A Machine Learning Approach to Monitor Air Quality from Traffic and Weather data

Claudio Rossi
LINKS Foundation, Italy
claudio.rossi@linksfoundation.com

Alessandro Farasin
Politecnico di Torino, Italy
& LINKS Foundation, Italy
alessandro.farasin@polito.it

Giacomo Falcone
LINKS Foundation, Italy
giacomo.falcone@linksfoundation.com

Carlotta Castelluccio
Microsoft, Italy
carlotta.castelluccio@microsoft.com

Abstract.

Knowing the amount of air pollutants in our cities is of great importance to help decision makers in the definition of effective strategies aimed at maintaining a good air quality, which is a key factor for a healthy life, especially in urban environments. Using a data set from a big metropolitan city, we realize the uAQE: urban Air Quality Evaluator, which is a supervised machine learning model able to estimate air pollutants values using only weather and traffic data. We evaluate the performance of our solution by comparing the predicted pollutant values with the real measurements provided by professional air monitoring stations. We use the predicted pollutants to compute a standard Air Quality Index (AQI) and we map it into a set of five qualitative AQI classes, which can be used for decision making at the city level. uAQE is able to predict the AQI class value with an accuracy of 0.8.

Keywords: Air Quality · Machine Learning · BRNN · Weather · Traffic

1 Introduction and Related Works

Air pollution introduces into the atmosphere chemicals, particulates, or biological materials that causes discomfort, disease, or death to humans and to other living organisms alike. More than 5.5 million people worldwide are dying prematurely every year as a result of air pollution exposure [1]. This fact confirms that air pollution is one of the world's largest environmental health risks. Most of these deaths are occurring in rapidly developing economies, e.g., China and India, but also in European metropolitan cities, e.g., Naples, Turin and Milan, which have an air pollution index among the highest ones according to recent rankings [1].
Road transport is one of the main causes of air pollutants emissions, accounting for the 14% of the total emissions in European countries [2]. In recent years, advanced after-treatment technology (Particulate Matter traps)

In: I. Chatzigiannakis, P. Ciampolini, S. Hanke, G. Mylonas: Proceedings of the BRAINS Workshop, Rome, Italy, 13-Nov-2019, published at http://ceur-ws.org

[1] http://www.numbeo.com/pollution/rankings.jsp
[2] http://www.eea.europa.eu

have been implemented by car manufacturers in response to increasingly tight European emission policies. Consequently, emissions of the main regulated pollutants from road transport, i.e., Nitrogen Oxides (NO_x) and Particulate Matter (PM), have been reduced despite the increased vehicle activity. Specifically, urban NO_x emissions from traffic has been reduced by 16% between 2000 and 2010, mainly due to the introduction of the Euro 4 and the Euro 5 standards for passenger cars (both petrol and diesel), which were applied from early 2005 and late 2009, respectively [3].
Other human activities having a strong impact on air quality are industrial processes, farming, heat and air conditioning, and other types of transport (trains, airplanes,etc.).
It is a well-known fact that weather phenomena have a strong impact on air pollutants because once pollutants are emitted into the air, they propagate into the atmosphere according to weather conditions, e.g., turbulence mixes pollutants into the surrounding air, and wind carries them away from the source location. Conversely, when the air near the surface of the earth is cooler than the air above (a phenomenon called temperature inversion) there is very little air mixing. Since cool air is heavy, it will not to move up to mix with the warmer air above. Thus, any pollutants released near the surface will get trapped and build up in the cooler air layer.
Municipalities struggle to predict the effect of traffic policies, e.g., total traffic block, stop of most pollutant vehicles, on the air quality because there is a lack of easy-to-use tools that can estimate the air pollution taking into account also the meteorological predictions. Furthermore, the availability of air quality measurement stations in a city is very limited due to economic constrains. A professional station requires a non negligible investment (about 200k € per installation) and it has a high maintenance cost (about 30k€ per year) [2].
Because of its importance, the estimation of the air quality has been subject to some studies. In [2], Microsoft researchers proposed a semi supervised learning approach able to predict PM_{10} and Nitrogen Dioxide (NO_2) emissions at an higher spatial resolution with respect to the one achieved by the installed air quality sensors by coupling other data sources such as traffic flows, the structure of the road network, meteorological conditions and point of interest locations. Their solution is complementary to ours, and it can can be used to improve the spatial resolution of the uAQE. Other relevant studies include the [4] and [5], which present a set of learning methods able to predict NO_x concentrations from past observations and weather conditions. In [6], the authors studied Delhi's $PM_{2.5}$ concentrations and its correlation with the vehicular traffic and with the weather conditions. However, the proposed model makes several empirical assumptions and it includes parameters specific to the city of Delhi. Hence, it cannot be re-used for our purpose.
To help decision maker in keeping under control the air quality we propose uAQE: urban Air Quality Evaluator, which is a set of supervised machine learning model able to predict air pollutants values in a urban environment using only weather and traffic data. We train our models with data taken from Milan, building one model for each air pollutants. Our work is different from all the above mentioned approaches because we aim to predict pollutants without requiring data from air quality stations. Note that we train one model for each air pollutants, namely Nitrogen Dioxide (NO_2), Ozone (O_3), Carbon Monoxide (CO), Benzene (C_6H_6), Total Nitrogen (N_2), Particulate Matter (PM_{10}), Sulfur Dioxide (SO_2), Particulate Matter ($PM_{2.5}$), Black Carbon (BC), and Ammonia (NH_3). We present the accuracy of each model using the pollutants as measured by professional air stations. Following a regional standard, we use the predicted pollutants to compute an Air Quality Index (AQI) which is then mapped it into a set of five qualitative classes that are used to manage air quality policies at city level. We finally asses the classification accuracy achieved by uAQE obtaining a value of 0.8.

