Technical report on application of SVMs to estimate PAHs maps in the urban area of Rome – Action 5.5

Authors

A. Pelliccioni [1], A. Cristofari [1]

[1] INAIL Dipartimento Installazioni di Produzione e Insediamenti Antropici

June, 2014

1 Executive Summary

Deterministic models developed in Action 4.5 (FARM bc and FARM) are used to forecast Polycyclic Aromatic Hydrocarbons (PAHs) and benzo[a]pyrene (BaP) exposure. Sometimes, these models must be corrected to fit with data inferred by monitoring stations.
In this work, Machine Learning methods have been used to forecast atmospheric pollution, trying also to improve the results achieved by FARM. In particular, Support Vector Machines (SVMs) have been applied. They represent one of the most widely used approach among Machine Learning methods.
Some experimental data concerning the urban area of Rome were available: 184 actual PAHs measurements (from Action 3.3) and daily samples of one year air quality data (from Actions 3.4, 4.1, 4.4, 4.5).
Starting from this data, SVM methods have been applied to build models able to forecast PAHs and BaP exposure.
It's important to highlight that the estimates produced by FARM bc have been used as SVM inputs. This choice has turned out to be crucial for the SVM performance.
The SVM models have been built and, then, they have been validated with a test set. They show much better performances than those achieved by FARM bc and FARM fc.
The same models have been applied to construct daily PAHs and BaP exposure maps.
Since actual measurements were not available for each day and for each pixel of the region, new performance indices have been introduced. Their role is to assess the maps not using a comparison between predicted and observed values. These indices show very promising values.
The same SVM models have been applied for the scenario analysis and new daily maps have been built considering new emission factors (from Action 7.1) and, consequently, new FARM outputs.
In conclusion, the SVM models seem to show excellent results in the reproduction of experimental data and in generalization, improving those achieved by FARM. Since FARM bc outputs have been used as SVM inputs, the SVM models seem also to apply a non-linear corrections to FARM bc estimates. Finally, the SVM models have produced very congruent PAHs and BaP exposure maps.

2 Introduction

The goal of Action 5.5 is to apply Machine Learning methods to build Polycyclic Aromatic Hydrocarbons (PAHs) exposure maps on the urban area of Rome. These maps can be considered an improvement of the results achieved in Action 4.5, which were obtained by means of a deterministic air dispersion model (Flexible Air quality Regional Model (FARM; Gariazzo et al., 2007)).
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 (Action 3.3): 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, hereinafter) and on BaP.

Machine Learning methods have been applied for estimating PAHs concentrations. There exist many Machine Learning algorithms. Among them, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) represent the two main approaches.

The usefulness of these methods lies in their capability to produce good predictions for new samples (never used during the training phase).

In literature, ANNs methods were used to forecast ozone and primary pollutants concentrations (Comrie, 1997).

Here, the results obtained by applying Support Vector Machines methods are given.

SVM methods (Vapnik, 1995) were rarely used for air dispersion modeling. They are a class of Supervised Machine Learning methods, developed to face with classification and regression problems.

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. SVMs can provide non-linear output functions.

There exist different SVM methods. In this work, the so called ε-SVR methods (Vapnik, 1998) have been applied, using LIBSVM software (Chang and Lin, 2011).

The dataset that has been used contained one year air quality data concerning the urban area of Rome. Data were collected between June 1st, 2011 and May 30th, 2012. The region of interest was an area 60 km × 60 km centered on the city of Rome, as shown in Figure 1 (the area within the blue square).

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.

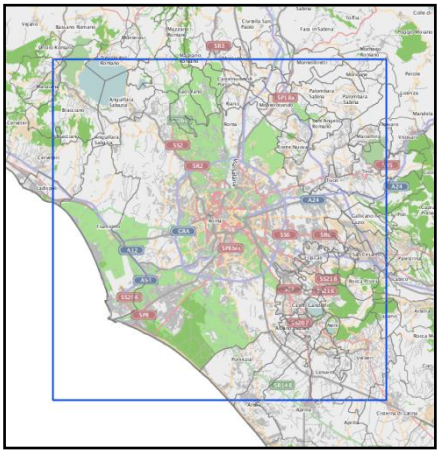

Figure 1. The region of interest around the urban area of Rome (within the blue square)

3 Materials and Methods

3.1 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 (from Action 3.4), pollutants emissions (from Action 4.1 and 4.4) and the outputs produced by base case FARM model (called FARM bc). 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).

Furthermore, also the dates (day and month) have been considered as input variables. Since the dates are periodic variables, they have been normalized within a circumference of one year and then decomposed into their sine e cosine components.

