Estimation of PAHs concentration fields in an urban area by means of Support Vector Machines

A. Pelliccioni [1], A. Cristofari [1], C. Silibello [2], M. Gherardi [1], A . Cecinato [3],
M. Lamberti [1]

[1] Inail Research, Via Fontana Candida 1, 00040 Monteporzio Catone (RM), Italy

[2] ARIANET Srl, via Gilino 9, 20128 Milan, Italy

[3] CNR- IIA, via Salaria km 29.3, 00015 Monterotondo Stazione, Italy

Abstract:
Epidemiological studies about health effects of air quality are often based on data inferred by monitoring stations, and the issue of constructing pollutants exposure maps is crucial for improving such studies. The Polycyclic Aromatic Hydrocarbons (PAHs) exposure in urban areas is the major goal of the EXPAH LIFE+ Project, so an integrated approach, based on measurements and modeling techniques, has been applied to simulate PAHs concentration in the metropolitan area of Rome in a period of one year (June 2011 - May 2012). Support Vector Machines (SVMs), which are a class of Machine Learning methods, have been applied. After a feature selection process, the SVM has been trained and tested with blind samples, showing very significant results. Then, the same SVM has been used for building PAHs daily exposure maps. Here, being not available the actual measurements, new indices have been considered for assessing the maps. All the outputs produced by the SVM have been also compared with those obtained by two applications of chemical transport models (FARM bc and FARM fc). The overall results suggest the applicability of SVM methods in estimating daily and annual PAHs exposure in urban areas.

Keywords: *SVM, PAH, PM2.5, Urban maps, health*

1. Introduction

Polycyclic Aromatic Hydrocarbons (PAHs) are pollutants linked to combustion processes. They can be considered relevant for health problems in high density urban areas.

In literature, some intelligent methods have been used to forecast ozone and primary pollutants (Comrie 1997) concentrations. However, Support Vector Machines (SVMs) methods have been rarely applied for air dispersion modeling.

PAHs exposure in urban areas is the major goal of the EXPAH LIFE+ Project (www.ispesl.it/expah). Many field campaigns has been conducted in the urban area of Rome. The project has more targets: one of them is to construct PAHs exposure maps, starting from the measurements and from the results obtained by an air dispersion model. For this aim, an integrated approach, based on measurements and SVMs methods, has been applied to reconstruct daily PAHs concentration maps. These maps may be used to estimate short and long term exposure.

SVMs are a class of supervised machine learning methods (SMLM), a branch of the artificial intelligence, developed by Vapnik in the ‘90s (Vapnik, 1995) to face with classification and regression problems (they have been later extended even to other problems). Given a system with a fixed number of inputs and outputs and a set of samples (the training set), the purpose of SMLM is to build a virtual machine able to *learn* the rules (supposed to be unknown) that link the outputs to the inputs of the system from the samples of the training set and to provide an analytic description of them. SVMs are able to find non-linear relations, by using kernel functions. The usefulness of these methods lies in their capability to produce good predictions once new samples (not used during the training phase) are available. In this work we have applied the so-called ε-SVR methods (Vapnik, 1998) using the LIBSVM software (Chang and Lin, 2011).

The dataset contains one year air quality data, which concern the urban area of Rome and some field six-eight days' campaigns distributed over all the seasons.

The period between June 1st, 2011 and May 30th, 2012 has been considered, and the region of interest is an area 60 km × 60 km centered on the city of Rome and divided into 3600 pixels (each one 1 km × 1 km), as shown in Figure 1a.

