Procedia Computer Science

Volume 80, 2016, Pages 313–323

ICCS 2016. The International Conference on Computational Science

Wind-sensitive Interpolation of Urban Air Pollution Forecasts

Lidia Contreras and Cèsar Ferri

DSIC, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain.
liconoc@upv.es,cferri@dsic.upv.es

Abstract

People living in urban areas are exposed to outdoor air pollution. Air contamination is linked to numerous premature and pre-native deaths each year. Urban air pollution is estimated to cost approximately 2% of GDP in developed countries and 5% in developing countries. Some works reckon that vehicle emissions produce over 90% of air pollution in cities in these countries. This paper presents some results in predicting and interpolating 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. We compare several regression models able to predict the levels of four different pollutants (NO, NO_2, SO_2, O_3) in six different locations of the city. Wind strength and direction is a key feature in the propagation of pollutants around the city, in this sense we study different techniques to incorporate this factor in the regression models. Finally, we also analyse how to interpolate forecasts all around the city. Here, we propose an interpolation method that takes wind direction into account. We compare this proposal with respect to well-known interpolation methods. By using these contamination estimates, we are able to generate a real-time pollution map of the city of Valencia.

Keywords: Machine Learning, Urban Air Pollution, Spatial Interpolation

1 Introduction

Air pollution is one of the factors with major impact on the health of people. Exposure to ambient air pollution increases the risk of suffering respiratory diseases, such as pneumonia, or chronic, such as lung cancer or cardiovascular diseases [21]. A recent work [20] relates structural changes in the brain to long-term exposure to ambient air pollution.The SOER 2015 report [19] concludes that although the atmosphere in Europe has improved in the last decades, there are significant traces of the most harmful contaminants. The report estimates that in 2011, 430.000 Europeans died prematurely because of pollution. In this context, citizens of urban

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.343

agglomerations must try to reduce their 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 study the prediction of urban air pollution in real-time by employing historical data. We concentrate on four pollutants (NO, NO_2, O_3, SO_2). For that reason we employ data from the city of Valencia in Spain. Valencia is a medium size urban agglomeration (around 1.000.000 inhabitants). The city provides an open data site with information about: traffic data, noise levels and air pollution... Data about pollutant levels need to be verified and it is published with a delay of three hours. This delay can represent a problem since risky high levels of pollution are not detected in real-time. Moreover, the network of sensors is limited (six in the city of Valencia). Considering these restrictions, we address the problem of producing real-time predictions of the levels of pollution all around the city. We will study the performance of the predictions of different techniques for building regression models that are trained using features that represent traffic intensity, persistence of pollutants and meteorological parameters. We also analyse how the direction of wind affects the level of pollution and how to use that information in order to increase the accuracy of the prediction models.

Additionally, we address how to interpolate predictions and, in this way, we are able to show the approximate concentration of pollutants all around the city. For that reason we analyse popular spatial interpolation methods [10] such as Inverse Weighting Distance (IDW) or Kriging. These methods are static in the sense that they do not consider context conditions of the points to interpolate. Meteorological parameters (specially wind condition) clearly affects the way in which the pollutants are dispersed around the city. We propose a new method that uses wind information in order to improve the interpolation of urban pollution. In this aspect, we consider this technique a wind-sensitive interpolation approach. Experiments using a cross validation methodology show that this new method get better forecasts in comparison with well-known methods, specially when the interpolation is computed using enough information. This paper can be considered an extension of [13]. That work was focused on presenting the *Airvlc* application for real-time forecasts of air pollutants. In the models of [13] we did not consider wind direction in the learning models and interpolation techniques were not studied.

The paper is organised as follows. Section 2 details the process of data collection of pollution particles and some factors that affect the generation or dispersion of these pollutants. We also include some experiments in learning regression models for predicting the pollutant concentrations and some results on including wind direction in the models. We study some interpolations methods in Section 3. Finally, Section 4 closes the paper with a discussion of the main conclusions and some plans for future work.

2 Prediction of urban pollution

2.1 Data collection

The historical pollution data for this work has been obtained from the open data web of the Generalitat Valenciana[1]. The following particles are studied in this work:

- **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 whose effects on the body are abnormal DNA, lipids and proteins.

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

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.