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, deriving from Action 3.3, 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.

So, the 184 samples used to construct the model have been built computing the mean values for each input variable in the related periods. In particular, daily means have been computed for the scalar variables, while wind direction and wind speed have been decomposed into their horizontal and vertical components.

3.2 Monitoring stations and input variables selection

Two different problems have been faced to achieve the goal of the action: 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 (16 out 26, corresponding to 124 samples) were located within the urban area, while some of the remaining 10 testing stations (corresponding to 60 samples) were located far away from the city. The stations location is shown in Figure 2 (some stations are overlapped because they belong to the same pixel): blue dots refer to training stations, red dots refer to testing stations. This choice about the stations location had the purpose to make test results rather robust. In fact, they are expected to provide a significant assessment of the model behavior when it is applied to samples representing different conditions from those of the training samples.

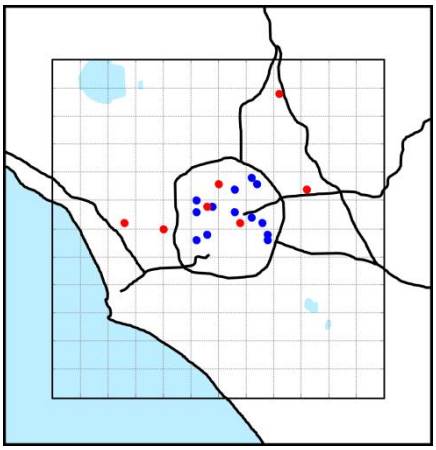

Figure 2. The location of the stations used for training (in blue) and for testing (in red)

Regarding the second issue, a preliminary analysis has been first conducted to study the domains of each variable and their distributions over the whole year and in the whole area.
These preliminary results have had an important role to characterize the possible presence of outliers.
The values assumed by the input variables in the whole dataset have been compared with those assumed by the same variables only in the training set. The comparison is summarized in Table 1.
While the minimum values are quite similar (except for temperature), the maximum training values tend to be much smaller than those computed over the whole dataset (especially for wind speed, precipitations and emissions). The global domains of some variables are much larger than those observed in the training phase. This observation will justify some choices about the normalization adopted, as explained below.

	TRAINING SET SAMPLES					ALL SAMPLES				
	min	max	avg	σ	σ/avg	min	max	avg	σ	σ/avg
Wind Speed (m/s)	0.44	7.21	2.89	1.14	0.39	0.29	16.27	3.36	1.58	0.47
P (mbar)	972.1	1030.2	1005.8	9.4	0.01	869.0	1032.3	998.4	19.4	0.02
Rain (mm)	0.00	40.00	1.78	5.16	2.90	0.00	91.81	1.22	4.82	3.96
RH (%)	39.82	98.36	78.83	13.68	0.17	24.21	100.00	79.09	14.25	0.18
T (°C)	1.02	25.57	12.55	6.42	0.51	-10.75	30.87	14.24	6.96	0.49
TCC (Arbitrary Units)	0.00	8.92	3.03	2.10	0.69	0.00	10.00	2.52	2.23	0.89
PAH emissions (g/s)	0.0001	0.0097	0.0032	0.0030	0.92	0.0000	0.0195	0.0006	0.0014	2.40
BaP emissions (g/s)	0.0000	0.0029	0.0010	0.0009	0.92	0.0000	0.0059	0.0002	0.0004	2.36
PAH FARM bc (ng/m^3)	0.17	32.78	6.81	7.07	1.04	0.00	38.28	2.23	3.18	1.43
BaP FARM bc (ng/m^3)	0.03	9.88	2.04	2.15	1.06	0.00	11.52	0.66	0.98	1.49

Table 1. Comparison of values assumed by the input variables in the training set and globally