2 Input data

Our data has been collected in the city of Milan during two months (Nov.- Dec. 2013), and it contains three distinct data categories:

- **Weather**: we have six different weather stations placed within the city limit. Each station has a unique *ID*, *type*, *location*, and it features a set of co-located sensors. Each sensor measures a different meteorological phenomena. This information has been obtained thanks to ARPA (Agenzia Regionale per la Protezione dell'Ambiente) [3].
- **Traffic**: through fixed video cameras already installed for traffic access control at 52 locations in the central area of Milan (*Cerchia dei Bastioni*) the local authority obtained the plate number of transiting vehicles, from which the vehicle characteristic could be extracted from the official database, i.e., the *Motorizzazione civile*, which holds the information of all Italian vehicles. Note that we received anonymized data, i.e., with

[3] http://ita.arpalombardia.it/ITA/qaria/doc_RichiestaDati.asp

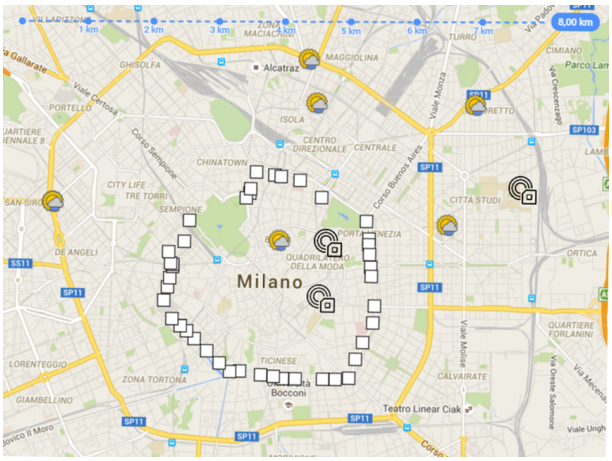

Fig. 1. Sensors locations map. White squares are fixed cameras (traffic gates), meteo icons represent weather station, and black icons are air stations.

hashed plate numbers and with no information about the vehicle owner. Therefore, only the technical details of each vehicle has been made available to us. These data have been provided as open data by the city of Milan.

– **Air**: we take the measurements of three different air stations located within the city limits. Each station features multiple co-located sensors, each of which measures a single air pollutant. Also these measurements are directly provided as open data by ARPA, who is the official source of this kind of data.