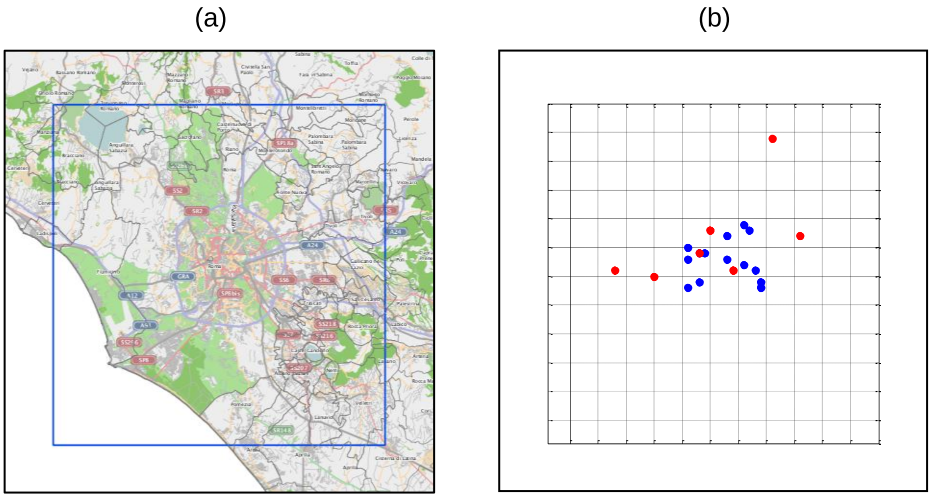

Figure 1. The region of interest around the urban area of Rome (a), and the location of the stations used for training (in blue) and stations used for testing (in red) (b).

Three kinds of variables have been initially considered: meteorological variables (wind direction, wind speed, pressure, precipitations, relative humidity, temperature and total cloud cover), pollutants emissions and the outputs of the base case FARM model (an air dispersion model based on a deterministic approach for pollutant modeling, see later).

For each variable, hourly values were available for each pixel and for each day of the period. Furthermore, also the dates (day and month) have been considered as input inputs.

PAHs concentrations measurements were available in different locations of the area and in different periods and they have been used as target values for the SVMs.

Almost all PAHs measurements referred to intervals of 2-10 days.

The goal of our work is to build an SVM to forecast daily pollutants concentrations on the basis of the values of the input variables.

For this reason, input variables have been first converted from hourly into period mean values, for being congruent with the relative measurements.

Two different problems have been faced. The first concerned which variables to use as model inputs. Generally, for machine learning methods, it is well known that using only a subset of the original variables could lead to better performances, because some of them could not contain any information. So, a feature selection process is often necessary for improving the model.

The second problem concerned the choice of monitoring stations to represent better the urban pollutant dispersion, that is which stations to use for training.

As regards the first issue, after some elaborations, the following variables have been chosen to be used as input variables: date, wind direction, wind speed, precipitations, total cloud cover and base case FARM outputs.

The use of the outputs of deterministic models as input variables for intelligent methods was first developed in Pelliccioni et al. (2001 and 2003) and the theoretical explanation can be found in Pelliccioni and Tirabassi (2006 and 2008).

As for the choice of the monitoring stations, all the stations chosen for the training (16 out 26) are located within the urban area, while some of the remaining 10 testing stations are located far away from the city, so they can provide a strong model generalization. The location of the stations is shown in Figure 1b (some stations are overlapped because they belong to the same pixel): blue dots refer to training stations, red dots refer to testing stations.

The SVM has been built following two steps: the training phase (where the machine has been effectively built with the samples of the training set), and a testing phase (where the model performances have been assessed with the samples of the test set).

The choice to select the training and the testing stations inside and outside the urban area, respectively, makes SVM results rather robust. Them, to build PAHs maps for every day of the year, the same SVM has been applied for each pixel of the area. In order to the evaluate the SVM performances, the comparison with the base case FARM (FARM bc) and the corrected FARM (FARM fc) are given in results.

2. Measurements and methods

- 2.1 PAHs characterization

Our concern was focused on carcinogenic PAHs compounds, namely 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). PAHs are known to accumulate on fine and ultra-fine fractions of airborne particulates; thus their measurements could be carried out as soon as gravimetric determinations of $PM_{2.5}$ were accomplished. $PM_{2.5}$ was collected daily from air by means of medium- or low-volume samplers, the formers adopted outside of schools, and the latters at all other indoor and outdoor locations: according to a preliminary test conducted in summer 2011, the two approaches were equivalent within 16% for all target PAHs. The sampling procedures refer to the modified NIOSH method 5515i. Fine particulate matter was collected on a PTFE filter by means of active sampling by achieving the selection of PM2.5 with a cyclone selector. The PM samples were gathered to form weekly pools, then PAHs were extracted with organic solvent, cleaned-up through column chromatography on alumina and determined through GC-MSD (SIM); the identification and quantitation of PAHs compounds were performed by means of isotopic PAHs as internal standards. The benzofluoranthene isomers were variously separated in all samples; their sum (BFs) was often taken in account for sake of homogeneity. The whole procedure is described extensively by Romagnoli et al. (2014); there, the whole of PAHs concentrations and behaviors is also discussed. The extended uncertainty was also derived, found to range from 8.8 % to 13.9 % for B(a)A and B(b)F+ B(j)F+ B(k)F, respectively. Besides the seven above mentioned congeners, the PAHs characterization was extended, whenever feasible, to chrysene (CH) and benzo[ghi]perylene (BPE), known to be mutagenic.