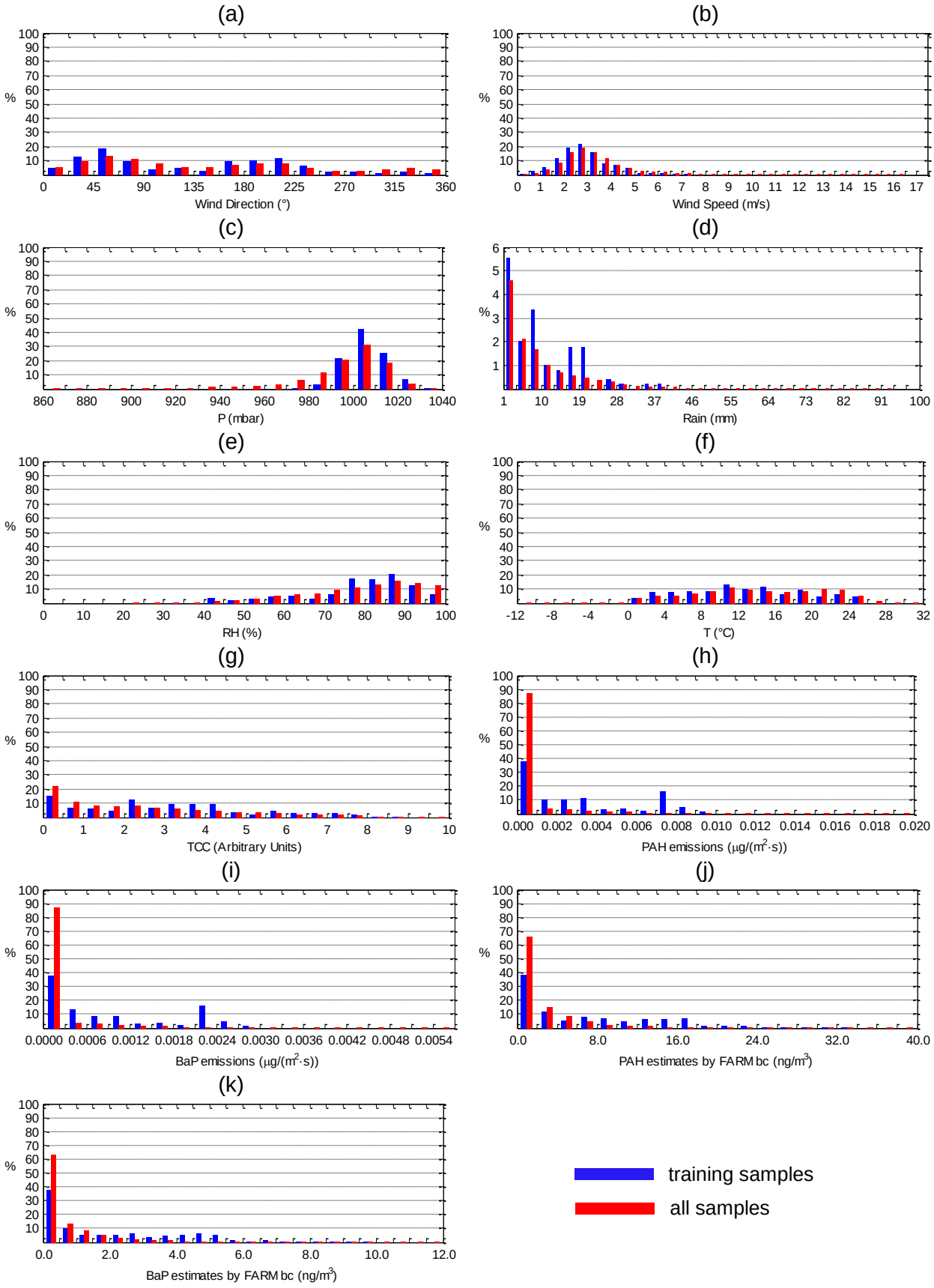

Figure 3. Histograms of wind direction (a), wind speed (b), pressure (c), precipitations (values higher than 1 mm; d), relative humidity (e), temperature (f), total cloud cover (g), PAH emissions (h), BaP emissions (i), PAH estimates by FARM bc (j), BAP estimates by FARM bc (k)

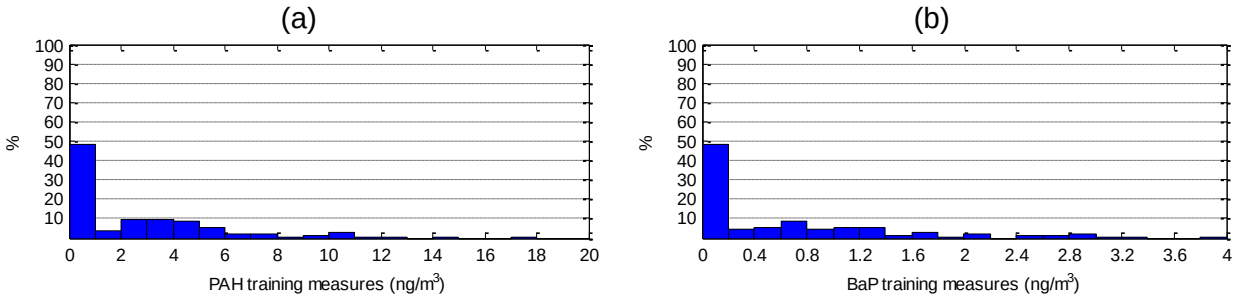

Figure 4. Histograms of PAH output target values (a) and BaP target values (b) of the training set

