STUDIA UBB CHEMIA, LVIII, 4, 2013 (p. 41-52)
(RECOMMENDED CITATION)

Dedicated to Professor Liviu Literat
On the occasion of his 85th birthday

METAMODELING LEVEL OF POLLUTION BASED ON OPERATING PARAMETERS OF A THERMO POWER STATION

CĂLIN I. ANGHEL[a,*], DANIELA J. CHELARU[a]

ABSTRACT. In process industries objectives as improving performance, reducing pollutant emission or predicting feasible operating conditions requires analysis based on complicated mathematical models and procedures. For quickly but reliable assessments besides of cumbersome approaches of potential support are so-called metamodeling techniques. The paper presents a metamodeling procedure belonging to artificial intelligence in a minimax approach able to assess predictions, trend and correlations to establish a safety-operating domain. The proposed procedure was compared with a robust total least squares regression based on principal component analysis. Numerical experiments are related to dependencies between the level of pollution and operating parameters of a thermo power station.

Keywords: *metamodeling, minimax probability, regression, classification, prediction, harmful level of pollution, operating parameters.*

INTRODUCTION

Many industrial activities produce pollutant emissions that affect the quality of environment and human health. Improving process quality and performance, predicting trends or feasible operating domain are important targets in process industries. Generally solving such problems requires complicated models, expensive analysis and simulation together with cumbersome correlations of many operating parameters and pollutant emissions concentration. In order to reach a high level of accuracy often these analyses becomes computational burden. To address such a challenge,

[a] *Universitatea Babeş-Bolyai, Facultatea de Chimie şi Inginerie Chimică, Str. Kogălniceanu, Nr. 1, RO-400084 Cluj-Napoca, Romania, * canghel@chem.ubbcluj.ro*

approximation-empirical or metamodeling techniques are often used. These models of reduced order developed for expensive simulation process in order to improve the overall computation efficiency are also known as “surrogate” models [1-3]. A metamodel replaces a true functional relationship $g:\Re^n \mapsto \Re$ and know values $y_i = g(x_i)$ at some selected input variables usually called sampling points ($\boldsymbol{X} = \{x_1, . . . , x_m\}, \boldsymbol{X} \in \Re^n$), by an empirical mathematical expression $\tilde{g}(x)$ that is much easier to evaluate. Thus, “surrogates” of the objectives functions can replace the original functional relationship as, $g(x) = \tilde{g}(x) + \varepsilon(x)$. Based on correlated input-output values, parameters of the model are fit to approximate the original data in a best possible way. Among the well-known metamodeling techniques, it can be mentioned: response surface, radial basis function, Gaussian process also known as kriging, high dimensional model representation, artificial neural network, genetic algorithms, support vector machine and many others. The main goal of this paper is to implement a metamodel-based approach able to represent the behaviour of pollution level according some operating parameters into a unit plant. For unity and generality the level of pollution due to the pollutant emission will be assess by the harmful level of pollution. The metamodel based on artificial intelligence methods in a minimax manner is able to assess predictions, trends and dependencies between pollutant emission concentrations and operating parameters. Much more the metamodel can be used into a first step for approximation of optimal domain in which a unit plant can be safely operated. Comparative numerical experiments with a powerful total least squares regression based on principal component analysis (TLS-PCA) are presented. The numerical examples are related to some pollutant emission concentrations expressed by a harmful level of pollution and some operating parameters of an industrial thermo power station. The implementation of the procedure and numerical experiments were developed as a user-friendly computer application in MATLAB language. The results point out the ability of proposed procedure at least into: (1) predicting parameters of interest to facilitate monitoring purposes, (2) approximating trends and correlated dependencies between pollutant emission concentrations and operating parameters to ensure feasible performance.

