proper public sources of data. Using actual precipitation or UV radiation readings instead of a proxy scale could improve estimator results.

The meager impact of traffic emissions on pollution levels is a remarkable outcome of the analysis that should be observed cautiously. Vehicular exhaust chemicals have been proven to be the major contributors to Madrid pollution levels. However, other factors such as its flat topography, the absence of zones with concentration of high buildings and its dry, atmospherically stable conditions contribute to a global background pollution that affects in similar ways to different areas. As global pollution increases daily with contributions from all city traffic, industry, heating systems and others, the local contributions decrease relatively to the volume of general accumulated concentrations of pollutants. Thus, only locations supporting heavy traffic produce a contribution of traffic-related pollutant chemicals largely enough to impact on local concentrations of the pollutants under study.

The overall results of this study suggest that countermeasures to reduce pollution based on restricting traffic for short periods of time could have a modest impact if meteorological conditions to mitigate the current accumulated pollution do not occur. The particular climatic characteristics of Madrid and the increasing road traffic lead to the belief that long-term measures such as permanent low-emissions zones, campaigns for promoting the use of public transportation or policies favoring the widespread adoption of electric vehicles could help containing city-wide pollution issues in a more effective manner.

5. Acknowledgements

This work has been funded in part by the Basque Government under the ELKARTEK program (BID3A project, grant ref. KK-2015/0000080).

[1] "Informe de calidad y evaluación Ambiental," tech. rep., Ministerio de Agricultura, Alimentación y Medio Ambiente, Madrid, 2012.

[2] A. Monzón and M. J. Guerrero, "Valuation of social and health effects of transport-related air pollution in Madrid (Spain)," *Science of the Total Environment*, vol. 334-335, pp. 427–434, 2004.

[3] P. Salvador, B. Artíñano, D. G. Alonso, X. Querol, and A. Alastuey, "Identification and characterisation of sources of PM10 in Madrid (Spain) by statistical methods," *Atmospheric Environment*, vol. 38, no. 3, pp. 435–447, 2004.

[4] V. Valverde, M. T. Pay, and J. M. Baldasano, "Ozone attributed to Madrid and Barcelona on-road transport emissions: Characterization of plume dynamics over the Iberian Peninsula," *Science of the Total Environment*, vol. 543, pp. 670–682, 2016.

[5] M. Brauer, G. Hoek, P. Van Vliet, K. Meliefste, P. H. Fischer, A. Wijga, L. P. Koopman, H. J. Neijens, J. Gerritsen, M. Kerkhof, J. Heinrich, T. Bellander, and B. Brunekreef, "Air pollution from traffic and the development of respiratory infections and asthmatic and allergic symptoms in children," *American Journal of Respiratory and Critical Care Medicine*, vol. 166, no. 8, pp. 1092–1098, 2002.

[6] M. Zuurbier, G. Hoek, M. Oldenwening, K. Meliefste, P. van den Hazel, and B. Brunekreef, "Respiratory effects of commuters' exposure to air pollution in traffic.," *Epidemiology (Cambridge, Mass.)*, vol. 22, no. 2, pp. 219–227, 2011.

[7] G. Hoek, R. M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, and J. D. Kaufman, "Long-term air pollution exposure and cardio- respiratory mortality: a review.," *Environmental Health: A Global Access Science Source*, vol. 12, no. 1, p. 43, 2013.

[8] O. Raaschou-Nielsen, Z. J. Andersen, M. Hvidberg, S. S. Jensen, M. Ketzel, M. Sørensen, S. Loft, K. Overvad, and A. Tjønneland, "Lung cancer incidence and long-term exposure to air pollution from traffic," *Environmental Health Perspectives*, vol. 119, no. 6, pp. 860–865, 2011.

[9] IARC, *Air pollution and cancer.* No. 161, 2013.

[10] DGT, "Anuario Estadístico General," tech. rep., 2012.

[11] P. Salvador, B. Artíñano, M. M. Viana, A. Alastuey, and X. Querol, "Multicriteria approach to interpret the variability of the levels of particulate matter and gaseous pollutants in the madrid metropolitan area, during the 1999-2012 period," *Atmospheric Environment*, vol. 109, pp. 205–216, 2015.

[12] European Parliament and Council of the European Union, "Directive 98/69/EC of the European Parliament and of the Council of 13 October 1998 relating to measures to be taken against air pollution by emissions from motor vehicles and amending Council Directive 70/220/EEC," 1998.

[13] P. A. Kassomenos, H. A. Flocas, S. Lykoudis, and A. Skouloudis, "Spatial and temporal characteristics of the relationship between air quality status and mesoscale circulation over an urban Mediterranean basin," *Science of the Total Environment*, vol. 217, no. 1-2, pp. 37–57, 1998.

[14] P. A. Kassomenos, S. Vardoulakis, A. Chaloulakou, A. K. Paschalidou, G. Grivas, R. Borge, and J. Lumbreras, "Study of PM10 and PM2.5 levels in three European cities: Analysis of intra and inter urban variations," *Atmospheric Environment*, vol. 87, pp. 153–163, 2014.

[15] J. Kukkonen, M. Pohjola, R. S. Sokhi, L. Luhana, N. Kitwiroon, L. Fragkou, M. Rantamäki, E. Berge, V. Ødegaard, L. H. Slørdal, B. Denby, and S. Finardi, "Analysis and evaluation of selected local-scale PM10 air pollution episodes in four European cities: Helsinki, London, Milan and Oslo," in *Atmospheric Environment*, vol. 39, pp. 2759–2773, 2005.

[16] S. Vardoulakis and P. Kassomenos, "Sources and factors affecting PM10 levels in two European cities: Implications for local air quality management," *Atmospheric Environment*, vol. 42, no. 17, pp. 3949–3963, 2008.

[17] M. Aldrin and I. H. Haff, "Generalised additive modelling of air pollution, traffic volume and meteorology," *Atmospheric Environment*, vol. 39, no. 11, pp. 2145–2155, 2005.

[18] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: A review," *Atmospheric Environment*, vol. 37, no. 2, pp. 155–182, 2003.

[19] J. Hansen, M. Sato, R. Ruedy, G. A. Schmidt, and K. Lo, "Global Temperature in 2014 and 2015," 2015.

[20] "Resumen de la Calidad del Aire 2015," tech. rep., Ayuntamiento de Madrid, Madrid, 2016.

[21] F. Serrato, "Una ciudad en vilo por la polución," dec 2015.

