Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peaks

Wani W. Tamas, Gilles Notton, Christophe Paoli, Marie-Laure Nivet, Cyril Voyant

Aerosol and Air Quality Research, 16: 405–416, 2016

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2015.03.0193

Hybridization of Air Quality Forecasting Models Using Machine Learning and Clustering: An Original Approach to Detect Pollutant Peaks

Wani Tamas[1], Gilles Notton[1], Christophe Paoli[1,2*], Marie-Laure Nivet[1], Cyril Voyant[1,3]

[1] *University of Corsica - Pasquale Paoli, UMR CNRS 6134 SPE, 20250 Corte, France*
[2] *Galatasaray University, Department of Computer Engineering, TR-34357 Istanbul, Turkey*
[3] *CHD Castelluccio, radiophysics unit, BP85 20177 Ajaccio, France*

ABSTRACT

This paper presents an original approach combining Artificial Neural Networks (ANNs) and clustering in order to detect pollutant peaks. We developed air quality forecasting models using machine learning methods applied to hourly concentrations of ozone (O_3), nitrogen dioxide (NO_2) and particulate matter (PM_{10}) 24 hours ahead. MultiLayer Perceptron (MLP) was used alone, then hybridized successively with hierarchical clustering and with a combination of self-organizing map and k-means clustering. Clustering methods were used to subdivide the dataset, and then an MLP was trained on each subset. Two urban sites of Corsica Island in the western Mediterranean Sea were investigated. These models showed a good global precision (Index of Agreement reaching 0.87 for O_3, 0.80 for NO_2 and 0.74 for PM_{10}). Considering it is particularly important than forecasting model used on an operational basis correctly predict pollution peaks, a sensitivity analysis was performed using Receiver Operating Characteristic curves (ROC curves). It allowed to evaluate the behaviour and the robustness of the models for high concentration situations. The results show that for PM_{10} and O_3, hybrid models made of a combination of clustering and MLP outperform classical MLP most of the time for high concentration prediction. An operational tool has been built with the models presented in this paper, and is used for air quality forecasting in Corsica.

Keywords: Air quality forecasting; ROC curve; Multilayer perceptron; Clustering.

INTRODUCTION

Air quality is a major concern, both for public health and environment preservation. In France, Air Quality Monitoring Agreed Associations (AQMAA) are in charge of main ground level pollutant monitoring.

Air quality forecasting is an important part of AQMAA's missions, allowing the anticipation of pollution peak formation. Different air quality forecasting techniques have been developed in recent years (Zhang *et al.*, 2012a) and two families of models can be distinguished. On the one hand, deterministic models operate by modelizing all the physicochemical mechanisms responsible of the evolution of air quality. On the other hand, statistical models must learn the underlying relationships between the different variables related to air quality to make their predictions. The first family of models, frequently called Chemical Transport Models (CTM), use similar principles to Numerical Weather Prediction (NWP) models. They can offer predictions with a good spatial definition and help scientists understand and validate mechanisms of atmospheric pollution. A state of the art in CTM research is presented in Zhang *et al.* (2012b). Their construction demands a considerable research effort, large computing resources and an available emission inventory. The second family, statistical models, need a large amount of data and various preprocessing operations to be operational. The precision of statistical models can outperform CTM's, but they produce only local predictions. One of the stakes in CTM research is the hybridization with statistical models, which are used to post-process CTM outputs in order to take into account available observations to improve forecasting. We can cite the work on PM_{10} forecasting made with the CTM CHIMERE (its aerosol module presented in Bessagnet *et al.*, 2004) and linear regression models by Konovalov *et al.* (2009).

Those last decades, various types of statistical models have been applied to air quality forecasting. Among them, Artificial Neural Networks (ANNs) has been particularly used in research. ANNs show good results when used as time series forecasting models (Zhang, 2012). Their applications in atmospheric sciences were reviewed in the late 90's by Gardner and Dorling (1998), through the model of the MultiLayer Perceptron (MLP), a type of ANN known for its universal approximator ability (Hornik *et al.*, 1989).

* Corresponding author.
Tel.: 90-212-227-44-80; Fax: 90-212-259-20-85
E-mail address: cpaoli@univ-corse.fr

Various studies using MLP can be found in the literature (Kolehmainen *et al.*, 2000; Perez and Reyes, 2002; Kukkonen *et al.*, 2003; Dutot *et al.*, 2007). Thus, our preliminary work focused on ozone (O_3) concentration forecasting one hour ahead (Paoli *et al.*, 2011). Then a work on *h + 24* hourly O_3 concentration forecasting with MLP was initiated (Tamas *et al.*, 2014) and showed a good reliability (IA reaching 0.88), in the same order of magnitude than previous studies found in the literature as Coman *et al.* (2008). As clustering of training data appeared to offer an improvement of ANN predictive models abilities (Davis and Bouldin, 1979; Lu *et al.*, 2006; Poggi and Portier, 2011), we decided to apply such a method to improve our capicity to detect pollutant peacks.

In this work, we first built a *h + 24* MLP model for each of the three major pollutants in two urban sites: ozone (O_3), nitrogen oxides (NO_2), and particulate matter (PM_{10}, particles below 10 µm in diameter). Predictors were pollution and weather measurements and outputs from the NWP model AROME from Météo-France, the French national meteorological service. Those models were trained and evaluated on independent test sets, showing a good precision. After the first results, we focused on the pollution peak prediction ability and no longer on the global performances.

In Corsica, a French island in western Mediterranean Sea where this study take place (see Fig. 1), pollution peaks mainly occur when external sources bring pollutants over the island. PM_{10} high level episodes are often linked with Saharan dust events in addition to local sources. Other typical high PM_{10} events are due to stable meteorological conditions like thermal inversion causing the stagnation of locally emitted particles. High O_3 levels may also be linked to transport events. Old air masses can come from the south of mainland France or from the highly industrialized Po valley in the north of Italy.

We investigated two clustering methods to separate the data into several subsets, in order to isolate the different weather patterns likely to favor pollution peaks. The first method was based on SOM and k-means clustering and the second on hierarchical clustering.

After the clustering step, an MLP was trained on each cluster to obtain an MLP specialized on each weather pattern. Each trained MLP was evaluated on the part of test set corresponding to its cluster. The hybrid model was made of all those MLPs, each being used when the data corresponded to its cluster. The behaviour and the robustness of the resulting models was studied with a focus on high concentration situations and compared to the classical MLP using Receiver Operating Characteristic curves (ROC curves, see Fawcett, 2006). Those curves allow the comparison of threshold overrun detection rate for every threshold.

The next section will present the data used in this study. Then we will introduce the MLP based forecasting model before focusing on our clustering approaches. The global results of all the models will be shown, followed by an evaluation with ROC curves focused on peak forecasting abilities. A conclusion with associated perspectives will be discussed.

MATERIALS AND METHODS

Air Quality in Corsica

Corsica Island is located in the Western Mediterranean Sea, in the south of France, west of Italy and north of Sardinia Island. This mountainous island (average and maximum altitudes of 568 meters and 2710 meters), with a small industrialization, has a population of 310000 inhabitants for an area of 8680 km^2. The air constituents (four regulated pollutants) are monitored by the approved association Qualitair Corse, using a network of 9 monitoring stations, mainly deployed around the two largest cities, Ajaccio and Bastia. We built our models to forecast the concentrations measured in the two urban stations (Canetto and Giraud), so that the predictions are representative of air state around the urban population.

The main pollutant emissions in the island are due to energy production industry (mainly fuel), traffic (road, sea and air), domestic heating, waste incineration and agriculture.

Fig. 1. Corsica Island in Mediterranean Sea.

