Airvlc: An application for real-time forecasting urban air pollution

Lidia Contreras Ochando LICONOC@UPV.ES
Universitat Politècnica de València. Spain

Cristina I. Font Julián CRIFONJU@EI.UPV.ES
Universitat Politècnica de València. Spain

Francisco Contreras Ochando FRACONOC@GMAIL.COM
Universitat Politècnica de València. Spain

Cèsar Ferri CFERRI@DSIC.UPV.ES
DSIC. Universitat Politècnica de València. Spain

Abstract

This paper presents Airvlc, an application for producing real-time urban air pollution forecasts for the city of Valencia in Spain. Although many cities provide air quality data, in many cases, this information is presented with significant delays (three hours for the city of Valencia) and it is limited to the area where the measurement stations are located. The application employs regression models able to predict the levels of four different pollutants (CO, NO, PM2.5, NO2) in three different locations of the city. These models are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters such as wind speed and temperature. We compare different learning techniques to get the better performance in the prediction of pollutants. According to our experiments, ensembles of decision trees (Random Forest) outperforms the rest of methods in almost all of our tests. Airvlc incorporates the best regression models and, by a distance-weighted combination of the predictions, is able to generate a real-time pollution map of the city of Valencia. The application also includes a warning system for sending notifications to users when a nearby risk pollution concentration is detected.

Proceedings of the 2^{nd} International Workshop on Mining Urban Data, Lille, France, 2015. *Copyright ©2015 for this paper by its authors. Copying permitted for private and academic purposes.*

1. Introduction

Air pollution can have important impact (short and long-term) on the health of people. For instance, urban air pollution increases the risk of suffering respiratory diseases such as pneumonia, or chronic, such as lung cancer or cardiovascular disease (World Health Organisation, 2015). A recent work (Wilker et al., 2015) relates long-term exposure to ambient air pollution to structural changes in the brain.The SOER 2015 report (The European Environment Agency , 2015), with data about the European Union countries' air quality in 2011, concludes that although the atmosphere in the continent has improved in the last decades, there are significant traces of the most harmful contaminants. In fact, in 2011, the report estimates that 430.000 Europeans died prematurely because of pollution.

Although some governments are introducing restriction policies that limit the use of vehicles (main source of pollution in most cases), only in Europe, important cities such as Paris, Naples, Moscow, Milan or Barcelona still report significant levels of urban pollution in 2015 (The European Environment Agency , 2015). In this context, it is important for citizens of urban agglomerations to reduce the exposition to urban air pollution as much as possible. This is especially relevant for high risk population such as: kids, elderly people, asthmatics or people suffering respiratory diseases.

In this work we present an application that predicts urban air pollution in real time by employing historical data. The application is based on the city of Valencia in Spain. This city can be considered a medium size urban agglomeration (around 1.000.000 inhabitants). The city provides an open data site containing real-time information about the city in different aspects such as traffic data, noise sensors, pollen

sensors... Although different sensors of urban pollution air are included in the site, this information needs to be carefully verified and it is published with a delay of three hours. This delay can represent a problem since risky high levels of pollutions are not detected in real-time. Additionally, the network of sensors is limited (six in the city of Valencia).

Considering these limitations, we have developed an application able to display in real-time foreseeable levels of pollution in a wide number of points of the city. The application is based on the predictions of regression models that are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters.

The paper is organised as follows. Section 2 details the process of data recollection of pollution particles and the factors that affect the generation, concentration or dispersion of these pollutants. Experiments in learning regression models for predicting the pollutant concentrations are included in Section 3. The Airvlc application is detailed in 4. Related works are discussed in Section 5. Finally, Section 6 closes the paper with a discussion of the main conclusions and some plans for future work.

