Study of PM_{10} Pollution Using Signal Analysis Methods

K. Siwek, S. Osowski, B. Swiderski

Warsaw University of Technology

International Symposium on Theoretical Electrical Engineering - ISTET'11, Klagenfurt

Schedule

- Problem statement
- The analysis of the pollution dependencies
- Predicting systems
- Results of numerical experiments
- Conclusions

Problem statement

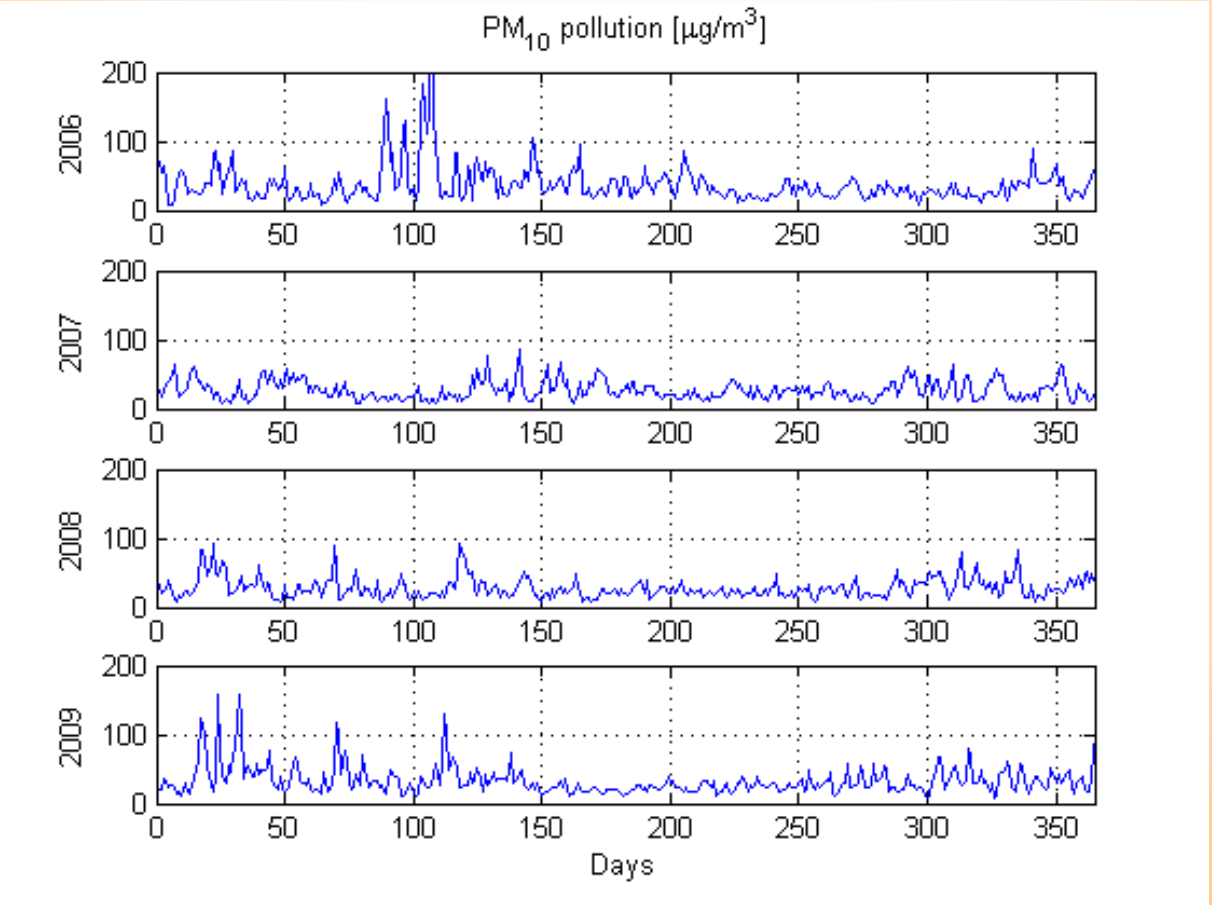
One of the most important component of the pollution is the ambient particulate matter (PM) of the diameters up to 10μm (PM_{10}) and 2.5μm ($PM_{2.5}$).

The main source of PM is the vehicular traffic, smoke from the coal heating system and the dust of the streets generated by the circulation.