The locations of weather stations, air stations and fixed cameras are shown in Fig. 1.

The **weather station data** contain wind direction (degree), wind speed (m/s), temperature (Celsius degree), relative humidity (%), precipitation (mm), global radiation ($\mu W/m^2$), net radiation ($\mu W/m^2$), and atmospheric pressure (hPa). The **traffic data** include each vehicle passage at each gate, for which the location and the timestamp of each passage is known. For each passage, the vehicle characteristics are given, namely the European emission standard category (EURO category from 1 to 6), the vehicle type (i.e., bus, freight, transport, people transport or not available), the fuel type (i.e., petrol, diesel, electric, LPG, hybrid or missing), the presence of the Diesel Particle Filter (DPF) and the vehicle length expressed in mm.

The **air pollution data** contain ten different agents: NO_2 ($\mu g/m^3$), NH_3 ($\mu g/m^3$), C_6H_6 ($\mu g/m^3$), SO_2 ($\mu g/m^3$), BC ($\mu g/m^3$), CO ($\mu g/m^3$), N_2 (ppb), PM_{10} ($\mu g/m^3$), $PM_{2.5}$ ($\mu g/m^3$), O_3 ($\mu g/m^3$).

We compute the hourly air quality index as defined by Piedmont index AQI because is the only example of operational use of an air quality index in Italy[4]. The AQI uses only three pollutants, namely NO_2, PM_{10}, O_3, and it is formulated as follows:

$$I_{PM_{10}} = \frac{V_{med24h_{PM_{10}}}}{V_{rif_{PM_{10}}}} \times 100 \tag{1}$$

$$I_{NO_2} = \frac{V_{maxh_{NO_2}}}{V_{rif_{NO_2}}} \times 100 \tag{2}$$

$$I_{8h_{O_3}} = \frac{V_{max8h_{O_3}}}{V_{rif_{8h_{O_3}}}} \times 100 \tag{3}$$

$$I_{AQI} = \frac{I_{PM_{10}} + max(I_{NO_2}, I_{O_3})}{2} \tag{4}$$

We observe that in the considered data set O_3 never exceeds the maximum value established for preserving human health (i.e., $120\mu g/m^3$), whereas NO_2 exceeds its hourly maximum value (i.e., $200\mu g/m^3$) only in few cases ($< 5\%$). Conversely, PM_{10} exceeds the the daily maximum value (i.e., $50\mu g/m^3$) in 50% in the cases.

We map the computed AQI in the five classes defined by the Piedmont region, namely Optimal ($0 \leq AQI < 50$), Good ($50 \leq AQI < 75$), Fair ($75 \leq AQI < 100$), Average ($100 \leq AQI < 125$), Not Very Healthy ($125 \leq AQI <$

[4] http://www.arpae.it/cms3/documenti/aria/IQA.pdf

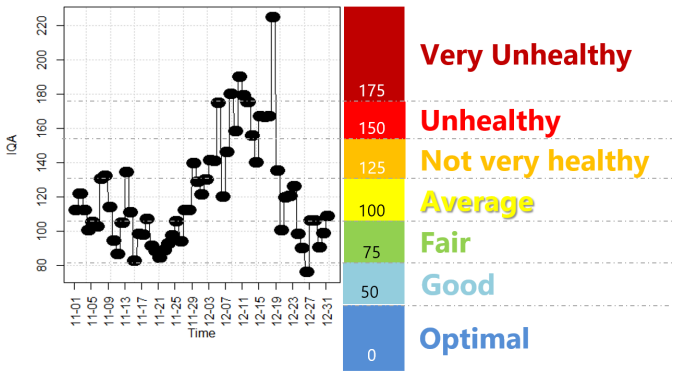

Fig. 2. Temporal evolution of the Air Quality Index computed in the considered time frame.

150), Unhealthy ($150 \leq AQI < 175$), Very Unhealthy ($AQI \geq 175$). In our data set, we observe that there are no AQI values in the Optimal level and very few in the Good one, while the most part of values ($\approx 80\%$) fall between Fair and Not Very Healthy levels. We show in Fig. 2 the temporal evolution of the AQI.