- 2.2 Farm air dispersion model

The concentration fields are produced by an an Air Quality Modelling System (AQMS) that is routinely used by the Lazio Region Environmental Protection Agency (ARPA Lazio) to produce air quality forecasts, to assess air quality and to evaluate the impact of different emission control strategies over the region and

Rome urban area. The AQMS is based on the Flexible Air quality Regional Model (FARM; Gariazzo et al., 2007) and includes subsystems used to:

- reconstruct flows and related turbulence parameters
- apportion data from the emission inventories to grid cells
- calculate the air quality indicators required by the EC directives

FARM employed the SAPRC-99 (Carter, 2000) chemical mechanism and the aero3 modal aerosol scheme from the CMAQ framework (Binkowski, 1999; Binkowski and Roselle, 2003). In Silibello (2012) and Silibello *et al.* (2013) are described respectively the upgraded version of FARM model, including PAHs chemistry, and its application to Rome metropolitan area (1 km resolution) to produce gridded pollution fields for epidemiological studies foreseen within the EXPAH Project. The comparison between observed and predicted PAHs concentrations has evidenced the capability of the modeling system to reconstruct PAHs concentration levels over Rome conurbation and to describe their seasonal variation. Nonetheless, an overestimation of observed concentrations is identified during colder periods when domestic heating is assumed to operate. This problem can be mainly attributed to the large uncertainty affecting PAHs emission estimates from the house heating sector due to the very large variation of emission factors depending on the fuel burned, and to the difficulty in distributing emissions within the urban texture. In order to provide more reliable data to the exposure model, estimated daily PAHs concentrations have been corrected by a factor f_m^c computed as the monthly average of the ratios between observed and predicted concentrations:

$$C_m^{Corrected}(t) = C_m^{Model}(t) \cdot f_m^c$$

$$f_m^c = \overline{\left(\frac{C_m^{Observation}(t)}{C_m^{Model}(t)} \right)}$$

Hereinafter, the base case FARM model and the corrected FARM model will be referred to as FARM bc and FARM fc, respectively.

- 2.3 Dataset characteristic

Meteorological and emission variables used for the SVM model are also the main input variables for the FARM model. In particular, meteorological data concern physical information on advection transport and turbulence diffusion.

FARM model takes also in account the chemical reaction between the involved compounds.

In order to build the maps, meteorological and emissions data are required for each point of the domain. For this purpose, meteorological field maps have been reconstructed by weather prediction RAMS model driven by ECMWF analyses (Silibello et al., 2013).

Emission maps have been mainly reconstructed starting from National Emission Inventory (hereafter ISPRA2005), characterized by province level resolution, and have been downscaled at municipal level resolution. Finally, in order to prepare model-ready emission inputs, the processing system Emission Manager (EMGR) has been used (Radice et al., 2012).

Traffic fluxes on Lazio Region road network have been estimated from AISCAT (http://www.aiscat.it/), ASTRAL (http://www.astralspa.it/) and ATAC (http://www.atac.roma.it/), and hourly PAHs emission maps have been computed using the COPERT IV methodology (Radice et al., 2012).

3. Results

Results are divided into two parts: the first one concerns the performances of the SVM model in the test phase and the second one concerns the maps obtained by applying the SVM.

SVM performances have been assessed by the comparison between the results obtained by FARM bc and FARM fc. Note that FARM bc outputs are also used as input variables of the SVM model and, consequently, the comparison should be done between SVM and FARM fc. However, FARM bc has been included in the comparison because it shows the systematic deviation of such model respect to the observed pollutant values.