2. Data collection

Different particles are associated with urban air pollution. In order to measure air contamination, pollutant parameters found in the lower levels of the troposphere are controlled. Air quality sensors measure concentrations of particles that have an anthropogenic origin and produce effects during or after the inhalation by humans. The historical pollution data for this work has been obtained from the open data web of the Generalitat Valenciana[1]. Following the recommendations of (The European Environment Agency , 2015), we concentrate on the following particles:

- **PM 2.5 (Suspended particles below 2.5 microns):** This parameter has been chosen because of its pollutant power. It is one of the most dangerous particles, since its size makes it almost unstoppable by the natural filters of the body. This fact means that the PM 2.5 are usually able to reach the pulmonary alveoli and in some cases, these particles are attached to these alveoli with a consequent reduction of lung capacity; in worst cases, the particles cross the alveolar membranes and reach the blood stream. Considering that PM 2.5 particles have its origin in anthropogenic activities (especially in the use of fuels in motor vehicles), it is not surprising that its atomic structure contains heavy metals, extremely toxic to the human body. Atmospheric conditions in the Mediterranean coast of Spain can influence the particle levels, due to lower rainfall and wind action with respect to other northern Europe countries, and the North African particles (Saharan dust), PM10 and PM2.5.

- **NO (Nitrogen monoxide):** Nitrogen monoxide is a highly unstable compound; it causes nitrogen dioxide by quickly reacting in the atmosphere. This instability makes the nitrogen monoxide a radical, namely, a high reactive power molecule, whose effects on the body are abnormal DNA, lipids and proteins. This kind of changes derives in the medium and long term as a greater chance of developing cancer.Its origin stems largely from vehicle engines.

- **NO2 (Nitrogen dioxide):** Nitrogen dioxide is not a directly generated pollutant, since its presence in the atmosphere is caused by the oxidation of nitrogen monoxide. In the presence of moisture, this compound results in nitric acid, and its inhalation, even in low concentrations, can cause lung tissue degradation, as well as can reduce the efficacy of the immune system, especially in children.

- **CO (Carbon monoxide):** Carbon monoxide is a primary pollutant. CO is toxic; it prevents oxygen transport by poisoning the blood, since it replaces the haemoglobin. People with cardiovascular and cerebrovascular problems could suffer heart attacks or strokes because of problems related to high concentrations of CO.

The distribution of air pollution is decisively influenced by climatic conditions. We have collected Climatological observations for the meteorological data of Valencia city from Meteorological Agency of the Government of Spain (AEMET)[2]. We consider the following parameters:

- **Temperature:** In an ordinary atmosphere situation, temperature decreases with altitude, favouring ascension of warmer (and less dense) air, and dragging contaminants upwards. In a situation of thermal inversion, a warmer layer of air is over the colder surface air and prevents the rise of this last (denser), so the contamination is confined and increases.

- **Humidity:** Humidity is a weather factor to be considered; in its presence, nitrogen dioxide derives in nitric acid, harmful to human health.

- **Wind speed:** Strong winds can disperse pollutants and transport them away from their emission point.

[1] http://www.cma.gva.es/cidam/emedio/atmosfera/jsp/historicos.jsp

[2] http://www.aemet.es/

- **Precipitations** Precipitations wash contaminants and can dissolve substances and gases.

The two main sources of pollution in developed countries are motor vehicles and industry. Vehicles release large amounts of nitrogen oxides, carbon oxides, hydrocarbons and particulates when burning gasoline and diesel. Therefore, we need to measure the level of traffic in the city in order to predict the air pollution. For this purpose, the City of Valencia provides a network of sensors (electromagnetic coils) that measure the intensity of traffic (Vehicles/hour) in the city. This data can be found in the open data site of the Valencia City Council[3].

3. Experiments

With all the selected parameters, we have built datasets aimed to predict the concentration of pollutants from the intensity of traffic and weather parameters. Concretely, we have collected data for a period of two years (2013 and 2014). Data was collected every 60 minutes, 24 hours a day during those two years. Although Valencia city has six stations for the detection and measurement of air pollution, three of them have not sufficient data for the analysed period and were discarded. In this way we collected data from these stations: *Molì*, *Avd Francia* and *Pista de Silla*. These three stations are located inside the urban agglomeration, and thus most of the pollutants measured in the sensors should be generated by urban activities (mainly traffic). For each one of these stations, we create a dataset with the level of the pollutants measured and parameters that can affect these measurements, we concentrate on traffic level (measured by electromagnetic coils), weather conditions. In order to measure the traffic related to each air pollution station, we average the traffic intensity of the closest six traffic measurament sensors. This is a simplification since, certainly, all the traffic of the city has effect on the measured level of all the stations in the city.

