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

Ibai Laña\textsuperscript{a}, Javier Del Ser\textsuperscript{a,b,c,*}, Ales Padró\textsuperscript{a}, Manuel Vélez\textsuperscript{b}, Carlos Casanova-Mateo\textsuperscript{d}

\textsuperscript{a}OPTIMA Unit. TECNALIA. P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain
\textsuperscript{b}Dept. of Communications Engineering. University of the Basque Country UPV/EHU. Alameda Urquijo S/N, 48013 Bilbao, Spain
\textsuperscript{c}Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Spain
\textsuperscript{d}Department 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\textsubscript{2}, O\textsubscript{3} and PM\textsubscript{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,

*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.

Preprint submitted to Atmospheric Environment September 13, 2016
traffic and pollution historic data in roadside and background locations of
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/km²) 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 con-
gestion 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, NO₅ and CO emissions are related to traffic in more than
80% in the city [2], 48% of PM₁₀ mass was proven to be contributed by vehicle
emissions [3], and 65% of tropospheric O₃ formation is on account of traffic-
related precursors [4]. This close relationship between traffic and pollution
comes along with severe health implications: indeed, worldwide epidemiolog-
ical 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, NO₂, CO and PM₁₀ have featured a de-
creasing trend in Madrid [11] as a result of the pollution abatement policies
promoted by the European Parliament (Directive 98/69/EC [12]). The im-
plementation 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 vehi-
cle 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 consumption [11]. On the other side, despite this NO\textsubscript{x} reduction an upward trend is found in tropospheric O\textsubscript{3} concentration in the last decade [4]. O\textsubscript{3} is formed within a complex photochemical process that requires, among others, anthropogenic and natural sources of NO\textsubscript{x} and Volatile Organic Compounds (VOC), collectively referred to as ozone precursors, enhanced by favorable meteorological conditions (high temperatures and strong solar radiation).

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. Precipitation can help dissipating heavier pollutants, while wind can help dispersing lighter ones [13, 14, 15, 16]. The literature also evinces that the lack of wind and precipitation combined with high-pressure atmospheric conditions curb pollutant dispersion, and high UV radiation levels unchain the formerly O\textsubscript{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 the 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: levels of NO\textsubscript{2} exceeded the 200 \(\mu g/m^3\) limit up to 95 times during 2015 in El Pilar district. Considering all air quality stations in the city, the NO\textsubscript{2} level limit has been exceeded an average of 23 times during the year. Likewise, levels of tropospheric O\textsubscript{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. On the contrary, PM\textsubscript{10} and CO levels remained 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]; although in some cities they do have an impact in pollutant levels [25], in other locations their relevance is milder. 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 the previous literature a great part of prediction models designed for this purpose hinge on neural networks [35]. Variations in the neural network model and improvements in the pre-processing of input data are introduced by [36] to
enhance its 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 benzopyrene 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 1 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$_x$) are directly related to road traffic, and particulate matters (PM$_{2.5}$ and PM$_{10}$) are also appreciably contributed by vehicle emissions [1, 44]. Ozone (O$_3$) levels depend mainly on meteorological conditions, but changes in NO$_x$ emissions strongly influence O$_3$ trends [45]. Therefore, the O$_3$ concentration embodies a good indicator of traffic-related pollution, specially in locations where cloudy weather conditions are infrequent as it 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. Regardless the latter condition, we have also included a suburban station (Casa de Campo) in the study in order to reinforce the analysis with a location relatively far from any local traffic source. The selected stations are depicted in Figure 1 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 Vía*, 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 Vía* 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.

• **Casa de Campo** (SU-CC): Suburban “super-station” in a large park in western Madrid. The closest roads are located at 1.6 km, which limits the impact of local traffic emissions in the measurements recorded in this site. Thus, models built for this station will be compared to those obtained for stations with closer local traffic sources.

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 three of them. The rest of stations provide only part of these pollutant agents, with several of them missing, not being useful for comparison purposes. Readings of the selected pollutants are published by the stations in µg/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 (flow, occupancy, load, level of service, speed) of inherent utility for different traffic management purposes.

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, a measure of vehicles per hour in a certain point. 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\textsuperscript{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. 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

\textsuperscript{1}Except for UB-FA, which is 520 meters from the closest ATR.
following data: Precipitation, ambient temperature, dew point temperature, wind speed and cloud cover.

