uAQE: Urban Air Quality Evaluator

uAQE: Urban Air Quality Evaluator

Claudio Rossi[1], Alessandro Farasin[1,2], Giacomo Falcone[1], and Carlotta Castelluccio[3]

[1] LINKS Foundation, Italy; {name.surname}@linksfoundation.com
[2] Polytechnic of Turin, Italy; {name.surname}@polito.it
[3] Microsoft, Italy; {name.surname}@microsoft.it

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 · Environment · Weather · Traffic

1 Introduction and Related Works

Air pollution is the introduction into the atmosphere of 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., Milan or Turin, which have an air pollution index among the highest ones according to recent rankings [4].

Road transport is one of the main causes of air pollutants emissions, accounting for the 14% of the total emissions in European countries [5].

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

[4] http://www.numbeo.com/pollution/rankings.jsp
[5] http://www.eea.europa.eu

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 [3] and [4], which present a set of learning methods able to predict NO_x concentrations from past observations and weather conditions. In [5], 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-makers 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 a big metropolitan city, i.e., 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 dellAmbiente). The **weather data set** contains 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). [6].

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 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. The **traffic data set** includes 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.

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 **air pollution data set** contains the measured values of the aforementioned pollutants: 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[7]. The AQI uses only three pollutants, namely NO_2, PM_{10}, O_3, and it is formulated as $I_{AQI} = \frac{I_{PM_{10}} + max(I_{NO_2}, I_{O_3})}{2}$, where $I_{PM_{10}} = \frac{Vavg24h_{PM_{10}}}{Vref_{PM_{10}}} \times 100$, $I_{NO_2} = \frac{Vmaxh_{NO_2}}{Vref_{NO_2}} \times 100$, $I_{O_3} = \frac{Vmax8h_{O_3}}{Vref8h_{O_3}} \times 100$. $Vref$ are reference values, while $Vavg24h$, $Vmaxh$, $Vmax8h$ means values averaged over the last 24 hours, hourly maximum, maximum over the last 8 hours, respectively.

We observe that in the considered data set O_3 never exceeds the maximum value established for preserving human health (i.e., $Vref8h_{O_3} = 120\mu g/m^3$),

[6] http://ita.arpalombardia.it/ITA/qaria/doc_RichiestaDati.asp
[7] http://www.arpae.it/cms3/documenti/aria/IQA.pdf

whereas NO_2 exceeds its hourly maximum value (i.e., $Vref_{NO_2} = 200\mu g/m^3$) only in few cases ($< 5\%$). Conversely, PM_{10} exceeds the the daily maximum value (i.e., $Vref_{PM_{10}} = 50\mu g/m^3$) 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 ($125 \leq AQI < 150$), 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 ($\approx 80\%$) falls between Fair and Not Very Healthy levels.

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 merge the three data categories previously described (weather, traffic, air) at hourly resolution because this is the maximal temporal resolution of both weather and air data. We average the measurements produced by different sensor of the same type in the same hour, while we count the hourly passages at all gates of vehicles. Specifically, we count the total hourly passages of each EURO category, type, fuel, DPF availability. In order to perform all data manipulations, we use *R* and the *plyr* library, which provides data aggregation operators. We fill few missing values ($< 1\%$) in the air and weather data by polynomial interpolation using the spline function of the *zoo* library.

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, global radiation, net radiation, atmospheric pressure.

Additionally, 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 $fp(t)$ that at each time instant c is equal to the sum of $f(t)$ over the last x hours ($fp(t) = \sum_{t=c-x}^{t=c} f(t)$). We evaluate increasing values of x starting from 1 and and we empirically find the best value to be 12. Studying the correlation between weather and pollutants we notice that temperature, relative humidity, precipitation wind speed, and atmospheric pressure are the most significant ones, having an average absolute 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.

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

Agent	Unit	BRNN 9 Neur. $\mu(\varepsilon)$	$\delta(\varepsilon)$	GLM $\mu(\varepsilon)$	$\delta(\varepsilon)$	Comparison $\Delta[\mu(\varepsilon)]$	$\Delta[\delta(\varepsilon)]$
NO_2	$\mu g/m^3$	12.45	10.28	21.02	20.58	41%	50%
O_3	$\mu g/m^3$	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%
C_6H_6	$\mu g/m^3$	39.85	64.44	98.92	145.89	60%	56%
N_2	ppb	27.33	29.12	56.27	81.32	51%	64%
PM_{10}	$\mu g/m^3$	13.65	14.35	34.56	34.62	61%	59%
SO_2	$\mu g/m^3$	18.64	20.67	38.58	42.68	52%	52%
$PM_{2.5}$	$\mu g/m^3$	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%
NH_3	$\mu g/m^3$	26.95	52.95	41.83	119.12	36%	56%

4 Pollutants prediction and and evaluation of AQI

We implement the machine learning models using the *caret* package. In particular, we test several machine learning algorithms for regression, including the Generalized Linear Model (GLM), the Random Forest (RF), the Support Vector Machines (SVM), and the Artificial Neural Networks (ANN). We test 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 train all models with a 5-fold cross-validation and we compute the model performances in terms of mean squared error for pollutants (regression problem).For brevity, we report only the results of the GLM, which we consider as the baseline, and of the ANN, which is the model that performs best.

Artificial Neural Networks (ANNs) [7] are inspired by biological nervous systems such as the human brain, which process information through a large number of highly interconnected processing elements (neurones). ANNs can be used in several applications, such as pattern recognition or data classification, and they are a supervised machine learning technique.

Specifically, we choose a particular type of ANN, namely the BRNN model (Bayesian Regularization of Neural Networks) [6], because it is more robust than standard back-propagation networks and it can reduce the need for lengthy cross-validation. The main model parameter is the number of neuron n to be used. In order to define the optimal n, we incrementally evaluate 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 improvement of the average absolute 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 compute the AQI value, and then map into the classes described in Section 2 (i.e. Optimal, Good, Fair, Average, Not Very Healthy, Unhealthy, Very Unhealthy).

Our model predicts AQI with a class accuracy of 0.8, which we evaluate as satisfactory, 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 a qualitative air quality classed 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. 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.
4. 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.
5. 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.
6. F. Burden,D. Winkler,"Bayesian regularitazion of neural networks", *PubMed*, 2008
7. C. Stergiou, D. Siganos ,"Neural networks", *Imperial College London*, 2011.