As reported in Table 1, the SVM model provides much better results than the other two models. On particularly, while FARM bc tends to overestimate (slope = 2.0) and FARM fc model tends to underestimate (slope = 0.78), the SVM model avoids both of these distortions (slope = 0.96), with also a better correlation (R^2 = 0.93 against an average of $R^2 \approx$ 0.82).

Table 1. Comparison of PAHs test results between SVM, FARM bc and FARM fc.

	MAE ($\mu g/m^3$)	R^2	Slope	Interc.	FB	NMSE	r	CV	IOA
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

With regard to the daily exposure maps constructed by the SVM, note first that the model has been built (and tested) for reproducing not daily, but period concentrations, so a little forcing was necessary.

Moreover, taking account of the locations of the training samples in the whole area shown in Figure 1b, an extrapolation has been carried out, so a generalization capability has been required by the SVM.

Generally, for large area simulations, not all pixels are covered by measurements. For that reason, it is difficult to test the maps deriving by air dispersion results. In such a way, indirect performance indices may be introduced.

In our case, the following indices have been developed: $\mathrm{R_{neg}}$ measures the percentage of negative values, $\mathrm{R_{U-NU}}$ indicates the percentage of days where pollutants concentrations is lower in the urban than in a non-urban area.

The choice of these indices lies in the observation that negative concentrations are forbidden and that pollutants concentrations are higher in the urban than in a non-urban area.

To define R_{U-NU}, three pixels have been fixed: one on the sea (South-West of the area), one on the lake (South-East of the area) and one in the center of Rome. Then the daily model outputs have been compared. Only those days where the output in the city is major than 1 and the difference between the concentration over the sea (or the lake) and the concentration in the city is more than 0.2 have been counted.

The following values have been obtained: R_{neg} = 0, R_{U-NU} = 3.29% comparing the city with the sea, and R_{U-NU} = 2.74% comparing the city with the lake.

A comparison between the daily estimates produced by FARM bc and by SVM at these representitave pixels is reported in Figure 2a,b,c, while a comparison between the daily estimates produced by the SVM model in the city and over the sea is reported in Figure 3.

The analysis of these figure evidences the congruent behavior of the SVM model and its generalization capability. It produces estimates generally lower over the lake and over the sea than in the city, even though only urban samples have been used for training.

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

The resulting maps are shown in Figure 4a, 4b and 4c.

All the maps produce higher values in the urban area than outside. However, while the maps obatined 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 4a, 4b and 4c, the mean values over all the area are 2.23 μg/m^3, 0.98 μg/m^3 and 1.78 μg/m^3 for FARM bc, FARM fc and SVM, respectively.

