Modelling the time fluctuation of indoor air formaldehyde concentrations: variability structure identification and forecasting using non linear models

Rachid Ouaret, Anda Ionescu, Olivier Ramalho, Yves Candau, Viorel Petrehus, Lucille Labat

Topic B6: Prediction and measurement

MODELLING THE TIME FLUCTUATION OF INDOOR AIR FORMALDEHYDE CONCENTRATION: VARIABILITY STRUCTURE IDENTIFICATION AND FORECASTING USING NONLINEAR MODELS

Rachid OUARET[1, 2,*], Anda IONESCU[2], Olivier RAMALHO[1], Yves CANDAU[2], Viorel PETREHUS[3], Lucille LABAT[1]

[1] University Paris-Est, Scientific and Technical Centre for Building, Marne-la-Vallée, France.
[2] CERTES, University Paris-Est Créteil Val de Marne, France.
[3] Technical University of Civil Engineering Bucharest, Romania.

[*] *Corresponding email*: rachid.ouaret@cstb.fr

Keywords: formaldehyde, clustering, forecasting, SOTA, SETAR

SUMMARY

This study explores the possibility to forecast formaldehyde (HCHO) concentration from past observations in an office. A monitoring campaign of HCHO was performed during 96 days, with a short time-step (1 minute). The monitored formaldehyde time series exhibited, in particular, abrupt changes and structural breaks in variance, that cannot be modelled implicitly using simple models. To overcome this problem, hybrid model for forecasting HCHO time series (characterized by nonlinear and nonstationary behaviour) was used. The periodicities of the data were modelled by Fast Fourier Transform (FFT). A Self-Exciting Threshold AutoRegressive (SETAR) model was used to model the FFT component. SETAR models are typically designed to accommodate the nonlinear features and can explain two regimes in a time series. The residuals of the FFT component subtracted from raw data was modelled using a second SETAR model. The output of the two models were summed and compared to raw test data.

INTRODUCTION

The indoor concentration of formaldehyde (HCHO) fluctuates over time and remains difficult to model because of a number of different emission sources (including reactivity) and its sensitivity to temperature, humidity and ventilation changes. The concentration dynamic and its relation to sources and meteorological variables are complex as they involve the dependency of sources and sinks as well as other parameters such as temperature, humidity, air flow but also to the concentration of oxidants in the air (i.e. radicals, peroxides, ozone, oxygen). While the outdoor air research has experience in long monitoring studies with advanced statistical modelling at various time scales, the field of indoor air quality lacks monitoring studies of pollutant concentrations with a short time-step. In particular, the temporal variability of indoor formaldehyde fluctuations remains unknown. Identification of predominant scales of temporal variability of HCHO concentration is a prerequisite to understand the heterogeneity of indoor source emissions (type and intensity). Information on the magnitude of variation and underlying mechanism is important both to improve our understanding of HCHO dynamics and key characteristics of source emission.

The indoor concentration of a pollutant, e.g. formaldehyde, can be modelled by two different manners: physical models or statistical models. The physical models are based on general mass balance integral-differential equations. These models are commonly used in experimental chambers, where all parameters are controlled and do not vary in time. The use of such models in real environments require the knowledge of all contributing sources and their emission parameters, as well as the sink effects, and how all these parameters fluctuates over time. On the opposite, statistical models require less a priori knowledge, i.e. information can be retrieved from a concentration time-series by inverse modelling and provide insights on the contributing sources and possible forecast.

The objective of this study is to explore the statistical methods to forecast indoor formaldehyde concentration using past observations. The variability structure of the past observations is the key parameter to achieve this goal. It is determined through the analysis of abrupt point changes and clustering of similar trends in the data, which oriented us towards the use of a hybrid SETAR model (Self-Exciting Threshold Autoregressive) to predict future concentrations.