[22] F. Ferreira, P. Gomes, H. Tente, A. C. Carvalho, P. Pereira, and J. Monjardino, "Air quality improvements following implementation of Lisbon's Low Emission Zone," *Atmospheric Environment*, vol. 122, pp. 373–381, 2015.

[23] C. Holman, R. Harrison, and X. Querol, "Review of the efficacy of low emission zones to improve urban air quality in European cities," 2015.

[24] G. Titos, H. Lyamani, L. Drinovec, F. J. Olmo, G. Močnik, and L. Alados-Arboledas, "Evaluation of the impact of transportation changes on air quality," *Atmospheric Environment*, vol. 114, pp. 19–31, 2015.

[25] J. M. Baldasano, M. Gonçalves, A. Soret, and P. Jiménez-Guerrero, "Air pollution impacts of speed limitation measures in large cities: The need for improving traffic data in a metropolitan area," *Atmospheric Environment*, vol. 44, no. 25, pp. 2997–3006, 2010.

[26] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pp. 1027–1036, 2014.

[27] J. L. Reyna, M. V. Chester, S. Ahn, and A. M. Fraser, "Improving the accuracy of vehicle emissions profiles for urban transportation greenhouse gas and air pollution inventories," *Environmental Science and Technology*, vol. 49, no. 1, pp. 369–376, 2015.

[28] H. Frey, N. Rouphail, A. Unal, and J. Colyar, "Emissions Reduction Through Better Traffic Management: An Empirical Evaluation Based Upon On-Road Measurements," tech. rep., 2001.

[29] B. Barratt, R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, I. Mudway, and P. Wilkinson, "Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme," *Atmospheric Environment*, vol. 41, no. 8, pp. 1784–1791, 2007.

[30] B. Ando, S. Baglio, S. Graziani, and N. Pitrone, "Models for air quality management and assessment," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 3, pp. 358–363, 2000.

[31] P. Mlakar and M. Boinar, "Perceptron Neural Network - Based Model Predicts Air Pollution," in *Intelligent Information Systems*, pp. 345–349, 1997.

[32] R. H. Keeler, *A machine learning model of Manhattn air pollution at high spatial resolution*. PhD thesis, 2014.

[33] G. Ibarra-Berastegi, J. Saenz, A. Ezcurra, A. Elias, and A. Barona, "Using neural networks for short-term prediction of air pollution levels," *2009 International Conference on Advances in Computational Tools for Engineering Applications*, pp. 498–502, 2009.

[34] I. González-Aparicio, J. Hidalgo, A. Baklanov, A. Padró, and O. Santa-Coloma, "An hourly PM10 diagnosis model for the Bilbao metropolitan area using a linear regression methodology," *Environmental Science and Pollution Research*, vol. 20, no. 7, pp. 4469–4483, 2013.

[35] J. Hopfield, "Artificial neural networks," *IEEE Circuits and Devices Magazine*, vol. 4, no. 5, pp. 3–10, 1988.

[36] K. S. Lei and F. Wan, "Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau," in *2010 IEEE International Conference on Automation and Logistics, ICAL 2010*, pp. 418–422, 2010.

[37] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[38] V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[39] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y. Wenjun, and D. Jin, "A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method," in *IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, pp. 176–181, 2015.

[40] E. Sahafizadeh and E. Ahmadi, "Prediction of Air Pollution of Boushehr City Using Data Mining," in *Second International Conference on Environmental and Computer Science, 2009. ICECS '09*, pp. 33 – 36, 2009.

[41] "Open Data Madrid." `http://datos.madrid.es/portal/site/egob`, 2016. [Online; accessed April 2016].

[42] J. Lau, W. T. Hung, and C. S. Cheung, "Interpretation of air quality in relation to monitoring station's surroundings," *Atmospheric Environment*, vol. 43, no. 4, pp. 769–777, 2009.

[43] H. Mayer, "Air pollution in cities," *Atmospheric Environment*, vol. 33, no. 24-25, pp. 4029–4037, 1999.

[44] S. K. Pandey, K. H. Kim, S. Y. Chung, S. J. Cho, M. Y. Kim, and Z. H. Shon, "Long-term study of NOx behavior at urban roadside and background locations in Seoul, Korea," *Atmospheric Environment*, vol. 42, no. 4, pp. 607–622, 2008.

[45] M. Escudero, A. Lozano, J. Hierro, J. del Valle, and E. Mantilla, "Urban influence on increasing ozone concentrations in a characteristic Mediterranean agglomeration," *Atmospheric Environment*, vol. 99, pp. 322–332, 2014.

[46] "AEMET Open Data Repository." http://www.aemet.es/en/datos_abiertos/catalogo, 2016. [Online; accessed April 2016].

[47] I. Annex 3, "3, meteorological service for international air navigation," *International Civil*, 2010.

[48] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[49] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics and Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[50] J. Tu, Z. G. Xia, H. Wang, and W. Li, "Temporal variations in surface ozone and its precursors and meteorological effects at an urban site in China," *Atmospheric Research*, vol. 85, no. 3-4, pp. 310–337, 2007.

[51] H. Zou, Y. Yue, Q. Li, and Y. Shi, "A spatial analysis approach for describing spatial pattern of urban traffic state," in *13th International IEEE Conference on Intelligent Transportation Systems*, pp. 557–562, 2010.

[52] W. Weijermars and E. van Berkum, "Analyzing highway flow patterns using cluster analysis," in *IEEE Intelligent Transportation Systems Conference*, pp. 831–836, 2005.

[53] I. Laña, J. Del Ser, and I. Olabarrieta, "Understanding Daily Mobility Patterns in Urban Road Networks using Traffic Flow Analytics," in *UMITS*, (Istanbul), 2016.

[54] G. Raducan and S. Stefan, "Characterization of traffic-generated pollutants in Bucharest," *Atmosfera*, vol. 22, no. 1, pp. 99–110, 2009.