We can see a summary of the three datasets in Table 1. This table includes averages and standard deviation for the three stations of the pollutant particles measured and the intensity of traffic associated with each station. If we analyse traffic intensity, *Avd Francia* is the busiest station, while the other two have similar values. With regard to pollution levels *Pista de Silla* station presents the maximum levels for three parameters. The only exception is PM2.5. This behaviour can probably be associated with the specific location of the stations: While *Pista de Silla* station is located in a the central part of the city, and therefore more vulnerable to the overall city pollution, the other two are in the suburbs of the city where external air streams can reduce the levels of pollutants.

We first study the weekly evolution of pollutants in the three stations. Figure 1 shows the evolution of the average of the four parameters of pollution analysed and the average traffic intensity for *Molì* station depending on the day of the week. Figure 2 presents the same plot for *Avd Francia* station and Figure 3 corresponds to *Pista de Silla* station. In order to make the values comparable in the plot we normalise each parameter by the maximum value of that parameter. The level of pollutants and traffic reach the maximum levels during the working days of the week for the three stations (Friday seems to be the worst day). We can clearly see the dependency of the four parameters of pollution on the traffic intensity level. During the weekend days, the level of traffic drastically descends and associated with this reduction the levels of pollutants significantly drop. Again, the exception is PM2.5. This behaviour can be caused because these particles can be generated by all types of combustion activities (motor vehicles, power plants, wood burning, etc.) and certain industrial processes (US Environmental Protection Agency , 2015).

We have performed a similar analysis considering the evolution of pollutants, traffic intensity and meteorological variables during a day (humidity and wind). Figure 4 shows the evolution of the daily average of these parameters for *Molì* station depending on the hour of the day . Figure 5 corresponds to *Avd Francia* station and Figure 6 to *Pista de Silla* station. Again, we normalise each parameter by the maximum value of that parameter. If we observe traffic intensity, we can discover in all the three plots a similar behaviour, there are three peaks in traffic intensity corresponding to the hours where workers travel to their work places (around 9 am), lunch time (around 2 pm) and an evening period (around 8 pm). In the three stations the maximum of pollution parameters is found at the same period of the first peak in traffic intensity (around 9 am). In the second peak of traffic intensity (around 2 pm) the levels of pollutants does not follow the increase in traffic. In fact, after the maximum period around 9 am, pollutants decrease their levels until around 4 pm where they change the behaviour and start an increasing of the values. The second peak in pollutant values is found around 9 pm. Our intuition with respect to this behaviour is that wind disperses part of the pollutant in the most sunny hours. Valencia is in the Mediterranean coast and in this city it is easy to find (especially in summer) sea breezes. These kind of winds are created over bodies of water (usually sea or big lakes) near land due to differences in air pressure created by their different heat capacity. This phenomenon can be detected in the plots if we observe the increase in wind strength during the midday hours. Finally, we observe a strange and different behaviour of the CO particle in *Molí* station. For this pollutant there is a second peak in the midday period.

[3] http://www.valencia.es/ayuntamiento/DatosAbiertos.nsf/

This behaviour probably corresponds to an extra source of pollution that needs to be further studied.

As stated previously, we are interested in predicting pollution levels in real time. Since these levels are only made public with a delay of three hours, we need to produce a prediction model from real time features. We extract the following set of features from the data collected from different sources (detailed in the previous section):

- **Climatological features:** Temperature (Celsius degrees), Relative humidity (Percentage), Pressure (hPa), Wind speed (km/h), Rain (mm/h)
- **Calendar features:** Year, Month, Day in the month, Day in the week, Hour
- **Traffic intensity features:** Traffic level in the surrounding stations (vehicles/hour), traffic level 1, 2, 3 and 24 hours before
- **Pollution features:** Pollution level in the target station 3 and 24 hours before