METHODS

Measurement of HCHO

The indoor formaldehyde concentration was monitored every minute in an open-plan office occupied by 6-8 persons over 96 days, using an AL4021 analyzer (Aerolaser GmbH). The detection of formaldehyde is based on the liquid phase reaction of formaldehyde with acetyl acetone and ammonia that produces 3,5-diacetyl-1,4-dihydrolutidine which is measured by fluorescence at 510 nm. The device was regularly calibrated with liquid solutions of formaldehyde and the internal permeation source in the range 0 to 55 ppb. The detection limit was 0.36 ppb. The sampling point was placed in the exhaust ventilation duct.

The office is located in a suburban area 30 km east of Paris. The total area of the office is 132 m^2 with a volume of 397 m^3. A constant air flow rate is maintained through a mechanical exhaust ventilation system with air inlets provided in the windows. The total exhaust ventilation rate is 252 m^3/h. The indoor materials were carpet on the floor, painted walls and ceiling tiles. The furniture comprises typical L-shaped desks in melaminated particleboards and aluminium closets. A laser multifunction copier was in use in the office space.

Statistical methodology

The first part of this analysis is to explore HCHO variability structure. The main daily profiles were obtained by classification. Several classification techniques were employed: hierarchical classification (CAH) based on the Ward algorithm (Saporta, 1990), k-means (Saporta, 1990) and SOTA (Self Organizing Tree Algorithm) (Dopazo et al., 1997; Herrero et al., 2001). The first methods (CAH and k-means) are classical techniques, widely used for clustering. SOTA algorithm has been developed in biology and gene expression, but it can be used in other fields as well.

With the variability of HCHO concentration, the problem is to find the abrupt changes in the time series, which probably indicates a source of variability. We are interested in abrupt point changes detection from a statistical point of view. That corresponds to a problem of estimating the point at which the statistical properties of a sequence of observations change.

For this purpose, we apply the Pruned Exact Linear Time (PELT) method (Jackson et al. 2005). Unfortunately, the choice of the appropriate penalty of the algorithm optimization depends on many factors including the size of the changes and the length of segments, which is still an open question.

The presence of abrupt changes revealed by PELT method and present in some daily profiles obtained by classification justified the use of a particular autoregressive model to forecast HCHO from past observations, which is called Self-Exciting Threshold Autoregressive model or SETAR (Tong et al.1980, Tong 2011). A 2-stage procedure was employed. First, a Fast Fourier Transform (FFT) was applied to raw data. This first component was modelled by a first SETAR. The residuals between raw data and the FFT component were calculated and a second FFT was applied to the residuals. The results were further modelled by a second SETAR. The results of the two forecast procedures were summed. Before performing the forecast procedure, the initial raw database was split in two parts: the first one used for 'training' (model fitting) and a second one, independent, used for test. A comparison between forecast and raw data from the test ensemble was made using classical performance criteria. A general framework of the statistical procedure developed in this study is given in Figure 1.