In order to conduct a deeper analysis of the input data, the histograms representing their probability density functions have been computed. They are shown in Figure 3a,b,c,d,e,f,g,h,i,j,k. Blue histograms refer to the training samples, red histograms refer to all the samples of the dataset. As for precipitations, the samples with values lower than 1 mm are 87.5% of the whole dataset and 82.7% of the training set. So, for a good readability of the graphs, only samples with precipitations values higher than 1 mm have been reported.
Also the histograms of the output target values over the training set have been computed, and they are shown in Figure 4a (PAH) and in Figure 4b (BaP).
Firstly, the histograms in Figure 3 and in Figure 4 show that some values distributions are skew (precipitations above all), with small intervals containing the great majority of data.
Secondly, emissions and FARM outputs distributions are quite different in the training set than in the whole dataset. In particular, modal values of all samples are about twice than those of training samples.
Generally, for Machine Learning methods, if the different states of the system are well represented in the training samples, then the model could generalize easier. In our case, this condition is mostly verified, except for the extreme values (which are low frequency). Indeed, as observed above, the domains of some variables over the training set are much smaller than those observed in the whole dataset. For this reason, each variable has been scaled in $[-1\,,1]$ using as bounds not the minimum and the maximum values assumed by the variable in the training set, but the minimum and maximum values assumed by the variable in the whole dataset.
In particular, for each input variable, the following formula has been applied:

$$x_s^{(i)} = l + \left[(u-l)\frac{\left(x^{(i)} - \min^{(i)}\right)}{\left(\max^{(i)} - \min^{(i)}\right)}\right] \qquad (1)$$

where $x_s^{(i)}$ indicates the scaled value, $x^{(i)}$ indicates the original value, $l = -1$, $u = 1$, and $max^{(i)}$ and $min^{(i)}$ indicate the maximum and minimum value assumed by the variable $x^{(i)}$.
Regarding the choice of the input variables, a feature selection process is often necessary when using Machine Learning methods to improve the model performance. For that aim, different SVMs have been built with different input variables, according to a specific scheme.
Many simulations have been conducted before obtaining the best parameters for building the final SVM. 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 2 for PAH and in Table 3 for BaP. The following performance indices have been considered: the Mean Absolute Error (MAE), the slope and the R^2 values of the trend line passing through the origin, and the intercept of the trend line not forced to pass through the origin.
From the results of Table 2 and Table 3, it is evident how much the choice of the input variables is crucial for the model performances.

	INPUT VARIABLES				TEST RESULTS			
	Date	Meteo	Emissions	PAH estimates 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. PAH test results using different subsets of input variables

	INPUT VARIABLES				TEST RESULTS			
	Date	Meteo	Emissions	BaP estimates 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

Table 3. BaP test results using different subsets of input variables

Taking into account the considerations about the different domains of the input variables between the training samples and all the samples, a further device has been adopted in order to improve the results and to facilitate the extrapolation: it consists in a logarithmic transformation applied, before scaling, to the target values and to the most unbalanced input variables: precipitations, emissions, FARM bc and wind speed.

In this way, the global and the training domains of these variables become more similar. For instance, for wind speed variable the minimum value becomes $min^{(ws)} = ln(0.29) = -1.24$ and the maximum value becomes $max^{(ws)} = ln(16.27) = 2.79$. Doing the same operations, the training domain changes into $[-0.82\,,1.98]$. As explained above, each transformed wind speed value has been scaled in $[-1\,,1]$ according to the relation (1) and using $min^{(ws)}$ and $max^{(ws)}$. Consequently, each training wind speed value is now in $[-0.79\,,0.60]$ and each testing wind speed value is in $[-1\,,1]$.

Moreover, the feature selection process has been carried on, trying also to identify the most important meteorological variables.

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
- PAH/BaP estimates by FARM bc

These variables have been used in the final SVM model. Test results are shown in the next paragraph.

4 Results

Results are divided into three parts: in the first one the SVMs test performances have been compared with those achieved by FARM, the second one concerns the maps obtained by using the SVMs and the third one describes how the SVMs have been applied for a scenario analysis.

4.1 Test results

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. In particular, FARM fc differs from FARM bc by the application of a correction factor (see Action 4.5). FARM bc outputs are used as SVM input variables and, consequently, the comparison should be made between SVM and FARM fc. However, FARM bc has been included to show the systematic deviation of the outputs produced by that model with respect to the observed values. This fact is important to point out how much the SVM improves these results. Moreover, the presence of this deterministic variable among the model inputs is fundamental to connect deterministic and statistical information.