In France, four pollutants (NO_2, PM_{10}, O_3 and SO_2) are regulated and controlled. Two concentration thresholds exist and trigger reactions of the administration if exceeded (Table 1). The first threshold is an information threshold, leading AQMAAs to communicate toward authorities and population on atmospheric state when exceeded or forecasted. The second one is an alert threshold, its exceedance forces the authorities to take actions in order to reduce the emissions. In Corsica, O_3 and PM_{10} are the two pollutants causing pollution episodes, NO_2 levels being less problematic. SO_2 levels are particularly low, with average concentrations around 2 μg m^{-3} for an information threshold of 300 μg.m^{-3}. For that reason, SO_2 is not looked at in this study.

General statistics calculated for O_3, PM_{10} and NO_2 are shown in Table 2. As PM_{10} concentrations thresholds are calculated on a daily basis, values of 24-hour sliding average are also displayed for this pollutant.

Located on the seaside, Ajaccio and Bastia are both subject to coastal breezes. Bastia is located at the foot of the mountain range of Serra and is subject to valley and mountain breezes. These phenomenons influence local pollutants dynamics.

Meteorological data were provided by Météo-France. The outputs from AROME NWP model (Seity *et al.*, 2011) are used. This model has a 0.025° resolution, allowing a good representation of convective processes. For each station, the closest point of the AROME meshing output was used: for Canetto station (41.925N, 8.736E) we used the point with geographic coordinates: 41.925N, 8.725E and for Giraud station (42.698N, 9.446E) the point with coordinates: 42.7N, 9.45E. The meteorological parameters used in our model and produced by AROME are: Temperature (T), Atmospheric Pressure (AP), U and V wind components, Relative Humidity (RH), Precipitations (P), Nebulosity (N), Geopotential (G), Short-Wave and Long-Wave net Radiation (SWR and LWR). Those variables are given for various altitude levels. Within the atmospheric boundary layer, thermal inversion can appear and provoke pollutants stagnation. A variable describing the thermal inversion is thus a valuable input for our models. We calculated the thermal Inversion Layer Thickness (ILT) from temperature outputs available at various levels (2 m, 20 m, 50 m, 100 m, 250 m, 500 m, 750 m, 1000 m, 1250 m and 1500 m). If the temperature gradient is positive between two levels, the corresponding altitude difference is added to the ILT value. Boundary Layer Height (BLH) was available and is a key parameter for qualifying the ground-level atmospheric state. However, it was excluded of the dataset because it was one year shorter than the other variables. During a preliminary test, we found that models performed better without BLH in the dataset but with one more year to train the models.

Pollutant time series to forecast were thus that of O_3, NO_2 and PM_{10} measured in Canetto and Giraud. Input data were both endogenous and exogenous time series, exogenous being measures of other pollutants, meteorological measures and output prediction from AROME. All time series consisted of hourly averages.

Forecasting with Multilayer Perceptron

An MLP is a feedforward ANN with at least one hidden layer. MLP is known to be able to modelize any smooth function (Hornik *et al.*, 1989). Typically, MLP has one or two hidden layers and an output layer with as many neurons as the number of desired outputs. The predictors correspond to the input data of the MLP, provided to the input layer. The neurons of the first layer process the data and their output becomes the input of next layer's neurons. Each input x_i is multiplied by a specific weight w_i. The sum of all weighted inputs is added to a specific bias b and this sum becomes the argument of the activation function of the neuron that produces the output y_i (See Fig. 2). The weights and biases are the parameters of the MLP, and must be set during a supervised learning phase by a training algorithm. Levenberg-Marquardt Algorithm (LMA) was used to train our networks. During the learning phase, a training dataset was used, with input data and target data,

Table 1. Information and alert concentration thresholds in 2014 for O_3, PM_{10}, NO_2 and SO_2.

Pollutant	Average	Information threshold (μg m^{-3})	Alert threshold (μg m^{-3})
O_3	Hourly	180	240
PM_{10}	Daily	50	80
NO_2	Hourly	200	400
SO_2	Hourly	300	500

Table 2. Statistics on hourly O_3, NO_2 and PM_{10} concentrations and on 24-hour sliding average PM_{10} concentrations.

Station	Pollutant	Mean (μg m^{-3})	STD (μg m^{-3})	Min (μg m^{-3})	Max (μg m^{-3})	Missing values (%)
Canetto (Ajaccio)	O_3	58.25	29.20	0	166	1.58
	PM_{10}	24.23	11.28	0	165	5.93
	*PM_{10}**	*24.23*	*8.38*	*6.21*	*82.33*	*6.39*
	NO_2	21.48	16.29	0	128	1.77
Giraud (Bastia)	O_3	75.89	23.44	1	164	3.38
	PM_{10}	22.31	10.73	0	149	8.52
	*PM_{10}**	*22.31*	*8.04*	*4.50*	*78.92*	*8.53*
	NO_2	15.09	12.69	0	130	4.30

* 24-hour sliding average.

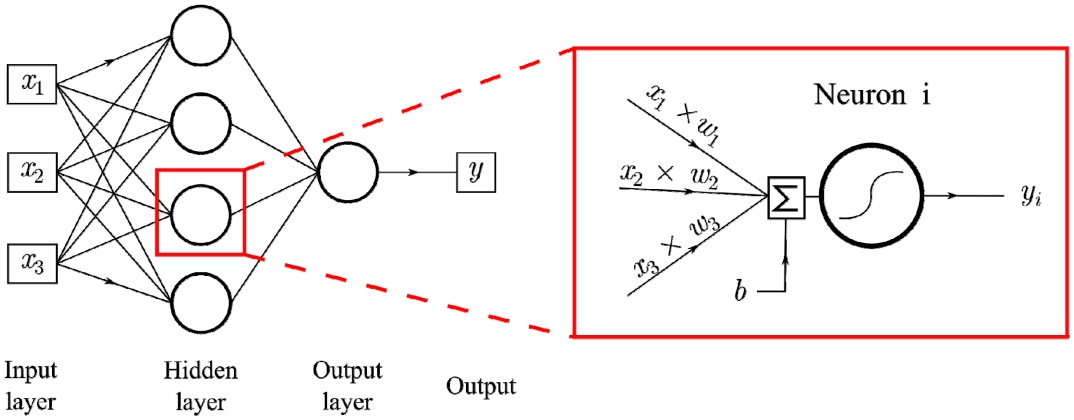

Fig. 2. Schematic view of a multilayer perceptron with focus on neuron's parameters.

and the LMA iteratively adapted all MLP's parameters in view to reduce the MSE between target data and MLP's output (Marquardt, 1963). Those parameters properly set allow the MLP to modelize underlying relationships between the predictors and the predictand. Our target (the predictand) was the concentration time series to forecast, shifted forward 24 hours. Thus, the MLP was trained to a *h* + *24* predictive model. Our network had one hidden layer of ten neurons with sigmoid activation function. Output neuron's activation function was a linear function.

The early stopping method was used to avoid overlearning, which leads to an over specialization of the network on training data and poor generalization abilities. Input data were divided into three subsets: the train set, the validation set and the test set. Three years of data were dedicated to the train set, one year for the validation set and one year for the test set. The MLP was trained using the training set, and at each iteration of the LMA, the MSE was calculated using the validation set. When validation MSE stopped decreasing for six consecutive iterations, the learning phase was stopped. The test set was then used to evaluate model's performances. Saving a full year for the test set allowed having all seasons equally represented.

Clustering Models

As observed in a previous study (Tamas *et al.*, 2014), the main difficulty encountered with MLP was to obtain good performances for high concentration episodes. Some authors use boosting, that is to say increase the frequency of such episodes in the training set (Kukkonen *et al.*, 2003; Paschalidou *et al.*, 2010), but it can lead to overfitting. Another way to improve the precision for high concentration is to build a forecasting model with the time series of maximum daily values of the pollutant as target (Corani, 2005; Lu *et al.*, 2006; Perez, 2012); but working with daily values does not bring information on air quality evolution during the day, which is useful for operational use.