With this goal we compare several regression learning techniques from R (R Core Team, 2015) in order to identify the technique that is able to better predict the levels of pollution. To test the prediction ability of different models, we learn the models using as training data the registers of 2013 and the first nine months of 2014. We test the models with the last three months of 2014. We use Mean Squared Error (MSE) as a performance measure. Concretely, we employ the following techniques for learning regression models (all of them with the default parameters, unless stated otherwise): Linear Regression (*lr*) (Hornik et al., 2009), quantile regression (*qr*) (Koenker, 2015) with *lasso* method, K nearest neighbours (*IBKreg*) with $k = 10$ (Hornik et al., 2009) , a decision tree for regression (*M5P*) (Hornik et al., 2009), Random Forest (*RF*) (Liaw & Wiener, 2002), Support Vector Machines (*SVM*) (Meyer et al., 2014) and Neural Networks (Venables & Ripley, 2002). In order to compare the predictive performance of these models, we also introduce three baseline models: A model that always predicts the mean of the train data (*TrainMean*), a model that always predicts the mean of the test data (*TestMean*), and a basic model that predicts the same value of the target pollutant 3 hours before (*X3H*).

Table 2 contains the MSE of the regression models for the prediction of the four target pollution levels of the *Molí* station. Results for *Pista de Silla* station and *Avd Francia* station are shown in Table 4 and 3 respectively. If we analyse these results, we can conclude that learned models are improving the performance of the basic baseline models in almost all cases. When we compare the learning techniques in the three tables, the ensemble of decision trees technique (random forest) is the best model in almost all of cases. These results are in concordance with (Singh et al., 2013) where ensembles of trees outperformed other approaches such as SVMs.

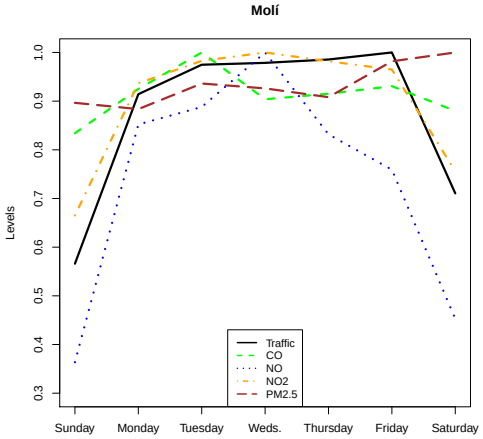

Figure 1. Average weekly traffic intensity and pollution parameters measured in Molí station.

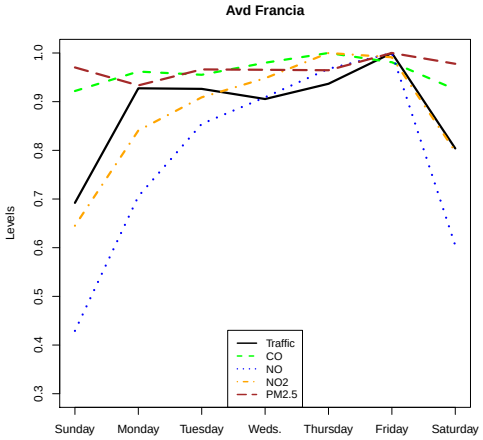

Figure 2. Average weekly traffic intensity and pollution parameters measured in Avd Francia station.

4. Airvlc

In the previous section we have analysed how to obtain real-time air pollution predictions from a given set of features. In this section we summarise Airvlc, a mobile app for Android and iOS and a web application[4]. This application generates from the regression models a map of the city of Valencia showing the predicted intensity of pollution levels. The application also allows the user to configure a set of automatic warnings every time a pollution threshold is reached near the position of the mobile device.

[4] http://airvlc.lidiacontreras.com/

Table 1. Averages and standard deviation of the three pollution detection sensors.