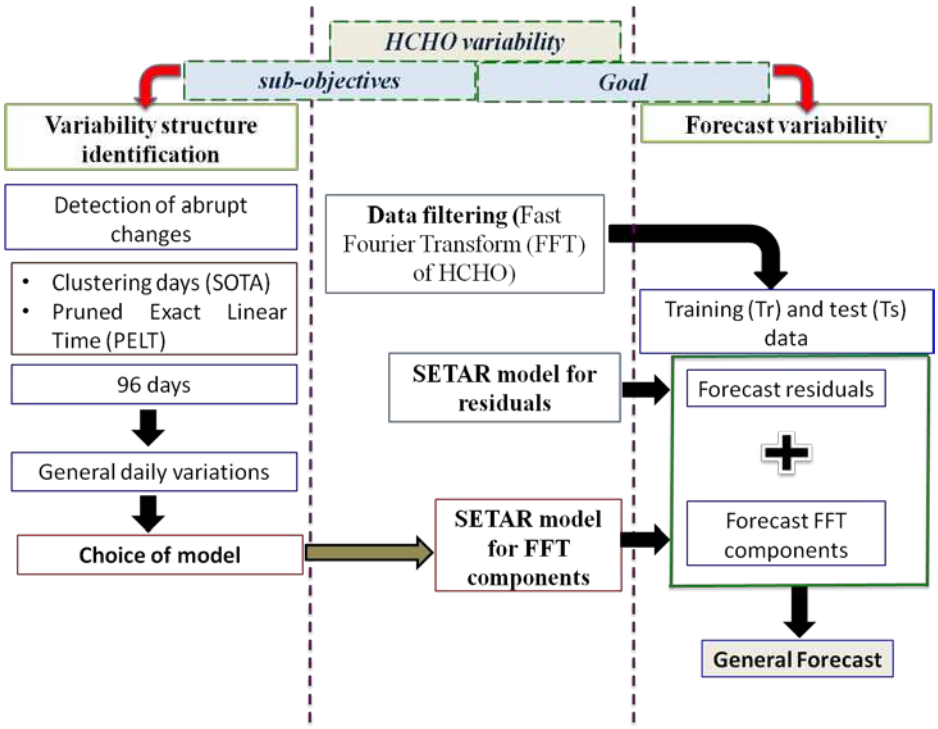

Figure 1. General framework for statistical analysis and forecast of HCHO concentrations.

RESULTS

HCHO time series

The concentration of HCHO monitored at 1-minute time step during 96 days (138,240 observations) in the open-plan office are presented in Figure 2. The indoor concentration of HCHO remains low with no value observed above 25 ppb. The measurements started on 2013-04-27 and ended on 2013-07-31. There are some gaps, especially during summer holidays. Long periods of missing data lead us to reduce the database used for forecast

purposes, in order to have a continuous time series. The final database used for prediction was 21 days long, from 2013-05-24 to 2013-06-14.

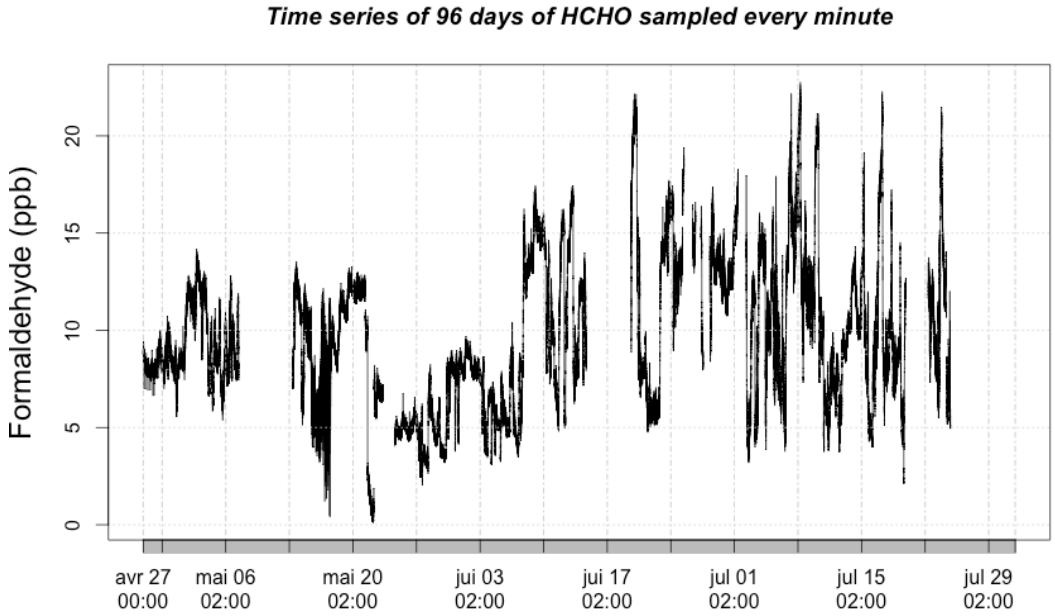

Figure 2: Time series of 96 days of HCHO sampled every minute

In order to define typical daily profiles, a matrix of 96 rows was designed, one row for each day. The mean daily profile over the 96 days is plotted in Figure 3a and the other daily profiles (median, minimum, maximum, quartiles and percentile 95) in figure 3b. The mean concentration of HCHO is around 10 ppb during night hours and decreased during occupancy hours at 8.8 ppb. This may be explained by higher air exchange rates during the day due to higher frequency of windows and doors openings, and thus maybe more ultraviolet radiation entering the office which increased the photolysis rate of formaldehyde.

