CUNOŞTINȚE FUNDAMENTALE PE BAZE INDUSTRIALE: APLICAȚIE. PARTEA a II-a – PREDICȚIE
KNOWLEDGE DISCOVERY IN INDUSTRIAL DATASETS: APPLICATION. PART II - PREDICTION

ZENO GHIZDĂVEȚ*
Universitatea POLITEHNICA Bucureşti, Str. G. Polizu nr. 1, cod 011061, sector 1, Bucureşti, România

Based on the results of a comprehensive analysis made in the first part of the article, here we are handling the issue of developing prediction models on a large industrial dataset. The outcome of the first part is to ascertain, through a careful analysis, which of the process parameters will be selected as regressors and whom will be predictors. In this second part of the article they will be used in several models and the results will be assessed by their efficiency. Though formally distinct, these two parts should be read together, as they are mutually linked together. In practice, continuous exchange of information should be considered between these two parts.

Pe baza rezultatelor analizei complexe din prima parte a articolului, în această lucrare sunt dezvoltate o serie de modele de predicție pentru o bază de date industrială de dimensiuni mari. Astfel, rezultatul general al primei părți este de a stabili pe baza analizei atente, care dintre parametri vor fi selectați ca mărimi dependente şi care independente. În această a doua parte a articolului, aceşti parametri vor fi utilizați în cadrul mai multor modele iar rezultatele vor fi evaluate prin prisma preciziei. Deşi sunt despărțite – în mod formal – aceste doua părți trebuie lecturate împreună, datorită conexiunilor intrinseci existente între ele. În mod practic trebuie considerată necesitatea unui flux continuu, bidirecțional, între aceste două părți.

Keywords: *time series, prediction, industrial processes*

The ultimate target of knowledge gathering is to be able to predict a future evolution of the system on study, by developing models based on existing data, relevant upon the considered dependant variable(s). To simulate this situation, the available dataset (given in the part I of the article) was split in two parts: a training set and a validation set. Validation sets' dimensions were even 192 or 350 data records, depending on the prediction application used and on the way validation records were extracted. For example, in *Multiple Linear Regression, Partial Least Squares Regression, ANN Function Approximation* and *ANN Time Series prediction* we used 192 consecutive validation records while in *ARIMA Time Series prediction*, *Decision Trees* and *Rule Induction* were used 350 validation records extracted with a given lag.

Some of these techniques – such as *Multiple Linear Regression*, *ANN Function Approximation* – were used in conjunction to analytical techniques, to augment their efficiency or to confirm the results while others were employed only to test their efficiency. A typical approach to improve prediction efficiency is to plot both observed and predicted values and to manually/automatically extract records that fall far away from the statistical cloud, yet care should be taken, as stated. For example, DIRM variable in **Case 2B** has not its best R^2 value however the mentioned plot has the best shape of the statistical cloud in figure 9a. At, somehow, the opposite side lies DCX prediction in **Case 2D**: the statistical cloud is not as balanced (yet it does not look too bad) especially in the range of the highest observed values (6 records could be easily assigned with low prediction accuracy) while the R^2 value is the best over all cases and, intriguingly, it depends on the existence of these six records.

Partial Least Squares Regression (see [1-3] for more details about the method and its applications) was used in **Case 2C** in two ways: first, by having a single target function that was DIRM or DCX and, secondly, a model with the two selected target functions (DIRM and DCX), in the same time. For the last variant (with two response-variables), the cross-validation method was used to enhance fitting; results from training were used to validate a new set of observations and these results, at their turn, were used to improve training efficiency and so on. Equations for all single response-variable cases are given in table 4. A particular attention will need the double output (two response-variables) *PLS regression*.

The general form of the PLS model is:

$$X = T \cdot P^T + E \quad (8)$$

$$Y = T \cdot B + F \quad (9)$$

where

* Autor corespondent/Corresponding author,
Tel. 0040 21 402 38 74, e-mail: zghizdavet@gmail.com

