airVLC: An application for visualizing wind-sensitive interpolation of urban air pollution forecasts

Lidia Contreras-Ochando
ICube,
Université de Strasbourg
CNRS, 300 Bd Sebastien Brant, BP 10413,
67412 Illkirch Cedex, France
Email: liconoc@upv.es

Cèsar Ferri
DSIC,
Universitat Politècnica de València
Camí de Vera s/n, 46022,
Valencia, Spain
Email: cferri@dsic.upv.es

***Abstract*—Air pollution has been identified as a major source of health problems for people living in cities. In this sense, it is important to identify the areas of the city that present high levels of pollutants in order to avoid them. airVLC is an application for predicting and interpolating real-time urban air pollution forecasts for the city of Valencia (Spain). We compare different regression models in order to predict the levels of four pollutants (NO, NO_2, SO_2, O_3) in the six measurement stations of the city. Since wind is a key feature in the dispersion of the pollution, we study different techniques to incorporate this factor in the models. Finally, we are able to interpolate forecasts all around the city. For this goal, we propose a new interpolation method that takes wind direction into account, improving well-known methods like IDW or Kriging. By using these pollution estimates, we are able to generate real-time pollution maps of the city of Valencia and publish them into a public website.**

I. Introduction

People exposed to urban air pollution are more vulnerable to suffering diseases such as pneumonia, lung cancer or cardiovascular diseases [1]. Air pollution is one of the factors with most impact in the health of people and animals, as well as in the decline of terrestrial and aquatic ecosystems [2]. Urban agglomerations and especially big cities must try to decrease their emissions. People in these cities must reduce their exposure to air contamination as much as possible. This is especially important for high risk population like kids, elderly people, pregnant women and people suffering respiratory diseases.

For these reasons, monitoring air pollution levels is an important aspect to consider. Many cities provide open data about the levels of several pollutants. However, these data has to be verified causing delays in the publish of the contamination levels, mainly in the case of pollutants of reduced dimensions. In Valencia, our case of study, this delay is about three hours, and it can represent a problem since risky high levels of pollution are not detected in real-time. Moreover, due to the high cost of the equipment and its maintenance, the sensor network tends to be limited (six sensors in the city of Valencia) leaving many zones without information.

Considering these restrictions, in this work we address the problem of producing real-time predictions of the levels of pollution by employing data about traffic intensity and meteorological information. We study the performance of predictions with different techniques for building regression models with three years of hourly historical data. We also analyse how to incorporate wind direction into the regression models, due to the influence that this feature has on pollutant dispersion. Spatial interpolation is useful to produce pollutions forecasts all around the city. Well-known spatial interpolation methods are static methods due to the fact they do not consider context conditions around the points to interpolate. Wind can produce obvious effects in the dispersion or concentration of pollution. For that reason, this is an important feature that have to be analysed and used in order to calculate the level of pollution, however not many models take it into account. We propose a new method that uses wind direction in order to improve the results of the interpolation of urban air pollution, considering this technique as a wind-sensitive interpolation approach. This method has obtained better results in comparison with well-known methods, mostly when we have enough information to interpolate. With our models, we are able to show the estimate concentration of pollutants for all the city, based on the prediction of six stations.

The experiments described in this paper and their results were presented at ICCS 2016 [3] and this application can be considered an extension of [4]. That work was focused on presenting the *airVLC* application for real-time forecasts of air pollutants. In the models of [4] we did not consider wind direction in the learning models and interpolation techniques were not applied.

The paper is organised as follows. Section II include some experiments in learning regression models for predicting pollutant concentrations and some results on including wind direction in the models. Also, we introduce interpolations methods. Section III describes the system architecture and shows some visualisations that will be included in the application. Related Work and Conclusions are presented in IV and V.