List of Tables

1 R^2 scores of 96 (+8) models 30
2 Wilcoxon p-values comparing pair of result sets. 31

List of Figures

1 Radial distribution of the road network around Madrid. . . . 32
2 Location of the 24 urban air quality stations deployed over Madrid: urban background (□), roadside traffic (☆) and Sub-urban (◯). . . . 33
3 Location of selected stations. . . . 34
4 Example of dataset instances. The last four target variables correspond to pollutant readings, and are used separately when the regression models are built. . . . 35
5 Comparison among six zones of the day-average traffic flow along the year. The X axis represents each one of the days in the sample, and is shorter than 365 days because no traffic data were available for the first 14 days of January and the whole December, and days with incomplete data were removed. 36
6 Average traffic per hour and day of the week in the RS-EA location. In this location, Sundays are similar to Saturdays, and both present more traffic by night and less by day than weekdays. . . . 37
7 Temperature and precipitation in Madrid during 2015. Temperature is shown as a line with a shaded region from its minimum to its maximum value. Precipitations are shown as the number of hours at which there were precipitations in each day. 38
8 A) Average month temperatures in 2015 compared with average month temperatures of the period 1995-2015. B) Monthly precipitations in 2015 compared with average month precipitations of the period 1995-2015. C) Monthly wind max speed in 2015 compared with average month max speed of the period 2012-2015. D) Monthly total UV radiation in 2015 compared with average month UV radiation of the period 2012-2015. . . 39
9 Daily averaged pollution levels for CO, NO, NO_2 and O_3 through 2015 for the 6 zones considered in this work. . . . 40
10 Daily averaged pollution levels for PM_{10} through 2015 comparing the two most antithetic traffic locations. . . . 41
11 Traffic and pollution levels through 2015 in RS-EA. . . . 42
12 Traffic and pollution levels through 2015 in UB-FA. . . . 43
13 Hourly average pollutant and traffic readings by day type and season in RS-EA. . . . 44

14 Feature importance of each variable for each dataset. Blank cells are part of datasets for which less ATR readings have been used due to the distance to the air monitoring station criteria. 45

Table 1: R^2 scores of 96 (+8) models

	① Temporal + Meteorology + Traffic						② Traffic + Meteorology					
	RS-EA	**RS-BP**	**RS-FL**	**UB-PC**	**UB-AS**	**UB-FA**	**RS-EA**	**RS-BP**	**RS-FL**	**UB-PC**	**UB-AS**	**UB-FA**
CO	**0.465**	**0.38**	**0.577**	0.474	**0.442**	0.143	0.325	0.281	0.384	0.306	0.205	0.304
NO	**0.35**	**0.326**	**0.437**	**0.397**	**0.316**	**0.391**	0.209	0.22	0.287	0.202	0.1	0.283
NO_2	**0.534**	0.489	0.494	**0.516**	**0.508**	0.528	0.355	0.342	0.332	0.297	0.265	0.371
O_3	**0.7**	0.724	0.714	**0.766**	**0.721**	**0.712**	0.538	0.578	0.554	0.596	0.515	0.587
PM_{10}	0,383	-	-	-	-	0,373	0.21	-	-	-	-	0.206
	③ Temporal + Traffic						④ Temporal + Meteorology					
CO	0.327	0.281	0.406	0.225	0.262	0.245	0.439	0.371	0.561	**0.479**	0.395	**0.427**
NO	0.214	0.198	0.27	0.166	0.176	0.193	0.311	0.321	0.43	0.404	0.264	0.376
NO_2	0.387	0.333	0.312	0.305	0.294	0.348	0.527	**0.49**	**0.498**	0.509	0.494	**0.543**
O_3	0.581	0.603	0.575	0.652	0.573	0.573	0.686	**0.726**	**0.718**	0.755	0.688	0.694
PM_{10}	0.223	-	-	-	-	0.258	**0.391**	-	-	-	-	**0.378**

Table 2: Wilcoxon p-values comparing pair of result sets.

	RS-EA	RS-BP	RS-FL	UB-PC	UB-AS	UB-FA
CO	0.05	0.57	0.05	0.50	0.01	0.16
NO	0.24	0.79	0.24	0.24	0.03	0.01
NO_2	0.38	0.95	0.87	0.24	0.02	0.87
O_3	0.07	0.20	0.20	0.01	0.01	0.16

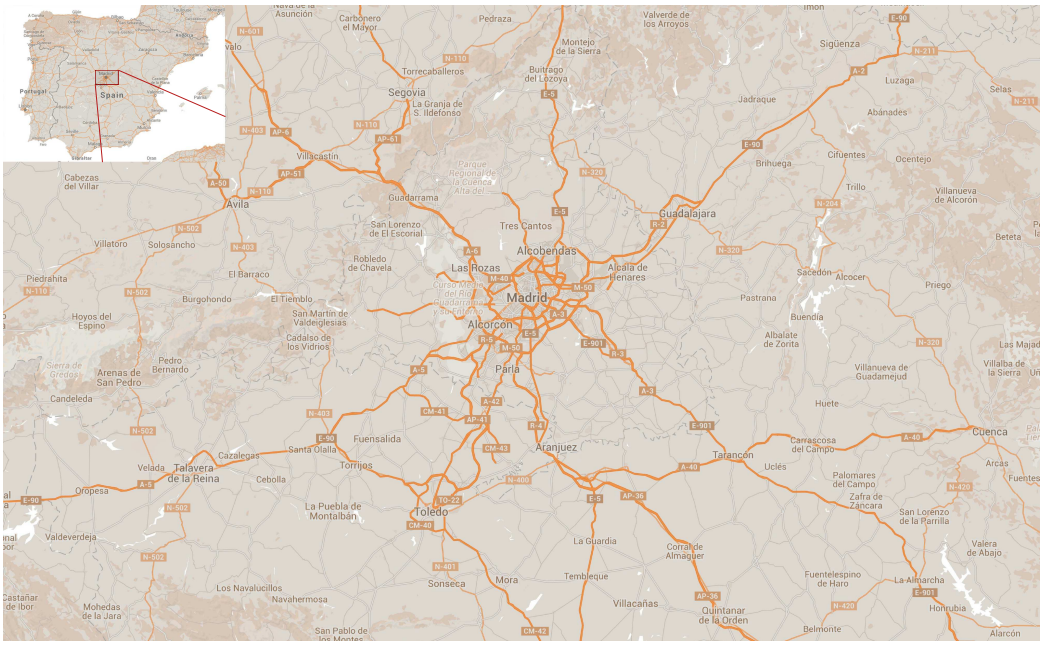

Figure 1: Radial distribution of the road network around Madrid.