	Traffic		CO		NO		NO2		PM2.5	
	ave	sd	ave	sd	ave	sd	ave	sd	ave	sd
Molí	442.333	339.489	0.116	0.093	8.642	20.311	26.608	21.003	10.650	6.926
Francia	631.569	431.412	0.185	0.122	9.092	19.271	27.840	23.992	7.909	4.235
Silla	484.722	298.768	0.228	0.187	23.559	33.024	45.631	25.376	8.309	6.020

Table 2. Results in MSE of different regression models for Molí Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.086	0.061	0.067	**0.057**	0.060	0.068	0.071	**0.057**	0.063	0.182
NO	30.202	28.739	36.516	25.200	29.805	27.821	25.555	**20.655**	25.870	32.944
NO2	19.918	19.914	25.258	19.680	17.370	15.683	31.242	14.877	**14.488**	32.152
PM2.5	8.803	8.803	8.634	6.889	6.564	6.674	7.248	**6.072**	6.135	13.089

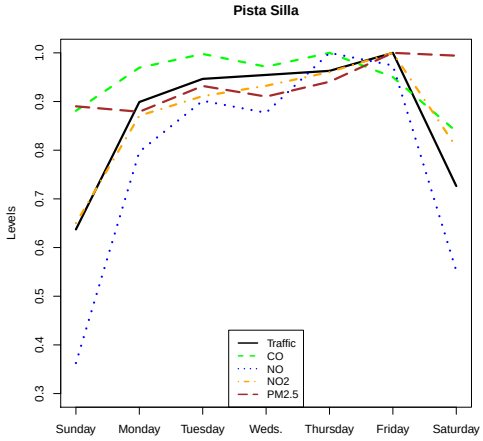

Figure 3. Average weekly traffic intensity and pollution parameters measured in Pista Silla station.

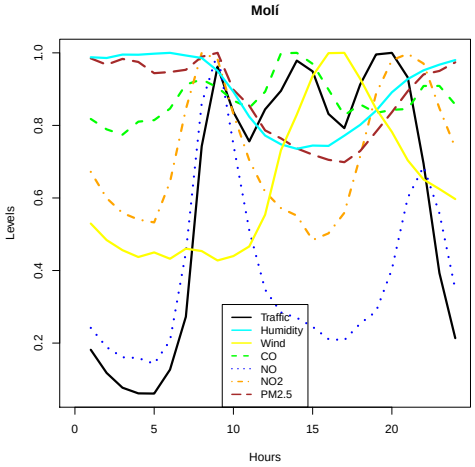

Figure 4. Average daily traffic intensity and pollution parameters measured in Molí station.

4.1. Contamination intensity map

Results of Section 3 show that random forest models obtain the best performance in most cases. Therfore, twelve random forest models are implemented in the Airvlc application. These models are able to predict every hour the level of the four analysed particles at the three pollution detection stations. We want, however, to predict pollution levels at the points of the city where the traffic is measured (1245 points around the city). For that purpose, given any of these points, we extract the features related to traffic intensity from the six nearest traffic sensors. The meteorological features are the same for all the city. The predictions of pollutants in that exact location is computed by the combination of the models corresponding to the three stations. The combination is weighted with respect to the distance of the target point with respect to the measurement stations giving more importance to the closest models. A simpler approach could be to learn a single model from the concatenation of the data from the three stations and then apply this in all the set of target points.

By computing the pollution predictions for a set of strategic and well-distributed locations we are able to estimate a real-time pollution map of the city. The map is generated with *Google Maps* technology. This map shows for each location its pollution level as a dot which colour varies among green, yellow and red depending on the calculated pollution level. If the user selects one of these dots, an extended window is opened where the exact predicted levels are shown. Figure 7 includes a screen-shot of the pollution map of the Airvlc application. The user can also select a second frame in the window of the Airvlc application where he/she can introduce a specific location and then the application computes the predicted pollution levels for that selection. An example of this process is included in Figure 8.

4.2. Risk levels

Figure 8 shows how the pollution levels are presented to users. However, showing just a concentration value of each parameter is not very useful for most users, since most of them are not experts in pollutants and they could not interpret correctly these numbers. In order to improve the comprehensibility of the predictions we have established three ranges of risk represented as speedometer: Low risk (green) corresponds to a measurement that is safe; Medium risk (yellow) when concentrations reach levels to cause harmful effects in people sensitive to air pollution exposure (kids, elderly people...); High risk (red) when concen-

Table 3. Results in MSE of different regression models for Avd. Francia Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.195	0.165	0.224	0.159	0.163	0.168	0.160	**0.153**	0.156	0.324
NO	35.493	33.634	44.955	32.992	36.049	34.092	30.350	**29.517**	33.364	38.262
NO2	23.443	20.900	27.162	16.299	16.718	18.494	23.782	**14.851**	19.100	41.929
PM2.5	3.721	3.718	3.879	3.326	**3.132**	3.523	4.655	3.214	3.265	7.974

Table 4. Results in MSE of different regression models for Pista Silla Station. The best prediction model is highlighted in bold.