3 Feature Construction

Our aim is to use a supervised machine learning approach to predict pollutants from traffic and weather data under the hypothesis that we do not have air sensors to directly measure air pollutants. We show the empirical cumulative distribution function (ecdf) of the considered weather data in Fig.3 and the distribution of the traffic features in Fig.4. For sake of readability, the ecdfs related to the air pollutants are shown in Fig.7, in the Appendix.
The aforementioned data categories, namely *weather, traffic and air pollutants, are merged at hourly resolution*, according to the maximal temporal resolution of both weather and air data. Then, *measurements produced by* different *sensors* of the *same type* in the *same hour are averaged*.
Finally, the *hourly passages at vehicle gates are counted*, according to four different sets: (i) the EURO class (EURO), (ii) the vehicle type (Vtype), (iii) the fuel type (Ftype), (iv) existence of Diesel Particulate Filter (DPF).
In order to perform all data manipulations, we use *R* and the *plyr* library, which provides data aggregation operators. As a final step, we filled a small percentage of missing values ($< 1\%$) in the air and weather data by polynomial interpolation using the spline function of the *zoo* library. Conversely, because of the remarkable percentage of NA values for each group in the traffic features and due to not available information, we filled missing values following the probability distribution of data.

The final feature set is composed by the following variables:

- **Time**: day of week (1-7), hour (1-24). This is to consider the regular patterns of human activities, which are framed within the day and within the week;
- **Hourly passages**: counts of total passages and aggregated count by EURO class, vehicle type, fuel type, existence of particulate filter. Because we compute the total passages, we remove one category from each aggregation to avoid creating features which are linear combination of other ones while reducing the number of total features;
- **Hourly weather phenomena averages**: wind direction, wind speed, temperature, relative humidity, precipitation, atmospheric pressure.

In order to consider the effect of the past on the current pollutants levels, for each traffic and weather feature $f(t)$ (wind direction excluded) we add another feature $fp(t)$ equal to the sum of $f(t)$ over the last x time slots ($fp(t) = \sum_{t=i}^{t=i-x} f(t)$). We evaluated increasing values of x starting from 1 and and we empirically found the best value to be 12.
Studying the correlation between weather and pollutants (Fig. 5) we noticed that temperature, relative humidity, precipitation wind speed, and atmospheric pressure are the most significant ones, having an average absolute