RESULTS AND DISCUSION

Theoretical foundation

The core of procedure is based on a novel type of pattern recognition machine developed in a minimax manner. The procedure casts both regression (numerical values as outputs) and classification problems (class labels as outputs) into a unified technique. Basic principles of these types of classification approach named minimax probability machine and regression approach named minimax probability machine regression were previously published [4-5]. Some depicted and proved advantages of the implemented procedure over many regression procedures must be mentioned: (a) avoids the specific problems such as over-fitting and local minima, (b) relative less influenced by outliers, (c) provides an explicit direct upper bound on the probability of misclassification of new data, without making any specific distribution assumptions (d) good generalisation ability. Detailed principles also a basic flowchart of implemented procedure and others were previously presented [6-8]. In the present paper only fundamentals principles will be presented.

For predictive purposes, a minimax regression approach was built as in Strohmann and. Grudic [5] by maximising the minimum probability of future predictions to be within some bound (±ε) of the true regression function. Starting from some unknown regression function formally expressed as $f : R^d \rightarrow R$ and

$$f \Rightarrow \mathrm{y} = f(z) + \rho \tag{1}$$

the task is to construct an approximation for **y** as $\hat{\mathbf{y}}$ such that for any

$$\mathrm{z} \in R^d, \ \hat{\mathrm{y}} = \hat{f}(z). \tag{2}$$

To avoid some mathematical limitations based on a kernel formulation minimax probability machine regression model will approximate this function not into a real Euclidean space but into a space of high dimension, named as feature space, by:

$$f \Rightarrow \hat{f} \Rightarrow \hat{\mathrm{y}} = \hat{f}(z) = \sum \beta_i K(z_i, z) + b_k \tag{3}$$

Here $K(z_i, z) = \Phi(z_i) \cdot \Phi(z)$ in the feature space is so-called kernel function satisfying Mercer's conditions. By this kernel function we simply map data from a real Euclidean space into a higher dimensional space named as feature space through a non-linear mapping function Φ(…). In this context, a kernel represents a legitimate inner product into a high dimensional feature space, that is basically a Hilbert space. The others, β_i are weighting coefficients and 'b_k' offset of the minimax regression model, obtained as outputs of the minimax probability machine regression from the learning data. The nonlinear regression function (eq.3) is only a formal basic function formulation. Because Φ(…) is done implicitly, all related computations would be carried by kernel function into a high dimensional feature space. Instead of '*d'* features now '*n*' features represent inputs vectors and the kernel map evaluates at all of the other training inputs. Generating two classes that are obtained by shifting the dependent variable $\pm\varepsilon$ the regression problem was reduce to a binary classification problem into features space. The regression surface is interpreted as being the boundary that separates the two classes, ***successfully*** and ***wrongly*** predicted. Into this feature space, a linear classifier-surface between the two classes of points corresponds to a high non-linear decision-hyper plane into original Euclidean input space. Therefore, a linear regression into the features space corresponds to a cumbersome and high non-linear regression into a real Euclidean space.

If a binary classifier is built to separate the two sets of points (***successfully*** and ***wrongly*** predicted), then finding a crossing point $\hat{y}$ at where the classifier separates these classes for some inputs named as $\mathbf{z} = (z_1,..,z_d)$, is equivalent to finding the output of the regression model for these inputs for any $z_i \in R^d$. Basically as was stated in minimax probability machine by Lanckriet [4] into a binary classification problem of **z** random vectors, with $\mathbf{z_1}$ and $\mathbf{z_2}$ denoting random vectors from each of two classes as $\mathbf{z}_1 \in \text{Class } 1$ and $\mathbf{z}_2 \in \text{Class } 2$, a hyper plane can separates these points, with maximal probability in respect to all distributions having mentioned means $\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2$ and covariance matrices $\sum \mathbf{z}_1, \sum \mathbf{z}_2$. This hyperplane

$$H(\mathbf{w},b) = \left\{\mathbf{z} \middle| \mathbf{w}^T \cdot \mathbf{z} = b\right\}, \ \text{where } \mathbf{w} \in \mathbf{R}^n \setminus \{0\} \text{ and } b \in \mathbf{R} \tag{4}$$

that separates the two classes of points with maximal probability with respect to all distributions must to obey the conditions:

$$\max_{\alpha, \mathbf{w} \neq 0, b} \alpha \quad s.t. \begin{cases} \inf\limits_{\mathbf{z}_1\left(\bar{z}_1, \Sigma z_1\right)} \text{Prob}\left\{\mathbf{w}^T \cdot \mathbf{z}_1 \geq b\right\} \geq \alpha \\ \inf\limits_{\mathbf{z}_2\left(\bar{z}_2, \Sigma z_2\right)} \text{Prob}\left\{\mathbf{w}^T \cdot \mathbf{z}_2 \leq b\right\} \geq \alpha \end{cases}. \tag{5}$$