Actually PM is of importance for an European policy (the new European Air Quality Directive EC/2008/50) defining the restrictions for the yearly and 24h averages PM_{10} concentrations.

The analysis of the pollution dependencies

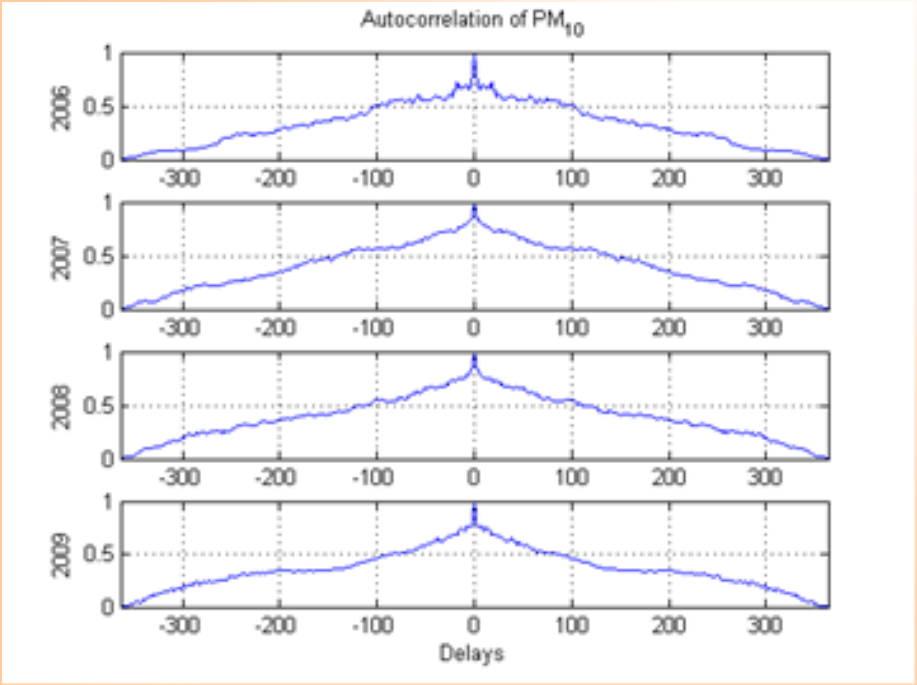
The most important difficulty in prediction of the next day pollution is great non-repeatable changes of their values from day to day.

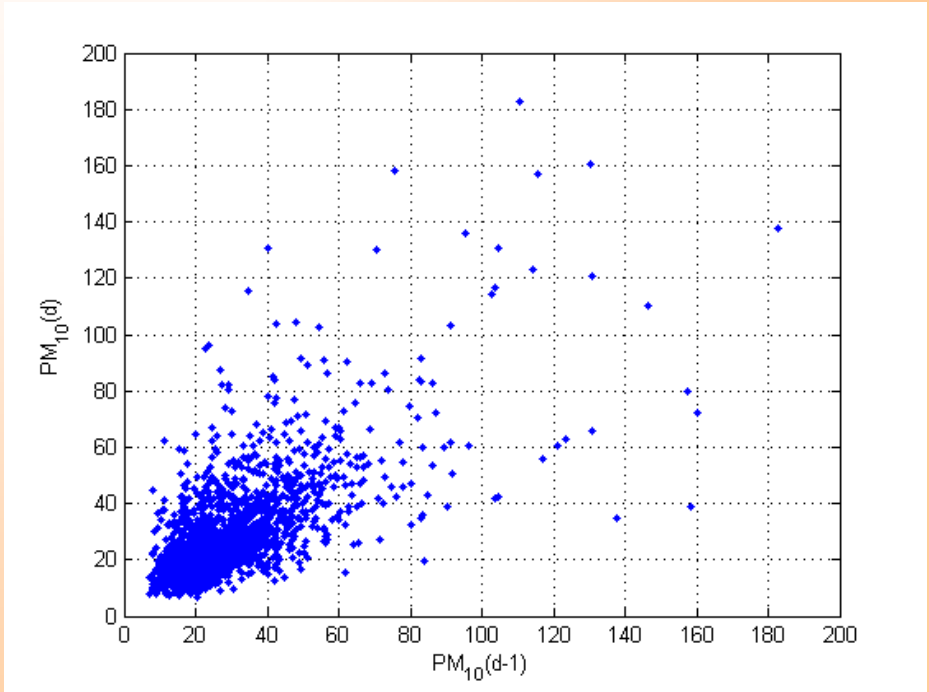


The mean values and std of PM10 pollution

	Mean	Std	Std/mean
2006	37.73	27.82	0.74
2007	27.11	13.59	0.50
2008	28.12	14.77	0.53
2009	33.78	21.35	0.63