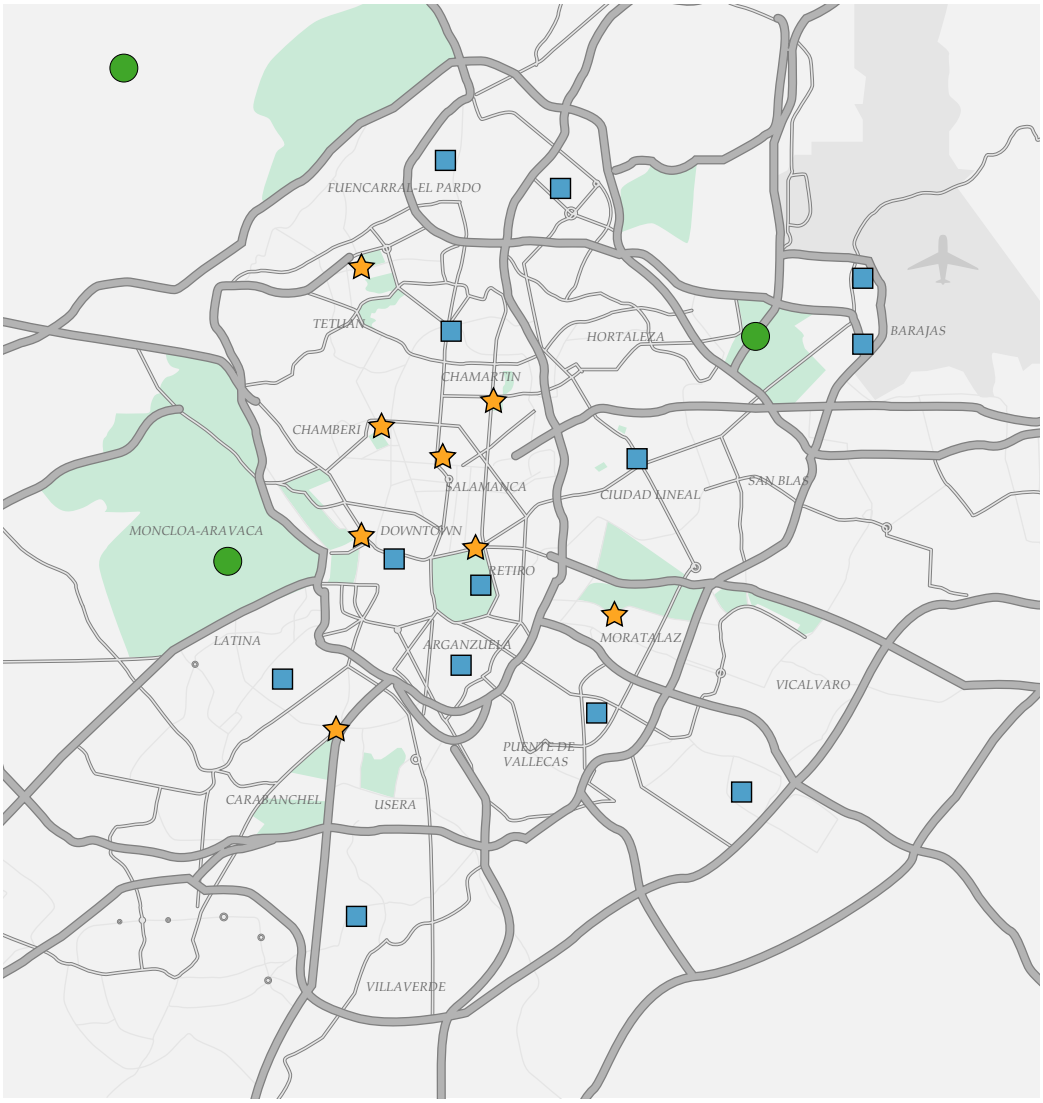

Figure 2: Location of the 24 urban air quality stations deployed over Madrid: urban background (□), roadside traffic (☆) and Suburban (◯).

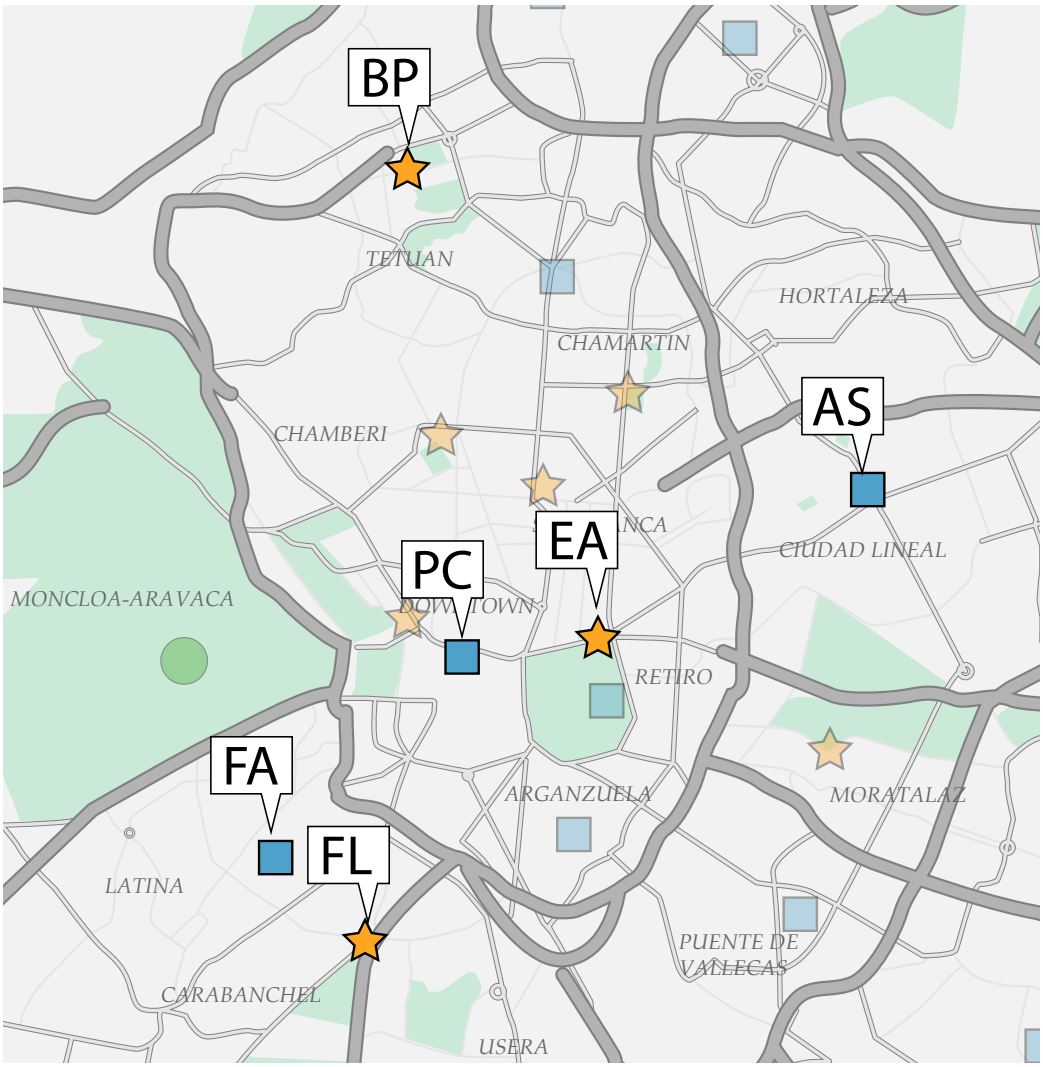

Figure 3: Location of selected stations.