Other indices have been considered, as well as those already introduced: Fractional Bias (FB), Normalized Mean Squared Error (NMSE), Correlation coefficient (r), Coefficient of Variation (CV), and Index of Agreement (IOA).

As shown in Table 4 and in Table 5, the SVM models provide much better results than the other two models on all indices, both for PAH and for BaP. In particular, while FARM bc tends to overestimate (slope = 2.0 and 2.23) and FARM fc tends to underestimate (slope = 0.78 and 0.71) the observed values, the SVM model avoids both of these deviations (slope = 0.96 and 0.94).

Further, the SVM produces higher correlation values (R^2 equals to 0.93 for PAH and 0.92 for BaP, against an R^2 average of 0.81 for FARM bc and 0.78 for FARM fc).

		MAE (ng/m^3)	R^2	slope	bias (ng/m^3)	FB	NMSE	r	CV	IOA
s_{17}	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 PAH concentrations

		MAE (ng/m^3)	R^2	slope	bias (ng/m^3)	FB	NMSE	r	CV	IOA
s_{18}	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

Table 5. Comparison between test results obtained by SVM, FARM bc and FARM fc to forecast BaP concentrations

4.2 Maps reconstruction

The models built for s_{17} and s_{18} 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 (Figure 2), 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 don't 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.

The reason why these indices have been chosen lies in the observation that negative concentrations are forbidden and that pollutants concentrations are generally higher in the urban area 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. This choice makes possible to verify if the model outputs are congruent with the expected concentrations. Plus, only samples of the urban area have

been used for building the SVM, so the outputs over sea and over lake can measure the model generalization capability.

To compute R_{U-NU} for PAH, only those days where the SVM output in the city is higher than 1 ng/m3 and the difference between the concentration over sea (or lake) and the concentration in the city is more than 0.2 ng/m3 have been counted.

As will be shown below, a strong linear relation between PAH and BaP concentrations holds. According to this relation, R_{U-NU} has been computed for BaP counting only those days where the SVM output in the city is higher than 0.2492 ng/m3 and the difference between the concentration over sea (or lake) and the concentration in the city is more than 0.04984 ng/m^3.

4.2.1 PAH exposure maps

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.