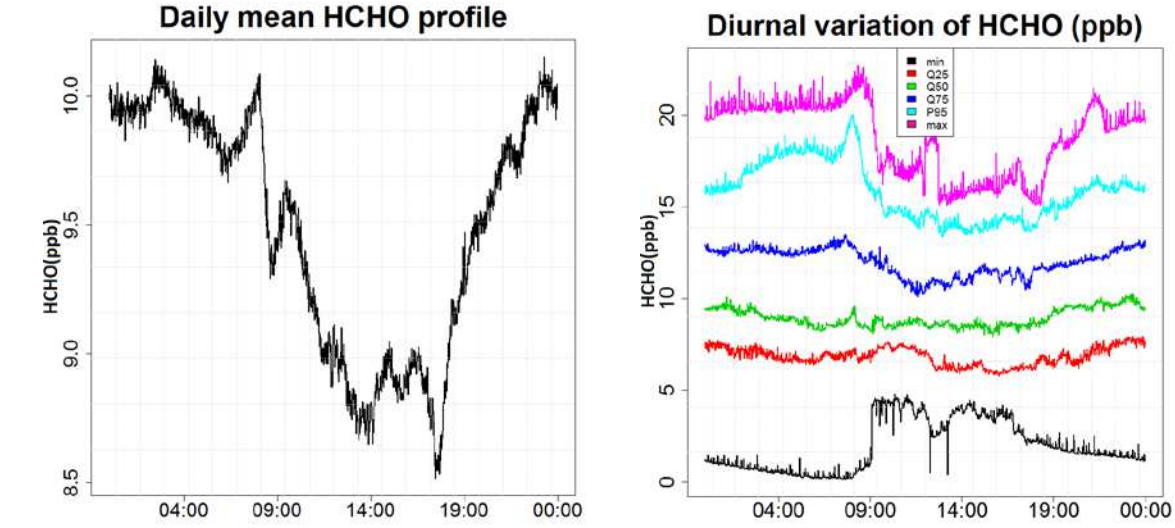

Figure 3: **(A)** Daily mean HCHO profiles for 96 days **(B)** Daily profile of HCHO maximum, minimum, mean, median, quartiles (Q25, Q75) and percentile (P95) for 96 days.

Classification

The 96 daily time series were considered as individuals and different clustering algorithms were applied in order to reveal the main patterns. Several classifications methods (HAC, k-means, SOTA) were used and the results were similar. SOTA classification results are presented in Figure 4.

The interpretation of the different clusters should be done using some exogenous variables. During the same campaign, other pollutants (ozone, nitrogen oxides) and climatic parameters were monitored. A complementary analysis that integrates all these parameters will be an important part of our future work.

There are some 'flat' profiles, characterized by a small variation amplitude (cluster 1), but also profiles with abrupt changes (especially clusters 3 and 4).