- **NO_2 (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.

- **SO_2 (Sulphur dioxide):** It is a toxic gas primarily produced for sulphuric acid manufacture. SO_2 emissions are related to acid rain and atmospheric particulates. Inhaling SO_2 is associated with respiratory diseases and premature death.

- **O_3 (Ozone):** Ozone is not released directly by specific sources. This pollutant is generated by sunlight acting on NOx and Volatile organic compounds (VOC) in the air. Exposure to ozone significantly reduces lung function and induces respiratory inflammation. It can also produce symptoms such as chest pain, coughing, and pulmonary congestion.

The main sources of pollution in developed countries are motor vehicles and industry. It is useful to measure the level of traffic in a city in order to predict air pollution. The City of Valencia provides a network of sensors (electromagnetic coils) that measure the intensity of traffic (Vehicles/hour). This information can be found in the open data site of the Valencia City Council[2]. Meteorological conditions influence severely in the generation and distribution of air pollutants. 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. Strong winds can disperse pollutants and transport them away from their emission point. We have collected Meteorological observations of Valencia city from Meteorological Agency of the Government of Spain (AEMET)[3].

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. Valencia city has six stations for the detection and measurement of air pollution, although not all the stations measure the same parameters. 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 calendar features and weather conditions. Concretely, we extract the following set of features for each station:

- **Meteorological conditions:** Temperature (Celsius degrees), Relative humidity (Percentage), Pressure (hPa), Wind speed (m/s), 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) and traffic level 3 hours before

- **Pollution features:** Pollution level in the target station 3 hours before

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

Additionally, given the particular behaviour of the set of pollutants analysed and in light of that some of these pollutants can derive from others in some cases we add extra features. Precisely, for predicting NO_2 we include NO and SO_2 3 hours before, for predicting NO we also use NO_2 and SO_2 3 hours before, and finally, for forecasting O_3 we include NO and NO_2.

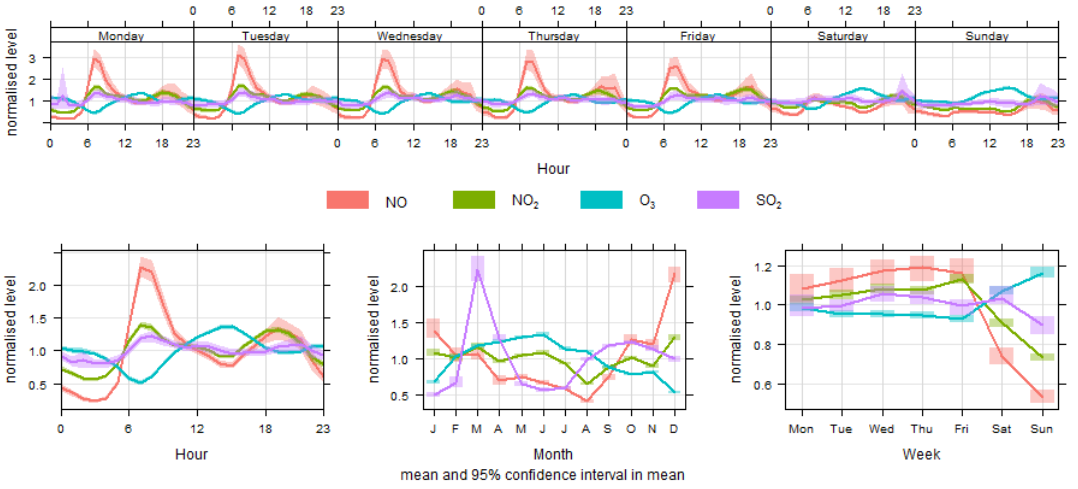

Figure 1: Distribution of the average level of four pollutants (NO, NO_2, O_3, SO_2) in *Pista* station depending on hour of the week (top), hour of the day (bottom left), month (bottom centre), week day (bottom right).

We can see a summary of the datasets in Table 1. Averages and standard deviation for the analysed stations of the pollutant particles measured and the intensity of traffic associated with each station (average traffic per hour measured the traffic sensors closer than 1km to the station) are included in this table. If we analyse traffic intensity, *Pista* and *Viveros* are the busiest stations. With regard to pollution levels *Pista* station presents the maximum levels for all the parameters except O_3. This behaviour can probably be associated with the specific situation of the station. *Pista* is located in a the central part of the city surrounded by busy streets, and therefore vulnerable to the overall city pollution. Note that some stations do not measure all the pollutants.

