4.3 CALCULATION OF PAH MAPS USING SVMS FOR AN URBAN AREA

A. Pelliccioni [1], A. Cristofari [1], S.E. Haupt [2]
[1] Inail Research, Rome Italy
[2]National Center for Atmospheric Research, Boulder, CO USA

1. INTRODUCTION

Polycyclic Aromatic Hydrocarbons (PAHs) are pollutants linked to combustion processes. They are considered potential causes of health problems in high density urban areas. Quantifying PAHs exposure in urban areas is the major goal of the EXPAH LIFE+ Project (www.ispesl.it/expah). Many field campaigns have been conducted in the urban area of Rome.

A major target of this project was to construct PAHs exposure maps starting from the measurements and from the results obtained by an air dispersion model. To meet this goal, an integrated approach based on measurements and machine learning 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 to address classification and regression problems (they have been later extended to other problems).

SVMs are capable of finding non-linear relations by using kernel functions. The usefulness of these methods lies in their capability to produce good predictions once new samples are available. In the work we have applied the so-called ε-SVR methods using the LIBSVM software. In the literature, some intelligent methods have been previously used to forecast ozone and primary pollutants concentrations. However, Support Vector Machines (SVMs) methods have been rarely applied for air dispersion modeling.

We consider a dataset that contains one year of air quality data for the urban area of Rome over the time period June 1st, 2011 and May 30th, 2012. In addition, we employ field data from six-eight days’ campaigns distributed over the seasons. 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).

Three kinds of variables have been initially considered: meteorological variables (wind direction, wind speed, pressure, precipitations, relative humidity, temperature and total cloud cover), pollutant emissions and the outputs of the base case FARM model (an air dispersion model based on a deterministic approach for pollutant modeling, described below). For each variable, hourly values were available for each pixel and for each day of the period. In addition, the dates (day and month) are included as inputs.

PAH concentration measurements were available in different locations of the area for different periods and they have been used as target values for the SVM. Almost all PAH measurements referred to intervals of 2-10 days. The initial effort was to build an SVM to forecast daily pollutant concentrations on the basis of the values of the input variables.

Two problems have been addressed. The first concerned which variables to use as model inputs. Generally, for machine learning methods, it is typical that a subset of the original variables will lead to the best performance, because some of them may not contain applicable information. Thus, a feature selection process is necessary for optimizing the model. The second problem concerns the choice of monitoring stations to best represent the urban pollutant dispersion; that is, which monitoring stations to use for training.

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

As for the choice of the monitoring stations, all 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 overlap 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 performance has 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. Then, to build PAH maps for every day of the year, the same SVM has been applied for each pixel of the area. In order to evaluate the SVM performance, comparison with the base case FARM (FARM bc) and the corrected FARM (FARM fc) are provided in results.

2. MEASUREMENTS AND METHODS

The PAH 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

Corresponding author address: Armando Pelliccioni, Inail Research, Via Fontana Candida 1, 00040 Monteporzio Catone (RM), Italy; email: a.pelliccioni@inail.it

region and Rome urban area. The AQMS is based on the Flexible Air quality Regional Model (FARM) 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 employs the SAPRC-99 chemical mechanism and the aerosol scheme from the CMAQ framework. The comparison between observed and predicted PAH concentrations has evidenced the capability of the modeling system to reconstruct PAH concentration levels over Rome conurbation and to describe their seasonal variation. 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 PAH 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. Hereinafter, the base case FARM model and the corrected FARM model will be referred to as FARM bc and FARM fc, respectively.

The meteorological and emission variables used for the SVM model are also the main input variables for the FARM model. In order to build the maps, meteorological and emission data are required for each point of the domain. For this purpose, meteorological field maps have been reconstructed by the by the numerical weather prediction model, RAMS, driven by ECMWF analyses.

Emission maps have been mainly reconstructed starting from National Emission Inventory (ISPRA2005), characterized by province level resolution, and have been downscaled at municipal level resolution.

3. RESULTS

Results are divided into two parts: the first portion describes the performance of the SVM model in the test phase and the second one deals with the maps obtained by applying the SVM.

SVM performance has been assessed by comparing the results obtained by FARM bc with those of 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 as a baseline comparison because it shows the systematic deviation of such a model with respect to the observed pollutant values.

As reported in Table 1, the SVM model provides much better results than the other two models. In particular, 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 $R^2 \approx 0.82$).

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 average concentrations, so a little forcing was necessary to make an appropriate comparison.

Generally, for large area simulations, not all pixels are covered by measurements. For that reason, it is difficult to test the maps derived by air dispersion results. Thus, indirect performance indices should be introduced.

In our case, the following indices have been developed: $\mathrm{R_{neg}}$ measures the percentage of negative values, $\mathrm{R_{U\text{-}NU}}$ indicates the percentage of days where the pollutant concentrations are 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 pollutant 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 greater than 1 and the difference between the concentration over the sea (or the lake) and the concentration in the city is greater 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 the SVM at these representitave pixels is reported in Figure 1a,b,c. Analysis of these figures provides evidence of 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 by computing the the yearly estimates average for each pixel. The resulting maps are shown in Figures 2a, 2b and 2c. 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 difference.

The mean values over all maps 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 further confirmation 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).