We chose to investigate the precision gain for high concentration by specializing an MLP into each weather class, those classes being determined by a clustering process. It allows to separate, by unsupervised learning, different typical weather episodes, during which relationships between predictors and predictand may be different. Hybrid models consisted of the successive assignation to a cluster followed by the prevision using the proper MLP. Two clustering approaches were investigated: a hierarchical clustering and a SOM mapping followed by a k-means clustering. The clustering was applied on the dataset comprising outputs from AROME and pollution observations. This dataset was different from these used as MLP's input though the variables was the same because of a different lag choice. The clustering dataset used the *h* + *24* prediction from AROME (observed time series were not lagged).

Hierarchical clustering is an iterative method gathering data points in clusters using a distance metric representing their dissimilarity. In an agglomerative hierarchical clustering process, each data point is first assimilated to a group. A distance metric must be defined, that represents the dissimilarity between groups. A criterion is chosen that uses the metric to select at each iteration the two groups to be gathered into a new group. The process continues until a chosen number of final groups is reached. We used the euclidean distance as metric with Ward criterion for the clustering (Ward, 1963). Ward criterion is based on intergroup inertia. At each iteration of the algorithm, two groups are gathered together in order to maximize intergroup inertia *I*:

$$I = \frac{1}{m}\sum_{i=1}^{k} m_i \left\| g - g_i \right\|^2 \quad (1)$$

with *m* the sample size and g it's centroid, and *k* clusters indexed by *i*, with a size m_i and a centroid g_i.

The second clustering approach was based on SOM followed by a k-means algorithm. SOM were used as a first dimensional reduction step leading to a faster clustering. SOM are artificial neural networks with one group of interconnected neurons. Each neuron has *n* parameters, assimilated to a position in an n dimensions space, and is connected to its neighbours. The SOM is trained on the sample of dimension *n*, and for each data point, the neuron with the closest position moves closer to the data point, dragging with him neighbour neurons. At the end of the

training, neurons cover the space occupied by the sample, and each data point may be classified as belonging to the group of the closest neuron.

After this first step, we applied the k-means algorithm to trained SOM's neurons positions. The clusters are defined by their centroid g_i, and each data point belongs to the cluster having the nearest centroid. The number of desired groups k is given, and the algorithm seeks the k centroids g_i in order to minimize the intragroup sum of squares:

$$D = \sum_{i=1}^{k} \sum_{x \in Q_i} \|x - g_i\|^2 \quad (2)$$

with x the data points, Q_i the i^{est} cluster and g_i its centroid.

It is difficult to choose the appropriate number of clusters and to evaluate the quality of a clustering process dedicated to subdivide a training set for an MLP. Increasing the number of clusters may reduce the intragroup distance, but it would also decrease cluster sizes, lowering MLP's training potential. We chose to experiment every number of clusters between two and five for each experiment, retaining the model with the best forecasting abilities with the test set.

Choice of Predictors

Data used as predictors were pollutant variables and meteorological variables, from September 2009 to June 2014. The pollutant variables were PM_{10}, NO_2, O_3 concentration timeseries, along with meteorological measures (HR, P, T) and meteorological previsions from AROME (AP at sea level, T and RH at 2m,U and V at 10m, G at 800 hPa, ILT between ground level and 1500m, P, N, SWR and LWR). Those variables represent different phenomena related with the pollutant concentration. Three different prevision horizons were used (*h + 15*, *h + 20* and *h + 24*) to provide information about the weather evolution before the prediction horizon of the MLP.

Various preprocessing operations precede the learning phase. First, data points with missing values were deleted. Then, input and target time series were normalized (centred and reduced). This ensures that variables significance is not affected by their range or their unit. Annual and daily profiles of variables were computed, and time series presenting a periodic compound (input or target) were transformed into stationary time series, by subtracting their daily and annual mean values. Thus, a Principal Component Analysis (PCA) was performed on the entire input dataset, and the Principal Components (PCs) were used as input for the MLP. It is known to improve the precision of the predictive model (Sousa *et al.*, 2007). The PCA also allow the reduction of the amount of input variables. PCs are hierarchied by decrising corresponding eigenvalue. It is possible to discard some of the PCs with the less eigenvalue. Only the PCs with the higher eigenvalue, which accounts for the majority of the variability in the data, were selected as input of the MLP.

We used the same variables for the clustering dataset, but the horizon of predicted variables was *h + 24* only. Those variables at this horizon represent weather conditions when pollutant concentration must be forecasted. Some weather conditions are known to be responsible of high pollution event. Particularly, wind component and geopotential at 800 hPa bring informations about transport events. ILT represent the amplitude of the thermal inversion responsible of pollutant stagnation. Solar radiation is closely linked to O_3 photochemical formation. MLP trained with a cluster representative of a pattern linked with high concentration should have his detection rate increased.

As the unsupervised dimension reduction process of PCA is closely related to the unsupervised learning of k-means clustering (Ding and He, 2004), using PCs of the data helps the k-means clustering algorithm to find appropriate centroids. PCs of normalised data were used to perform the clustering.

RESULTS

Models Global Performances

Three 24 hour ahead forecasting models were built for each pollutant (PM_{10}, O_3 and NO_2) in Canetto station (Ajaccio) and Giraud station (Bastia). Each model consisted of an MLP of one hidden layer with ten neurons. MLPs that were trained with the full training set (without clustering) are referred as fMLP. The hybrid models formed of various MLPs each trained with data subsets after a hierarchical clustering are referred to as hMLP. The hybrid models formed of MLPs trained with data subsets after the SOM/k-means clustering are referred to as kMLP (see Fig. 3). For the two clustering processes, data was clustered before the separation between the tree datasets (train, validation and test sets). It means that the test set of each MLP was the part of the global test set that belonged to its cluster. To evaluate models with clustering, each of their MLP's outputs was merged to form the global test output.

Data used for learning and evaluation of MLPs covered years from 2009 to 2014. A full year was dedicated to the test set, another full year was used as validation set and the rest formed the train set. The initialisation of weights and biases by Nguyen-Widrow algorithm (Nguyen and Widrow, 1990) comprising a random component, all the trained models are different and their precision varies. Each model was therefore independently trained and evaluated six times. The variation of precision is more important for models with clustering that have smaller training sets. The models with N clusters is constituted of N MLPs, each one being the best of the six trained. To evaluate the complete model, each data point of the test set is assigned to its cluster, and the corresponding MLP produce the test output for this point. Models were trained to fit the 24 hour ahead shifted concentration time series. Their precision was evaluated with error indexes: Root Mean Square Error (RMSE), normalized Root Mean Square Error (nRMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE), Index of Agreement (IA) and correlation coefficient (R), reported in Table 3. We have:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2} \quad (3)$$