Figure 1 represents the distribution of the average level of four pollutants (NO, NO_2, O_3, SO_2) depending on different calendar factors[4]. It is easy to see that there are direct correspondence between the level of measured pollution and some of these factors. For instance, most of pollutants reach the lowest levels during weekend days and summer months when traffic is not intense. We can observe a peak in March in SO_2, this is a local phenomenon caused by the *fallas* traditional celebration that concludes around midnight on March 19th with the combustion of hundreds of cardboard monuments. We can also see a negative correlation between O_3 and the other three pollutants (specially NO, NO_2). This can be explained if we consider that part of the urban O_3 generation occurs when nitrogen oxides (NOx) and other compounds react in the atmosphere in the presence of sunlight. This effect can be observed in the high levels of ozone

[4]This figure has been generated using the R Openair Library [2].

around midday and in summer. The strange behaviour of high values in urban O_3 has been detected in several cities. In [3], the authors defend that "the primary cause of the higher O_3 on weekends is the reduction in oxides of nitrogen (NOx) emissions on weekends in a volatile organic compound (VOC)-limited chemical regime".

Station	Traffic		NO		NO_2		O_3		SO_2	
	ave	sd	ave	sd	ave	sd	ave	sd	ave	sd
Moli	13357	10302	10.3	19.9	28.2	20.7	46.9	25.5	2.4	2.1
Pista	34606	24462	23.5	33.8	45.1	27.0	47.1	25.0	3.8	3.8
Francia	20037	14277	8.5	18.8	26.7	23.2	50.2	25.2	2.3	2.3
Viveros	33214	24746	9.7	21.5	29.4	24.2	45.2	28.5	2.7	2.5
Bulevar	11352	8555	12.8	27.5	29.2	22.1	48.5	28.3	2.2	2.1
UPV	11987	8938	7.6	17.9	24.3	24.1	56.3	27.3	2.2	3.1

Table 1: Averages and standard deviation of the pollution detection sensors for the available stations.

2.2 Experiments

We use several regression learning techniques from R [14] in order to identify the technique that is able to better predict the levels of pollution. We build the models using as training data the registers of 2013 and the first nine months of 2014, and test the models with the last three months of 2014. Concretely, we employ the following techniques for learning regression models (all of them with the default parameters, unless stated otherwise): Linear Regression (*lr*) [5], quantile regression (*qr*) [9] with *lasso* method, K nearest neighbours (*IBKreg*) with $k = 10$ [5] , a decision tree for regression (*M5P*) [5], and Random Forest (*RF*) [11]. In order to compare the predictive performance of the regression models, we 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*). Root Mean Squared Error (RMSE) is used as performance measure.

Table 2 contains the RMSE of the regression models for the prediction of the four target pollution levels of the *Molí* station[5]. When observing these results, we can conclude that machine learning models are able to improve the performance of the basic baseline models in almost all cases. Comparing learning techniques, ensembles of decision trees technique (Random Forest) is the best model in almost all of cases. Given these results, Random Forest models will be applied in the following experiments in this work.

Machine learning has been widely used for predicting pollution levels. A seminal work in this area with neural networks is [22]. Neural networks models have been widely employed in this field, a review of these approaches can be found in [8]. A more related work is [7]. 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. The paper compares the predicted NO and NO_2 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. Our comparison of regression techniques obtains similar conclusions to the work presented in [17]. In this study, principal components analysis (PCA) is performed to identify air pollution sources. From the extracted

[5] Results for the other five stations are shown in http://www.dsic.upv.es/~flip/pollutionexp.pdf.

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.

	TrainMean	TestMean	X3h	LR	qr	IBkreg	M5p	RF
NO	30.41	28.45	33.76	25.32	30.29	27.33	23.55	**19.72**
NO2	20.80	20.80	20.72	15.52	15.47	15.04	16.51	**14.42**
O3	30.33	21.81	18.32	14.91	14.98	15.88	12.52	**11.79**
SO2	1.65	1.14	1.25	1.13	1.25	1.26	1.18	**1.04**

Table 2: Results in RMSE of different regression models for Moli Station. The best prediction model is highlighted in bold.