Related to a minimax probability machine approach the classifier must to minimise this misclassification probability by an optimal separating hyperplane, named minimax probabilistic decision hyperplane. The implemented procedure operates on two data sets: (1) training (learning) data to establish the best model-choosing the best mapping function $\Phi(\ldots)$ and (2) testing to evaluate the errors.

The procedure is conducted in a crude manner, without outliers' detection. To ensure a good distribution of the data the simulations were realised based on data cyclic randomly divided into a number of distinct learning and testing subsets. The errors were estimated by testing rather than by calculation during the training steps (learning and testing) in order to build and estimate the model. To carry out the most basic testing method (simple testing) a random percentage of the database (10-30%) is set aside and used in testing step. The implementation was developed as a user-friendly computer application in MATLAB software and works in multiple cyclic steps ("*k*" cyclic experiments). The performance of procedure was investigated based on the following model validation metrics:

- relative error between the predicted (outputs) and the corresponding test t values

$$RE = \left(\frac{(Y_{predicted} - Y_{test})}{Y_{predicted}} \right) \times 100 \ [\%] \ . \tag{6}$$

- simple equivalent linear dependency between the predicted (outputs) and the corresponding test values:

$$Y_{predicted} = a \cdot Y_{test} + b \ , \tag{7}$$

where *a* and *b* represents the slope and intercept of the equivalent linear dependency model, respectively. Better predictions, means *a* index close to unity and *b* index close to zero value. The performance criterions are evaluated with all values reconverted into the original real Euclidian R^d space. To obtain generalisation the best models will be establish after a number of "*k*" cyclic experiments (simulations). Formally, the best model means model that performs best. It involves best kernel function, kernel parameters and outputs. Basically as previously mentioned Lanckriet, et al. [4] one typically has to choose manually or determine it by tenfold cross validation. This time we preferred a simple-empirical principle for setting the type of the kernel function. The best model over these "*k*" cyclic simulations and the corresponding output values emerged from the procedure was chosen as the *best model*. Long random trials ($k > 100$) do not get improved accuracy and predictions that are more reliable. According some statements [9,10] we limit the trials to $k = \leq 100$. The proper size and selection of the training set (randomly divided into learning and test subsets) is very important to increase the performance of the algorithm. Regarding this, there are no an agreed approached concerning the dimension and the selection of the training set. However, it is a commonly agreed idea, which states that training set must be sufficiently large compared with the number of features/variables.

Case study

This section presents a numerical application of proposed metamodeling technique, related to pollutant emission concentrations and operating parameters of a thermo power station. The acquired data consists of a multivariate set of pollutant emission concentrations measured on the top of an industrial stack of a thermo power station and operating parameters. The statistics of this data set are presented in Table 1. The data set contains 64 daily measured values of operating parameters as: temperature of gaseous releases (T_G) coke content of C, S, humidity, coke specific caloric power and pollutant emission concentrations for SO_2, NO, NO_X, and CO. A general level of pollution from a gaseous release source may be done reporting current values of pollutant emission to the critical concentration C_R representing the value of the 'dangerous concentration' of particular interest. This general level of pollution named harmful level of pollution (*HLP*) were assessed by the following basic relation [7]:

$$HLP = \sum_{1}^{n} C_i / C_{ai}. \tag{8}$$

where, C_i current pollutant concentration, C_{ai} admissible/critical pollutant concentration and *n* the number of considered pollutants.