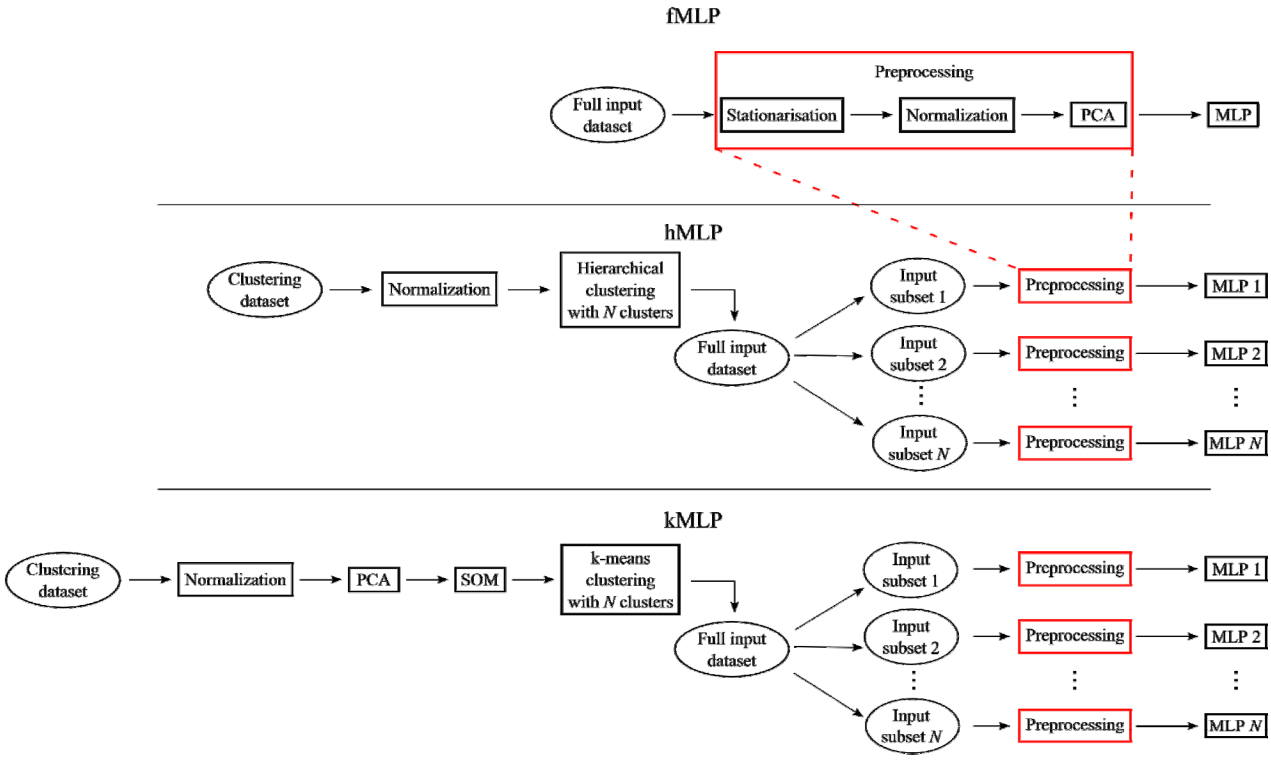

Fig. 3. Construction of forecasting models without clustering (fMLP), with hierarchical clustering (hMLP) and with SOM/k-means clustering (kMLP). The number of clusters is N.

$$nRMSE = \frac{RMSE}{\overline{o}} \tag{4}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|p_i - o_i\right| \tag{5}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}\left(p_i - o_i\right) \tag{6}$$

$$IA = 1 - \frac{\sum_{i=1}^{n}\left(p_i - o_i\right)^2}{\sum_{i=1}^{n}\left(\left|p_i - \overline{o}\right| + \left|o_i - \overline{o}\right|\right)^2} \tag{7}$$

$$R = \frac{\sum_{i=1}^{n}\left(p_i - \overline{p}\right)\left(o_i - \overline{o}\right)}{\sqrt{\sum_{i=1}^{n}\left(p_i - \overline{p}\right)^2 \sum_{i=1}^{n}\left(o_i - \overline{o}\right)^2}} \tag{8}$$

with n the sample size and i the sample index, o the observed variable and p the variable prediction.

IA (Willmott, 1982), ranging from 0 for the worst model to 1 for a perfect model, was preferred to the other indexes to be used as a criterion when a selection was necessary, for it can detect additive and proportional differences between the observed and predicted time series. Moreover, RMSE and MAE are more dependent to time series dynamics, and R is not adapted to rate models.

Those indexes are widely used but yet are insufficient to a proper evaluation, as they do not give informations about high concentration detection. This evaluation will be completed with proper tools (i.e., ROC curves) presented in next section.

Results of Table 3 show the indexes of performance of models. Error indexes of PM_{10} models are also displayed for 24-hour sliding average to correspond to the daily basis used in France for PM_{10} thresholds and alerts. The results of fMLP are quite good, reaching IA of 0.87 for O_3 at Canetto station. The precision of fMLP in Giraud station, where the dynamic of this pollutant is more complex, is lower (IA = 0.837) but RMSE and MAE are smaller, due to the little range of O_3 concentration variations in Bastia caused by nocturnal high concentrations. nRMSE is also lower than in Ajaccio, average concentrations in Bastia being higher for the same reason. We note that precision for O_3 forecasting is equivalent to that which was obtained in previous work (Tamas *et al.*, 2014) with models and data comparable to fMLP. However, we had used a heavier feature selection process, using mutual information, replaced here by the easier use of PCA on a large dataset. Precision of PM_{10} models is similar for the two stations, with an IA around 0.73 for hourly concentrations. NO_2 have a better IA in Canetto than in Giraud, reaching 0.80.

Evaluation of Threshold Overrun Detection with ROC Curves

While fMLP display better performance indexes for hourly prediction, we will see that hMLP and kMLP have

Table 3. Test set precision of *h + 24* forecasting models for O_3, PM_{10} and NO_2 hourly concentrations. 24-hour sliding averages values are also displayed for PM_{10} models. IA and R are dimensionless.

Station	Pollutant	RMSE ($\mu g\ m^{-3}$)	nRMSE ($\mu g\ m^{-3}$)	MAE ($\mu g\ m^{-3}$)	MBE ($\mu g\ m^{-3}$)	IA	R
Canetto	PM_{10}	7.40	37.02	5.77	1.63	0.736	0.599
	*PM_{10}**	*4.41*	*22.04*	*3.56*	*1.64*	*0.826*	*0.752*
	O_3	18.65	31.57	14.69	–0.38	0.870	0.766
	NO_2	12.10	55.84	8.58	0.60	0.805	0.669
Giraud	PM_{10}	7.49	38.89	5.70	1.14	0.728	0.606
	*PM_{10}**	*4.73*	*24.52*	*3.72*	*1.05*	*0.834*	*0.759*
	O_3	15.84	20.53	12.23	–1.65	0.837	0.743
	NO_2	10.90	71.97	7.30	0.49	0.735	0.592

MLP without clustering (fMLP).

Station	Pollutant	Retained num. of clusters	RMSE ($\mu g\ m^{-3}$)	nRMSE ($\mu g\ m^{-3}$)	MAE ($\mu g\ m^{-3}$)	MBE ($\mu g\ m^{-3}$)	IA	R
Canetto	PM_{10}	3	8.72	43.51	6.79	1.54	0.649	0.437
	*PM_{10}**		*4.53*	*22.59*	*3.65*	*1.52*	*0.826*	*0.728*
	O_3	4	24.22	38.69	19.03	0.03	0.779	0.613
	NO_2	4	17.57	85.26	12.82	0.75	0.591	0.326
Giraud	PM_{10}	2	7.91	41.36	6.02	1.01	0.696	0.545
	*PM_{10}**		*4.74*	*24.77*	*3.71*	*1.00*	*0.831*	*0.754*
	O_3	2	16.74	21.48	12.93	–1.57	0.826	0.711
	NO_2	2	12.34	82.23	8.52	0.68	0.655	0.464

MLP trained after hierachical clustering (hMLP). Models with two to five clusters were trained and tested, and we retained the number of cluster leading to the best Index of Agreement.

Station	Pollutant	Retained num. of clusters	RMSE ($\mu g\ m^{-3}$)	nRMSE ($\mu g\ m^{-3}$)	MAE ($\mu g\ m^{-3}$)	MBE ($\mu g\ m^{-3}$)	IA	R
Canetto	PM_{10}	4	8.28	41.32	6.48	1.94	0.681	0.499
	*PM_{10}**		*4.66*	*23.26*	*3.83*	*1.92*	*0.807*	*0.726*
	O_3	2	22.56	36.04	17.53	1.12	0.808	0.662
	NO_2	3	15.42	74.81	10.95	0.46	0.672	0.452
Giraud	PM_{10}	3	8.10	42.31	6.10	0.89	0.662	0.509
	*PM_{10}**		*5.00*	*26.09*	*3.90*	*0880*	*0.797*	*0.720*
	O_3	2	16.99	21.81	13.17	–1.23	0.818	0.699
	NO_2	2	11.96	79.79	8.21	0.615	0.642	0.474

MLP trained after k-means clustering (kMLP). Models with two to five clusters were trained and tested, and we retained the number of cluster leading to the best Index of Agreement.
* 24-hour sliding average.

better peak detection abilities, that is the interest of using clustering. We need a more appropriated evaluation method to measure these abilities and robustness. In Canetto, hMLP still reached the same IA than fMLP for PM_{10} 24-h sliding average.