Auto-correlation functions of PM_{10}





Cross-correlation coefficients for PM_{10} within different years

	2006	2007	2008	2009
2006	1	-0.198	-0.069	0.005
2007	-0.198	1	-0.082	0.017
2008	-0.069	-0.082	1	0.211
2009	0.005	0.017	0.211	1

Linearity or nonlinearity of the process

- The crucial problem that should be analyzed before starting to build the predictive model of the time series is the assessment of the linearity or nonlinearity of the process under modeling.
- Linearity may simplify the model, since the linear process can be described by the linear relations, much easier in modeling.

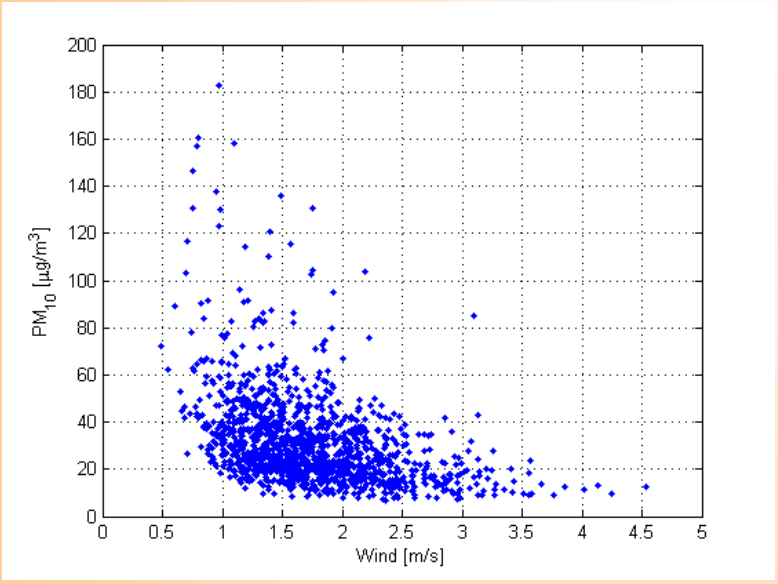
Hinich test of linearity

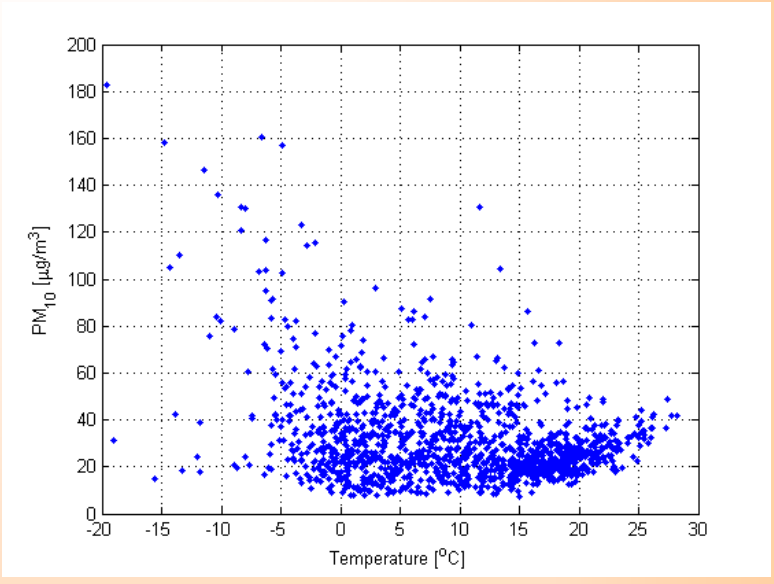
- The basic idea behind the test is checking if the third order cumulant (bispectrum and bicoherence) of the process is zero. If the bispectrum is not zero then the process is non-Gaussian (potentially nonlinear).
- In the case of a non-Gaussian and linear process, the bicoherence is a nonzero constant.
- In practice the so-called “probability of false alarm” (PFA), that is the probability that we will be wrong in assuming that the data have a nonzero bispectrum, has been implemented to test the Gaussianity. If this probability is small, reject the assumption of zero bispectrum and reject also the assumption of the Gaussianity of the process
- In the case of non-Gaussian process, the linearity test is applied, checking whether the squared bicoherence is constant for all frequencies. In practice the bicoherence is usually not flat. In testing for linearity or nonlinearity of the non-Gaussian processes we rely on comparison of the empirical and theoretical sample interquartile ranges. If their values are comparable the process is linear. In the case of high differences the process is regarded as nonlinear.