Table 1. The main values of pollutant emissions and operating parameters

Parameters (operating conditions)					
		Percentage content of coke analyse			*Coke specific*
	T_G [^oC]	*C [%]*	*S [%]*	*Humidity [%]*	*caloric power [kcal/kg]*
Range	269 ÷ 136	20.8 ÷ 19	2.88 ÷ 2.32	26.1 ÷ 23	1875÷1740
Mean value	200.1	20.3	2.50	24.8	1799
Standard deviation	55.18	0.43	0.112	0.78	36.34

Pollutant emission concentrations *[ppm]*				
	SO_2	***NO***	***NO_X***	***CO***
Range	726÷145	134÷54	208÷77	399÷70
Mean value	391.7	97.4	130	152.6
Standard deviation	158.5	22.23	34.52	73
Critical concentration	400	140	230	200

The critical values of polluting emissions are reported to industrial pollutant emissions for solid fuel elements in burning type II installations according to *Romanian HG- 541/2003*

According eqs.(8) an accepted level of pollution means *HLP* close to unity. Numerical application was developed according with this harmful level of pollution (*HLP*). After the model is generated on random training database, it is used to predict on random test database. In other words it is random validated. This time the kernel type that yields to the best performance (eqs.6-7) was an exponential radial basis function with standard width kernel (σ) tuned using 10-fold cross validation. The results were compared with those obtained by a well-known total least squares regression based on principal component analysis (*TLS-PCA*). To ensure a real comparison between these procedures, they were conducted to work on the same learn and test subsets. Table 2 presents the main conditions and results of procedures for the best model obtained after cyclic simulations.

Table 2. The main results and conditions of predictions for best metamodel

	TLS-PCA	**Minimax metamodel**
	Criterions of performance on test data set	
Range of relative errors RE (eq.6)	-140÷40 [%] based on Fig. 2	-5,377÷0 [%] based on Fig. 1
Coefficients of equivalent linear dependency (eq.7)	a = 0.754 b = 1.115 based on Fig. 2	a = 1 b = -0,029 based on Fig. 1
	Criterions of performance on unseen-validation data set	
Range of relative errors RE (eq.6)	-1441÷41 [%]	-5,377÷10 [%]
Coefficients of equivalent linear dependency (eq.7)	a = 1.918 b = -1.746	a = 0.992 b = 0,028

Formal kernel functions – exponential radial basis function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 \Big/ 2\sigma^2 \right)$$

Results related to test data reveal relative good performance for both procedures. Predicted and test harmful level of pollution (*HLP*) are in reasonable agreement because range of relative errors of both procedures presents a good adjustment. However *TLS-PCA* depicts a high range of relative errors than our procedure (Table 2 and Fig. 1-2). Results of predictions on unseen-validation data set (Table 2) are a little worse than those previously reported. Our metamodel retrieves this drawback by the range of relative errors less than 10% which suggest good generalisation capability. The dependency between predicted and corresponding test values illustrated based on simple equivalent linear dependency reveals also reasonable accurate results. Certainties of predicted values are established in a 95 % confidence interval. Regarding the robust regression technique, the results (Table 2) suggest a poor generalisation capability. By these reasons, only our metamodel was utilised into new correlated predictions to assess dependencies between the harmful level of pollution (*HLP*) and some operating parameters of a thermo power station (Fig. 3-4). Until now, all predictions reflecting dependencies and trends are correlated. Based on proved generalisation capability of our metamodel we extend a well-known principle of artificial neuronal network as in [11] to examine the effect of any individual input of interest on the output variable. This means to establish an uncorrelated functional dependency. This may be very difficult in reality or in some cases impossible to do in other way.