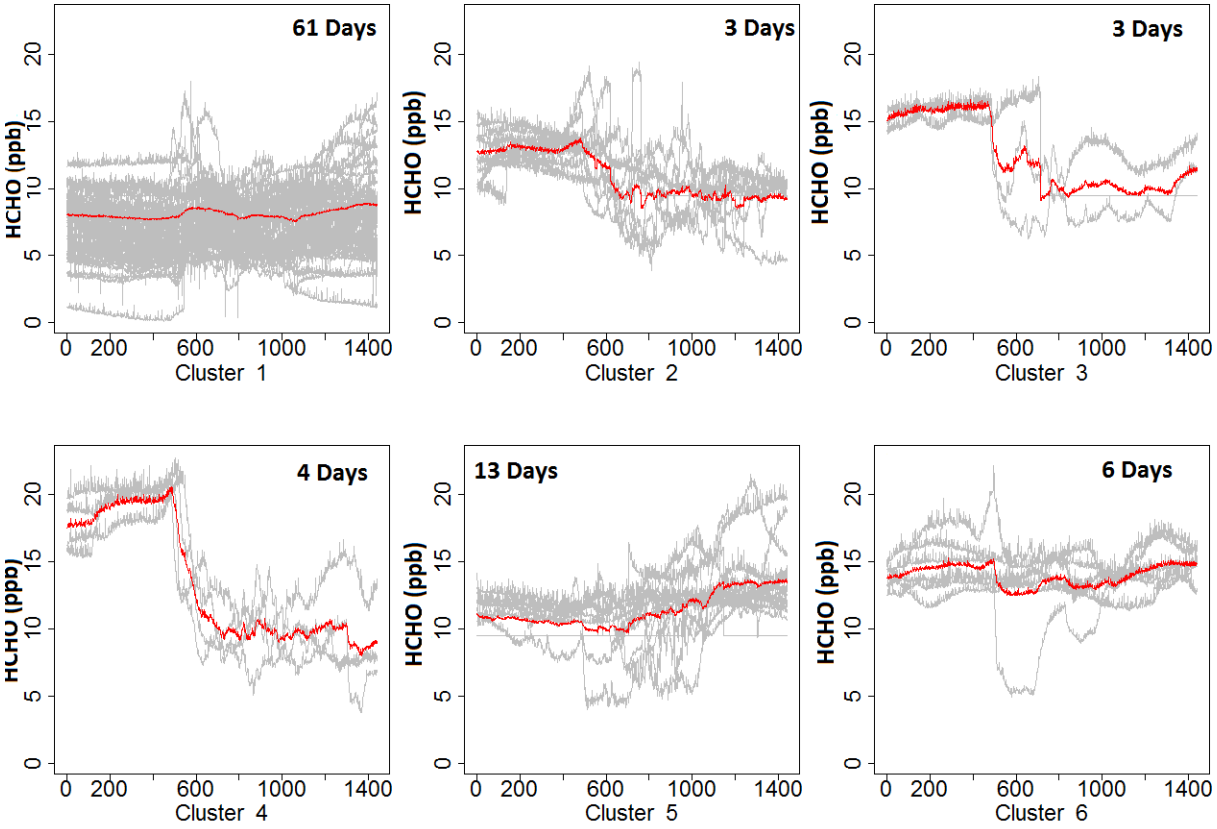

Figure 4: SOTA clustering of 96 daily time series of HCHO (0 to 1440 minutes).

Abrupt change detection

We detected some abrupt changes in the HCHO using the PELT algorithm. An example is given in Figure 5.