Table 4

Regression equations/*Ecuații de regresie liniară multiplă*

CASE/CAZ	EQUATION/RELAȚIE	Eq. No.
1	*DIRM = 55.8347 - 0.0843·CM - 9.3080·COB + 0.0562·CON - 1.7060·DCX + 0.0011·NO_xA - 2.1443·O_2B - 0.5517·PE - 0.0063·TAE - 0.0039·TAS - 0.0752·TGVRA*	(1)
2A	*DIRM = 54.4913 - 0.0749·CM - 9.3737·COB + 0.058·CON - 2.0249·DCX + 0.0008· NO_xA - 2.2338· O_2B - 0.4895·PE - 0.0067·TAE - 0.0028·TAS - 0.0726·TGVRA*	(2)
2B	*DIRM = 51.0212 - 0.0831·CM - 9.7369·COB + 0.0551·CON - 1.9750·DCX - 2.2040·O_2B - 0.5137·PE - 0.0708·TGVRA*	(3)
2C single target function and 10 predictors *o singură funcție obiectiv și 10 mărimi independente*	*DIRM = 46.15 - 0.1015·TGVRA + 3.3796E-03·TAS - 1.1974E-02·TAE - 7.3139·COB - 1.5120·O_2B + 2.6011E-04·NO_xA - 0.7642·PE + 0.0742·CON - 0.1325·CM* *DCX = 3.3374 + 1.43912E-02·TGVRA - 8.7356E-04·TAS + 2.0675E-03·TAE + 1.2211·COB + 0.1281·O_2B - 1.8267E-04· NO_xA + 0.1007·PE - 1.1710E-02·CON + 2.1667E-02·CM*	(4) (5)
2D	*DCX = 7.8853 + 0.0013·CM - 0.3906·COB + 0.0047·CON - 0.1195·DIRM + 0.0001· NO_xA - 0.2901· O_2B + 0.0418·PE - 0.0005·TAE - 0.0017·TAS + 0.0040·TGVRA*	(6)
3	*DIRM = 20.6193 - 0.1257·CM - 6.6091·COB + 0.0700·CON - 2.4376·DCX - 0.4111·PE*	(7)

X – predictors matrix, *Y* – response-variables matrix, *P* - X-Loadings, *T* – Scores, *E* – *X*-residuals, *F* – Y residuals. X-Loadings and Scores for that case are given in figures 10a and 10b.

About *Loadings* and *Scores*: *Loadings* describe the relationships between variables, while *Scores* describe the properties of the samples [1].

The same interpretation - related to the influence of the proximity as in *Principal Component Analysis* - could be used here as well (close variables or records are similar; values show the influence over the particular Principal Component, PC). Note that data records were grouped, before prediction (automatically) within six clusters. The obtained scores have a satisfactory grouping as well, confirming the efficiency of the prediction. No applicable interpretation of the X-loadings could be made on this case other that NO_xA, TAS and TAE are different from the others. In terms of X-loadings interpretation, the other six variables, by showing lower values of the PCs than this group of three (NO_xA, TAS and TAE), should be discarded from the model, because each of these variables are badly accounted for by the PC. This is not correct for all other case studies in our paper – being exactly the opposite – thus calling for the necessity of confirming the results by employing other techniques. A possible reason of why six variables have very low X-loadings could be that, by themselves (without the liaison offered by DIRM and DCX who are extracted as Y-variables this time), they have a small influence over the variability of the data. It could be a point of interest for a further work to study the influence – if any – of the order of magnitude of the variables, when knowing that NO_xA and TAS have the highest values while TAE and TGVRA are in the average range.

Artificial Neural Networks were used also for predicting both, one and two response variables, by considering the rest as independent variables, for all cases. Predicted values were plotted in figures