We will now consider forecasting models for their threshold overrun detection. The models can then be seen as binary classifiers, with possible outputs indicating "exeedance" or "no exeedance". It is possible to evaluate such models for a given threshold with contingency matrices. But it is useful to know model's behaviour for various thresholds, and for that reason we drew the ROC curves (presented by Fawcett (2006)) of our models. The True Positive Rate (TPR) and the False Positive Rate (FPR) of a model can be calculated for each threshold. The TPR is the rate of correctly predicted exceedances for total observed exceedances (between 1 for a perfect model and 0 for a totally defective model), and the FPR is the rate of predicted exceedances that were not observed (false alarm) for all situation when concentrations stay below the threshold (FPR is between 0 for a perfect model and 1 for a totally defective model). The ROC curve is drawn by plotting, for each threshold, the FPR on abscissa axis and the TPR on ordinate axis.

ROC curves of the best fMLP, hMLP and kMLP for PM_{10} and O_3 are shown in Fig. 4 for Canetto station and in Fig. 5 for Giraud station. PM_{10} results are shown for 24-hour sliding average and O_3 results for hourly averages, to correspond to official thresholds averages.

Remember that each MLP was trained six times, the best kMLP and hMLP are the combination of the best of those six MLPs for each cluster. For kMLP and hMLP, we built models with two to five clusters and show in Figs. 4 and 5 the best results, indicating the number of clusters.

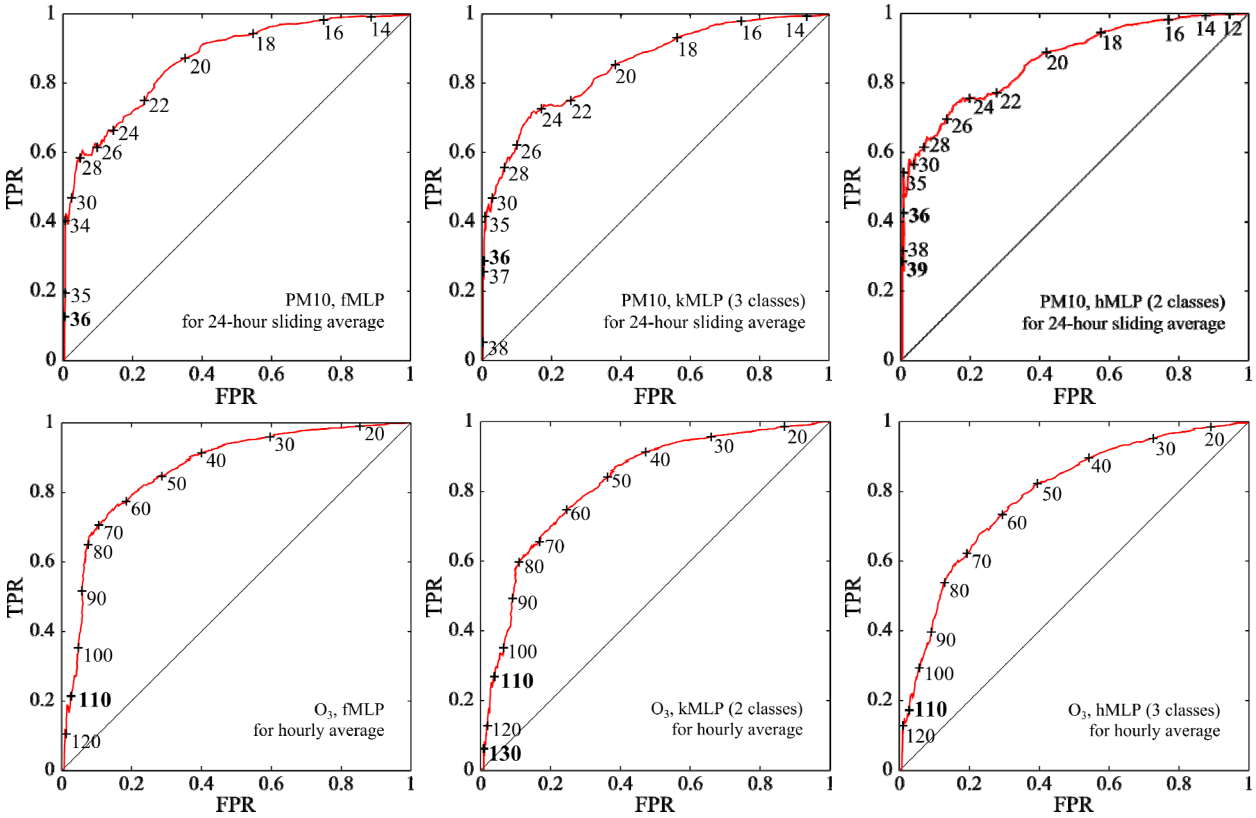

Fig. 4. ROC curves of 24 hour ahead forecasting models for PM_{10} and O_3 concentrations in Canetto station (Ajaccio). 24-hour sliding average is used for PM_{10} concentrations.