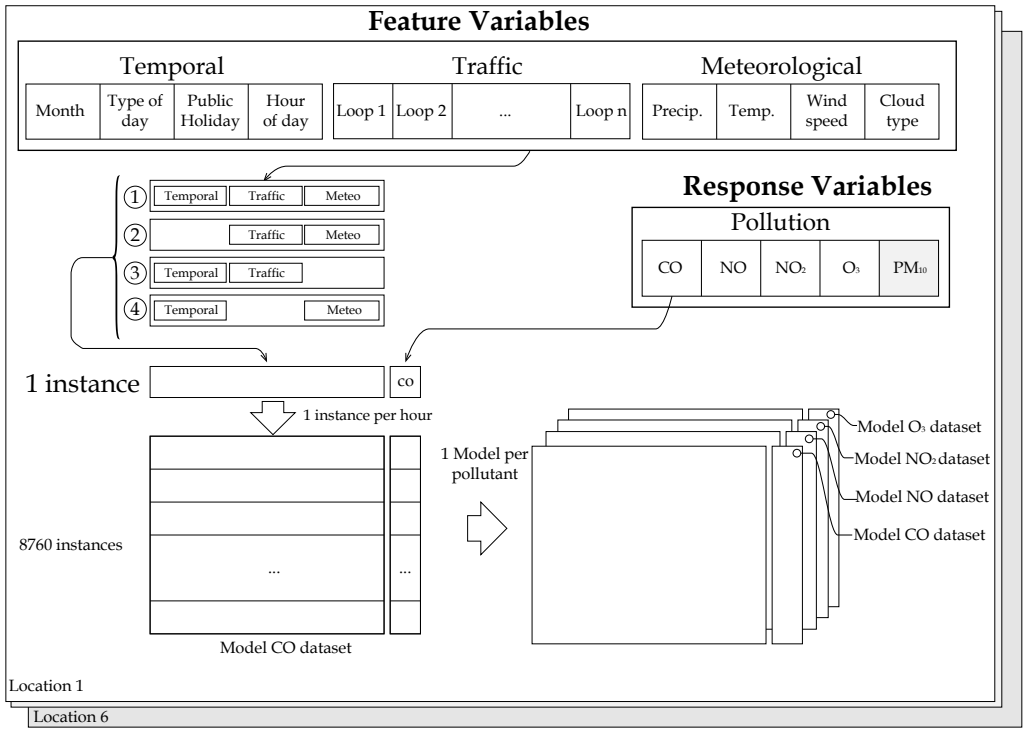

Figure 4: Example of dataset instances. The last four target variables correspond to pollutant readings, and are used separately when the regression models are built.

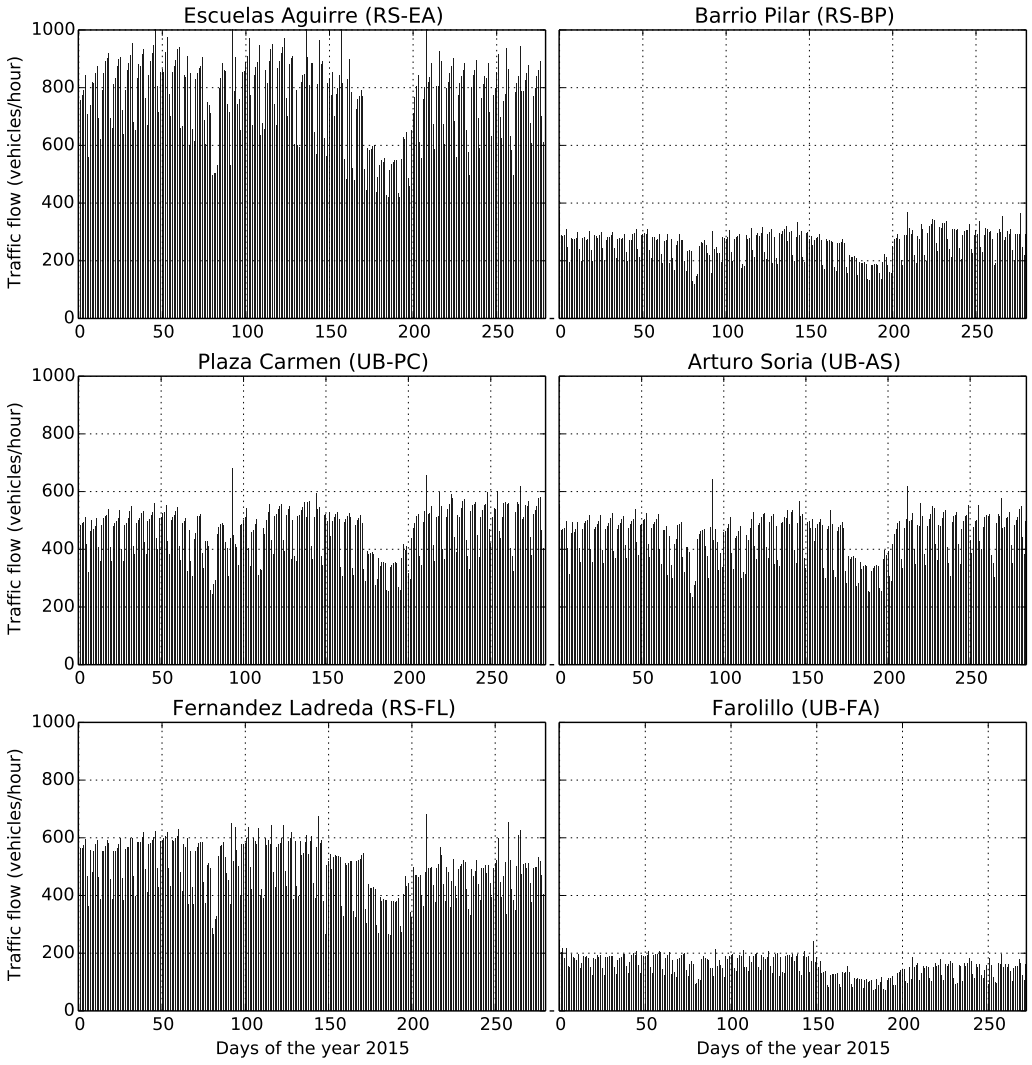

Figure 5: Comparison among six zones of the day-average traffic flow along the year. The X axis represents each one of the days in the sample, and is shorter than 365 days because no traffic data were available for the first 14 days of January and the whole December, and days with incomplete data were removed.