2.2.1 Models with wind direction

Previous methods only take wind strength into account in order to build machine learning models. Wind direction can contribute significantly in the dispersion of pollutants. For instance, if we consider quarters in the shore of coastal cities, when wind is coming from sea, the levels of pollutants are drastically lower than when winds is coming from dense populated areas of the city. This behaviour can be seen in Figure 2 (generated with library [2]). These plots show the average of NO and NO_2 pollutants depending on wind speed and direction, and they clearly show how these factors correlate with respect to the level of these pollutants.

A simple way of using wind direction component is to consider the pair attributes sine and cosine of the angle defined by wind direction. In our case, this method has obtained poor results. Therefore we have adopted a different approach of including wind direction: we modify the area where we select sensors for the traffic measurement according to wind direction. We use the idea that traffic is generating many of the pollutants that we are trying to predict, and these pollutants are dispersed according to wind speed and direction. We study two different versions of this idea: in the *dir* method we consider the traffic that is generated in the radius of 1km from the sensor but only considering the traffic sensors that are in the windward circular sector of 30º; and in the *wdir* method is similar to *dir* method but now we use the windward in order to weight traffic sensors, and in this way we give more importance to traffic measures in the circular sector of 30º. In Table 3 we compare the results in RMSE of the six stations (with the same methodology of the previous section) using three scenarios to incorporate wind direction: *nd* (wind direction is not used), *dir* method and *wdir* method. The results show that performance of the methods depends drastically on the pollutant. *dir* and, specially, *wdir* methods are able to improve the prediction performance in particles directly related to traffic emissions (SO_2, NO and NO_2) probably because these techniques of modelling wind direction are based on the selection of traffic measures according to the direction of the wind stream. O_3 is not directly related to traffic emissions, and in this case, the *dir* and *wdir* methods are generally not able to enhance the *nd* baseline method. In any case we can also observe a wide variety of performance depending on the station since every station is located in a different environment with specific features (city centre, residential area, coast shore...).

Wind direction has rarely been incorporated into land-use regression models. [4] identifies 25 land-use regression studies and only two incorporate wind direction in the predictive models. [1] studied the use of wind fields to improve the prediction of air pollution in Toronto. Wind direction fields were constructed from 38 weather stations, and these features were significantly useful for NO_2 prediction. Another approach is [18]. Here, the authors apply land use regression

(LUR) integrating wind speed, wind direction and cloud cover/insulation to estimate hourly NO and NO_2 concentrations.

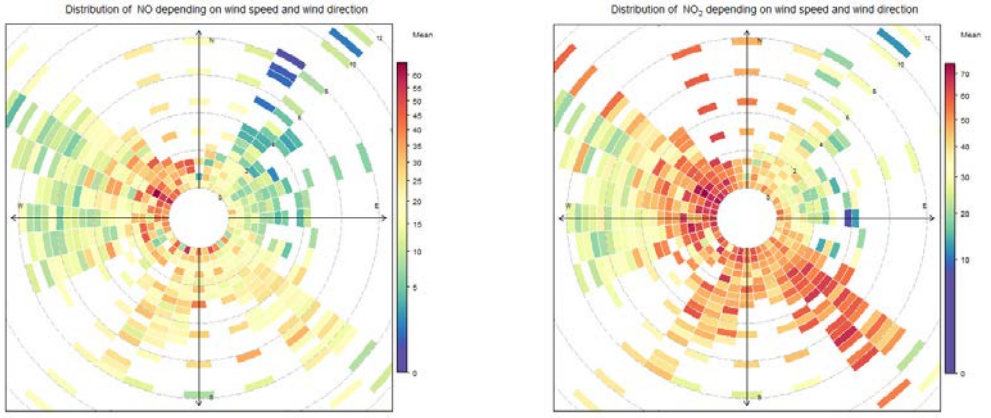

Figure 2: Distribution of NO and NO_2 pollutants in Pista station depending on wind speed and wind direction.

