16th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
8-11 September 2014, Varna, Bulgaria

PAHS URBAN CONCENTRATIONS MAPS USING SUPPORT VECTOR MACHINE

Armando Pelliccioni[1], *Andrea Cristofari*[1], *Mafalda Lamberti*[1], *Claudio Gariazzo*[1]
[1]INAIL-DIPIA, Roma, Italy.

Abstract: The studies about health effects are often based on data inferred by monitoring stations. For this purpose, pollutants exposure maps are crucial for evaluating health effects. Studying the Polycyclic Aromatic Hydrocarbons (PAHs) and the Benzo(a)Pyrene (BaP) exposure in urban areas is the major goal of the EXPAH LIFE+ Project. An integrated approach, based on measurements and modeling techniques, was applied to simulate PAHs and BaP levels in the Rome metropolitan area. Field campaigns of PM2.5 with PAHs content were performed for the period June 2011 - May 2012, and a chemical transport model (FARM) was applied to reconstruct PAHs urban concentrations. In this work, Machine Learning methods have been applied to forecast atmospheric pollution, trying also to improve the results achieved by FARM. In particular, Support Vector Machines (SVMs) have been used. They represent one of the most common approaches among Machine Learning methods. Starting from the experimental data, SVM methods have been applied to build models able to forecast PAHs and BaP exposure. The SVM models seem to show excellent results in the reproduction of experimental data and in generalization, improving those achieved by FARM. Finally, the SVM models have produced very congruent PAHs and BaP exposure maps.

Keywords: *Support Vector Machine, PAH, BaP, exposure levels, urban air pollution*

INTRODUCTION

Epidemiological studies about health effects of air quality are often based on data inferred by monitoring stations. The issue of constructing pollutants exposure maps is crucial for improving such studies. The study of the Polycyclic Aromatic Hydrocarbons (PAHs) and the Benzo(a)Pyrene (BaP) exposure in urban areas is the major goal of the EXPAH LIFE+ Project (www.ispesl.it/expah). An integrated approach, based on measurements and modeling techniques, was applied to simulate PAHs and BaP levels in the Rome metropolitan area. PAHs are pollutants linked to combustion processes and they can be considered relevant for health problems, especially in high density urban areas. The following PAHs compounds were measured during some previous field campaigns: benz[a]anthracene (BaA), benzo[b]fluoranthene (BbF), benzo[j]fluoranthene (BjF), benzo[k]fluoranthene (BkF), benzo[a]pyrene (BaP), indeno[1,2,3-cd]pyrene (IP) and dibenz[a,h]anthracene (DBA).

In particular, this work has focused on the overall PAHs (which will be referred to as PAH, henceforth) and on BaP. Supervised Machine learning methods have been applied for estimating their concentrations. The usefulness of these methods lies in their capability to produce good predictions for new samples (never used during the training phase). The purpose of Supervised Machine Learning methods is to build a virtual machine able to learn the rules (supposed to be unknown) linking the outputs to the inputs of a system from a set of samples. The training phase consists in an adaptive process which provides an analytic description of the output function.

Support Vector Machines (SVM; Vapnik, 1995) are a class of Machine Learning methods and they have been used in this work. They were rarely used for air dispersion modeling.

The dataset that has been used contained one year air quality data concerning the urban area of Rome. Data were collected between June 1, 2011 and May 30, 2012. The region of interest was an area 60 km × 60 km, centered on the city of Rome.

SVMs have been used to build a model able to forecast PAH and BaP concentrations starting from data measured by urban stations. Then, this model has been applied to construct daily maps.