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 – using the so-called coefficient of determination $R^2$ and the Mean Fractional Bias (MFB) metric – and variable importances obtained for each dataset. The $R^2$ 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 $\hat{y} = \{\hat{y}_n\}_{n=1}^N$ corresponding to the true instances $y = \{y_n\}_{n=1}^N$ will be given by

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

where $\bar{y} = \sum_{n=1}^N y_n / N$. The MFB score provides a symmetrical measure bounded in the range [-2, +2] that builds upon the concept of bias, which measures the tendency of a model to over- or under-predict [50]. A performance region is usually defined in the [-0.5, +0.5] interval that indicates a good performance of the model [51]. The best value of this score (MFB = 0) means that there is no bias between the predicted and the observed value. This score is computed over the predicted and observed values of the target variable as

$$ \text{MFB} = \frac{2}{N} \sum_{n=1}^N \frac{\hat{y}_n - y_n}{\hat{y}_n + y_n}, $$

An statistically meaningful estimation of the first-order statistics (mean, standard deviation) of the above scores will be computed using $K$-fold shuffled cross-validation over the datasets 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 2.

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, 3 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 3, 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. SU-CC location is not considered in these graphs, as this location has no local traffic.

As expected, Figure 3 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. UB-AS, holding more traffic, only exceeded the level 18 times (below the 20-times alert level). 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 [52]. 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 4 for the traffic captured in the RS-EA location. This traffic characterization is usual in the literature related to traffic modeling [53, 54] and temporal features were found to have a decisive relevance in long-term traffic forecasting [55]. 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 5 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 6A), 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 6B).

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 6C and 6D). The maximum wind speed attained in December is the most different recorded data (46 km/h vs average 64 km/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, the selected 6 stations are urban background and roadside, and they are placed in locations with dissimilar characteristics; different levels of pollution and traffic are expected. Figure 7 shows the distribution of CO, NO, NO2, and O3 through the year. General seasonal trends are visible in the figure. O3 reaches its maximum in summer months (35-120 µg/m³) when meteorological conditions facilitate its formation, whereas minimums (5-25 µg/m³) are found in winter months (less solar radiation, shorter days), coherently with other studies carried out in southern Europe [45]. NO2 levels remain stable through the year, closer to the 50 µg/m³ line in roadside traffic stations and to 40 µg/m³ in the urban background stations. Peaks attained by the CO and NO pollutants coincide when heating systems are active and ozone plunges.

PM10 levels are only available in 3 stations, and represented in Figure 8. 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 PM10 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 PM10, traffic is also a proven relevant source of this pollutant [3]. Furthermore the SU-CC station, far from traffic influence, presents lower yet similarly shaped PM10 levels through the year that constitute a background concentration over which traffic in the rest of locations is contributed. In light of the PM10 levels recorded in the rest of locations, traffic related share of this pollutant concentration can be concluded to be scarcely contributed
by local sources, and represents mainly a background traffic contribution, which is in line with the conclusions drawn in [11]. On the basis of this evidence, and not having PM$_{10}$ measurements for all the selected areas, this paper will focus on 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 9. 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 can be a priori expected, there is no apparent relation with traffic. When traffic plummets in August, NO and CO levels are maintained. The NO concentration even peaks over 30 µg/m$^3$ on August 26th, the 9 daylight hours with overcast cloud coverage, might be behind the low levels of O$_3$ (under 50 µg/m$^3$). In Figure 5 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 9 to the NO peak and O$_3$ valley after day 50. Traffic was about the same as in the previous week, but pollution increased. When examining the same data for UB-FA (the most different location), a similar detachment is found as shown in Figure 10. 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 11 traffic levels are plotted by day divided by working and non-working days, summer (April to October) and winter (November to March) months. Working and non-working days separation affects the traffic levels, and season separation impacts on both traffic and pollution levels. By inspecting these figures further, a closer relation between traffic and pollutants is discovered: when traffic starts in the first hours of the morning, specially in working
days, the formation of NO and CO is triggered. After 10 AM, and particularly in summer, the 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 112 datasets (Figure 2) and evaluated by applying shuffled cross-validation with 10 folds. For instance, the first dataset is created with temporal, traffic and meteorological variables as features, and CO measurement as the target variable in the RS-EA location. One data sample is created for every hour of the year, resulting in 8760 samples that are later cleansed by removing those with missing attributes. For the cross-validation, the 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, whereas the Mean Fractional Bias (MFB) of each model is obtained to evaluate the performance of the model.