Station	NO			NO_2			O_3			SO_2		
	nd	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*	*nd*	*dir*	*wdir*
Moli	19.72	20.50	**19.54**	**14.42**	15.78	14.50	11.79	**11.67**	12.01	1.038	**1.037**	1.046
Pista	32.98	33.90	**32.48**	17.63	**16.73**	17.55	14.48	14.92	**14.41**	1.76	**1.74**	1.75
Francia	22.65	**23.53**	22.64	13.53	14.57	**13.44**	**12.61**	13.00	12.76	2.23	2.25	**2.22**
Viveros	26.38	26.46	**25.96**	**12.79**	13.49	12.99	13.30	13.50	**13.29**	**1.17**	1.19	1.18
Bulevar	28.79	30.21	**28.46**	**15.26**	18.18	15.77	11.37	**11.21**	11.42	0.92	1.14	**0.91**
UPV	26.21	26.32	**26.18**	18.71	19.17	**18.51**	12.84	**12.44**	12.89	0.90	0.91	**0.85**

Table 3: RMSE of the Random Forest regressors for the pollutants NO, NO_2, O_3 and SO_2 depending on the method to model wind direction: *nd* (wind direction is not used), *dir* (direction is used to select traffic sensors), *wdir* (similar to *dir* method, but nearby traffic sensors are given more relevance). The best prediction model is highlighted in bold.

3 Interpolation of predictions

In the previous section we have analysed how to obtain real-time air pollution predictions from a given set of features. Our objective in this section is to interpolate predictions all around the city in order to be able of forecasting the concentration of pollutants in locations that are not close to the pollutant measurement station. Spatial interpolation [10] tries to predict values for cells in a raster from a limited number of sample data points. Spatial interpolation can be used to forecast unknown values (eg. elevation, chemical concentrations, noise levels..) for any geographic point data in the raster. Formally, given a set of N known sample data points in the study region $\mathcal{D}$. The set of N known data points are a list of tuples: $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}, x \in \mathcal{D}, y \in \mathbb{R}$. A spatial interpolation method is a function u, such that $u(x) : x \rightarrow \mathbb{R}, x \in \mathcal{D}$. We study the following interpolation techniques:

- **Mean**: A baseline method where we always predict the average of all the N known points ($\overline{y}$). Formally, $u(x) \rightarrow \overline{y}$.

- **Inverse Distance Weighting (IDW)**: The values of unknown points are computed using a weighted average of known points. The weights are estimated using distances among the target point and known points. Here we used the well-known *Shepard's method* [16] with power parameter $p = 1$.
- **Local Inverse Distance Weighting (LIDW)**: A different method for Inverse Distance Weighting. This version assigns greater influence to values closest to the interpolated point compared to IDW and $p = 1$, We define $d(x_a, x_b)$ as the Euclidean distance in $\mathcal{D}$ between points x_a and x_b. Then $u(x_i) = w(x_1, x_i) * y_i + ..w(x_N, x_i) * y_N$, where $w(j,k) = (D_{tot}(k) - d(j,k))/(D_{tot}(k) * (|N| - 1))$, and $D_{tot}(k) = \sum(d(k, x_i)), \forall x_i \in N$.
- **Wind Sensitive LIDW**: A modification of LIDW that takes into account wind direction in such a way that we increase the weights of the known points that are windward. We define a windward circular sector of 30º from the point to predict, and the weights of the stations located in that sector are increased by a factor $\alpha = 1.5$.
- **Kriging**: In Kriging the surrounding measured values are weighted to produce a predicted value for an unknown point. Weights are based on the distance between the known points, the prediction locations, and the overall spatial arrangement among the known points. Here we use the R implementation of [15].

In order to evaluate the interpolation methods, from the six available stations, we establish three different settings. A) 3 stations as known points versus 3 stations as unknown points (20 possible combinations); B) 4 stations as known points versus 2 stations as unknown points (15 possible combinations); C) 5 stations as known points versus 1 stations as unknown points (6 possible combinations). Table 4 includes the average RMSE of the unknown points with respect the real value for the three different settings and the studied pollutants. Here, we use the whole dataset, i.e. hourly measures for years 2013 and 2014. If we compare the results of the interpolation methods in this table, Kriging and Wind Sensitive LIDW obtain the best performance. In general, Kriging interpolates better with few known points and Wind Sensitive LIDW shows better performance when it can use more information to interpolate. The exception to this behaviour is O_3 where interpolation methods cannot improve the mean baseline.