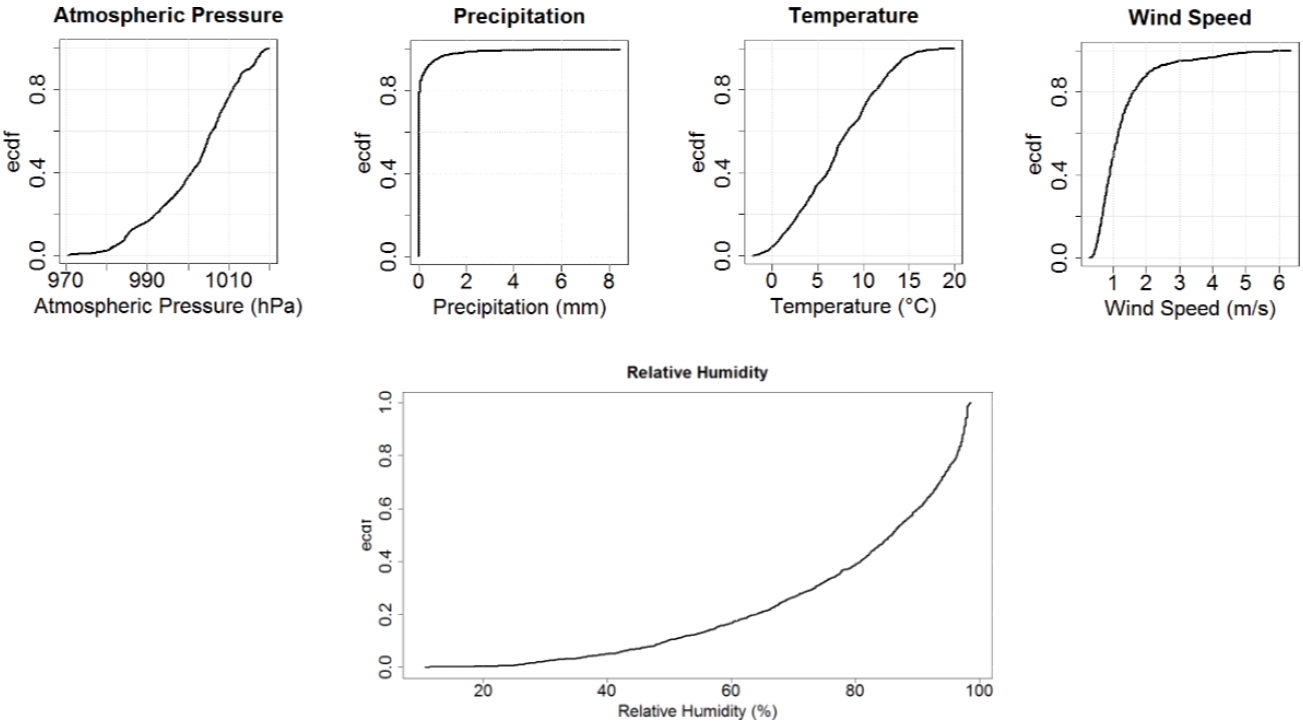

Fig. 3. Weather features Ecdf Graphs

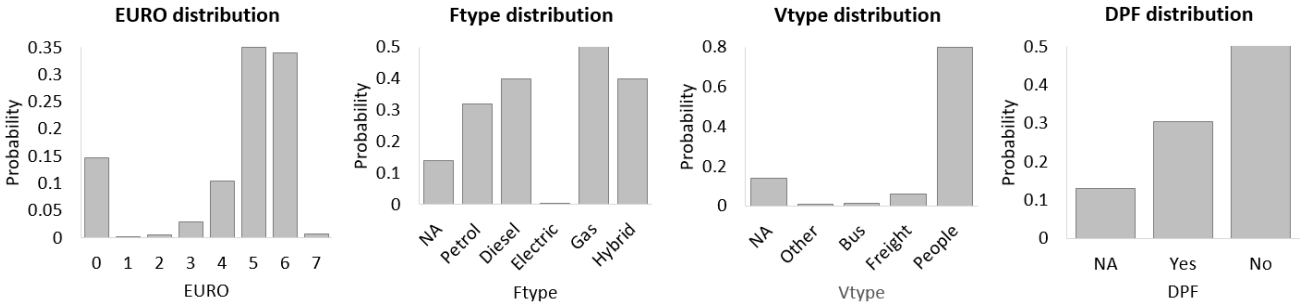

Fig. 4. Traffic Features Distribution

correlation of 0.40, 0.20, 0.13, 0.51, 0.51 with the pollutants considered in the AQI computation, respectively. Conversely, all traffic features results less correlated and they are not shown for brevity.

4 Pollutants prediction and and evaluation of AQI

We implemented several machine learning models using the *caret* package. In particular, we tested algorithms for regression, including Generalized Linear Model (GLM), Random Forest (RF), Support Vector Machines (SVM) and Artificial Neural Networks (ANN). We tested all algorithms with default hyper-parameters and with the same random seed, uniformly selecting in time the 70% of the samples as training set, and leaving the remaining 30% for the test set. We trained all models with a 5-fold cross-validation and we computed the model performances in terms of mean squared error (MSE) for pollutants. For brevity, we report only the results of the GLM, which we considered as the baseline and of the ANN, which is the model that performed best. Artificial Neural Network (ANN) [8] are inspired by biological nervous systems such as the human brain, which process information through a large number of highly interconnected processing elements (neurons). ANNs can be used in several applications, such as pattern recognition or data classification, and they are a supervised machine learning technique.
Specifically, we chose a particular type of ANN, namely the BRNN model (Bayesian Regularization of Neural