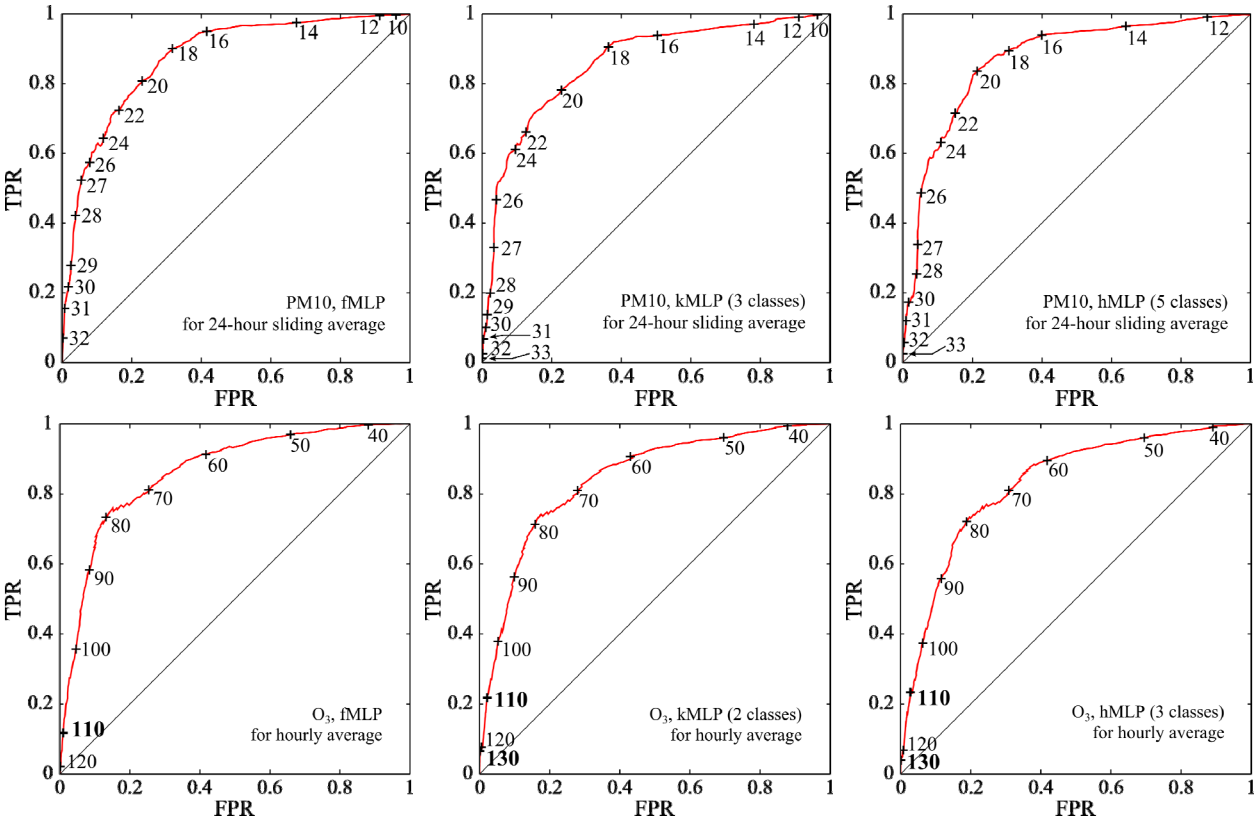

Fig. 5. ROC curves of 24 hour ahead forecasting models for PM_{10} and O_3 concentrations in Giraud station (Bastia). 24-hour sliding average is used for PM_{10} concentrations.

Numbers on the curves indicate the thresholds in μg m^{-3}. When models with clustering have better TPR for high thresholds than fMLP, those values are highlighted. Global performances in Table 3 showing better results for fMLP can also be noted on ROC curves in term of area under the curve. But the interest of the clustering is not to improve global performances, but to increase forecast precision for high concentration events. On a ROC curve, we will not focus on the area under the curve, but we will pay attention to the bottom-left part of the graph, evaluating the models for thresholds corresponding to the highest concentration values of the dataset. In this part of the curve, FPR is logically low as models will hardly make a false alarm with a high threshold. The objective is to increase the TPR for high thresholds to have a model that does not miss high concentration events, particularly when those events are rare and have little presence in the data of training set, which is the case here. We underlined some threshold values to emphasise the cases where models with clustering perform better for high levels.

For PM_{10} and O_3, models with clustering can show better detection for high concentration events than fMLP. In Fig. 4 showing Canetto results, the two clusters PM_{10} hMLP have slightly better behaviour than fMLP for thresholds higher than 30 μg m^{-3}, as O_3 kMLP for thresholds higher than 100 μg m^{-3}. At Giraud station, O_3 model is also improved by clustering, hMLP having the best detection rates. For those pollutant, PM_{10} models in Bastia is the only case where hybrid models do not improve high level detection.

Scatter plots of fMLP and of the best hybrid model (considering ROC curves) are shown in Fig. 6 for Canetto station and in Fig. 7 for Giraud station. PM_{10} scatter plots display some line patterns, due to the 24-sliding average used which smooths concentration evolution. On those curves, we can see the detection improvement for high concentration values. The global precision degradation due to the subdivision of training sets with hybrid models is also visible, with more scattered points for medium and low concentration values. This scattering echoes the lowering of IA observed with hybrid models. For medium concentrations fMLP appear to be more appropriated. This seems to be the consequence of the subdivision of dataset, reducing the size of training set for the MLPs of hybrid models. But for high concentrations, an improvement of peak detection is observed with hybrid model. Both should be used on operational basis, with focus on hybrid model for the peak forecasting.

NO_2 results are not plotted, being systematicly poorer with hybrid models than with fMLP. This lack of performances may come from the data used for the clustering, mainly meteorological and including variable outputs over the atmospheric boundary layer. Such a dataset brings both local and mesoscale informations, that are needed to explain

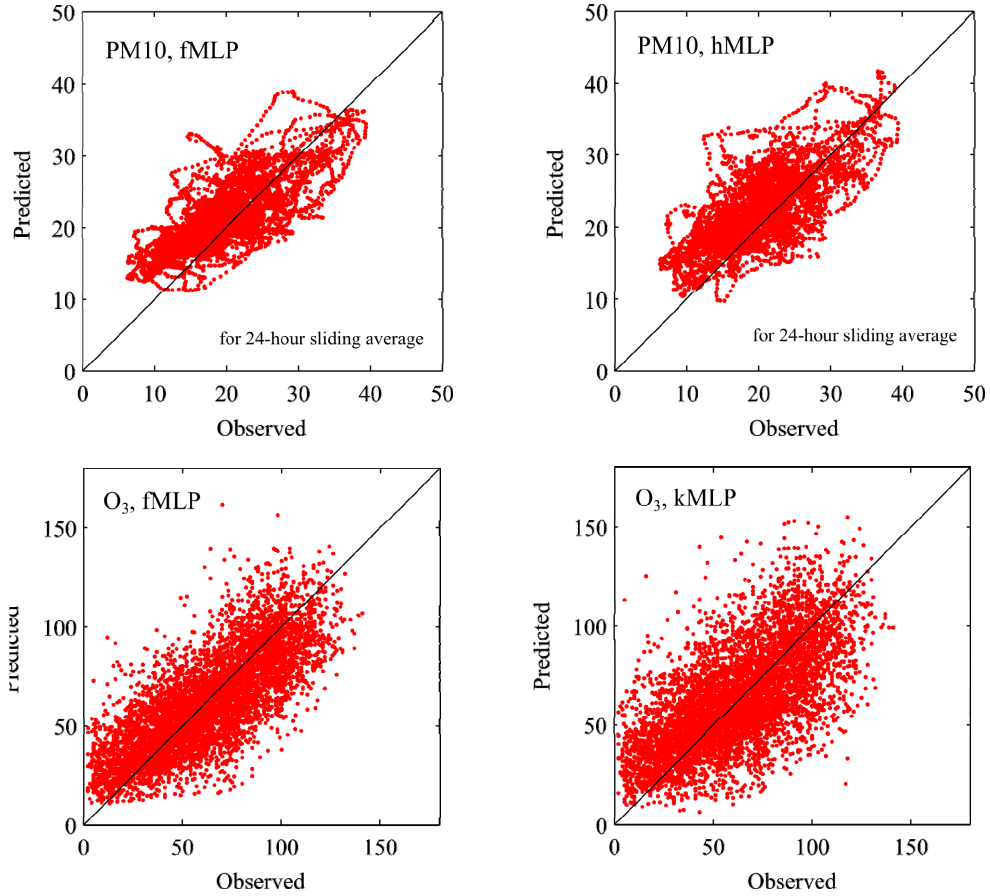

Fig. 6. Scatter plots of fMLP and the best hybrid model in Canetto station, in μg m^{-3}.