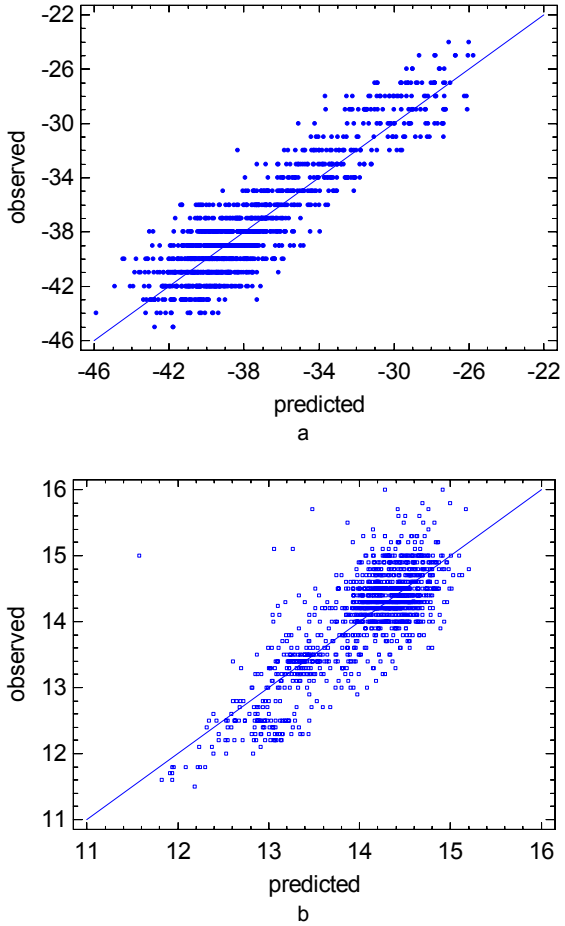
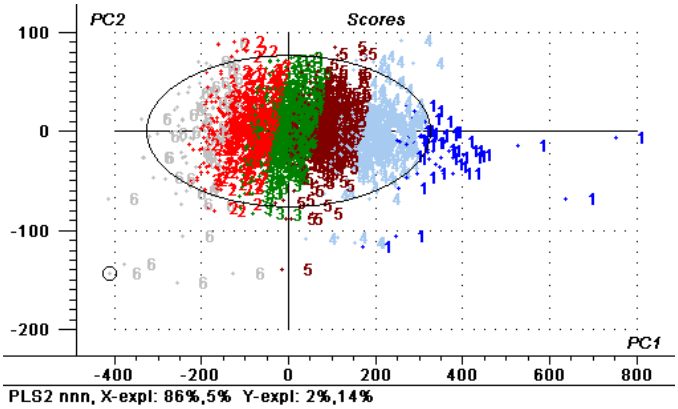
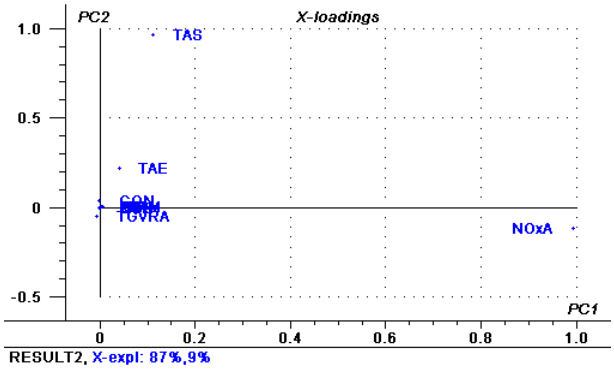


Fig. 9 - Actual and *Multiple Regression* predicted values of: a) DIRM, **Case 2B**; b) DCX, **Case 2D**/Valori de predicție cu *Metoda Regresiei Multiple* şi valori măsurate pentru: a) DIRM, **Cazul 2B**; b) DCX, **Cazul 2D**

11a to 11f along with measured values, for a 95% confidence level. It could be spotted easily that the 95% confidence level is higher when predicting DIRM and DCX in the same time (figures 11d and 11e) than the other ones corresponding to one response-variable prediction (figures 11b, 11c and, respectively, 11f).