2. MATERIALS AND METHODS

2.1 DATASET CHARACTERISTICS

The region of interest was divided into 3600 pixels (each one 1 km × 1 km) and three kinds of variables have been initially considered: meteorological variables, pollutants emissions and the outputs produced

by deterministic models (FARM bc and FARM fc models). All these values were on a daily basis and were available for each pixel and for each day (totally, there were 60*60*365 = 1314000 daily samples).
Meteorological variables included wind direction, wind speed, pressure (P), precipitations (Rain), relative humidity (RH), temperature (T) and total cloud cover (TCC).
In this work, the so-called mix models methodology has been applied: it consists in considering deterministic air dispersion forecasts as input variables. The use of the outputs of deterministic models as input variables for intelligent methods was first developed in (Pelliccioni et al., 2003) and the theoretical explanation can be found in (Pelliccioni and Tirabassi, 2006 and 2008). In our case, the presence of deterministic information (from FARM bc) among the input variables it's fundamental for the model performance, as will be shown below.
Moreover, 184 actual PAH and BaP measurements have been used as output target values for the SVM. These measurements came from some campaigns distributed over all the seasons and in different sites of the area. They were on a daily basis and most derived from 2-10 days campaigns.

2.2 MONITORING STATIONS AND INPUT VARIABLES SELECTION

Two different problems have been faced to create the maps: the first one concerned the best choice of the monitoring stations representing the urban pollutant dispersion, the second one concerned which variables (i.e. features) to use as model inputs and how to treat them.
The SVM model had to be applied to build maps of the whole area. So, the monitoring stations to use in training and testing phases have been selected so that the model could be effectively assessed both for urban and non-urban conditions. All the stations chosen for training were located within the urban area, while some of the remaining 10 testing stations were located far away from the city.
Regarding the second issue, a preliminary analysis has been first conducted to study the probability density distributions and the domains of each variable. In this way, it has been possible to choose an appropriate normalization. Moreover, for Machine Learning methods, it is well known that using only a subset of the original variables could lead to better performances. So, a features selection procedure has been carried out, in order to select the subset of features providing the best test results.
In particular, different simulations have been conducted, using different subsets of features. In each simulation, SVMs have been built following two steps: the training phase (where the machine has been effectively built with the samples of the training set), and the testing phase (where the model performance has been assessed with the samples of the test set).
The results are shown in Table 1 for PAH and in Table 2 for BaP.
From the results of Table 1 and Table 2, it is evident how much the choice of the input variables is crucial for the model performances.
Combining all the strategies described, the subset of input variables with the best test results has been finally selected: date, wind direction, wind speed, precipitations, total cloud cover, and PAH/BaP estimates by FARM bc. These variables have been used in the final SVM model.

Table 1. PAH test results using different subsets of input variables

	INPUT VARIABLES				TEST RESULTS			
	Date	Meteo	Emissions	PAH by FARM bc	MAE (ng/m^3)	R^2	slope	bias (ng/m^3)
s_1		x			0.64	0.86	0.91	0.23
s_2		x	x		0.61	0.88	1.01	0.01
s_3		x	x	x	0.51	0.91	0.90	-0.01
s_4		x		x	0.45	0.90	0.92	0.05
s_5	x	x			0.49	0.88	0.88	0.04
s_6	x	x	x		0.46	0.89	0.95	0.07
s_7	x	x	x	x	0.53	0.90	0.93	0.00
s_8	x	x		x	0.44	0.90	0.90	0.07

Table 2. BaP test results using different subsets of input variables

	INPUT VARIABLES				TEST RESULTS			
	Date	Meteo	Emissions	BaP by FARM bc	MAE (ng/m^3)	R^2	slope	bias (ng/m^3)
s_9		x			0.18	0.86	0.85	0.05
s_{10}		x	x		0.15	0.87	0.88	0.03
s_{11}		x	x	x	0.17	0.88	0.88	0.00
s_{12}		x		x	0.14	0.88	0.86	0.00
s_{13}	x	x			0.13	0.90	0.87	0.01
s_{14}	x	x	x		0.16	0.87	0.86	0.01
s_{15}	x	x	x	x	0.15	0.88	0.84	0.01
s_{16}	x	x		x	0.12	0.89	0.87	0.01

3. RESULTS

As explained in the previous paragraph, SVMs performances have been validated with the samples of the test set. These results have been compared with those obtained by two deterministic models: FARM bc and FARM fc.