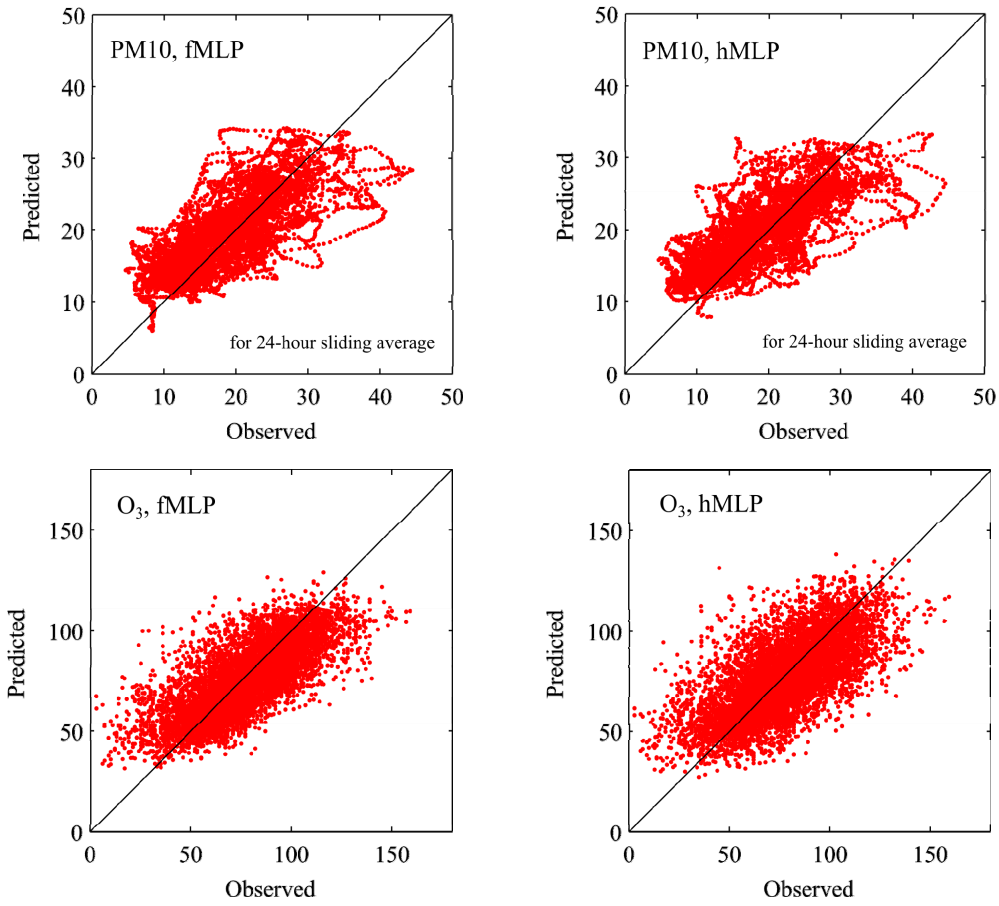

Fig. 7. Scatter plots of fMLP and the best hybrid model in Giraud station. 24-hour sliding average is used for PM_{10} concentrations.

both PM_{10} and O_3 concentration. Saharan dust event are the major cause of PM_{10} events in Corsica and O_3 transport events from France and Italy are also common. O_3 is a secondary pollutant and is more influenced by meteorological conditions than primary pollutants. On the other side, NO_2 levels are mainly due to local anthropogenic emissions. Those hybrid methods with this clustering dataset seem not to be adapted for this pollutant.

Those ROC curves have to be analysed knowing the pollutant statistics of Table 2, not only describing the test set but all available data. Information threshold are hardly reached, only PM_{10} is responsible for threshold exeedances.

Fig. 8 shows the most important PM_{10} event of the test set, and the better performances of the hMLP model are visible. For example, the 18/03/12 peak is underestimated by fMLP and forecasted by hMLP.

No clear relationship could be found between the nature of the clustering (hierarchical or SOM/k-means based, number of clusters) and the high concentration levels detection performances, but both kMLP and hMLP appear to improve those abilities in several situations for PM_{10} and O_3. Increasing the number of cluster may improve the clustering, but will decrease the amount of data for each MLP's learning. It should be noted that the best models in term of global performance indexes (Table 3) may not be the best for high concentration detection, and that point emphasizes the interest of the use of ROC curves for evaluation. The number of cluster to obtain the best high concentration detection can even be different for hourly concentrations and for 24-hour sliding averages, as it is for PM_{10} kMLP. With this average, TPR can be strongly improved by clustering, underlying the benefit of this method for operational forecasting. kMLP and hMLP improve high concentration detection rate in the majority of investigated cases. We suggest the use of both those clustering methods, and an evaluation with ROC curves in addition to classical indexes to identify the better architecture for each situation. In those conditions, an improvement of pollution peak detection can be waited. For operational use, ROC curves also bring information on the behaviour of the model, helping the forecaster interpret its outputs.

CONCLUSIONS

In order to detect pollutant peaks, we developed an original approach combining ANNs and clustering. First we built three different MLP-based models to forecast hourly concentration of PM_{10}, O_3 and NO_2 24 hours ahead, with Principal Components of endogenous and exogenous data as inputs. Data came from Corsica Island in the western Mediterranean Sea, and consisted of air quality and meteorological measurements and outputs from AROME

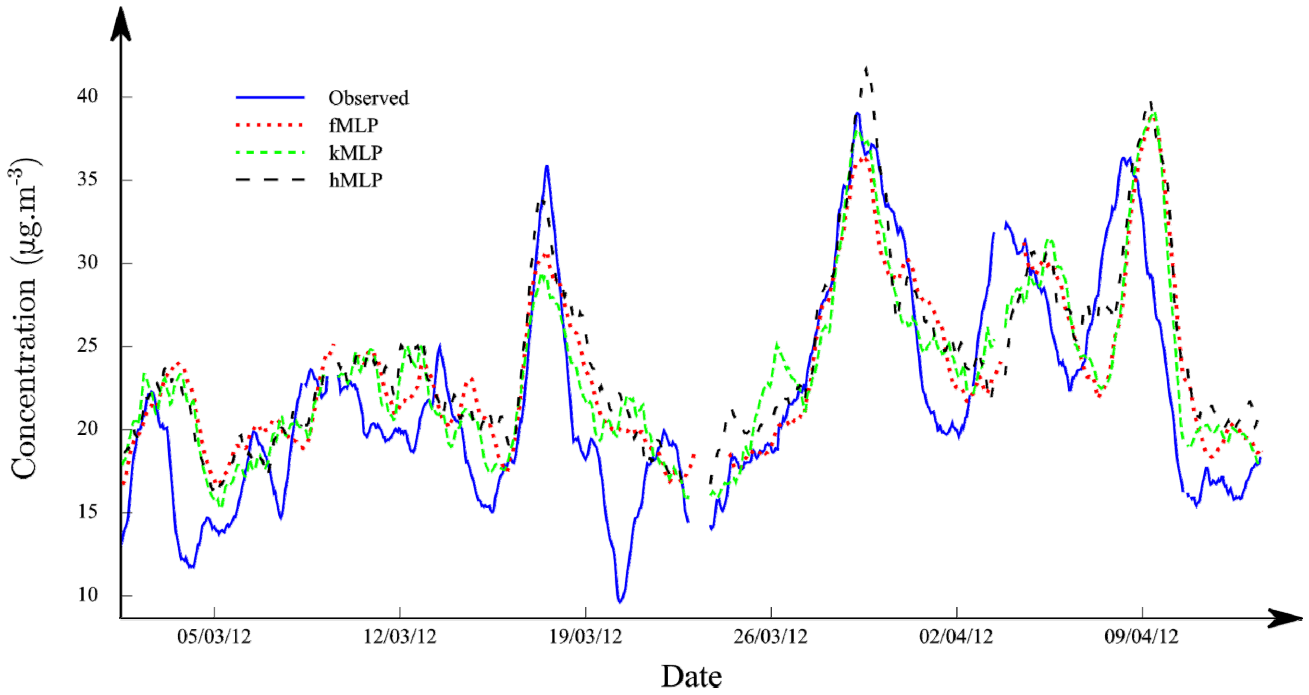

Fig. 8. Observed and predicted timeseries of PM_{10} concentration in Canetto with fMLP and hybrid models.

NWP model. Five years of data were available and divided into the train set, the validation set and the test set. A first model used the full dataset to be trained (fMLP), and was evaluated giving a good global precision (IA of 0.87 for O_3, 0.74 for PM_{10} and 0.80 for NO_2).

Two hybrid models were also built, combining MLP and clustering methods. The two investigated clustering approaches were hierarchical clustering using euclidean distance with Ward criterion (hMLP), and k-means clustering coupled with Self Organisation Map (kMLP).

Hybrid models had lower global precision in term of IA, but showed better ability to correctly forecast high concentration events. This ability is the most important for operational air quality forecasting, and the evaluation of models with ROC curves was useful to describe their behaviour and robustness for various concentration thresholds. We suggest the use of such curves for sensitivity analysis in studies relative to air quality forecasting, which are often limited to an evaluation on the full test set with error indexes that do not distinguish high concentration events from other situations.