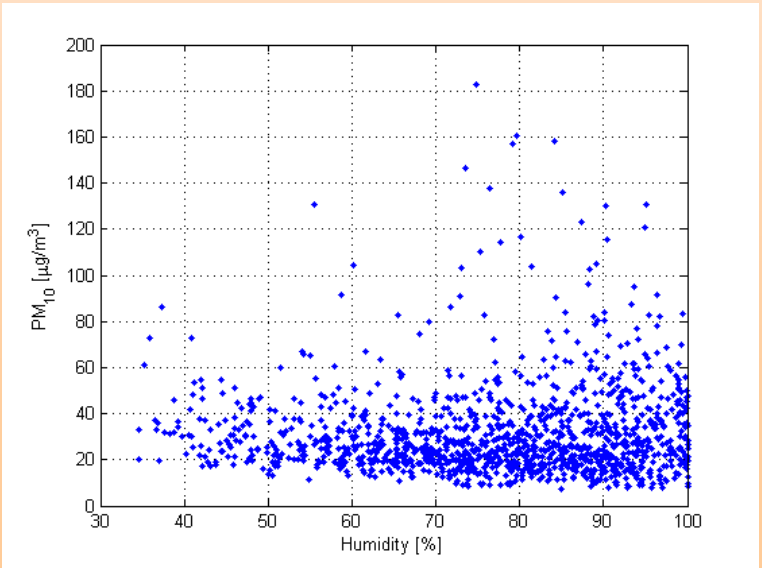
Results of Hinich test

- In checking the nonlinearity we have applied the function glstat.m of Matlab.
- After performing it on the PM_{10} pollution data we have got PFA=0, which means that the assumption of Gaussianity should be rejected.
- In checking linearity and nonlinearity of the process we compared in this test the estimated (R_e) and theoretical (R_t) values of the interquartile ranges.
- In our case we have got estimated value of R_e= 99.1574 and theoretical value R_t= 22.4438. The ratio R_e/R_t=4.41. Such value corresponds to weak nonlinearity of the process. It means that both linear and nonlinear methods might be applied in building the model of pollution.