	TrainMean	TestMean	X3h	lr	qr	IBkreg	M5p	RF	SVM	NN
CO	0.278	0.222	0.304	0.221	0.227	0.221	0.235	**0.218**	0.268	0.278
NO	49.149	46.232	60.353	39.863	43.524	42.760	44.150	**36.332**	52.438	58.798
NO2	23.135	23.122	30.167	20.861	19.031	18.487	25.972	**16.722**	23.313	49.699
PM2.5	6.911	6.660	7.119	5.663	5.342	5.750	7.189	**5.339**	7.368	11.061

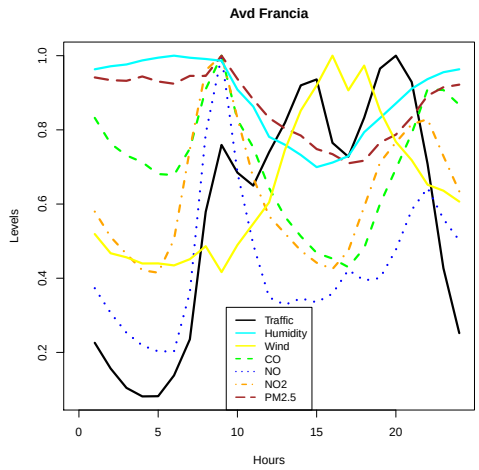

Figure 5. Average daily traffic intensity and pollution parameters measured in Avd Francia station.

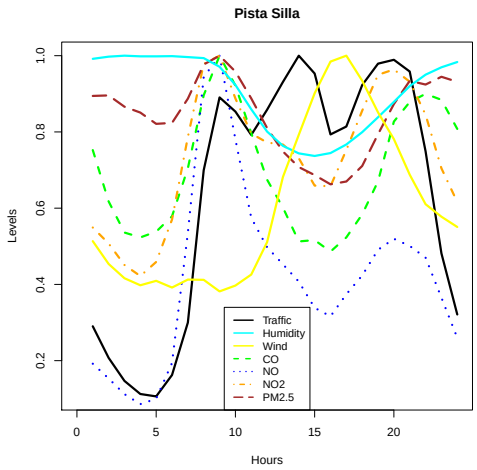

Figure 6. Average daily traffic intensity and pollution parameters measured in Pista Silla station.

trations can cause acute and chronic effects to anyone, especially those with sensitivity.

The ranges of risk shown by the application from the predicted values of the four pollutants are based on the recommendations of the Directive 2008/50/EC (European Comission, 2008). The variable as NOx (oxides of nitrogen) refers to NO or NO2, since the normative establishes the same limits for both levels.

- **Green level:** [NOx] $<$ 14.0 μg/m3 $\wedge$ [CO] $<$ 30.0 mg/m3 $\wedge$ [PM 2.5] $<$ 7.5 μg/m3.
- **Yellow level:** We establish medium risk (yellow level) if the levels do not satisfy the conditions of the green level and the red level.
- **Red level:** [NOx] $\geq$190.0 μg/m3 $\vee$ [CO] $\geq$ 55.0 mg/m3 $\vee$ [PM 2.5] $\geq$ 25.0 μg/m3

4.3. Risk warnings

Airvlc mobile application can be configured to send warnings to users if the device is near to a zone (200 meters approximately) where a high risk level is predicted. These warnings can be personalised by the user in different ways. For example, the user can establish personal limits for warnings or modify the range of distance for the detection of high risk levels of pollutant concentration. Obviously, the user needs to allow the application to know the actual GPS location of the device

In the case of the web application, given that here it is more complex to know the exact location of the user, we adopt a different strategy. We are working in an automated warning system where the user needs to fix a set of areas, and then the system sends an electronic email whenever a dangerous situation (high risk level by default) is detected.