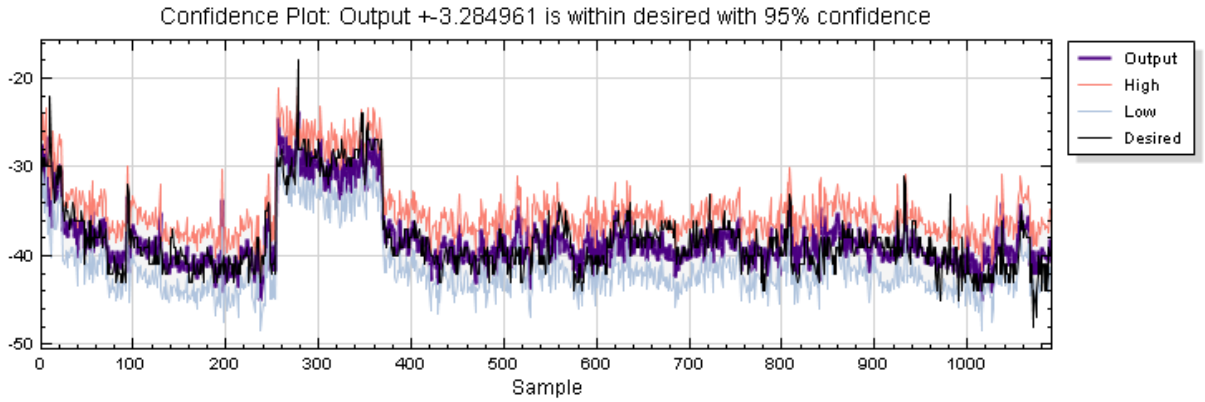


a

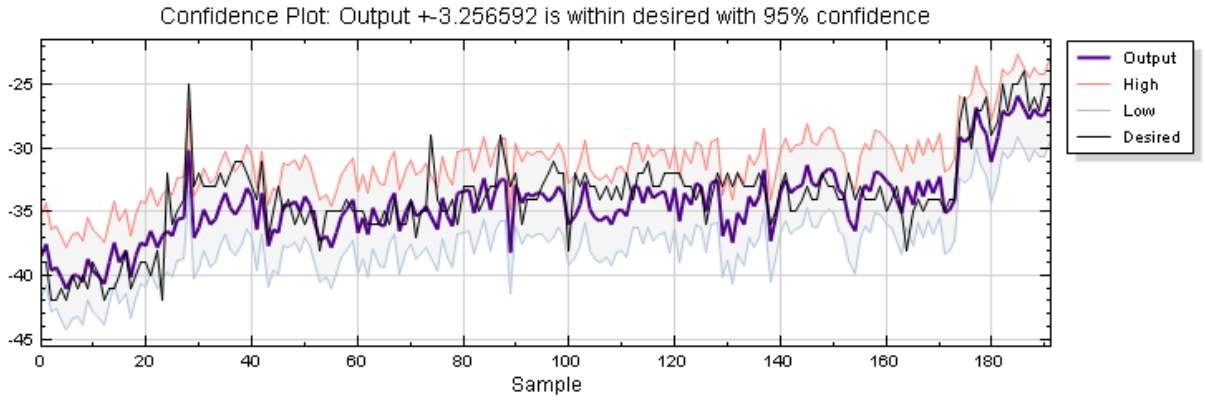


b

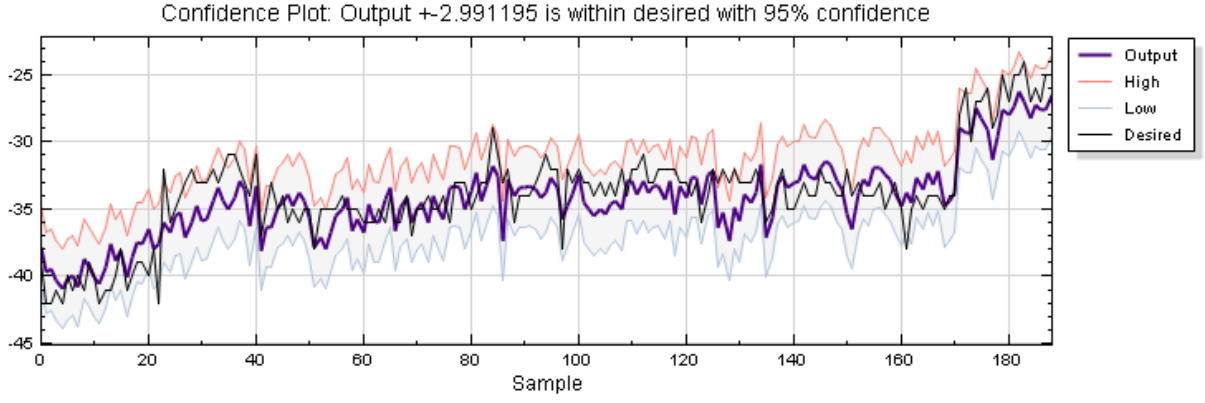
Fig. 10 - Scores, a) and X-Loadings, b), for **Case 2C**, two response-variables PLSR/ Scoruri, a) şi X-Încărcări, b), pentru **Cazul 2C**, regresie PLS cu două funcții obiectiv.