As shown in Table 3 and in Table 4, the SVM models provide much better results than the other two models on all indices, both for PAH and for BaP.

Table 3. Comparison between test results obtained by SVM, FARM bc and FARM fc to forecast PAH concentrations

		MAE (ng/m^3)	R^2	slope	bias (ng/m^3)	FB	NMSE	r	CV	IOA
s_A	SVM	0.37	0.93	0.96	-0.04	-0.06	0.15	0.96	0.37	0.98
	FARM bc	2.34	0.83	2.00	0.57	0.57	1.90	0.91	1.66	0.75
	FARM fc	0.61	0.80	0.78	0.25	-0.09	0.43	0.90	0.60	0.94

Table 4. Comparison between test results obtained by SVM, FARM bc and FARM fc to forecast BaP concentrations

		MAE (ng/m^3)	R^2	slope	bias (ng/m^3)	FB	NMSE	r	CV	IOA
s_B	SVM	0.11	0.92	0.94	-0.03	-0.10	0.20	0.96	0.41	0.98
	FARM bc	0.78	0.78	2.23	0.22	0.67	2.75	0.89	2.15	0.68
	FARM fc	0.18	0.76	0.71	0.08	-0.13	0.63	0.88	0.69	0.91

The models built for s_A and s_B have been applied to all the 1314000 daily samples to construct daily PAHs maps. Note that the model has been built for reproducing daily concentrations representing 2-10 days periods. So a little forcing was necessary to build daily maps. Moreover, taking account of the training samples locations, a generalization capability has been required by the SVM.

Generally, when building large area maps, not all pixels are covered by measurements and it is difficult to test them. In this work, indirect performance indices have been introduced in order to overcome this difficulty. Their role is to assess the model results using criteria that do not need the comparison between predicted and observed values. Here, the following indices have been developed:

- R_{neg} measures the percentage of negative values;
- R_{U-NU} indicates the percentage of days where pollutants concentrations is lower in the urban than in a non-urban area.

To define R_{U-NU}, three pixels have been fixed: one over sea (South-West of the area), one over lake (South-East of the area) and one in the center of Rome. They have been selected because representing urban and non-urban conditions. To compute R_{U-NU} for PAH, only those days where the SVM output in the city is higher than 1 ng/m^3 and the difference between the concentration over sea (or lake) and the concentration in the city is more than 0.2 ng/m^3 have been counted.

As for daily PAH exposure maps, the following index values have been obtained: $R_{neg} = 0$, $R_{U-NU} = 3.29\%$ comparing city with sea, and $R_{U-NU} = 2.74\%$ comparing city with lake.

A comparison between the daily PAH estimates produced by FARM bc and by SVM at the three representative pixels (city, sea, lake) pointed out the congruent behavior of the SVM model and its generalization capability. It produces estimates mostly lower over the urban area than outside, even though only urban samples have been used for training. Moreover, no remarkable difference exists between the different seasons for the estimated concentrations over sea and over lake.

In order to evaluate the annual average exposure, the daily maps can be used to build the annual average exposure maps, just computing the annual average estimates for each pixel.

The resulting maps are shown in Figure 1a, 1b and 1c.

All the maps produce higher values in the urban area than outside. However, while the maps obtained by FARM bc and by FARM fc are strongly related, the maps produced by SVM show a slight shape different. Still referring to the maps illustrated in Figure 1a, 1b and 1c, the mean PAH values over all the area are 2.23 ng/m^3, 0.98 ng/m^3 and 1.78 ng/m^3 for FARM bc, FARM fc and SVM, respectively.

These maps seem to provide a further confirm of the results obtained in the test phase, where the estimates produced by the SVM model are between those obtained by FARM bc (that tends to overestimate) and those produced by FARM fc (that tends to underestimate).