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

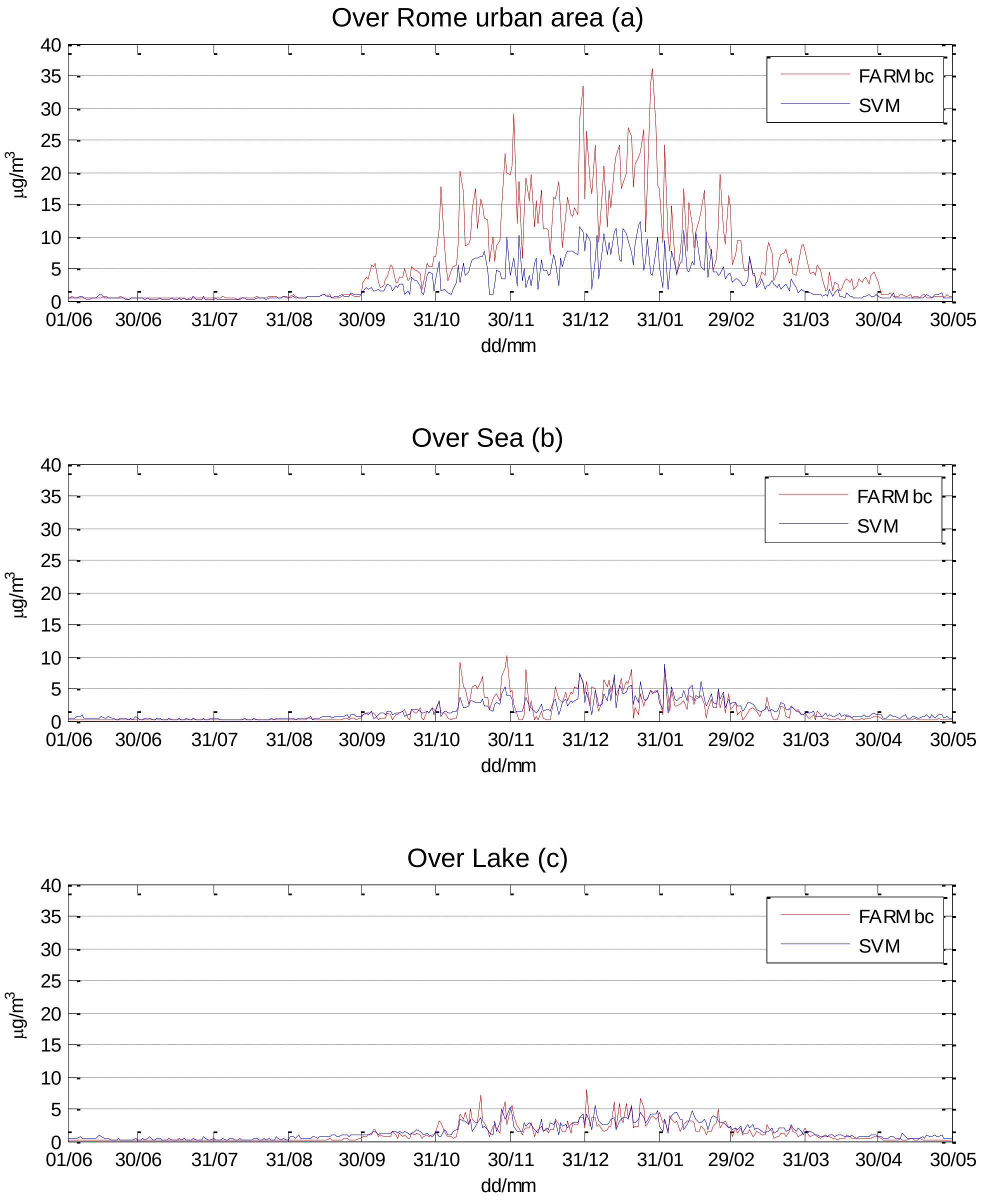

Figure 2. Comparison between outputs produced by SVM and FARM bc model over Rome urban area (a), sea (b) and lake (c).

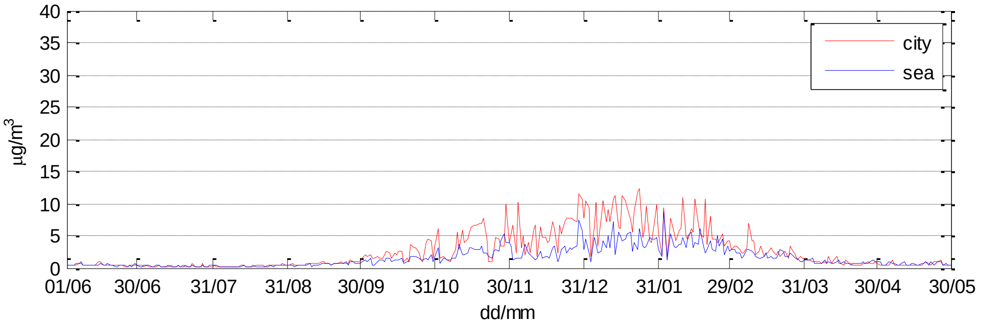

Figure 3. Comparison between the daily estimates produced by the SVM model in the city and over the sea.