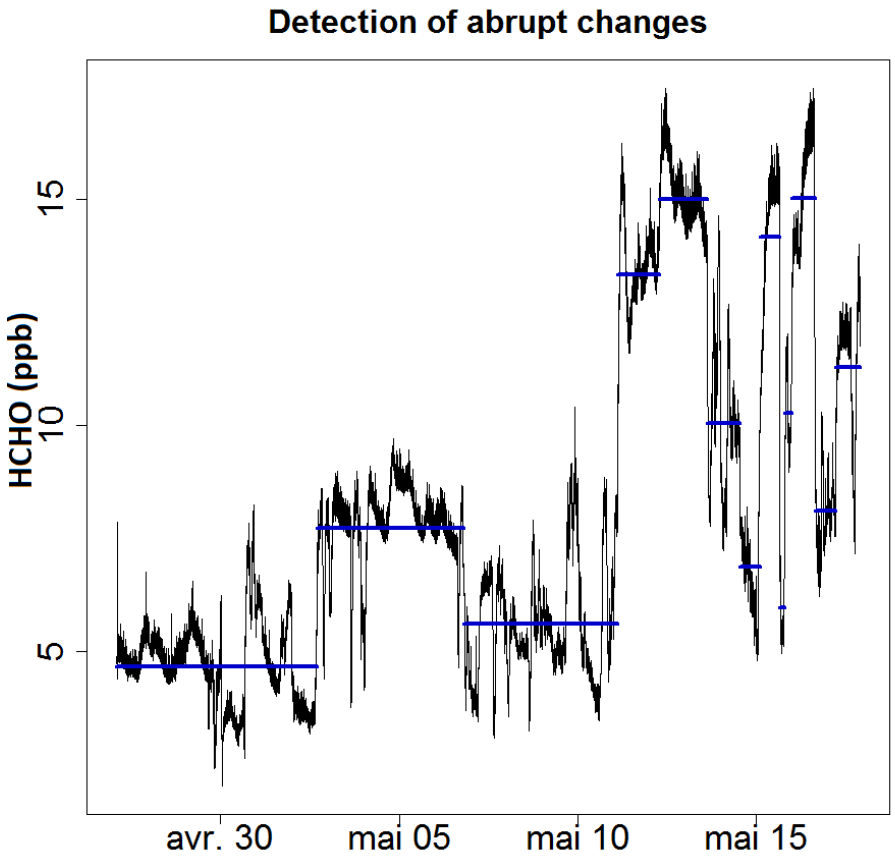

Figure 5: Example of detecting abrupt change using the PELT algorithm.

Forecast

In order to take into account this type of variations (abrupt changes), we selected a SETAR model to forecast HCHO concentrations. The database comprises 21 days with no missing data (or very short gaps that could be interpolated), which corresponds to 30,240 observations. A training sample of 27,600 observations was created and also a test sample with the remaining 2,640 observations. The model was fitted on the training dataset and evaluated on the test set.

Figure 6a presents the training set and figure 6b the test set. In the upper plot, raw HCHO concentrations are plotted in black and the concentration filtered using the FFT, in red. The residuals obtained as the difference between raw data and FFT component are given in the lower plot, in black. A second FFT component was calculated for theses residuals and plotted in red. A first SETAR model was fitted to the first FFT component (calculated from the raw data) and a second SETAR model was fitted to the second FFT component (calculated from the residuals). Two forecast sets were calculated using the two SETAR models and then summed; the result is plotted in blue on Figure 6b. The forecast (blue line) can be compared to raw test data (in black). In addition, a FFT of the raw test data was plotted in green.

The forecast model fits quite well the reconstructed FFT component (global trend + residuals) up to 900 min (15 hours) and was satisfactory during the 44 hours of the test set (less than 2 ppb difference). The model is thus useful to predict the general variation of HCHO but unsufficient to predict fast variations change. In order to achieve that, other informative data must be input in the model. That is part of the second step of our work.