Table 1 shows the averaged $R^2$ and MFB values of the 112 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 the parameters of the model were tuned for a better performance of the model, the obtained $R^2$ scores are in general low (under 0.7). The MFB scores were in general under 0.3, which reveal a fair performance of the models except for the NO concentrations. However, if predicting pollutant levels were the main purpose of these models the input features to each 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 also required; being a future instant would make necessary to predict the traffic and meteorological conditions at that point. Relying in predicted values as inputs to the predictive model would probably lead to worse results.
Anyhow, presented 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 the particular location or the predicted pollutant, the best \(R^2\) 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. The results obtained for the SU-CC location behave in a similar way for CO, NO, \(NO_2\) and \(O_3\): models with the three types of inputs and fed with temporal and meteorology information obtain the best performances, with barely significant differences among them. This represents a first evidence of a low relevance of local sources of traffic emissions, as models from all kind of locations – with acute differences in local traffic – perform in a very similar fashion. A posterior analysis will delve into this statement.

\(PM_{10}\) is also a pollutant influenced by traffic, but as seen in Section 3.3, it is affected mainly by general contributions, not local. The performance scores reported for the models in the SU-CC location bolster this idea: a much lower direct traffic influence helps this pollutant to be more predictable. Local traffic emissions in RS-EA and UB-FA may introduce small-scale variations that make it more difficult to predict. The relevance of traffic sources in the total \(PM_{10}\) concentration levels in Madrid could be addressed in depth in a future research. Other obtained results buttress further this observation: Table 1 (\(1\) and \(4\)) contain the best \(R^2\) scores, being very similar. Although \(1\) comprehends most of the best \(R^2\) scores obtained, and \(4\) all of best MFB scores (in bold type), the differences for both metrics 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 for the \(R^2\)
scores in order to confirm this conjecture and to shed light on the statistical
significance of such performance gaps. Table 2 shows the 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 significantly. 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 7
and 3 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. A deeper analysis, left out of the scope of this study,
should 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), the Wilcoxon test 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 12, 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 [55] 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 NO₂ levels are highly driven by the hour of day, as its production is linked to O₃ and the latter exhibit day and night cycles. Wind speed is a good NO₂ 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). The importance of ATR readings is in general low, which not only validates the R² scores and the analysis, but also finds that in all cases traffic levels relevance for predicting CO, NO or NO₂ double their relevance for predicting O₃.

Although in Table 1 O₃ 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 O₃ 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 noisy 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 effects on particular sites are limited and stringently linked to localities. Other research contributions such as [44, 56, 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. An additional suburban station was also analyzed to account for the influence of regional background PM values on the local measurements.

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).


List of Tables

1. $R^2$/ MFB scores of 112 (+12) models. .................................. 31
2. Wilcoxon p-values comparing pairs of $R^2$ result sets. ........... 32
List of Figures

1. Radial distribution of the road network around Madrid, and location of the 24 urban air quality stations deployed over Madrid: urban background (☐), roadside traffic (⋆) and suburban (○). Selected stations are tagged. 33

2. Example of dataset instances. The last four target variables correspond to pollutant readings, and are used separately when the regression models are built. 34

3. 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. 35

4. 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. 36

5. 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. 37


7. Daily averaged pollution levels for CO, NO, NO₂ and O₃ through 2015 for the 6 zones considered in this work. 39

8. Daily averaged pollution levels for PM₁₀ through 2015 comparing the two most antithetic traffic locations, and the control location SU-CC. 40

9. Traffic and pollution levels through 2015 in RS-EA. 41

10. Traffic and pollution levels through 2015 in UB-FA. 42

11. Hourly average pollutant and traffic readings by day type and season in RS-EA. 43
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.
Table 1: $R^2$/ MFB scores of 112 (+12) models.