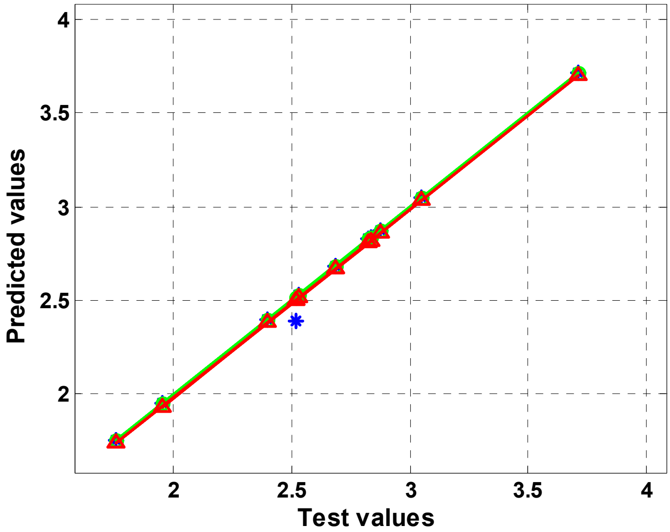


Figure 1. The performance of metamodel on test set.
red ∇ **-** simulated values; blue ∗ - test/predicted values;
green **o** – ideal hypothetical simple equivalent linear dependency

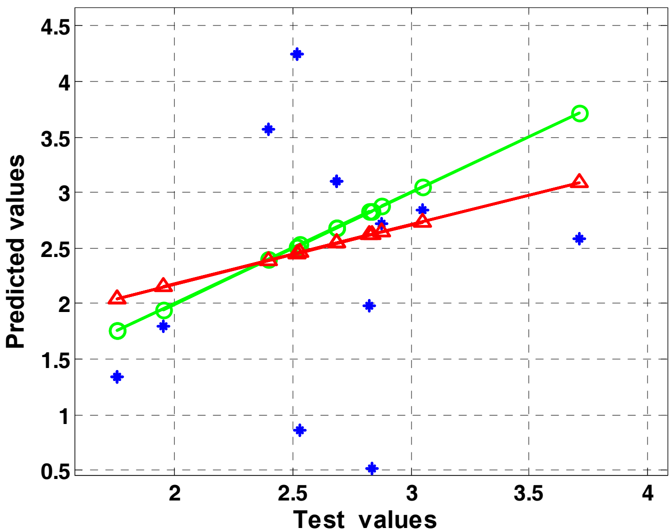


Figure 2. The performance of robust regression technique on test set.
red ∇ **-** simulated values; blue ∗ - test/predicted values;
green **o** – ideal hypothetical simple equivalent linear dependency

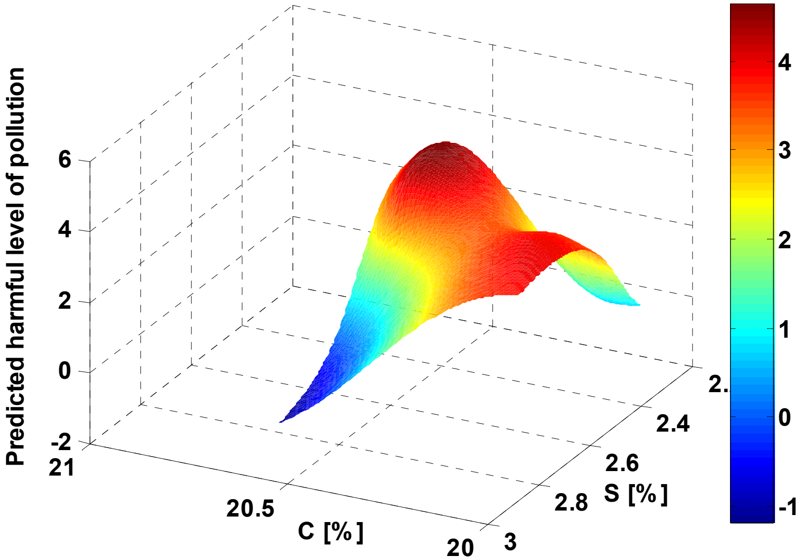


Figure 3. Correlated predicted dependencies between HLF and (C, S) content.