Relations between PM_{10} and meteorological parameters



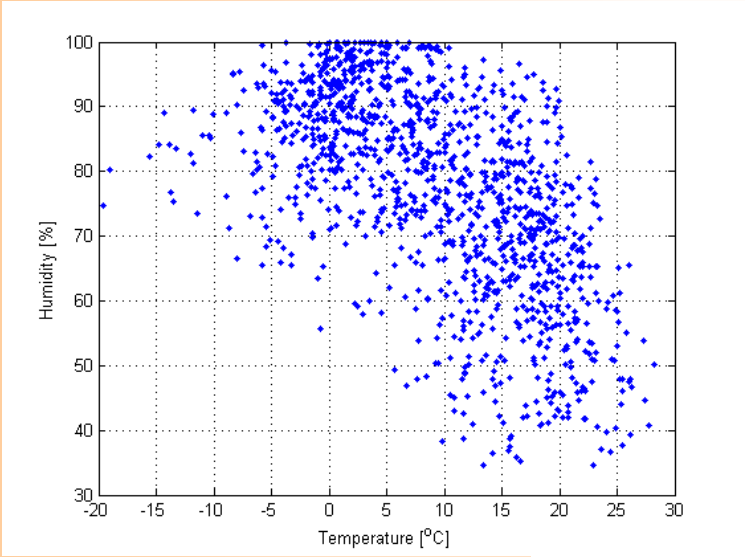


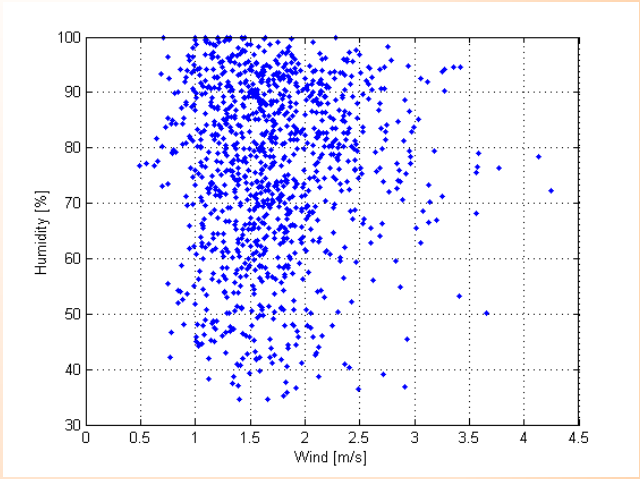


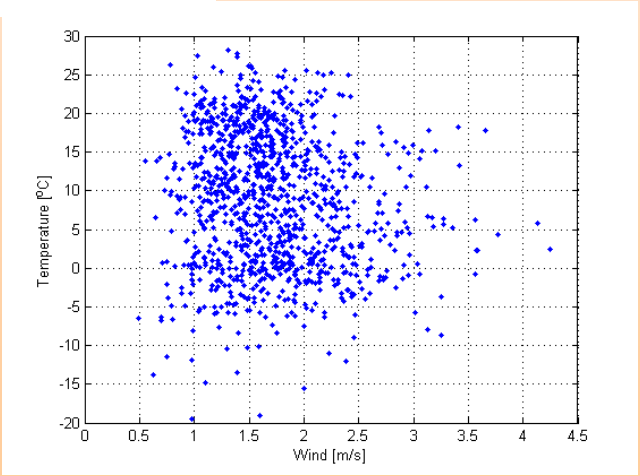
The correlations between meteorological parameters

	Temperature	$Wind_x$	$Wind_y$	Humidity	PM_{10}
Temperature	1	-0.034	-0.021	-0.601	-0.235
$Wind_x$	-0.034	1	-0.008	0.024	0.003
$Wind_y$	-0.021	-0.008	1	0.055	-0.041
Humidity	-0.601	0.024	0.055	1	-0.024
PM_{10}	-0.235	0.003	-0.041	-0.024	1

Relations between meteorological parameters







Multi-collinearity checking

The F-test testing the joint hypothesis that all coefficients of linear equation are all equal zero

$$PM_{10} = \alpha_0 + \alpha_1 temp + \alpha_2 wind_x + \alpha_3 wind_y + \alpha_4 hum + \varepsilon$$

where ε is iid ~$N(0,1)$, we have tested the null hypothesis H_0:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

against the alternative one H_1:

$$H_1 : \alpha_1 \neq 0 \vee \alpha_2 \neq 0 \vee \alpha_3 \neq 0 \vee \alpha_4 \neq 0$$

The results *Fstat* = 47.25 , *p_value*<0.001 vote against null hypothesis.

T-test of significance of each variable

We test if each meteorological variable is significant.

- Null hypothesis:

$$H_0 : \alpha_i = 0$$

- Alternative hypothesis:

$$H_1 : \alpha_i \neq 0$$

- Four tests, each for : temperature, $wind_x$, $wind_y$, humidity.