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CO</strong></td>
<td>0.465/0.107</td>
<td>0.38/0.109</td>
<td>0.577/0.098</td>
<td>0.474/0.143</td>
<td>0.442/0.134</td>
<td>0.143/0.209</td>
<td>0.297/0.144</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>0.35/0.338</td>
<td>0.326/0.472</td>
<td>0.437/0.312</td>
<td>0.397/0.416</td>
<td>0.316/0.304</td>
<td>0.391/0.423</td>
<td>0.317/0.355</td>
</tr>
<tr>
<td><strong>NO$_2$</strong></td>
<td>0.534/0.076</td>
<td>0.489/0.148</td>
<td>0.494/0.077</td>
<td>0.516/0.139</td>
<td>0.508/0.084</td>
<td>0.528/0.126</td>
<td>0.532/0.218</td>
</tr>
<tr>
<td><strong>O$_3$</strong></td>
<td>0.7/0.044</td>
<td>0.724/0.076</td>
<td>0.714/0.053</td>
<td>0.766/0.052</td>
<td>0.721/0.061</td>
<td>0.712/0.111</td>
<td>0.694/0.047</td>
</tr>
<tr>
<td><strong>PM$_{10}$</strong></td>
<td>0.383/0.159</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.373/0.173</td>
<td>0.482/0.061</td>
</tr>
</tbody>
</table>

**Temporal + Meteorology + Traffic**

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CO</strong></td>
<td>0.325/0.148</td>
<td>0.281/0.138</td>
<td>0.384/0.117</td>
<td>0.306/0.171</td>
<td>0.205/0.162</td>
<td>0.304/0.22</td>
<td>0.153/0.149</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>0.209/0.377</td>
<td>0.22/0.538</td>
<td>0.287/0.342</td>
<td>0.202/0.454</td>
<td>0.1/0.371</td>
<td>0.283/0.509</td>
<td>0.167/0.5</td>
</tr>
<tr>
<td><strong>NO$_2$</strong></td>
<td>0.355/0.1</td>
<td>0.342/0.184</td>
<td>0.332/0.096</td>
<td>0.297/0.165</td>
<td>0.265/0.112</td>
<td>0.371/0.16</td>
<td>0.382/0.249</td>
</tr>
<tr>
<td><strong>O$_3$</strong></td>
<td>0.538/0.057</td>
<td>0.578/0.091</td>
<td>0.554/0.0623</td>
<td>0.596/0.062</td>
<td>0.515/0.085</td>
<td>0.587/0.134</td>
<td>0.581/0.054</td>
</tr>
<tr>
<td><strong>PM$_{10}$</strong></td>
<td>0.21/0.22</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.206/0.233</td>
<td>0.393/0.101</td>
</tr>
</tbody>
</table>

**Traffic + Meteorology**

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CO</strong></td>
<td>0.439/0.099</td>
<td>0.371/0.097</td>
<td>0.561/0.091</td>
<td>0.479/0.138</td>
<td>0.395/0.124</td>
<td>0.427/0.193</td>
<td>0.307/0.132</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>0.311/0.311</td>
<td>0.321/0.422</td>
<td>0.43/0.287</td>
<td>0.404/0.391</td>
<td>0.264/0.273</td>
<td>0.376/0.347</td>
<td>0.317/0.28</td>
</tr>
<tr>
<td><strong>NO$_2$</strong></td>
<td>0.527/0.066</td>
<td>0.49/0.134</td>
<td>0.498/0.072</td>
<td>0.509/0.129</td>
<td>0.494/0.078</td>
<td>0.543/0.108</td>
<td>0.518/0.188</td>
</tr>
<tr>
<td><strong>O$_3$</strong></td>
<td>0.686/0.042</td>
<td>0.726/0.077</td>
<td>0.718/0.052</td>
<td>0.755/0.052</td>
<td>0.688/0.06</td>
<td>0.694/0.097</td>
<td>0.695/0.04</td>
</tr>
<tr>
<td><strong>PM$_{10}$</strong></td>
<td>0.391/0.183</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.378/0.192</td>
<td>0.479/0.071</td>
</tr>
</tbody>
</table>