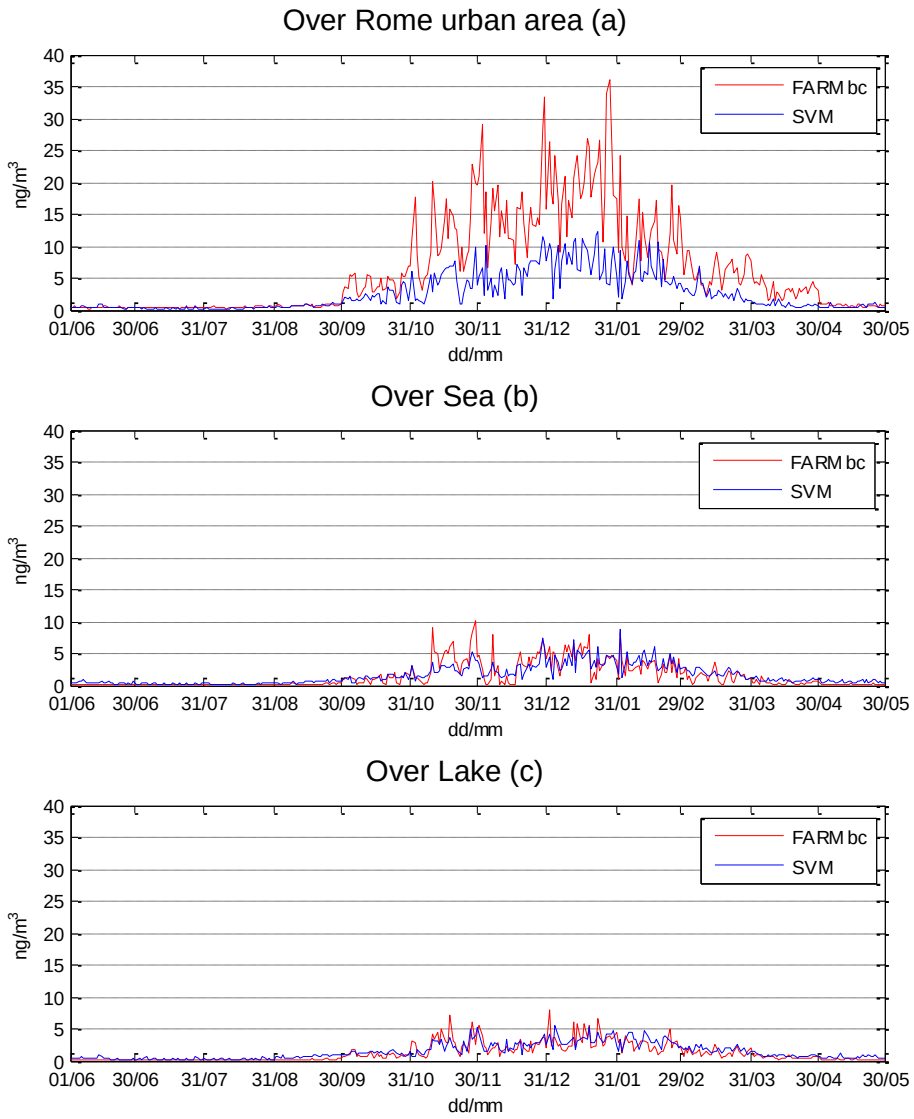

Figure 5. Comparison between PAH estimates produced by SVM and FARM bc model over Rome urban area (a), sea (b) and lake (c)

A comparison between the daily PAH estimates produced by FARM bc and by SVM at the three representative pixels (city, sea, lake) is shown in Figure 5a,b,c.
The analysis of these graphs points 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 6a, 6b and 6c.
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 6a, 6b and 6c, 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 (Table 4), 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).

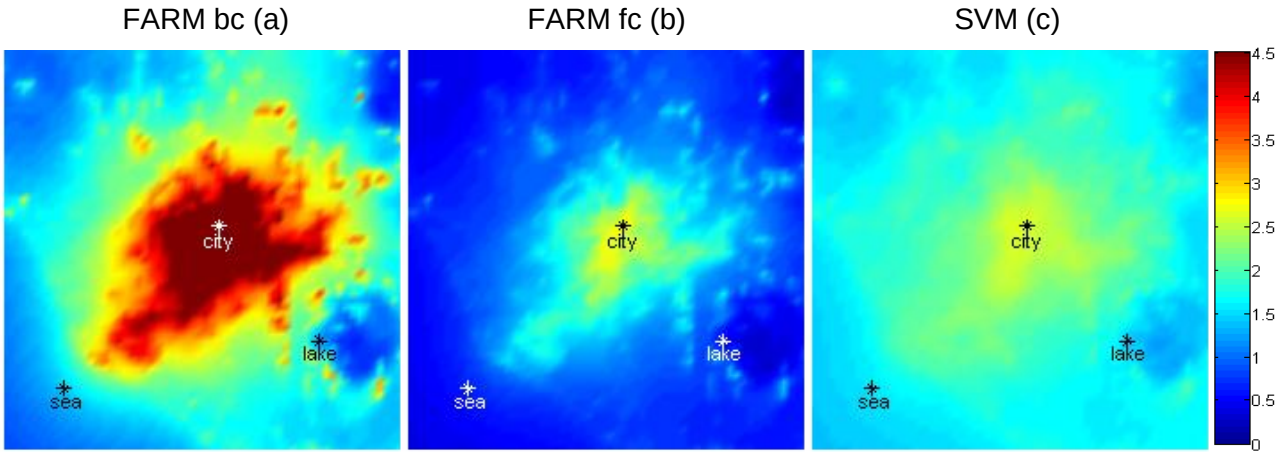

Figure 6. Annual mean PAH exposure maps by FARM bc (a), FARM fc (b) and SVM (c), in ng/m3

4.2.2 BaP exposure maps

As for BaP daily exposure maps, the SVM model built for s18 has been applied following the same steps just described for PAH.
BaP annual maps are shown in Figure 7a,b,c.