BaP annual maps are shown in Figure 2a,b,c. The mean BaP values over all the area are 0.66 ng/m^3, 0.24 ng/m^3 and 0.47 ng/m^3 for FARM bc, FARM fc and SVM, respectively.

The following index values have been obtained: $R_{neg} = 2.04\%$, $R_{U-NU} = 6.30\%$ comparing the city with the sea, and $R_{U-NU} = 5.48\%$ comparing the city with the lake.

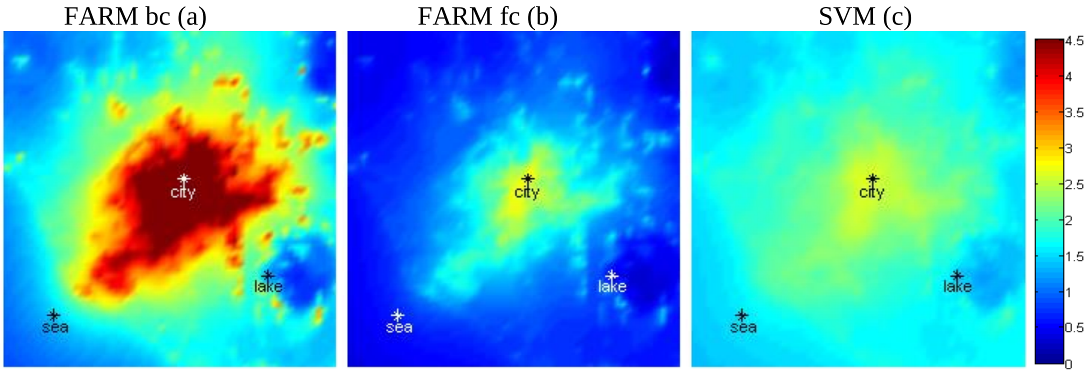

Figure 1. Annual mean PAH exposure maps by FARM bc (a), FARM fc (b) and SVM (c), in ng/m^3

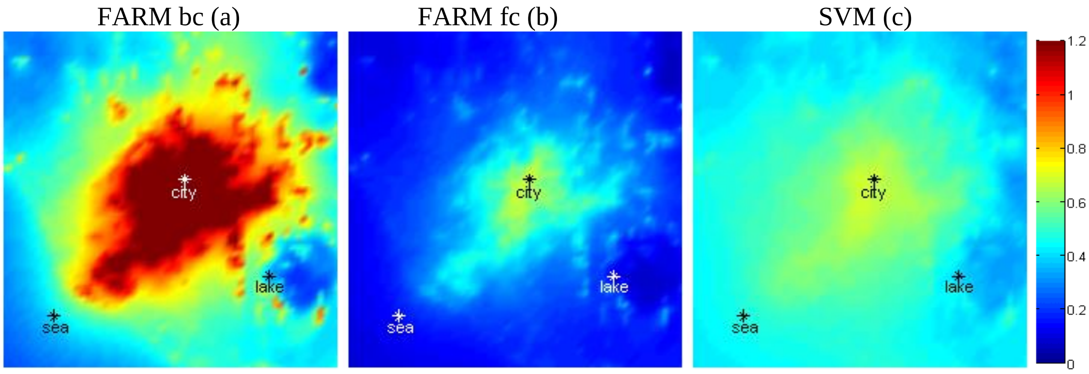

Figure 2. Annual mean BaP exposure maps by FARM bc (a), FARM fc (b) and SVM (c), in ng/m^3

4. CONCLUSIONS

Support Vector Machines (SVMs) are a class of Machine Learning methods, whose purpose is to use some samples to learn the rules that link the outputs to the inputs of a system through an adaptive learning process. These methods have been applied to forecast PAH and BaP concentrations on an area 60 km × 60 km, centered on the city of Rome over one year period, using one year air quality data and some actual measurements distributed over all the seasons.