II. Experiments

A. Data collection

In order to predict the concentration of pollution and given that we are not able to obtain data about pollution levels in real-time, we have built datasets using the traffic intensity and meteorological conditions in Valencia, in addition to the levels of pollutants three hours before (when data is already validated). These data can be downloaded from open data sources in real-time. Concretely, we have one dataset for each pollution measurement station with the following set of features:

- **Pollution level:** NO, NO_2, SO_2, O_3
- **Meteorological conditions:** Temperature, Relative humidity, Pressure, Wind speed, Rain
- **Calendar features:** Year, Month, Day in the month, Day in the week, Hour
- **Traffic intensity features:** Traffic level in the surrounding of the stations (1km) and traffic level 3 hours before
- **Pollution features:** Pollution level in the target station 3 hours before

B. Prediction

We have tested different regression learning techniques in order to determine the best technique for predicting the levels of pollution. All of our experiments has been carried out using R [5]. We have used historic hourly data of three years (2013, 2014 and 2015) and we have employed *data splitting* for time series[1], moving the test data from a period of one month until a period of four months. Concretely, we employ the following techniques for learning regression models (all of them with the default parameters, unless stated otherwise): Linear Regression (*lr*) [6], quantile regression (*qr*) [7] with *lasso* method, K nearest neighbours (*IBKreg*) with $k = 10$ [6], a decision tree for regression (*M5P*) [6], and Random Forest (*RF*) [8]. 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.

With the results [3] we can conclude that machine learning models are able to improve the performance of the basic baseline models in almost every case. Comparing these techniques, ensambles of decision trees (Random Forest) result the best models and they will be applied in the following experiments in this work.

1) Models with wind direction: Since wind direction can have an important effect in the dispersion or concentration of pollutants, we have modified the area where we select traffic sensors around the stations depending on this feature, due to the fact that traffic is generating most of the pollutants that we are trying to predict. 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, considering only the traffic sensors that are in the windward circular sector of 30°; and in the *wdir* method 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 30°.

The results depends drastically on the pollutant. However, *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).

C. Interpolation of predictions

Spatial interpolation [9] tries to predict values for cells in a raster from a limited number of sample data points and it can be used to forecast unknown values for any geographic point in the raster. We have studied the following interpolation techniques:

- **Mean**: A baseline method where we always predict the average of all the N known points.
- **Inverse Distance Weighting (IDW)**: The values of unknown points are computed using a weighted average of known points. Here we used the well-known *Shepard's method* [10] 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. [3]
- **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 by a factor $\alpha = 1.5$. [3]
- **Kriging**: In Kriging the surrounding measured values are weighted to produce a predicted value for an unknown points. Here we use the R implementation of [11].

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).

In this case, 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 we have more information to interpolate.

III. System overview

airVLC is a web application created with `PHP` that shows in different ways the predicted levels of the pollutants, generated using R, and the interpolation all around the city.

In the server side an R script is executed every hour and it calculates prediction and interpolation levels of the four pollutants in the following way:

1) Download traffic and meteorological data corresponding to the current hour from open data sources.

[1] http://topepo.github.io/caret/splitting.html#time

2) For each station in Valencia, calculate the prediction level of the four pollutants with Random Forest model and using *wdir* method.
3) For each cell of a grid (11000x11000 cells approximately), calculate the value of pollutants with Wind Sensitive LIDW method, using the predicted data of the six stations and create some static maps and gifs of the 24h evolution.

All the data are stored in a MySQL data base on an Apache server. Figure 1 represents the general architecture of the system and how it works.