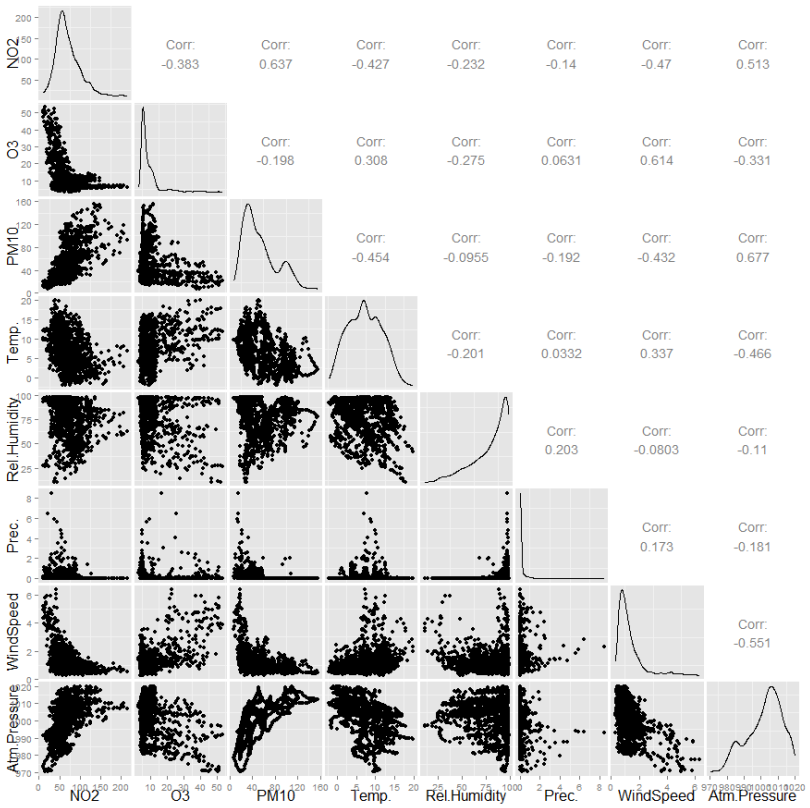

Fig. 5. Correlation between selected weather features and pollutants used for the AQI computation.

Table 1. Performance comparison between the GLM and the BRNN reporting average and standard deviation of the absolute error for each pollutant.

Agent	Unit	BRNN 9 Neur. $\mu(\varepsilon)$	$\delta(\varepsilon)$	GLM $\mu(\varepsilon)$	$\delta(\varepsilon)$	Comparison $\Delta[\mu(\varepsilon)]$	$\Delta[\delta(\varepsilon)]$
NO2	$\mu g/m^3$	12.45	10.28	21.02	20.58	41%	50%
O3	$\mu g/m^3$	22.47	20.06	46.47	45.09	52%	56%
CO	$\mu g/m^3$	10.99	10.18	19.31	15.59	43%	35%
C6H6	$\mu g/m^3$	39.85	64.44	98.92	145.89	60%	56%
N2	ppb	27.33	29.12	56.27	81.32	51%	64%
PM10	$\mu g/m^3$	13.65	14.35	34.56	34.62	61%	59%
SO2	$\mu g/m^3$	18.64	20.67	38.58	42.68	52%	52%
PM2.5	$\mu g/m^3$	13.28	13.88	32.55	31.87	59%	56%
BC	$\mu g/m^3$	18.47	24.21	35.91	66.30	49%	63%
NH3	$\mu g/m^3$	26.95	52.95	41.83	119.12	36%	56%

Networks) [7], because it is more robust than standard back-propagation networks and it can reduce the need for lengthy cross-validation. Bayesian regularization is a mathematical process that converts a nonlinear regression into a "well-posed" statistical problem in the manner of a ridge regression [7]. The main model parameter is the number of neurons n to be used. In order to define the optimal n, we incrementally evaluated the model accuracy starting with $n = 1$ and incrementing it in steps of 1 until 20. Therefore, we empirically find the best value of $n = 9$, after which the performance improvement can be considered negligible.
For each pollutant, we compare the BRNN model with the GLM performances, obtaining for the BRNN an

Table 2. Confusion Matrix of the Air Quality Index class computed with the pollutants estimated with the BRNN model.