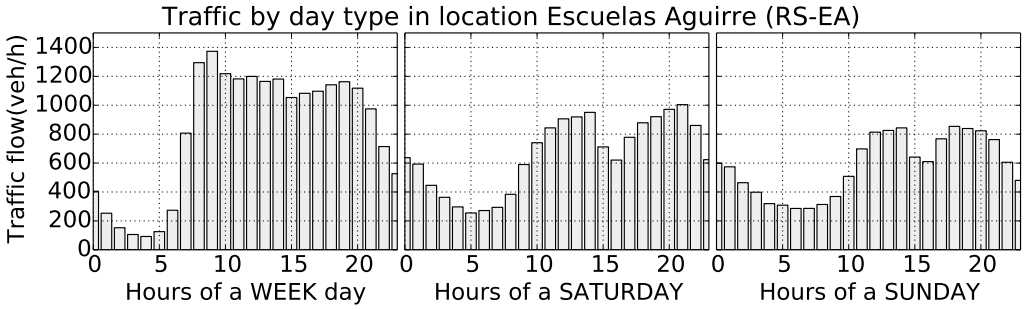

Figure 6: Average traffic per hour and day of the week in the RS-EA location. In this location, Sundays are similar to Saturdays, and both present more traffic by night and less by day than weekdays.

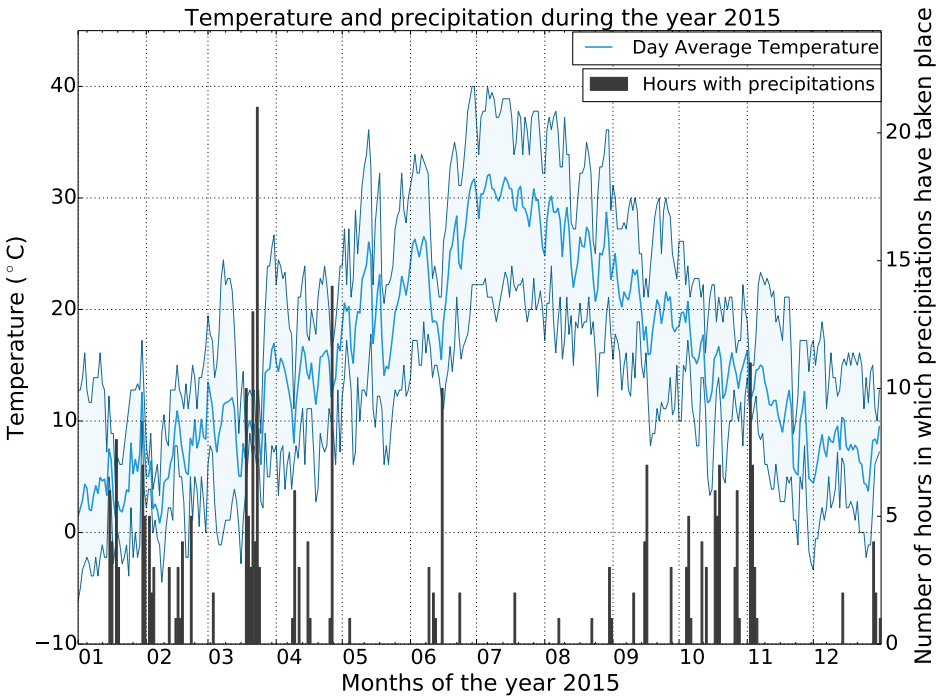

Figure 7: Temperature and precipitation in Madrid during 2015. Temperature is shown as a line with a shaded region from its minimum to its maximum value. Precipitations are shown as the number of hours at which there were precipitations in each day.

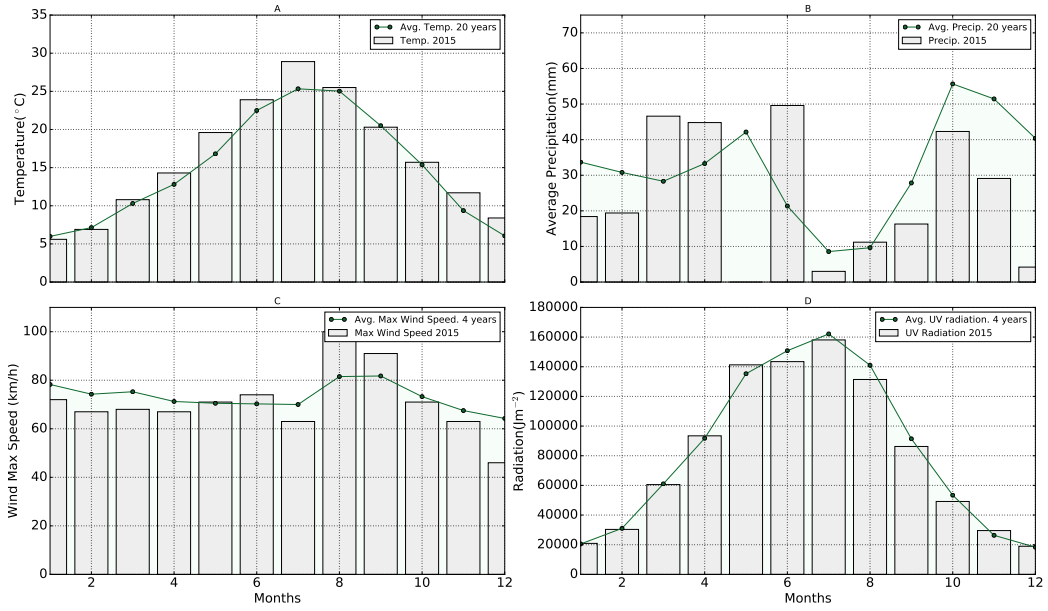

Figure 8: A) Average month temperatures in 2015 compared with average month temperatures of the period 1995-2015. B) Monthly precipitations in 2015 compared with average month precipitations of the period 1995-2015. C) Monthly wind max speed in 2015 compared with average month max speed of the period 2012-2015. D) Monthly total UV radiation in 2015 compared with average month UV radiation of the period 2012-2015.