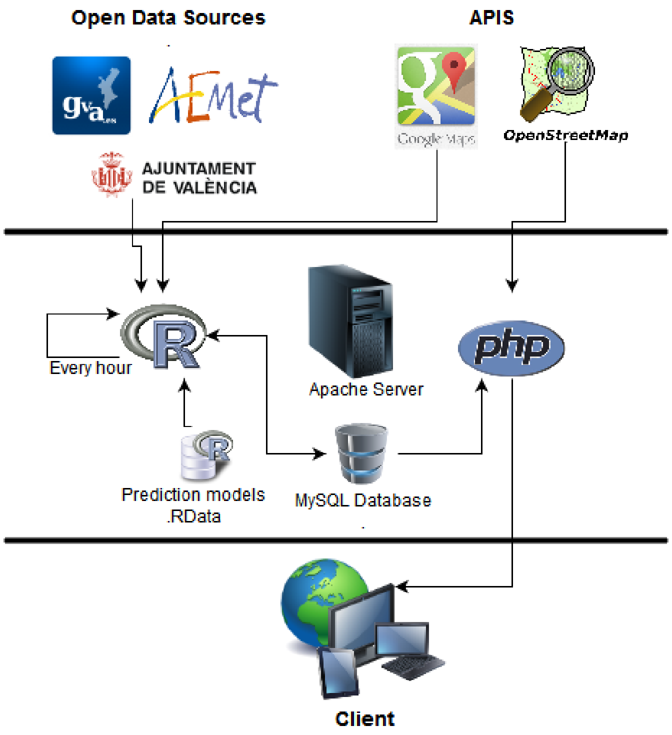

Fig. 1. General system architecture of *airVLC* application

The public website is created with PHP and jQuery, and uses javascript libraries such as leaflet[2] as well as plugins and services of Mapbox platform[3]. With these two libraries, the website shows OpenStreetMap maps with different interactive layers (density maps, cloropleth maps, etc.) showing the information in different ways. Figure 2 shows an example of a interpolation map for one pollutant (SO_2), created using R and Google Maps, that will be included at the website, besides the interactive maps created using javascript. The website also includes a data section to explore the last collected and calculated data.

Figure 3 represents a comparison between the real data and the predictions. These type of graphics have been done using the R Openair Library [12].

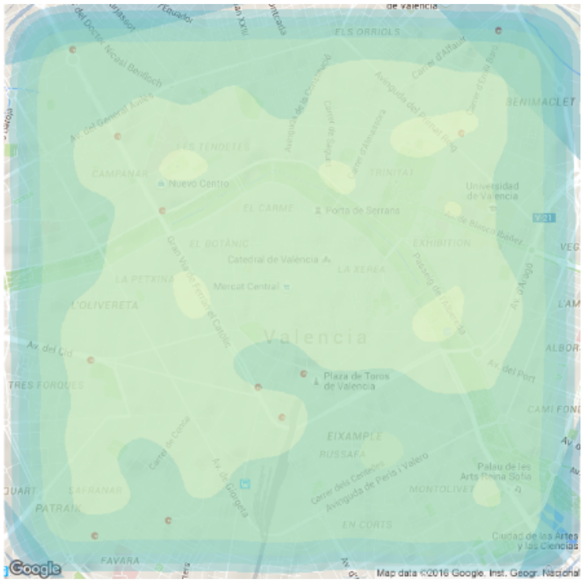

Fig. 2. Spatial interpolation of SO_2 by Wind Sensitive LIDW

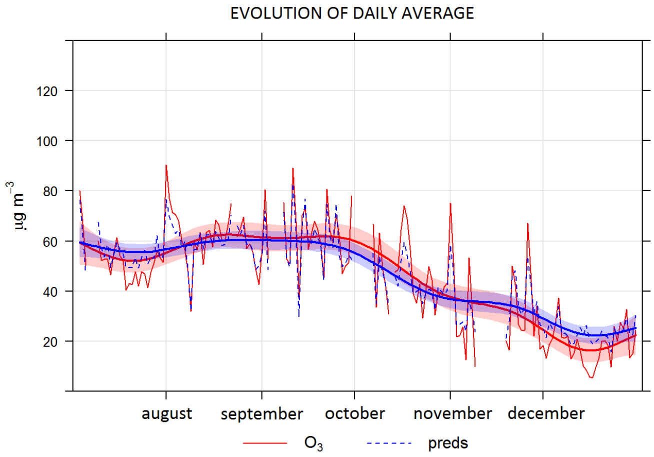

Fig. 3. Comparison between observed data and predicted data of O_3

In order to calculate the correct levels of pollutants and show them with different colours in the thematic maps using javascript, we adopted the standard hourly air quality index published in the Directives of the European Commission[4].

Since all the experiments have been ended, the R script is already programmed, downloading and storing data. However, the website is currently under development and we presume that it will be available by the end of September.

IV. Related Work

Machine learning has been widely used for predicting pollution levels, in particular by using neural networks [13] [14]. In [15] the authors propose a modelling system for predicting the traffic, emissions, and atmospheric dispersion of pollution in an urban area.

[2] Leaflet is an open-source JavaScript library for mobile-friendly interactive maps. See http://leafletjs.com/

[3] Mapbox is a full-featured design suite for creating custom map styles. See https://www.mapbox.com/

[4] http://ec.europa.eu/environment/air/quality/standards.htm

Our comparison of regression techniques obtains similar conclusions to the work presented in [16], in which principal components analysis (PCA) is performed to identify air pollution sources.