	Real					
Predicted	Good	Fair	Average	Not Very Healthy	Unhealthy	Very Unhealthy
Good	72	0	0	0	0	0
Fair	120	528	0	0	0	0
Average	0	69	312	0	0	0
Not Very Healthy	0	0	0	144	0	0
Unhealthy	0	0	0	72	120	0
Very Unhealthy	0	0	0	0	24	0
Overall Statics	Accuracy: 0.8049					

improvement of the average relative error between 36% and 61% over the GLM. The performance comparison is fully reported in Table 1.
Using the predicted values of NO_2, O_3 and PM_{10}, we computed the AQI value, and then we mapped it into the classes described in Section 2 (i.e. Optimal, Good, Fair, Average, Not Very Healthy, Unhealthy, Very Unhealthy). Finally, we computed the accuracy of the estimated AQI class, which is reported in Table 2.

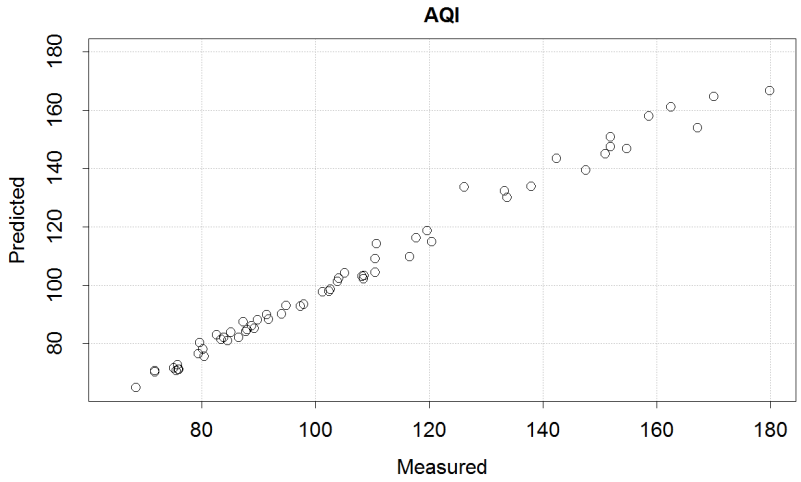

Fig. 6. Scatter Plot AQI measured/predicted

Our model predicts AQI with a class accuracy of 0.8, which we evaluate as satisfactory (see also the scatter plot in Fig.6), especially considering that the distance of the classification error is never greater than one, meaning that when the model predicts an erroneous class it is never beyond the adjacent one, e.g., the model can predict Fair instead of Good but it never predicts Average or any worst condition instead of Good.

5 Conclusion and Future works

In this paper we used traffic and weather data in order to predict the air pollution in a metropolitan city. We designed and implemented a set of machine learning models to predict single pollutants that we used to compute qualitative air quality classes based on a standardized Air Quality Index (AQI). The performance of our best model, (BRNN with 9 neurons), achieves an AQI class accuracy of 0.8.

Future works will include the evaluation of our approach on a bigger dataset, an improvement of the feature set, and the evaluation of several scenarios (e.g., including partial or complete traffic block, different weather conditions, etc.) in order to evaluate the impact of local traffic policies on the air quality.

References

1. J. Amos, "Polluted air cause 5.5 million deaths a year new research says", BBC NEWS, Science and Environment, 2016.
2. Y. Zheng,F. Liu,H. Hsieh,"U-air: When Urban Air Quality Inference Meets Big Data", Microsoft Research Asia, *ACM*, 2013
3. I. Sundvor, N. Castell Balaguer, M. Viana, X. Querol, C. Reche, F. Amato, G. Mellios, C. Guerreiro,"Road traffics contribution to air quality in European cities", *ETC/ACM*, 2012
4. I. Juhosa, L. Makrab, B. Ttha, "Forecasting of traffic origin NO and NO2 concentrations by Support Vector Machines and neural networks using Principal Component Analysis", Simulation Modelling Practice and Theory, vol. 16, no. 9, pp. 1488-1502, 2008.
5. R. Berkowicz , F. Palmgren, O. Hertel, E. Vignati, "A Study on Effects of Weather, Vehicular Traffic and Other Sources of Particulate Air Pollution on the City of Delhi for the Year 2015", Journal of Environment Pollution and Human Health, vol. 4, no. 2, pp. 24-41, 2016.
6. R. Gopalaswami,"Using measurements of air pollution in streets for evaluation of urban air quality meterological analysis and model calculations", Science of The Total Environment, vol. 189-190, pp. 259-265, 1996.
7. F. Burden,D. Winkler,"Bayesian regularitazion of neural networks", *PubMed*, 2008
8. C. Stergiou, D. Siganos ,"Neural networks", *Imperial College London*, 2011.