$$w_x = |w| \cos\varphi$$

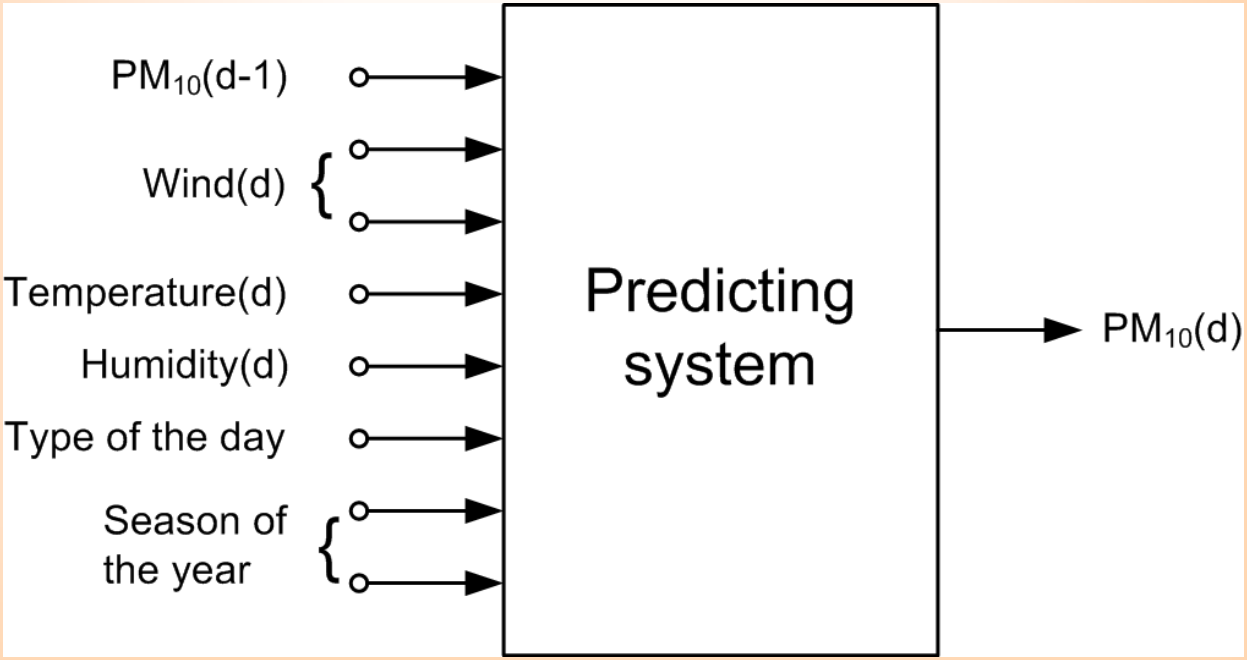
$$w_x = |w| \sin\varphi$$

Results of t-test

	t-statistics	p-value
Temperature	6.06	<0.0001
$Wind_x$	9.33	<0.0001
$Wind_y$	2.08	<0.0380
Humidity	4.74	<0.0001

These results suggest that there is no collinearity between the meteorological variables, since all p-values are below the threshold level 0.05.

Prediction of the next day pollution



$$\hat{P}(d) = f\left(\mathbf{w}, wind_x, wind_y, temp, hum, r, s, P(d-1)\right)$$

Measures of prediction quality

The mean absolute error (MAE)

$$MAE = \frac{1}{p}\left(\sum_{i=1}^{p}\left|d_i - y_i\right|\right)$$

The root mean squared error (RMS)

$$RMS = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left|d_i - y_i\right|^2}$$

The mean absolute percentage error (MAPE)

$$MAPE = \frac{1}{p}\left(\sum_{i=1}^{p}\frac{\left|d_i - y_i\right|}{d_i}\right)\cdot 100\%$$

Measures of prediction quality (cont.)

- *Correlation coefficient (R) of the observed and predicted data*

$$R = \frac{R_{yt}}{std(y)std(t)}$$

- *Index of agreement (IA)*

$$IA = 1 - \frac{\sum_{i=1}^{p}(t_i - y_i)^2}{\sum_{i=1}^{p}\left(|t_i - \bar{t}| + |y_i - \bar{t}|\right)^2}$$