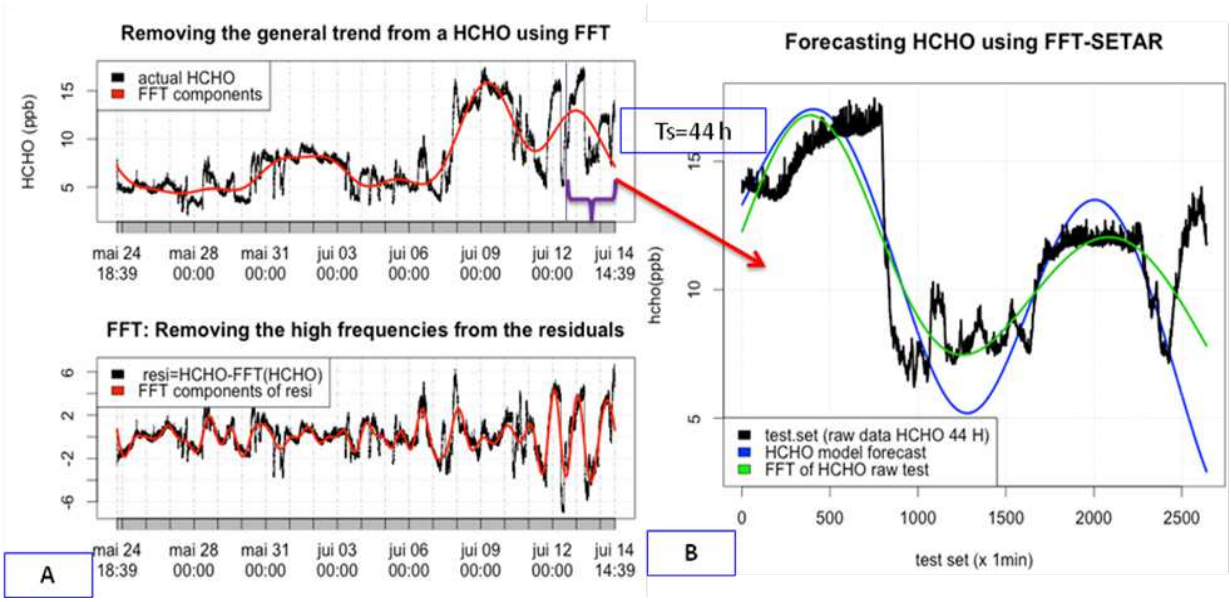

Figure 6. **(A)**: Training set and residuals (raw data and FFT). **(B)** Test set of 44 hours (raw data and forecast)

The criteria to evaluate the forecast performance on the filtered FFT components are given in table 1, with ME (mean error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MPE (Mean Percentage Error), MAPE (Mean Absolute Percentage Error), R^2 (determination coefficient).

With a forecasting horizon of 44 hours, the performance of the model is satisfactory to predict the general variation of formaldehyde concentration (FFT test set). When compared to the raw data set, more discrepancies appear as the model do not take into account high frequency variations of formaldehyde. However, the performance of the model is not bad with 66% of the total variance explained.

Table 1: Forecast performance criteria.

	Forecasting Horizon = 44 h			
Performance	1st FFT-SETAR (HCHO)	2nd FFT-SETAR	FFT test set and FFT-SETAR forecast	raw test set and FFT-SETAR forecast
ME (ppb)	-0.96	0.53	0.43	0.48
RMSE	1.15	1.65	1.58	2.57
MAE (ppb)	1.06	1.29	1.18	1.80
MPE (%)	-10.49	1.58	6.40	4.37
MAPE (%)	11.21	176.73	12.43	16.75
R^2	-	-	0.91	0.66

CONCLUSIONS

The analysis of HCHO time series revealed the presence of some abrupt changes that could be modelled using a SETAR model. The forecast performance of the general trend (derived from FFT filtering) was good for a horizon of 44 hours (almost 2 days ahead). The performance of

the model should be increased using exogenous variables if the statistical analysis reveals their importance in the ongoing work.

The classification permitted to obtain some diurnal profiles. These profiles should be further analysed using other pollutant concentrations and climatic parameters, as well as ventilation conditions.

REFERENCES

Dopazo, J. and J. M. Carazo (1997). "Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree." *J Mol Evol* **44**(2): 226-233

Herrero, J., Valencia, A, and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126-136.

Güler, N. and S. Koçer (2005). "Classification of EMG Signals Using PCA and FFT." *Journal of Medical Systems,* **29**(3): 241-250.

Jackson B., Sargle J.D., Barnes et al. (2005) An algorithm for optimal partitioning of data on an interval, *IEEE, Signal Process. Lett.*, **12**, pp. 105–108

Saporta G, Probabilités, analyse des données et statistique, Technip Editions, 1990

Tong, H., 1995, Non-Linear Time Series. A Dynamical System Approach, Clarendon Press, Oxford

Tong, H. & Lim, K. S. (1980) "Threshold Autoregression, Limit Cycles and Cyclical Data (with discussion)", *Journal of the Royal Statistical Society, Series* B, **42**, 245-292.

Tong, H. (2011). "Threshold models in time series analysis —30 years on (with discussions by P.Whittle, M.Rosenblatt, B.E.Hansen, P.Brockwell, N.I.Samia & F.Battaglia)". *Statistics & Its Interface*, **4**, 107-136.