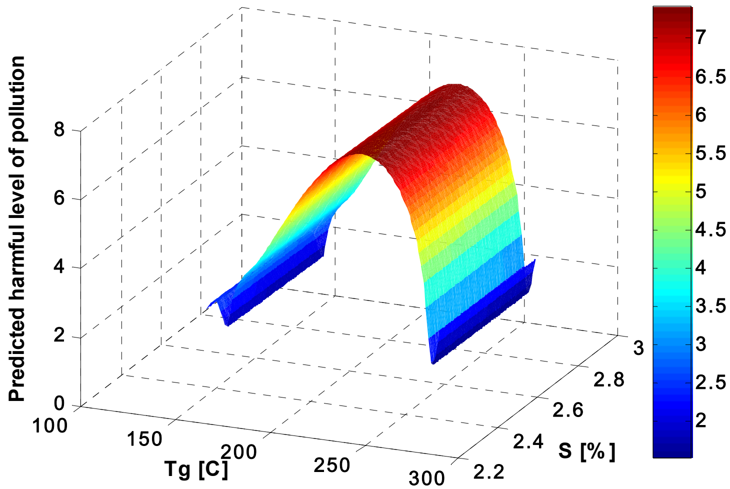
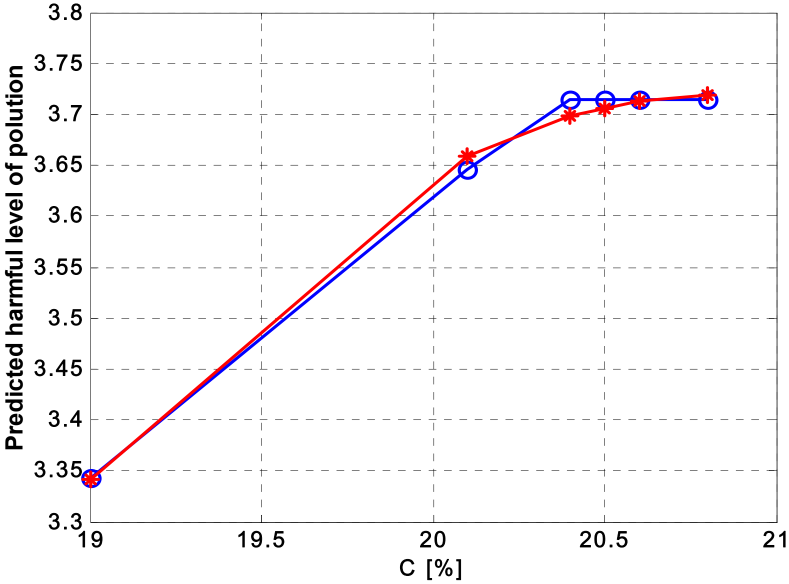
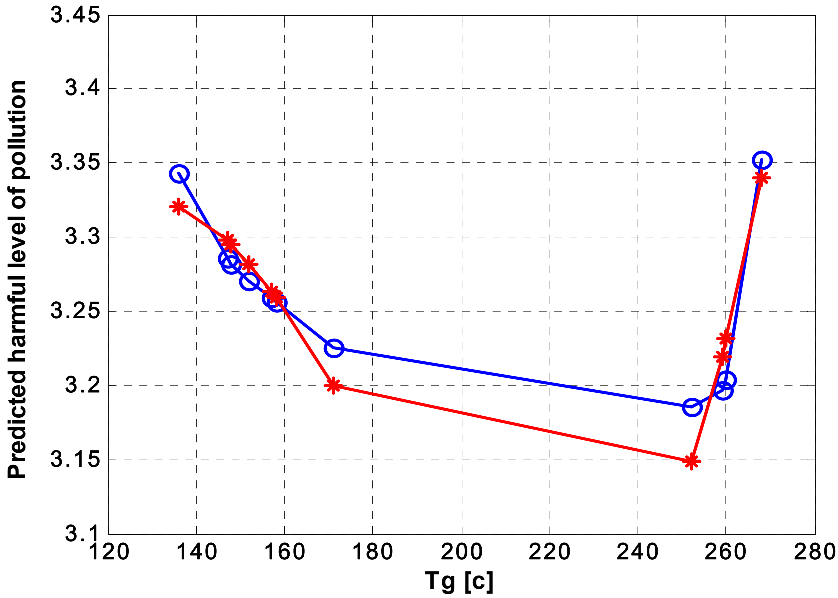


Figure 4. Correlated predicted dependencies between HLF and (T_G, S) content.



a.



b.