**Temporal + Traffic**

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CO</strong></td>
<td>0.327/0.109</td>
<td>0.281/0.115</td>
<td>0.406/0.094</td>
<td>0.225/0.127</td>
<td>0.262/0.141</td>
<td>0.245/0.205</td>
<td>0.063/0.147</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>0.214/0.325</td>
<td>0.198/0.44</td>
<td>0.27/0.292</td>
<td>0.166/0.426</td>
<td>0.176/0.309</td>
<td>0.193/0.435</td>
<td>0.161/0.377</td>
</tr>
<tr>
<td><strong>NO$_2$</strong></td>
<td>0.387/0.084</td>
<td>0.333/0.163</td>
<td>0.312/0.08</td>
<td>0.305/0.133</td>
<td>0.294/0.097</td>
<td>0.348/0.146</td>
<td>0.282/0.241</td>
</tr>
<tr>
<td><strong>O$_3$</strong></td>
<td>0.581/0.046</td>
<td>0.603/0.078</td>
<td>0.575/0.053</td>
<td>0.652/0.058</td>
<td>0.573/0.062</td>
<td>0.573/0.117</td>
<td>0.53/0.054</td>
</tr>
<tr>
<td><strong>PM$_{10}$</strong></td>
<td>0.223/0.221</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.258/0.216</td>
<td>0.255/0.153</td>
</tr>
</tbody>
</table>

**Temporal + Meteorology**

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CO</strong></td>
<td>0.311/0.321</td>
<td>0.422/0.43</td>
<td>0.287/0.404</td>
<td>0.391/0.264</td>
<td>0.273/0.376</td>
<td>0.347/0.317</td>
<td>0.28</td>
</tr>
<tr>
<td><strong>NO</strong></td>
<td>0.311/0.311</td>
<td>0.321/0.422</td>
<td>0.43/0.287</td>
<td>0.404/0.391</td>
<td>0.264/0.273</td>
<td>0.376/0.347</td>
<td>0.317/0.28</td>
</tr>
<tr>
<td><strong>NO$_2$</strong></td>
<td>0.527/0.066</td>
<td>0.49/0.134</td>
<td>0.498/0.072</td>
<td>0.509/0.129</td>
<td>0.494/0.078</td>
<td>0.543/0.108</td>
<td>0.518/0.188</td>
</tr>
<tr>
<td><strong>O$_3$</strong></td>
<td>0.686/0.042</td>
<td>0.726/0.077</td>
<td>0.718/0.052</td>
<td>0.755/0.052</td>
<td>0.688/0.06</td>
<td>0.694/0.097</td>
<td>0.695/0.04</td>
</tr>
<tr>
<td><strong>PM$_{10}$</strong></td>
<td>0.391/0.183</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.378/0.192</td>
<td>0.479/0.071</td>
</tr>
</tbody>
</table>
Table 2: Wilcoxon p-values comparing pairs of $R^2$ result sets.

<table>
<thead>
<tr>
<th></th>
<th>RS-EA</th>
<th>RS-BP</th>
<th>RS-FL</th>
<th>UB-PC</th>
<th>UB-AS</th>
<th>UB-FA</th>
<th>SU-CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>0.05</td>
<td>0.57</td>
<td>0.05</td>
<td>0.50</td>
<td>0.01</td>
<td>0.16</td>
<td>0.23</td>
</tr>
<tr>
<td>NO</td>
<td>0.24</td>
<td>0.79</td>
<td>0.24</td>
<td>0.24</td>
<td>0.03</td>
<td>0.01</td>
<td>0.71</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>0.38</td>
<td>0.95</td>
<td>0.87</td>
<td>0.24</td>
<td>0.02</td>
<td>0.87</td>
<td>0.51</td>
</tr>
<tr>
<td>O$_3$</td>
<td>0.07</td>
<td>0.20</td>
<td>0.20</td>
<td>0.01</td>
<td>0.01</td>
<td>0.16</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Figure 1: Radial distribution of the road network around Madrid, and location of the 24 urban air quality stations deployed over Madrid: urban background (□), roadside traffic (☆) and Suburban (○). Selected stations are tagged.
Figure 2: Example of dataset instances. The last four target variables correspond to pollutant readings, and are used separately when the regression models are built.
Figure 3: 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 4: 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 5: 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 7: Daily averaged pollution levels for CO, NO, NO\textsubscript{2} and O\textsubscript{3} through 2015 for the 6 zones considered in this work.
Figure 8: Daily averaged pollution levels for PM$_{10}$ through 2015 comparing the two most antithetic traffic locations, and the control location SU-CC.
Figure 9: Traffic and pollution levels through 2015 in RS-EA.
Figure 10: Traffic and pollution levels through 2015 in UB-FA.
Figure 11: Hourly average pollutant and traffic readings by day type and season in RS-EA.
Figure 12: 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.