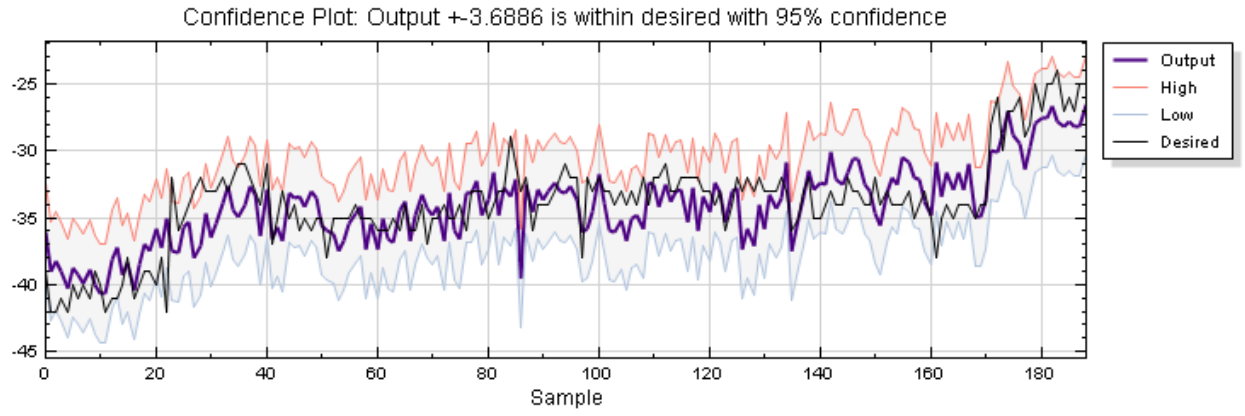
a



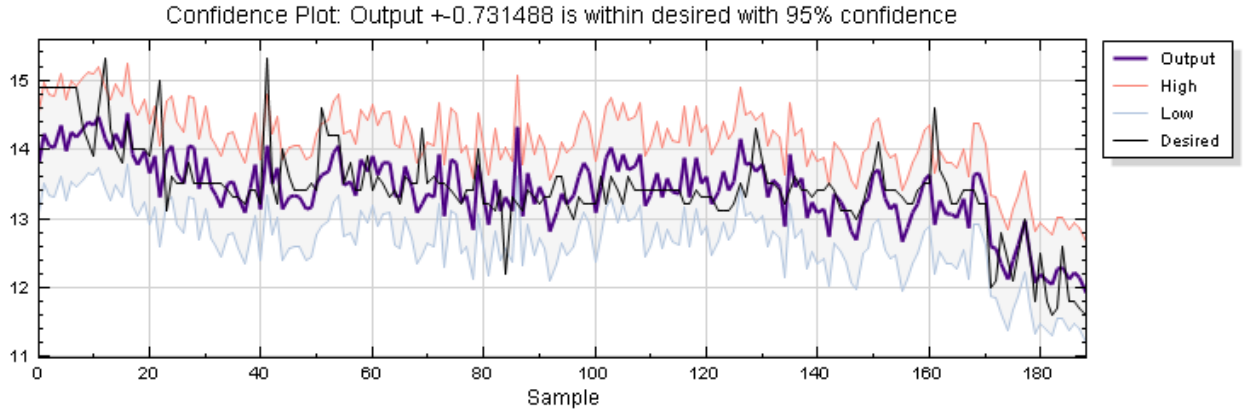
b



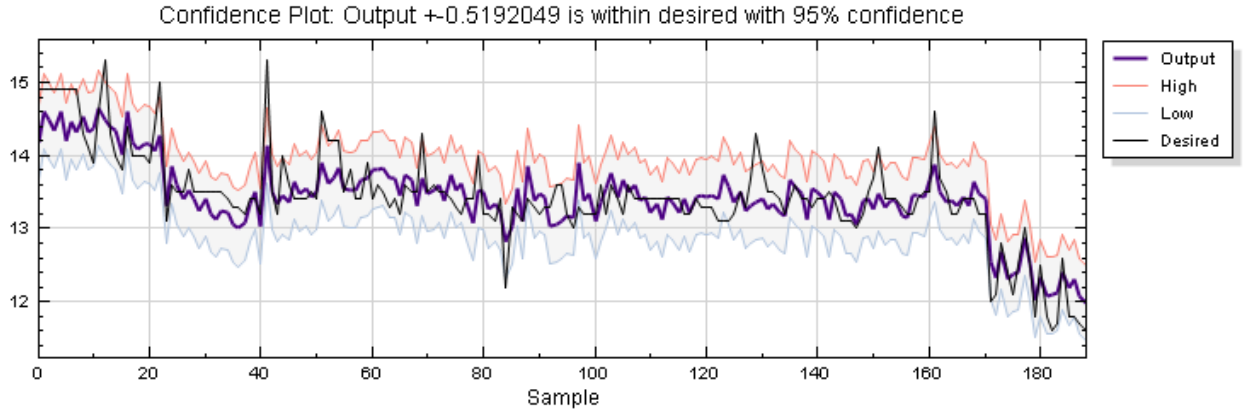
c



d



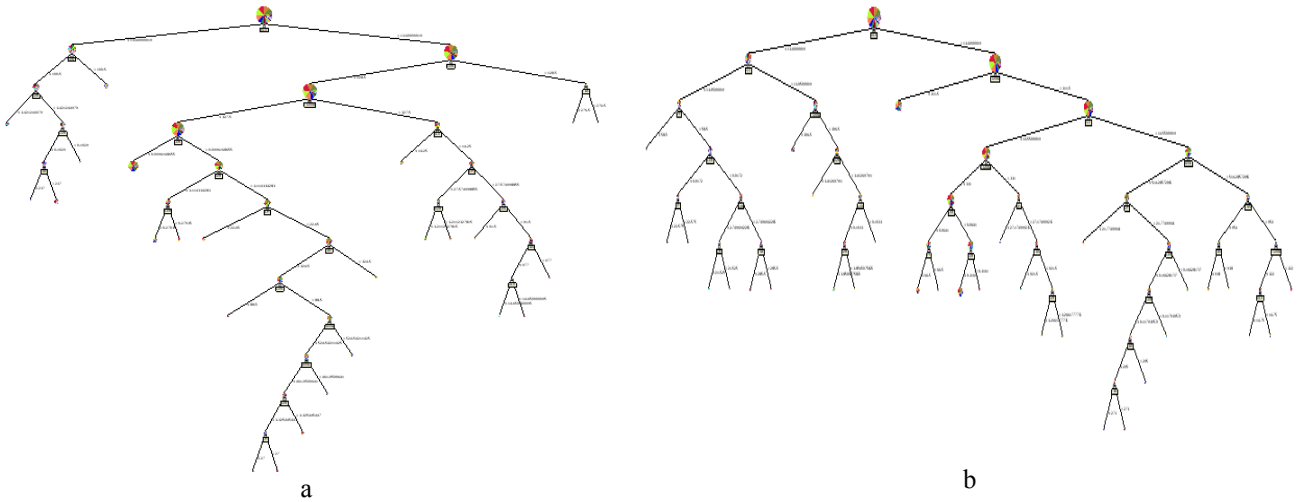
e



f

Fig. 11 - Confidence plot for: a) Case 1, DIRM; b) Case 1, DIRM; c) Case 2A, DIRM; d) Case 2C, DIRM; e) Case 2C, DCX; f) Case 2D, DCX, in ANN function finding, where a) refers to training sets; b) to e) refer to validation sets/*Rezultate de predicție în aproximarea funcțiilor cu ANN: a) Cazul 1, DIRM; b) Cazul 1, DIRM; c) Cazul 2A, DIRM; d) Cazul 2C, DIRM; e) Cazul 2C, DCX; f) Cazul 2D, DCX*

Decision trees (called Trees), ***Ensemble of Trees*** (called EOT) and ***Rules induction*** (called Rules) where used to predict DIRM in **Cases 1** and **2B** and DCX in **Case 2D**. Examples of obtained Trees are given in figure 12 while some examples of Rules are: ***RULE 