Wind direction has rarely been incorporated into regression models. [17] identifies 25 land-use regression studies where only two incorporate wind direction in the predictive models.

The most advanced example of these type of systems in Spain is CALIOPE[5]. This system makes forecasts of the pollution levels in Spain using data from every station in the country. CALIOPE uses Lineal Regression for predicting and Kriging for the spatial interpolation of its results.

V. Conclusion

High levels of pollution can increase the risk of suffering respiratory diseases and, in this way, decrease life expectancy. The detection of these levels in real-time or even in advance is a crucial point. In this work, we have studied techniques of machine learning for predicting the levels of four pollutants in real-time for the city of Valencia. Our results show that Random Forest is able to produce the best predictions in most of the cases. We have analysed how to incorporate wind features in order to improve these results. For this, we have proposed an approach where wind direction is used for dynamically selecting traffic emission sources. We also have proposed a new interpolation method that takes wind direction into account, obtaining better results compared to well-known spatial interpolation methods. Theses techniques have been incorporated into a web system, *airVLC*, that provide these predictions in a public and open way.

As future work, we propose different lines where we can continue this work. First, we plan to study the characteristics of each station and each pollutant separately. Second, we propose the application of local features in the target points when we use interpolation methods, e.g. nearby traffic level or altitude. Finally, we plan to apply the presented techniques in other cities in order to study if similar behaviours are observed.

References

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

[2] P. A. López Jiménez and V. Espert Alemany, *Dispersión de contaminantes en la atmósfera*.

[3] L. Contreras and C. Ferri, “Wind-sensitive interpolation of urban air pollution forecasts,” *Procedia Computer Science*, vol. 80, pp. 313 – 323, 2016, international Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187705091630758X

[4] L. C. Ochando, C. I. F. Julián, F. C. Ochando, and C. F. Ramirez, “Airvlc: An application for real-time forecasting urban air pollution,” in *Proceedings of the 2nd International Workshop on Mining Urban Data*, 2015, pp. 72–79. [Online]. Available: http://ceur-ws.org/Vol-1392/paper-10.pdf

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

[6] K. Hornik, C. Buchta, and A. Zeileis, “Open-source machine learning: R meets Weka,” *Computational Statistics*, vol. 24, no. 2, pp. 225–232, 2009.

[7] R. Koenker, *quantreg: Quantile Regression*, 2015, r package version 5.11. [Online]. Available: http://CRAN.R-project.org/package=quantreg

[8] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[9] J. Li and A. D. Heap, “A review of comparative studies of spatial interpolation methods in environmental sciences: performance and impact factors,” *Ecological Informatics*, vol. 6, no. 3, pp. 228–241, 2011.

[10] D. Shepard, “A two-dimensional interpolation function for irregularly-spaced data,” in *Proceedings of the 1968 23rd ACM National Conference*, ser. ACM ’68. New York, NY, USA: ACM, 1968, pp. 517–524. [Online]. Available: http://doi.acm.org/10.1145/800186.810616

[11] P. Ribeiro Jr. and P. Diggle, “geoR: a package for geostatistical analysis,” *R-NEWS*, vol. 1, no. 2, pp. 15–18, 2001. [Online]. Available: http://cran.R-project.org/doc/Rnews

[12] D. C. Carslaw and K. Ropkins, “openair — an r package for air quality data analysis,” *Environmental Modelling and Software*, vol. 27–28, no. 0, pp. 52–61, 2012.

[13] J. Yi and V. R. Prybutok, “A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area,” *Environmental Pollution*, vol. 92, no. 3, pp. 349–357, 1996.

[14] M. Khare and S. S. Nagendra, *Artificial neural networks in vehicular pollution modelling*. Springer, 2006, vol. 41.

[15] 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*, vol. 34, no. 22, pp. 3735–3743, 2000.

[16] K. P. Singh, S. Gupta, and P. Rai, “Identifying pollution sources and predicting urban air quality using ensemble learning methods,” *Atmospheric Environment*, vol. 80, pp. 426–437, 2013.

[17] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, “A review of land-use regression models to assess spatial variation of outdoor air pollution,” *Atmospheric Environment*, vol. 42, no. 33, pp. 7561 – 7578, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231008005748

[5] CALIOPE: http://www.bsc.es/caliope/