Figure 5. Uncorrelated functional dependency between harmful level of pollution (*HLP*) and some functional parameters.
blue **o** – predicted values of *HLP*; red ∗ - polynomial fitted values.

Analysing correlated dependencies (Fig. 3-4) between harmful level of pollution (*HLP*) and some operating parameters we can identify that correlated dependencies of T_G, S produce a level of pollution higher than those due to C, S. Based on these correlated dependencies and others many times it is possible to adjust some operational parameters to decrease the dangerous level of pollution.

Regarding the effect of an individual input of interest on the harmful level of pollution the solution involves the use of the best metamodel (that preserves the best non-linear interactions between all variables) to perform virtual experiments. These virtual experiments predict outputs variations (Fig. 5) with any individual input variable (keeping the values of the other input variables at constant values within their range-usually at their nominal-mean values in the training data). This time inputs of interest were chosen operating parameters as: coke content of C (Fig. 5a) and temperature of gaseous releases T_G (Fig. 5b). As a result, our metamodeling procedure can be promoted as a valid procedure for simple monitoring activities or quickly assessment of feasible operational parameters of a thermo power station into a reasonable level of pollution. It is clear by the values of harmful level of pollution (*HLP*) that basically thermo power station works in a wrong way. Therefore future developments based on optimisation tasks need to establish a feasible and safety operating domain.

CONCLUSIONS

The paper presents the feasibility of applying a novel metamodeling procedure to estimate dependencies of concentration of pollutant emissions generalised into a harmful level of pollution and operational parameters of a thermo power station without the relational physics or analytic being explicitly or known. By principle this metamodeling procedure, represents a pattern recognition and classification task. The procedure casts both regression and classification problems into a unified technique able to predict trends, correlations and various dependencies. The procedure performs highly complex mappings into high dimensional space on nonlinearly related data, by inferring subtle relationships between input and output parameters. Used it with an appropriate policy related to environmental protection it could leads to a drop from industrial pollutant emissions and hence to decrease the atmospheric level of pollution.

This procedure can be a promising alternative for engineers dealing with risk pollution to work out optimal feasible operating conditions according to the regulations and beyond this one to other engineering domains. Because of these, we consider opportune to promote our procedure as predictive tool towards real engineering applications at least as a simple preliminary investigation before any cumbersome operational optimisation of process.

REFERENCES

1. T. Chena, K. Hadinoto, W. YAN W, Y. MA, Computers & Chemical Engineering, **2011**, *35*, 502.
2. N.V. Queipoa, R.T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P.C. Tucker, *Progress in Aerospace Sciences*, **2005**, *41*, 1.
3. G.G. Wang, S. Shan, ASME Transactions, *Journal of Mechanical Design*, **2007**, *129*, 370.
4. G.R.G. Lanckriet, E.L. Ghaoui, C. Bhattacharyya, M.I. Jordan, *Journal of Machine Learning Research*, **2002**, *3*, 555.
5. T.R. Strohmann, G.Z. Grudic, in: S. Thrun, S. Becker, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, **2003**, *15*, MA: MIT Press, Cambridge, 769.
6. C.I. Anghel, Al. Ozunu, *Chemical Paper,* **2006**, *60*, 410.
7. C.I. Anghel, *Revista de Chimie*, **2006**, *57*, 773.
8. C.I. Anghel, M.V. Cristea, *Revista de Chimie*, **2010**, *61*, 87.
9. A. Bordes, S. Ertekin, J. Weston, L. Bottou, *Journal of Machine Learning Research*, **2005**, *6*, 1579.
10. Y-J. Lee, O.L. Mangasarian, RSVM: Reduced Support Vector Machines, Computer Sciences Department, University of Wisconsin, Madison, WI 53706, **2000**, *olvi@cs.wisc.edu.*
11. H.K.D.H. Bhadeshia, *Statiatical Analises. Data Mining*, **2009**, *1*, 296.