Appendix: Pollution agents Ecdf Graphs

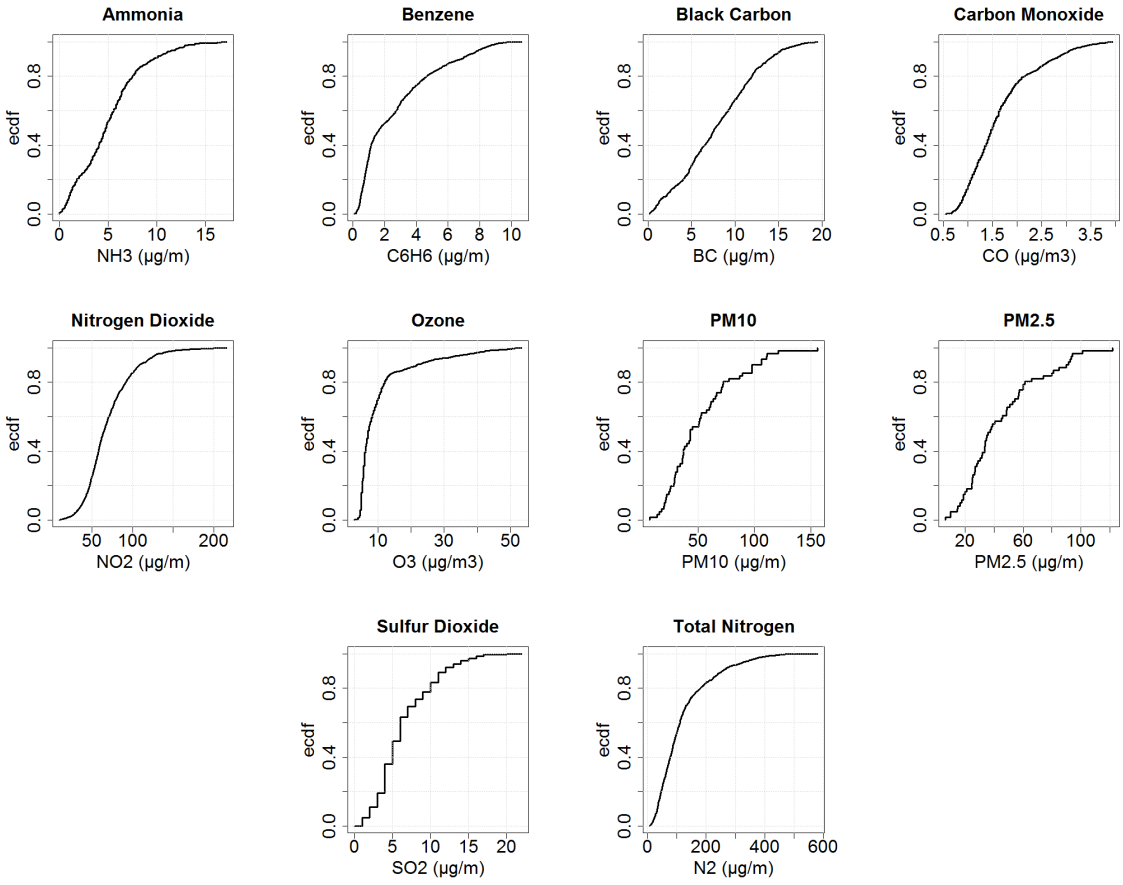

Fig. 7. Pollution Agents Ecdf Graphs