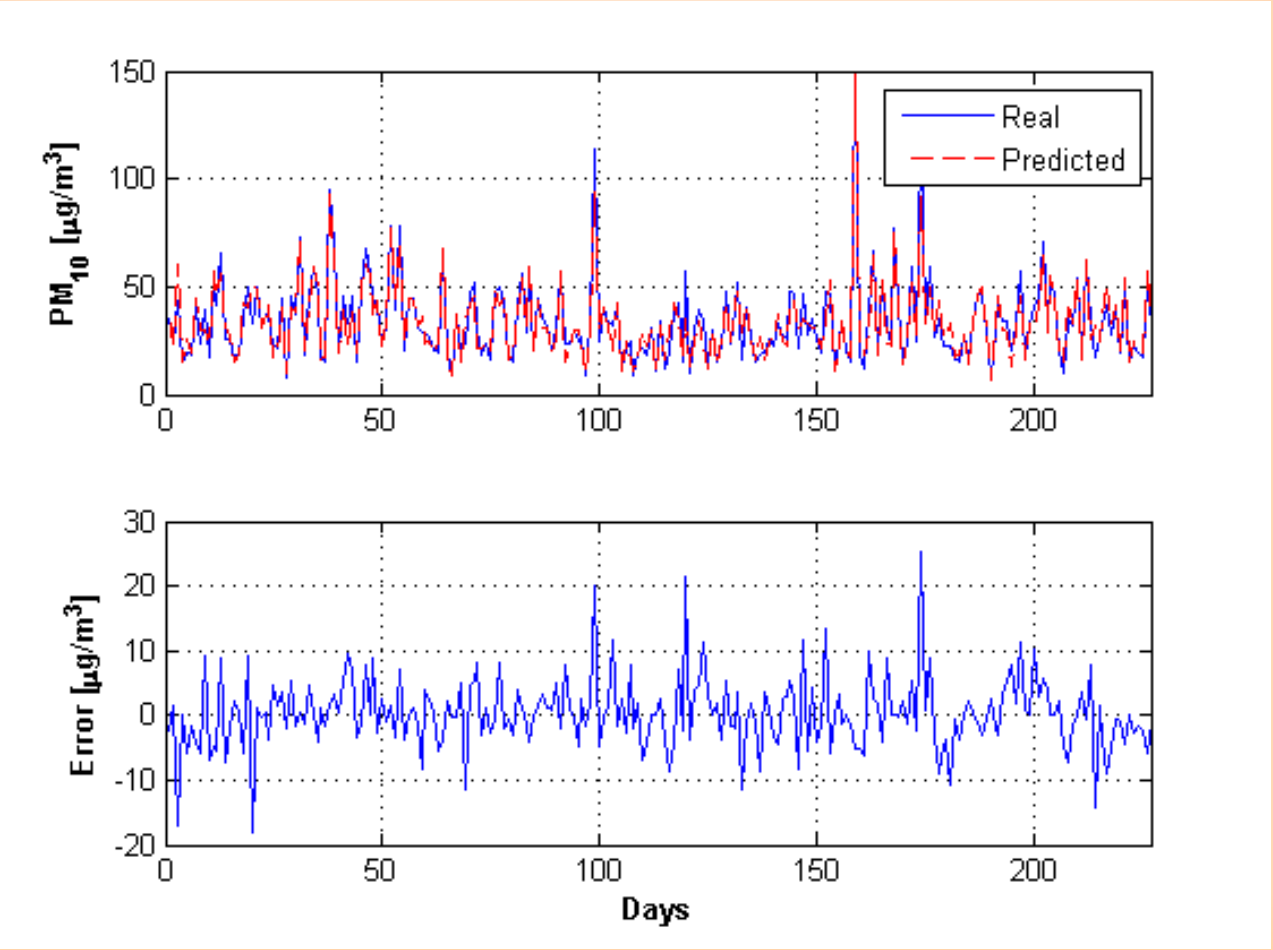
Predicting systems

- ARX linear model of N_a=4 and N_b=1
- Support Vector Machine for regression (SVR) of Gaussian kernel (C=1000, σ=0.8, ε=0.1)
- Procedure of 10-fold cross validation. In each session we have generated randomly the set of learning (500 out of available 871 samples) and testing (227 samples) data.

Statistical results of PM10 prediction

	SVR	ARX
MAE [μg/m^3]	8.66±0.78	10.56±0.99
RMSE [μg/m^3]	14.69±3.71	17.11±3.84
MAPE [%]	27.47±1.39	35.49±3.27
R	0.66±0.09	0.52±0.11
IA	0.73±0.07	0.67±0.11

Graphical results for PM_{10} prediction



Conclusions

- The results presented in this work have shown, that the PM10 distribution represents complex problem belonging to the weakly nonlinear process, not easy in modeling.
- The results of experiments have shown, that to obtain the highest quality of prediction results we should rather apply the nonlinear model of the process, better taking into account the complex relations between PM10 concentration and the basic atmospheric parameters.
- The proposed method has been tested on the data of the meteorological station situated in Warsaw. The obtained results of prediction are in sufficiently good agreement with the actual measurements made at the station.
- The presented approach offers also great potential in other area of modeling of time series.

Thank you