4. CONCLUSIONS

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

It was necessary to address several problems to obtain a good spatial reproduction of pollutant 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 issue, a feature selection was 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 over 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, the FARM bc model and FARM fc model show a tendency to overestimate and underestimate concentrations, respectively. The SVM model fits the data better than either with a higher correlation. It is important to underline the fact that the SVM uses FARM bc outputs as input variables and produces results that improve upon those obtained by the FARM bc model itself.

Thus, 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 at each point and for 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 pollutant concentrations are 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 PAH's spatial concentration.

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

	MAE (ng/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

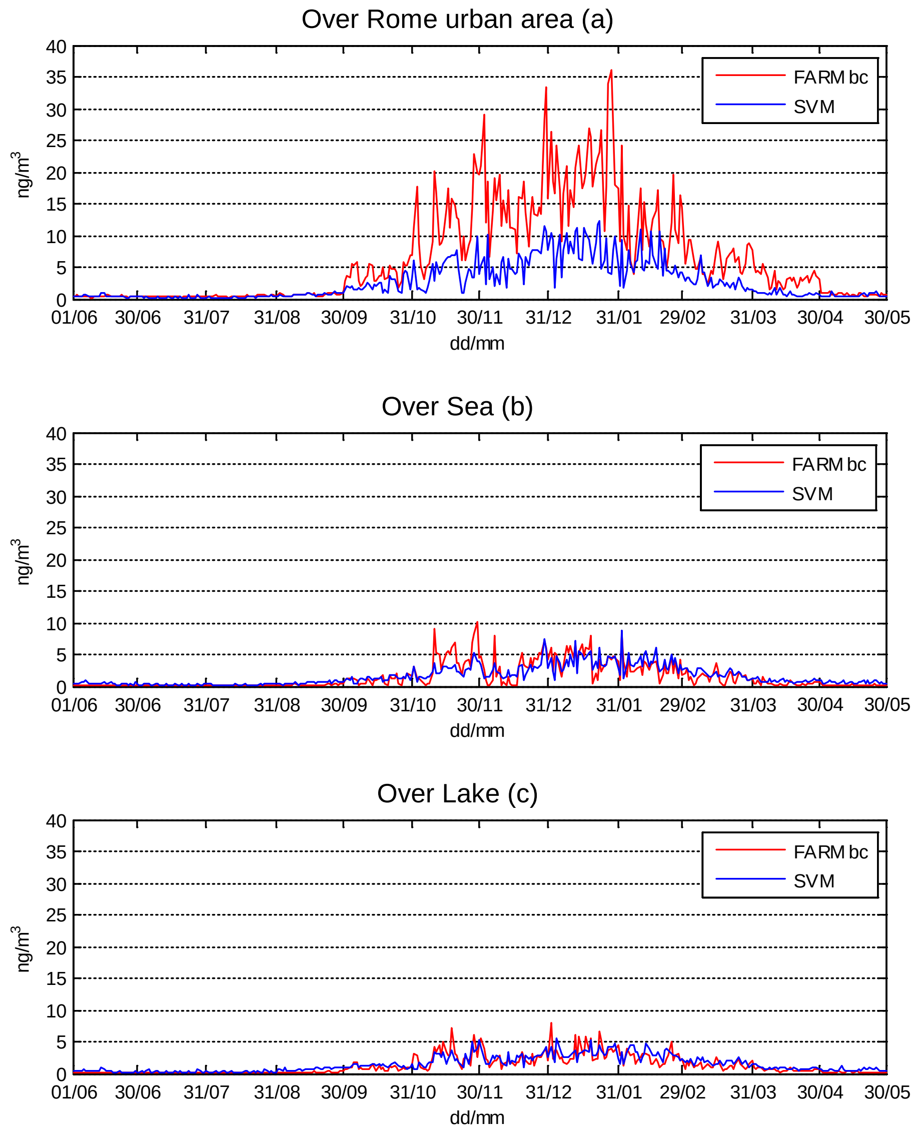

Figure 1. Comparison between outputs produced by SVM and FARM bc model over Rome urban area (a), sea (b) and lake (c).

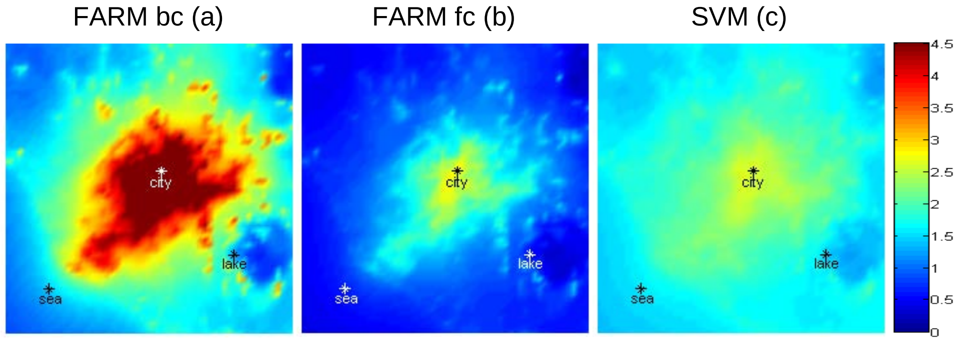

Figure 2. Mean PAHs maps by FARM bc (a), FARM fc (b) and SVM (c), in ng/m^3.