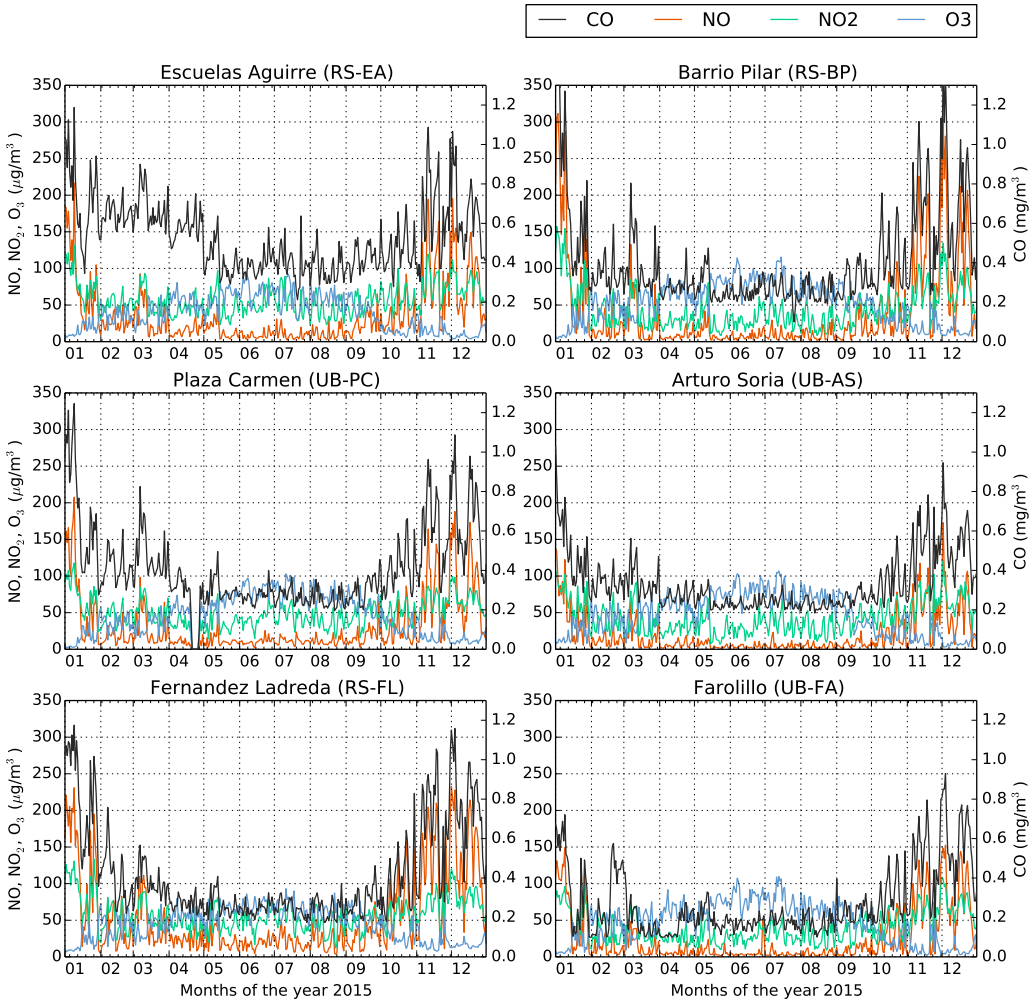

Figure 9: Daily averaged pollution levels for CO, NO, NO_2 and O_3 through 2015 for the 6 zones considered in this work.

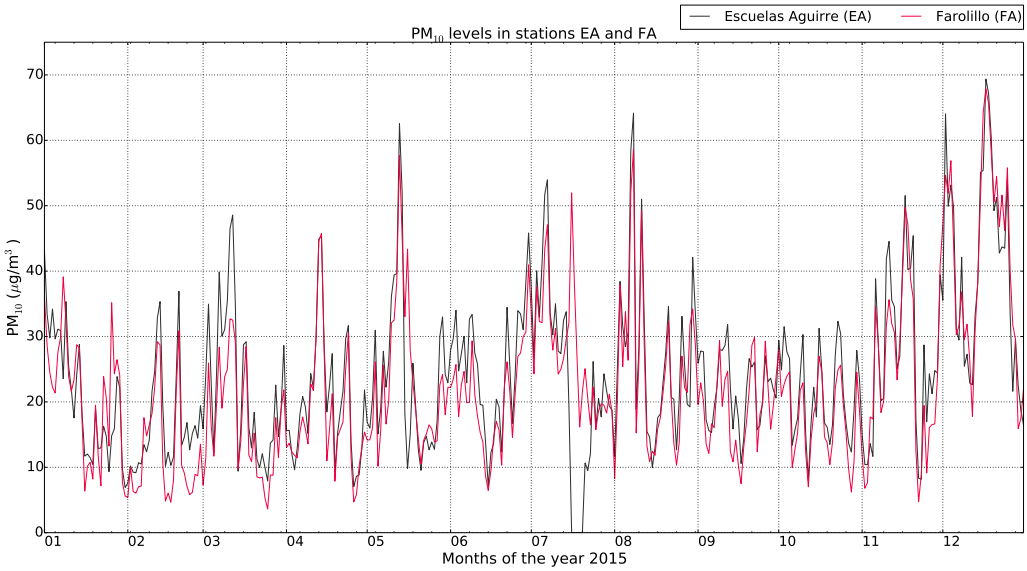

Figure 10: Daily averaged pollution levels for PM_{10} through 2015 comparing the two most antithetic traffic locations.

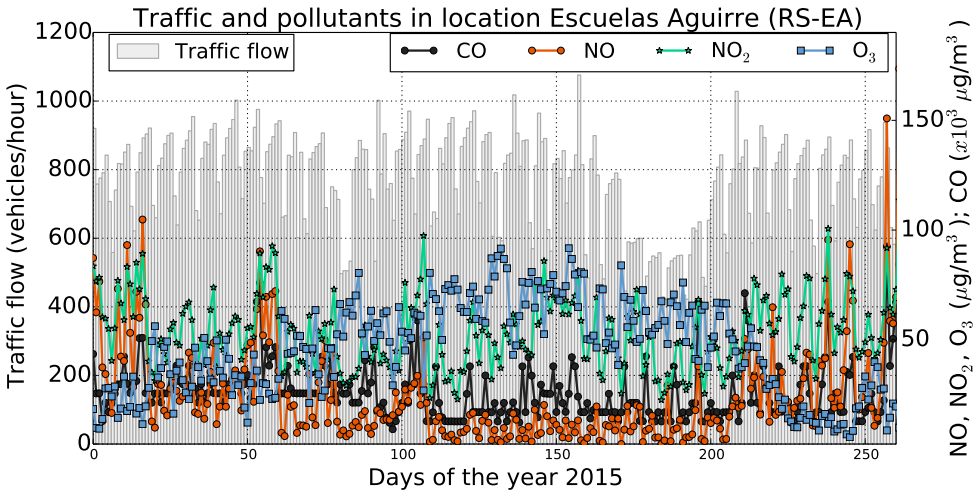

Figure 11: Traffic and pollution levels through 2015 in RS-EA.