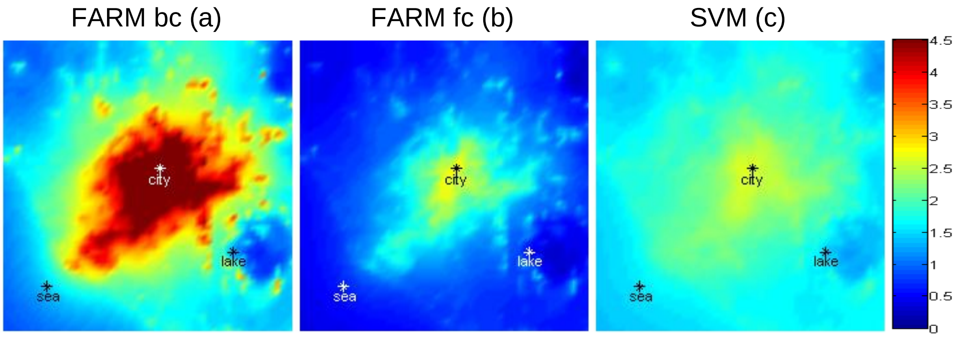

Figure 4. Mean PAHs maps by FARM bc (a), FARM fc (b) and SVM (c), in μg/m^3.

4. Conclusions

SVMs, which are a class of Machine Learning models, have been used to forecast PAHs concentrations on an area 60 km × 60 km centered on the city of Rome over one year. In environmental field, SVMs are rarely used for constructing maps.

It was necessary to face with many problems to obtain a good spatial reproduction of pollutants concentrations. In particular, the main issues dealt with the optimization of model inputs and with the reconstruction of daily maps.

With regard to the first problem, a feature selection has been conducted for choosing the best input model variables, that are: the date (day and month), wind direction, wind speed, precipitations, total cloud cover and the outputs produced by the FARM bc model (a deterministic air dispersion model).

The SVM has been trained and tested using some measurements available in different points of the area and in different periods of the year.

SVM test results have been compared with those obtained by the FARM bc model and the FARM fc model (which differs from FARM bc by the application of a correction factor). The SVM shows the best values for each criterion considered.

In particular, FARM bc model and FARM fc model show a tendency to overestimate and underestimate, respectively. The SVM model fits the data better and with a higher correlation.

It's important to underline that the SVM uses FARM bc outputs as input variables and produces results that improve those obtained by the FARM bc model itself.

So, the SVM seems to be able to apply a non-linear correction to the deterministic model.

The same SVM, trained for reproducing period concentrations, has been used to build daily exposure maps.

Generally, for constructing maps, it’s impossible to know the actual measurements in each point and in each day.

For this reason, it was necessary to introduce new indices for assessing the maps.

Since measurements can’t assume negative values and since pollutants concentrations are expected to be higher in urban areas than in non-urban areas, the new indices check whether these conditions are respected.

The indices measure the percentage of negative values and the percentage of days where pollutants concentrations is lower in the urban than in a non-urban area, respectively. The performances show values close to zero for the first one, and between 2.7% and 3.3% for the second one.

Finally, the overall results seem to confirm the capability of the SVM to reconstruct PAHs spatial concentration.

REFERENCES

Binkowski, F.S, Roselle, S.J., 2003. Models-3 community multiscale air quality (CMAQ) model aerosol component 1. Model description. Journal of Geophysical Research, 108 (D6), 4183.

Carter, W.P.L., 2000. Documentation of the SAPRC-99 Chemical Mechanism for VOC Reactivity Assessment. Final Report to California Air Resources Board, Contract 92-329 and 95-308, SAPRC, University of California, Riverside,CA.

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.

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.

Radice, P., Smith, P., Costa, M.P., D'Allura, A., Pozzi, C., Nanni, A., Finardi, S., 2012. 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

Pelliccioni, A., Tirabassi, T., 2001. Application of a Neural Net Filter to Improve the Performances of an air Pollution Model. Seventh International Conference on “Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes” in Belgirate (28th-31st May).179-182.

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

Silibello, C., 2012. EXPAH - ACTION 4.5: Integration of PAHs atmospheric processes within FARM model. ARIANET, R2012.1, Milano, Italy. http://www.ispesl.it/expah/documenti/R2012-01_ARIANET_EXPAH_A4.5.pdf

Silibello, C., D'Allura, A., Finardi, S., Radice, P., 2013. EXPAH - Technical report on FARM model capability to simulate PM2.5 and PAHs in the base case – Action 4.5. http://www.ispesl.it/expah/documenti/R2013-6_ARIANET_EXPAH_A4.5_final.pdf.

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

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