5. Related work

A wide number of works employs machine learning techniques or statistical approaches for predicting pollution levels. A classical work is (Yi & Prybutok, 1996). In this paper, the authors propose ozone prediction models. Specifically, they develop a neural network model for forecasting daily maximum ozone levels and compare it to previous approaches by regression, and Box-Jenkins ARIMA. The results show that the neural network model improves the performance of the regression and Box-Jenkins ARIMA models tested. Neural networks models have been widely

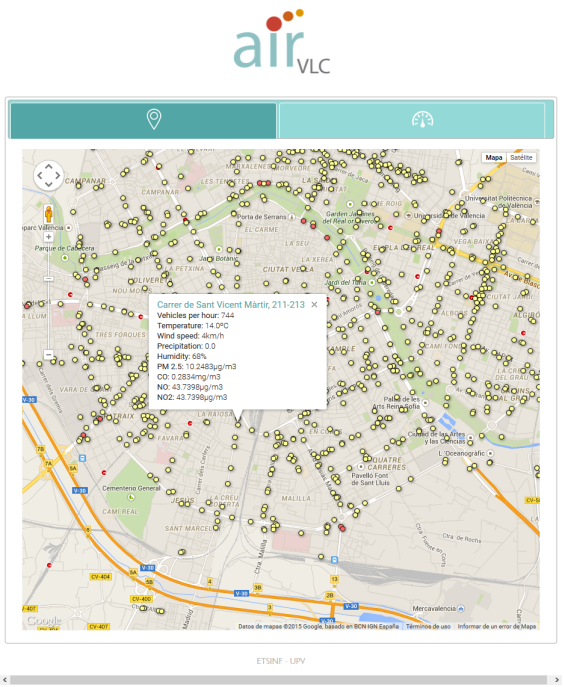

Figure 7. Airvlc application. Real-time pollution map.

Figure 8. Frame where the user can introduce specific locations to know the predicted levels of pollution.

employed in this field, a review of these approaches can be found in (Khare & Nagendra, 2006).

A more related work is (Karppinen et al., 2000a). Here the authors propose a modelling system for predicting the traffic volumes, emissions from stationary and vehicular sources, and atmospheric dispersion of pollution in an urban area. They employ four monitoring stations in the Helsinki metropolitan area in 1993. The paper compares the predicted NOx and NO2 concentrations with the results of an urban air quality monitoring network. The agreement of model predictions was better for the two suburban monitoring stations, compared with two urban stations. Some applications of these models are introduced in (Karppinen et al., 2000b). A similar work for the city of Izmir in Turkey is (Elbir, 2003). Here, the authors compare The CALMET meteorological model and its puff dispersion model CALPUFF for predicting dispersion of the sulphur dioxide emissions from industrial and domestic sources.

Another related work, and in this case very recent, is (Donnelly et al., 2015). This paper presents a model for real time air quality forecasts. The predictions are concentrated in nitrogen dioxide (NO2) and they are used to estimate air quality 48 hours in advance. The model is based on a multiple linear regression which uses linearised factors describing variations in concentrations together with meteorological parameters and persistence as predictors.

Our comparison of regression techniques obtains similar conclusions to the work presented in (Singh et al., 2013). In this study, principal components analysis (PCA) is performed to identify air pollution sources. From the extracted features, tree based ensemble learning models are induced to predict the urban air quality of Lucknow (India) together with the air quality and meteorological databases for a period of five years.

6. Conclusions

Air pollution can decrease life expectancy since contamination rises the risk of suffering respiratory diseases. Although policies motivating the reduction of emissions of pollutant particles have been introduced in the last years, many cities frequently still present risky levels of air pollution. In these situations, the reduction of the exposure to ambient air pollution is highly recommended. In this work, we have presented Airvlc, an application that predicts in real-time the levels of four dangerous pollutants in a wide set of points in the city of Valencia. The system is able to predict these pollution levels by applying regression models trained from data containing information traffic intensity, persistence of pollutants and meteorological parameters. Airvlc can be a useful tool for avoiding risky locations in terms of air pollution.