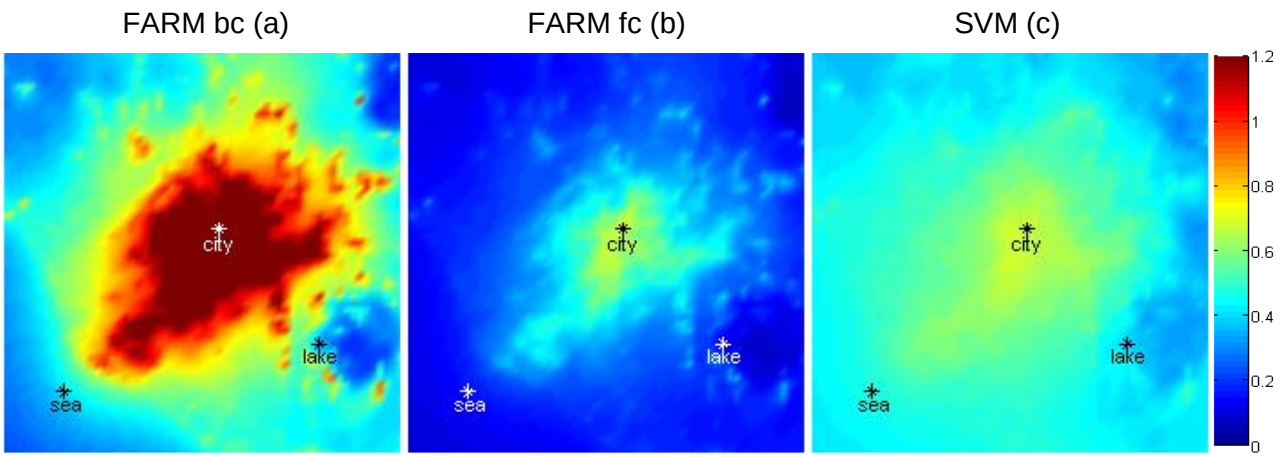

Figure 7. Annual mean BaP exposure maps by FARM bc (a), FARM fc (b) and SVM (c), in ng/m^3

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.
The performances get a little bit worse than those achieved for PAH.
As mentioned above, PAH and BaP concentrations seem to be linked very strongly by a linear relation. This relation has been calculated over the samples of the training set:

$$conc(BaP) = 0.2492 \cdot conc(PAH) \quad (2)$$

where PAH and BaP concentrations have been indicated by $conc(PAH)$ e $conc(BaP)$, respectively. The relation (2) shows a R^2 value equal to 0.97. Therefore, daily BaP exposure values could be obtained just multiplying each PAH estimate by 0.2492. By definition, R_{neg} and R_{U-NU} values would become identical to those for PAH. However, this new "model" had to be assessed by a test. Test results in Table 6 are obtained by multiplying the estimates produced in s_{17} by 0.2492.

		MAE (ng/m^3)	R^2	slope	bias (ng/m^3)	FB	NMSE	r	CV	IOA
s_{19}	BaP = 0.2492 · PAH by s_{17}	0.11	0.92	0.90	0.01	-0.09	0.19	0.96	0.40	0.98

Table 6. BaP test results obtained by multiplying PAH estimates (from s17) by 0.2492

Test results get slightly worse than those obtained in s_{18}, but are still very satisfactory and better than those produced by FARM bc and FARM fc.
In this way, daily BaP exposure maps having exactly the same shape and same index values of PAH maps have been obtained.

4.3 Application of the SVM model for the scenarios analysis

Another aim of Action 5.5 was to build forecasting maps using new hypothetical emissions factors (the so called scenarios analysis).
Two different future emissions scenarios (developed in Action 7.1) have been considered: the first one simulates the emissions for year 2020 with the current legislation, the second one simulates the emissions in the same year with the current legislation plus substitution of biomass with natural gas for house heating within Rome municipality.
All the sample have been simply updated with the new FARM values deriving from the new emissions factors. Then, the same SVMs built in the previous section have been applied to produce daily estimates.
Annual average PAH maps by SVM are shown in Figure 8a (the first scenario) and in Figure 8b (the second scenario). The mean values over the whole area are 1.86 ng/m^3 for the first scenario and 1.32 ng/m^3 for the second scenario.

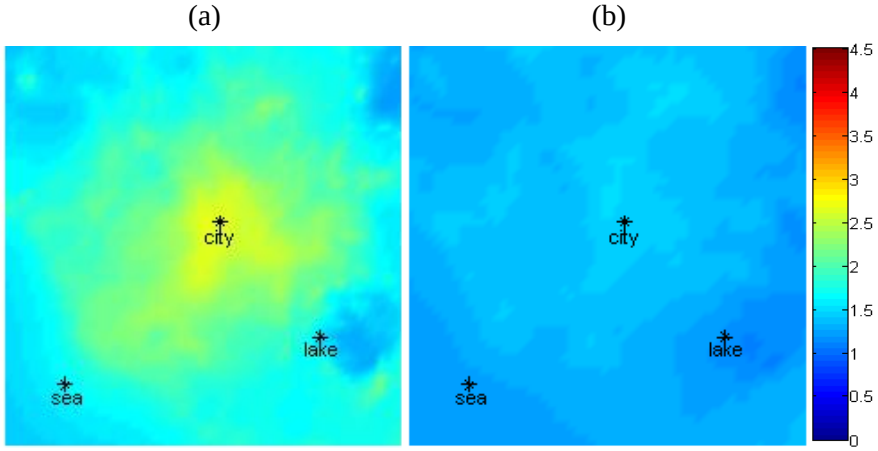

Figure 8. Annual mean PAH exposure maps by SVM for the first scenario (a) and for the second scenario (b), in ng/m^3.