	A)**3 Known - 3 Unknown 20it**					B)**4 Known - 2 Unknown 15it**					C)**5 Known - 1 Unknown 5it**				
	Mean	LIDW	IDW	Wind	Krig.	Mean	LIDW	IDW	Wind	Krig.	Mean	LIDW	IDW	Wind	Krig.
NO	22.15	19.55	17.67	**16.92**	19.94	18.13	17.76	17.89	**14.99**	19.89	17.13	16.72	14.76	**14.71**	22.98
NO_2	21.49	22.06	21.46	20.02	**19.99**	19.04	22.19	20.80	**17.31**	20.56	17.90	19.20	16.31	**16.09**	24.18
O_3	15.37	14.54	17.54	17.72	**13.16**	**12.32**	20.44	14.44	15.13	12.66	**12.03**	13.74	13.87	13.90	18.66
SO_2	3.05	3.23	2.95	2.91	**2.90**	3.19	2.96	3.05	**2.47**	3.42	2.97	2.90	**2.37**	2.47	3.42

Table 4: Comparison of five methods of spatial interpolation of four pollutants and three different settings. RMSE of of the unknown points with respect the actual value for the hourly measures of years 2013 and 2014. The best prediction model is highlighted in bold.

The application of spatial interpolation methods over the forecasts of the six pollution stations provides a way to estimate a real-time pollution heat map of the city. An example of these plots are included in Figure 3. Here we show the spatial interpolation of O_3 by Wind Sensitive LIDW (left) and Kriging (right) in the city of Valencia.

Some works have addressed the interpolation of air pollution forecasts. In [12] the authors compare Land-use regression (LUR) [4] and universal Kriging (UK) (a version of Kriging that assumes a general polynomial trend model). In their experiments with prediction models for NOx in Los Angeles (USA), the UK interpolation consistently outperformed LUR. The RIO method is presented in [6] as a interpolation model for air pollution. The method uses a β

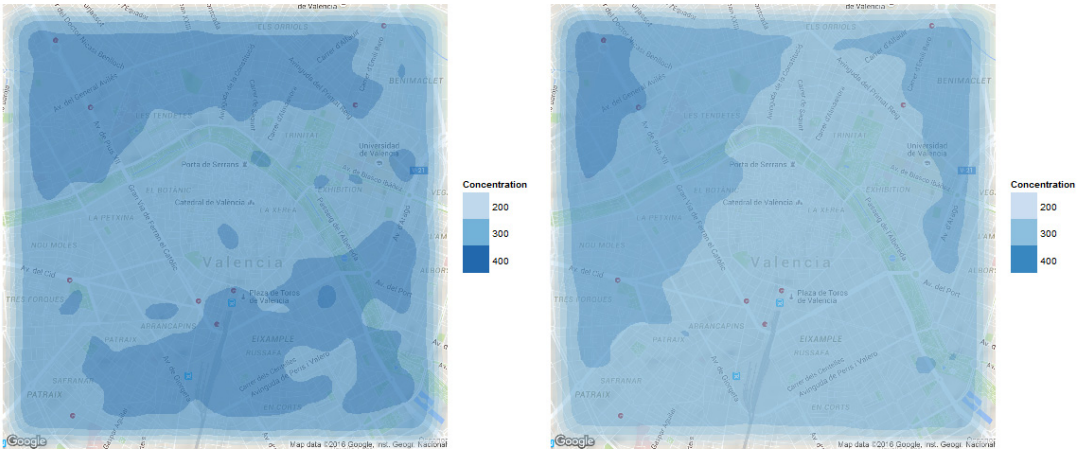

Figure 3: Spatial interpolation of O_3 by Wind Sensitive LIDW (left) and Kriging (right).

parameter that offers flexibility in the weighting between land use and air pollution levels. The experiments with O_3, NO_2 and PM_{10} in a cross-validation procedure show that RIO produces better results compared to IDW and Ordinary Kriging.