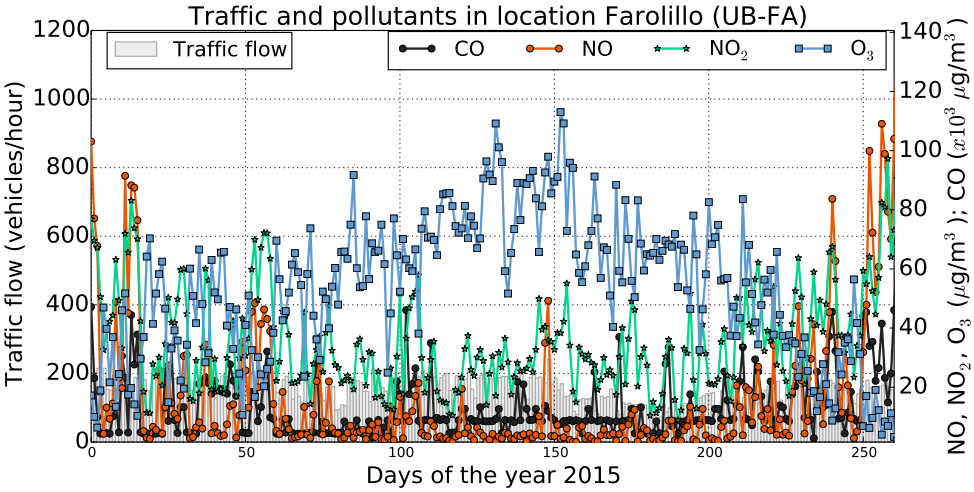

Figure 12: Traffic and pollution levels through 2015 in UB-FA.

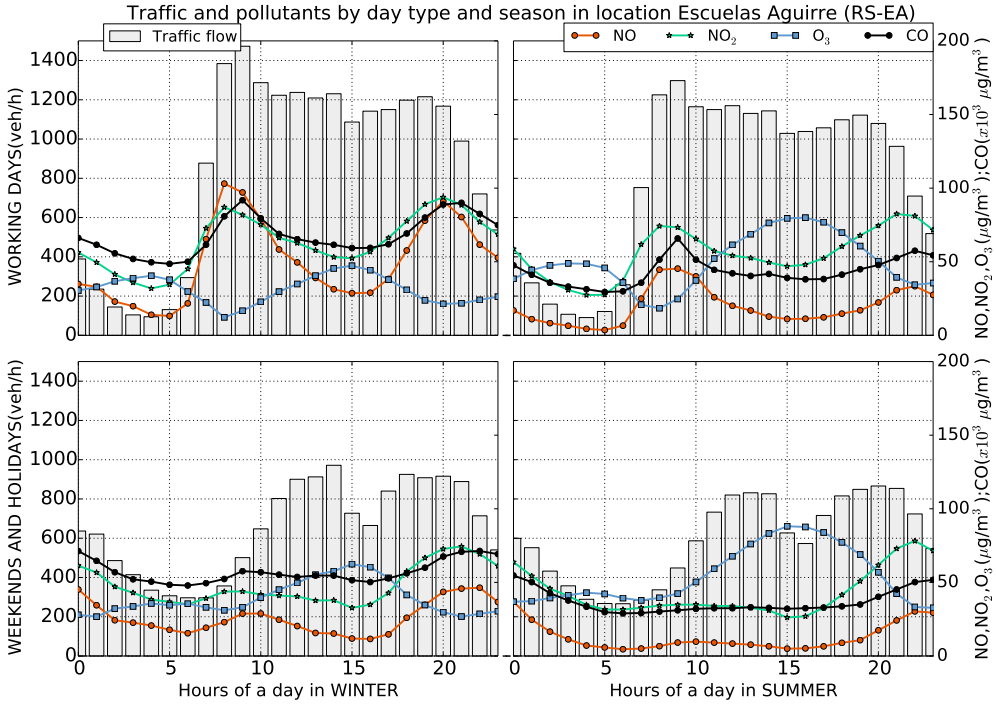

Figure 13: Hourly average pollutant and traffic readings by day type and season in RS-EA.

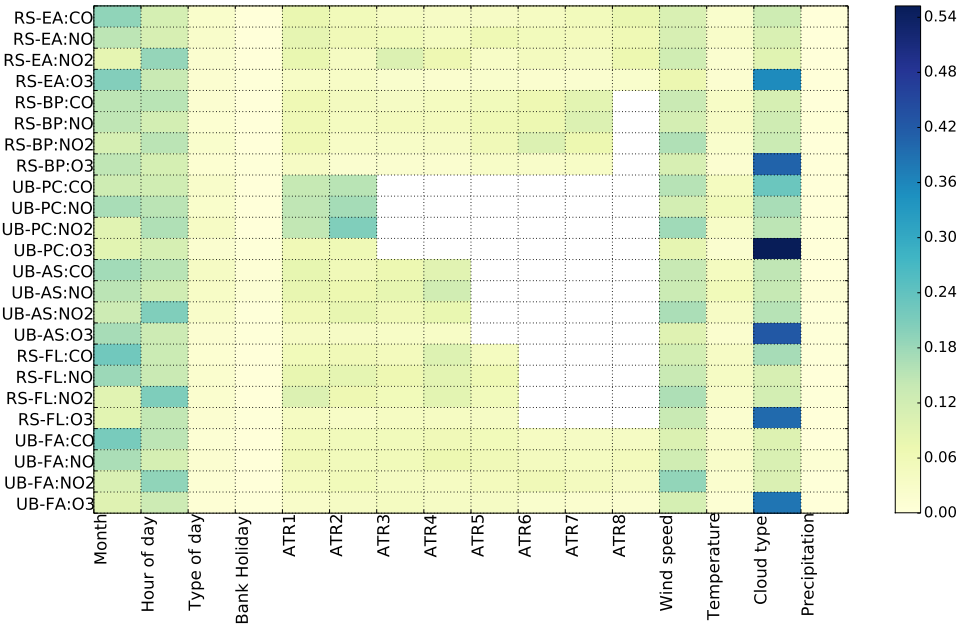

Figure 14: Feature importance of each variable for each dataset. Blank cells are part of datasets for which less ATR readings have been used due to the distance to the air monitoring station criteria.