Both hierarchical and SOM/k-means clustering approaches appeared to be efficient, depending on the situation. Their use generally increased the detection rate of high pollution events compared to the classical MLP for PM_{10} and for O_3. However, classical MLP still performed better than hybrid models in global performances. As clustering process reduce the size of training set, an improvement of hybrid models can be waited when more data will be available.

The results obtained lead to a continuation of our research effort using those methods. Hybrid models may be used to focus on the high concentration events, and classical MLP for air quality forecasting regardless to high pollution. Our perspective is to apply those processes on other datasets, with different pollutant patterns and from other regions of Earth. As these models need few computing resources, they seem adapted for AAQMAs with limited financial and human resources, on territories such as French Islands (Guadeloupe and Martinique Islands in the Caribbean Sea, and Réunion Island in the Indian Ocean). An operational model has been built following the template presented here, and is working at the local AAQMA, Qualitair Corse, to improve the forecasting of pollution events.

ACKNOWLEDGMENTS

We would like to thank Qualitair Corse and Météo-France, the two organisms that provided the data used in this study. This research has been funded by the Economic Development Agency of Corsica (ADEC) and by the Sustainable Planning and Urbanization Agency of Corsica (AAUC).

REFERENCES

Bessagnet, B., Hodzic, A., Vautard, R., Beekmann, M., Cheinet, S., Honoré, C., Liousse, C. and Rouil, L. (2004). Aerosol Modeling with CHIMERE - Preliminary Evaluation at the Continental Scale. *Atmos. Environ.* 38: 2803–2817, doi: 10.1016/j.atmosenv.2004.02.034.

Coman, A., Ionescu, A. and Candau, Y. (2008). Hourly Ozone Prediction for a 24-h Horizon Using Neural Networks. *Environ. Modell. Software* 23: 1407–1421, doi: 10.1016/j.envsoft.2008.04.004.

Corani, G. (2005). Air Quality Prediction in Milan: Feed-forward Neural Networks, Pruned Neural Networks and Lazy Learning. *Ecol. Modell.* 185: 513–529, doi: 10.1016/j.ecolmodel.2005.01.008.

Davis, D.L. and Bouldin, D.W. (1979). A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1: 224–227.

Ding, C. and He, X. (2004). K-means Clustering via Principal Component Analysis. Banff, Canada.

Dutot, A., Rynkiewicz, J., Steiner, F. and Rude, J. (2007). A 24-h Forecast of Ozone Peaks and Exceedance Levels using Neural Classifiers and Weather Predictions. *Environ. Modell. Software* 22: 1261–1269, doi: 10.1016/j.envsoft.2006.08.002.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognit. Lett.* 27: 861–874, doi: 10.1016/j.patrec.2005.10.010.

Gardner, M.W. and Dorling, S.R. (1998). Artificial Neural Networks (The Multilayer Perceptron) - A Review of Applications in the Atmospheric Sciences. *Atmos. Environ.* 32: 2627–2636.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* 2: 359–366, doi: 10.1016/0893-6080(89)90020-8.

Kolehmainen, M., Martikainen, H., Hiltunen, T. and Ruuskanen, J. (2000). Forecasting Air Quality Parameters Using Hybrid Neural Network Modelling. *Environ. Monit. Assess.* 65: 277–286.

Konovalov, I.B., Beekmann, M., Meleux, F., Dutot, A. and Foret, G. (2009). Combining Deterministic and Statistical Approaches for PM_{10} Forecasting in Europe. *Atmos. Environ.* 43: 6425–6434, doi: 10.1016/j.atmosenv.2009.06.039.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorlind, S., Chatterton, T., Foxall, R. and Cawley, G. (2003). Extensive Evaluation of Neural Network Models for the Prediction of NO_2 and PM_{10} Concentrations, Compared with a Deterministic Modelling System and Measurements in Central Helsinki. *Atmos. Environ.* 37: 4539–4550.

Lu, H.C., Hsieh, J.C. and Chang, T.S. (2006). Prediction of Daily Maximum Ozone Concentrations from Meteorological Conditions Using a Two-stage Neural Network. *Atmos. Res.* 81: 124–139, doi: 10.1016/j.atmosres.2005.11.007.

Marquardt, D.W. (1963). An Algorithm for Least-squares Estimation of Nonlinear Parameters. *J. Soc. Ind. Appl. Math*. 11: 431–441.

Nguyen, D. and Widrow, B. (1990). Improving the Learning Speed of 2-layer Neural Networks by Choosing Initial Values of the Adaptive Weights. Proc. of the Int. Joint Conference on Neural Networks, 3: 21–26.

Paoli, C., Notton, G., Nivet, M.L., Padovani, M. and Savelli, J.L. (2011). A Neural Network Model Forecasting for Prediction of Hourly Ozone Concentration in Corsica. 10th International Conference on Environment and Electrical Engineering (EEEIC), pp. 1–4.

Paschalidou, A.K., Karakitsios, S., Kleanthous, S. and Kassomenos, P.A. (2010). Forecasting Hourly PM_{10} Concentration in Cyprus through Artificial Neural Networks and Multiple Regression Models: Implications to Local Environmental Management. *Environ. Sci. Pollut. Res.* 18: 316–327, doi: 10.1007/s11356-010-0375-2.

Perez, P. and Reyes, J. (2002). Prediction of Maximum of 24-h Average of PM_{10} Concentrations 30 h in Advance in Santiago, Chile. *Atmos. Environ.* 36: 4555–4561.

Perez, P. (2012). Combined Model for PM_{10} Forecasting in a Large City. *Atmos. Environ.* 60: 271–276, doi: 10.1016/j.atmosenv.2012.06.024.

Poggi, J.M. and Portier, B. (2011). PM_{10} Forecasting Using Clusterwise Regression *Atmos. Environ.* 45: 7005–7014, doi: 10.1016/j.atmosenv.2011.09.016.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C. and Masson, V. (2011). The AROME-France Convective-Scale Operational Model. *Mon. Weather Rev.* 139: 976–991, doi: 10.1175/2010MWR3425.1.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M. and Pereira, M.C. (2007). Multiple Linear Regression and Artificial Neural Networks Based on Principal Components to Predict Ozone Concentrations. *Environ. Modell. Software* 22: 97–103.

Tamas, W., Notton, G., Paoli, C., Voyant, C., Nivet, M.L. and Balu, A. (2014). Urban Ozone Concentration Forecasting with Artificial Neural Network in Corsica. *Math. Modell. Civil Eng.* 10: 29–37, doi: 10.2478/mmce-2014-0004.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function." *J. Am. Stat. Assoc.* 58: 236–244, doi: 10.1080/01621459.1963.10500845.

Willmott, C.J. (1982). Some Comments on the Evaluation of Model Performance. *Bull. Am. Meteorol. Soc.* 63: 1309–1313.

Zhang, G.P. (2012). Neural Networks for Time-Series Forecasting. In *Handbook of Natural Computing*, Rozenberg, G., Bäck, T. and Kok, J.N. (Eds.), Springer Berlin Heidelberg, Heidelberg, Berlin, pp. 461–477.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. and Baklanov, A. (2012a). Real-time Air Quality Forecasting, Part I: History, Techniques, and current Status. *Atmos. Environ.* 60: 632–655, doi: 10.1016/j.atmosenv.2012.06.031.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. and Baklanov, A. (2012b). Realtime Air Quality Forecasting, Part II: State of the Science, Current Research Needs, and Future Prospects. *Atmos. Environ.* 60: 656–676, doi: 10.1016/j.atmosenv.2012.02.041.

Received for review, April 10, 2015
Revised, July 2, 2015
Accepted, August 21, 2015