4 Conclusions

Poor Air quality is one of the factors that can decrease life expectancy since contamination rises the risk of suffering respiratory diseases. The detection of risky levels in real time can reduce the exposure to ambient air pollution. In this work, we have studied machine learning methods that predicts in real-time the levels of four dangerous pollutants for the six pollution measurement stations in the city of Valencia. According to our experiments, the Random Forest technique is able to build the best forecast models in most of the studied cases. We have analysed how we can enrich these models by incorporating wind direction information. We have proposed an approach where wind direction is used for dynamically select the traffic emission sources. The results show that this approach is able to improve the performance of the predictions for the pollutants directly related to emissions by fuel combustion. Finally, we have proposed a new interpolation method based on LIDW (Local Inverse Distance Weighting) that takes wind direction into account. We have compared the novel technique with respect to well-known spatial interpolation methods such as Kriging or common IDW. The experiments show that our Wind Sensitive LIDW obtains a positive performance specially when there are significant number of known points to use in the interpolation.

As future work, we propose the application of local features of the target points in the interpolation methods, e.g. nearby traffic level or altitude. We also plan to apply the presented techniques in other cites in order to study if similar behaviours are observed.

Acknowledgments

This work has been partially 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) funded by MINECO in Spain (PCIN-2013-037), the EU (FEDER) and the Spanish MINECO under grants TIN 2015-69175-C4-1-R and TIN 2013-45732-C4-1-P and by Generalitat Valenciana under grant PROMETEOII/2015/013. We are also grateful to Ajuntament de València, InnDEA-València and specially to Ramón Ferri and Ruth López for their help in providing traffic data.

References

[1] M.A. Arain, R. Blair, N. Finkelstein, J.R. Brook, T. Sahsuvaroglu, B. Beckerman, L. Zhang, and M. Jerrett. The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies. *Atmospheric Environment*, 41(16):3453 – 3464, 2007.

[2] David C. Carslaw and Karl Ropkins. openair — an r package for air quality data analysis. *Environmental Modelling and Software*, 27–28(0):52–61, 2012.

[3] Jon M Heuss, Dennis F Kahlbaum, and George T Wolff. Weekday/weekend ozone differences: what can we learn from them? *Journal of the Air & Waste Management Association*, 53(7):772–788, 2003.

[4] Gerard Hoek, Rob Beelen, Kees de Hoogh, Danielle Vienneau, John Gulliver, Paul Fischer, and David Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561 – 7578, 2008.

[5] Kurt Hornik, Christian Buchta, and Achim Zeileis. Open-source machine learning: R meets Weka. *Computational Statistics*, 24(2):225–232, 2009.

[6] Stijn Janssen, Gerwin Dumont, Frans Fierens, and Clemens Mensink. Spatial interpolation of air pollution measurements using {CORINE} land cover data. *Atmospheric Environment*, 42(20):4884 – 4903, 2008.

[7] A Karppinen, J Kukkonen, T Elolähde, M Konttinen, and T Koskentalo. 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, 2000.

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

[9] Roger Koenker. *quantreg: Quantile Regression*, 2015. R package version 5.11.

[10] Jin Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors. *Ecological Informatics*, 6(3):228–241, 2011.

[11] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[12] Laina D Mercer, Adam A Szpiro, et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

[13] Lidia Contreras Ochando, Cristina I. Font Julián, Francisco Contreras Ochando, and Cèsar Ferri Ramirez. Airvlc: An application for real-time forecasting urban air pollution. In *Proceedings of the 2nd International Workshop on Mining Urban Data*, pages 72–79, 2015.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[15] P.J. Ribeiro Jr. and P.J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):15–18, 2001.

[16] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA, 1968. ACM.

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

[18] Jason G. Su, Michael Brauer, Bruce Ainslie, Douw Steyn, Timothy Larson, and Michael Buzzelli. An innovative land use regression model incorporating meteorology for exposure analysis. *Science of The Total Environment*, 390(2–3):520 – 529, 2008.

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

[20] Elissa H. Wilker, Sarah R. Preis, Alexa S. Beiser, Philip A. Wolf, Rhoda Au, Itai Kloog, Wenyuan Li, Joel Schwartz, Petros Koutrakis, Charles DeCarli, Sudha Seshadri, and Murray A. Mittleman. Long-Term Exposure to Fine Particulate Matter, Residential Proximity to Major Roads and Measures of Brain Structure. *Stroke*, April 2015.

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

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