The input variables initially considered are the following: date, meteorological variables, emissions and the outputs produced by FARM bc. Moreover, 184 actual PAHs concentration measurements were available and they have been used as SVM output targets. PAHs concentration measurements came from 2-10 days campaigns distributed over all the seasons and in different sites.

Different problems have been faced to obtain a good spatial reproduction of pollutants concentrations. The main issues concerned the choice of the monitoring stations to use for training and for testing, the way to scale the variables to make the model able to generalize, the choice of the best model inputs and the reconstruction of daily maps.

Once the SVM models have been built, they have also been applied to all the daily samples to build daily exposure maps.

Generally, when constructing maps, it's impossible to know the actual measurements in each point and in each day. For this reason, it's been necessary to introduce new indices to assess the maps. The choice of these indices is based on the observation that measurements can't assume negative values and pollutants concentrations are expected to be higher in urban areas than in non-urban areas. They measure the percentage of negative values and the percentage of days where pollutants concentrations is lower in the urban area than in a non-urban area, respectively.

As for PAH, the performances show values close to zero for the first index, and between 2.7% and 3.3% for the second one. As for BaP, the performances show values around 2% for the first index, and between 5.5% and 6.3% for the second one.

The overall results seem to confirm the SVM capability to reconstruct PAH and BaP spatial concentration and to produce realistic maps for both of them.

REFERENCES

EXPAH. Extended Technical Report on Indoor/Outdoor monitoring of PAHs, PM2.5 and its chemical components with ancillary measurements of gaseous toxicants in the frame of the EXPAH Project (Action 3.3) *http://www.ispesl.it/expah/documenti/Technical_Report_CNR_INAIL_2012h%20finale.pdf.*

EXPAH. Technical report on meteorological measurements carried out in urban and sub-urban areas of Rome in the frame of EXPAH project. Action 3.4. *http://www.ispesl.it/expah/documenti/Technical%20report%20on%20meteorological%20measurementsrev1.pdf.*

EXPAH. ACTION 4.1: Collection of raw emission inventories and their upgrading emission inventories and their upgrading emission inventories and their upgrading emission inventories and their upgrading. *http://www.ispesl.it/expah/documenti/R2011-13_ARIANET_EXPAH_A4.1.pdf.*

EXPAH. ACTIONS 4.3-4.4: Calculation and integration of traffic emissions with the updated Lazio Region inventory. Spatial, temporal and chemical disaggregation of the emission inventory. *http://www.ispesl.it/expah/documenti/R2012-05_ARIANET_EXPAH_A4.3-4_final.pdf.*

EXPAH. ACTION 4.5: Integration of PAHs atmospheric processes within FARM model. *http://www.ispesl.it/expah/documenti/R2012-01_ARIANET_EXPAH_A4.5.pdf.*

EXPAH. ACTION 7.1: Report on evaluation of policy and mitigation scenarios (revision). *http://www.ispesl.it/expah/documenti/R2013-14_ARIANET_EXPAH_A7.1_rev1.pdf.*

Gariazzo, C., Silibello, C., Finardi, S., Radice, P., Piersanti, A., Calori, G., Cecinato, A., Perrino, C., Nussio, F., Cagnoli, M., Pelliccioni, A., Gobbi, G.P., Di Filippo, P., 2007. A gas/aerosol air pollutants study over the urban area of Rome using a comprehensive chemical transport model. *Atmospheric Environment, **41**, 7286-7303.*

Pelliccioni, A., Tirabassi, T., Gariazzo, C., 2003. Coupling of Neural Network and Dispersion Models: a novel methodology for air pollution models. *Int. J. Environment and Pollution*, **20**, Nos 1-6, 136-146.

Pelliccioni, A., Tirabassi, T., 2006. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. *Environmental Modelling & Software*, **21,** 4, 539–546.

Pelliccioni, A., Tirabassi, T., 2008. Air pollution model and neural network: an integrated modelling system. *Il Nuovo Cimento C*, **31 C**, 3, 253-273.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. *Springer-Verlag, New York.*