As for BaP, the same proceeding has been carried out, applying both the model used for s_{18}, and the model used for s_{19}. Obviously, in the latter cases, the maps have exactly the same shape of those in Figure 8.
BaP maps obtained by applying the model used for s_{18} are shown in Figure 9. The mean values are 0.50 ng/m^3 for the first scenario and 0.36 ng/m^3 for the second scenario.

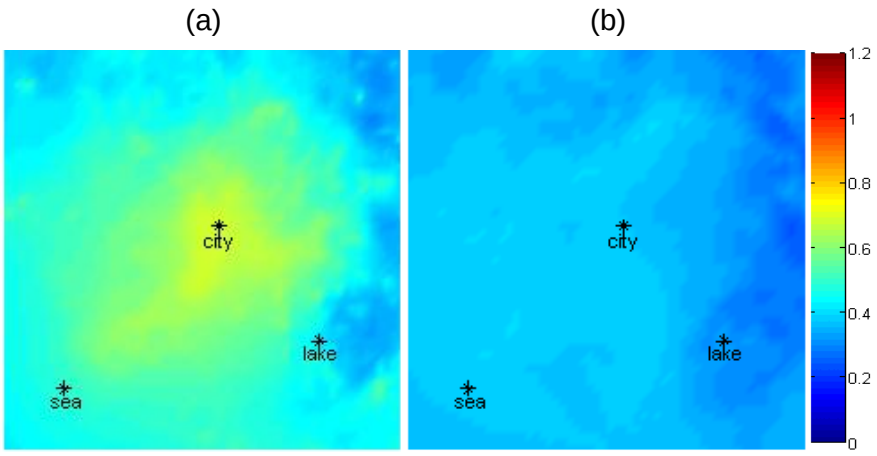

Figure 9. Annual mean BaP exposure maps by SVM for the first scenario (a) and for the second scenario (b), in ng/m^3.

5 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 were derived from Action 3.4, 4.1, 4.4, 4.5 and are the following: date (day and month), meteorological variables (wind direction, wind speed, pressure, precipitations, relative humidity, temperature and total cloud cover), emissions and the outputs produced by FARM bc (a deterministic air dispersion model). 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 for making the model able to generalize, the choice of the best model inputs and the reconstruction of daily maps.

With regard to the first problem, 16 stations (corresponding to 124 samples) have been selected in the urban area to train the model. The remaining 10 stations (corresponding to 60 samples) have been used just for testing. Most of the test station were located outside the city. This choice was justified by the fact that, since models had to be used to build daily maps, their spatial extrapolation capability needed to be effectively assessed.

With regard to the other issues, a logarithmic transformation has been applied to some input/output variables and, then, all of them have been scaled taking account of the overall values assumed.

Plus, a feature selection has been conducted for choosing the best input variables to use. The following input variables have been finally selected for the SVM model: the date (day and month), wind direction, wind speed, precipitations, total cloud cover and PAH/BaP estimates by FARM bc.

The SVMs have been trained and tested and the results have been compared with those obtained by FARM bc and FARM fc. The SVM models show the best values on each performance index both for PAH and for BaP.
In particular, while FARM bc and FARM fc show a tendency to overestimate and underestimate the actual measurements, respectively, the SVM models fit the data better and with a higher correlation.
It's important to underline that the SVMs use FARM bc outputs as input variables and produce results that improve those obtained by the FARM bc model itself. So, the SVMs seem to be able to apply a non-linear correction to the deterministic model.
The same SVMs have 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.
Since PAH and BaP measurements are strongly related linearly (as verified over the samples of the training set), good BaP estimates could also be obtained just multiplying PAH estimates by an opportune constant. Test results are still good and the daily maps have the same shape and the same index values of those concerning PAH.
Moreover, these models have been applied for building maps considering two different scenarios for year 2020 developed in Action 7.1. Simply updating the FARM bc values on the basis of the new emissions factors, new daily exposure maps have been built.
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.

6 References

Chang, C.-C, Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1-27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Comrie, R.S, 1997. Comparing neural network and regression models for ozone forecasting. Journal of the Air and Waste Management Association 47, 653-663.

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. Vol. 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.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.