As future work we propose the integration of the application in middleware platforms such as Fi-Ware[5], this could help to extend the applicability of the system to other cities or regions. We also are interested in the incorporation of additional features in order to improve the prediction models: wind direction, sand storms, forest wildfires and agricultural burnings... Finally, the use of the tool for the recommendation of routes that minimise the exposure to air pollution.

[5] http://www.fiware.org/

Acknowledgments

We thank the anonymous reviewers for their comments, which have helped to improve this paper significantly. We are also grateful to Ajuntament de València, InnDEA València and specially to Ramón Ferri, Ruth López and Paula Llobet for their help in providing traffic data. This work was supported by the REFRAME project, granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Ministerio de Economía y Competitividad in Spain (PCIN-2013-037). It also has been partially supported by the EU (FEDER) and the Spanish MINECO project ref. TIN2013-45732-C4-01 (DAMAS), and by Generalitat Valenciana ref. PROMETEOII/2015/013 (SmartLogic).

References

Donnelly, Aoife, Misstear, Bruce, and Broderick, Brian. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmospheric Environment*, 103:53–65, 2015.

Elbir, Tolga. Comparison of model predictions with the data of an urban air quality monitoring network in izmir, turkey. *Atmospheric Environment*, 37(15):2149–2157, 2003.

European Comission. Directive 2008/50/ec of the european parliament on ambient air quality and cleaner air for europe. http://ec.europa.eu/environment/air/quality/legislation/directive.htm, 2008.

Hornik, Kurt, Buchta, Christian, and Zeileis, Achim. Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232, 2009. doi: 10.1007/s00180-008-0119-7.

Karppinen, A, Kukkonen, J, Elolähde, T, Konttinen, M, and Koskentalo, T. A modelling system for predicting urban air pollution:: comparison of model predictions with the data of an urban measurement network in helsinki. *Atmospheric Environment*, 34(22):3735–3743, 2000a.

Karppinen, A, Kukkonen, J, Elolähde, T, Konttinen, M, Koskentalo, T, and Rantakrans, E. A modelling system for predicting urban air pollution: model description and applications in the helsinki metropolitan area. *Atmospheric Environment*, 34(22):3723–3733, 2000b.

Khare, Mukesh and Nagendra, SM Shiva. *Artificial neural networks in vehicular pollution modelling*, volume 41. Springer, 2006.

Koenker, Roger. *quantreg: Quantile Regression*, 2015. URL http://CRAN.R-project.org/package=quantreg. R package version 5.11.

Liaw, Andy and Wiener, Matthew. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL http://CRAN.R-project.org/doc/Rnews/.

Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas, and Leisch, Friedrich. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2014. URL http://CRAN.R-project.org/package=e1071. R package version 1.6-4.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org/.

Singh, Kunwar P, Gupta, Shikha, and Rai, Premanjali. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmospheric Environment*, 80:426–437, 2013.

The European Environment Agency . Soer 2015 — the european environment — state and outlook 2015. http://www.eea.europa.eu/soer, 2015.

US Environmental Protection Agency . Particulate matter (pm) regulations. http://www.epa.gov/airquality/particlepollution/index.html, 2015.

Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Wilker, Elissa H., Preis, Sarah R., Beiser, Alexa S., Wolf, Philip A., Au, Rhoda, Kloog, Itai, Li, Wenyuan, Schwartz, Joel, Koutrakis, Petros, DeCarli, Charles, Seshadri, Sudha, and Mittleman, Murray A. Long-Term Exposure to Fine Particulate Matter, Residential Proximity to Major Roads and Measures of Brain Structure. *Stroke*, April 2015. doi: 10.1161/strokeaha.114.008348. URL http://dx.doi.org/10.1161/strokeaha.114.008348.

World Health Organisation. Public health, environmental and social determinants of health. http://www.who.int/phe/health_topics/outdoorair/databases/health_impacts/en/, 2015.

Yi, Junsub and Prybutok, Victor R. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3):349–357, 1996.