15***: *IF TGVRA > 311.5 AND TGVRA ≤ 331 AND COB ≤ 0.0641 AND PE > 66.5 AND DCX > 13.60 AND DCX ≤ 14.55 THEN PREDICTED = '-40' (covered examples: 4% on training data, 5% on validation data).* ***RULE 16***: *IF TGVRA > 311.5 AND TGVRA ≤ 331 AND COB > 0.0641 AND COB ≤ 0.3161 AND DCX > 13.60 AND DCX ≤ 14.55 THEN PREDICTED = '-40' (covered examples: 15% on training data, 19% on validation data).* ***RULE 2:*** *IF COB ≤ 0.0172 AND PE > 58.5 AND DCX ≤ 13.05 AND CM ≤ 22.57 THEN PREDICTED = '-29' (covered examples: 2% on training data, 2% on validation data).* ***RULE 5***: *IF COB > 0.0172 AND O2 ≤ 2.74 AND PE > 58.5 AND DCX ≤ 13.05 AND CM > 21.52 THEN PREDICTED = '-31' (covered examples: 1% on training data, 1% on validation data).*



a b

Fig. 12 - Grown trees for a) **Case 1,** b) **Case 2B**/*Arbori obținuți pentru a)* ***Cazul 1****, b)* ***Cazul 2B.***

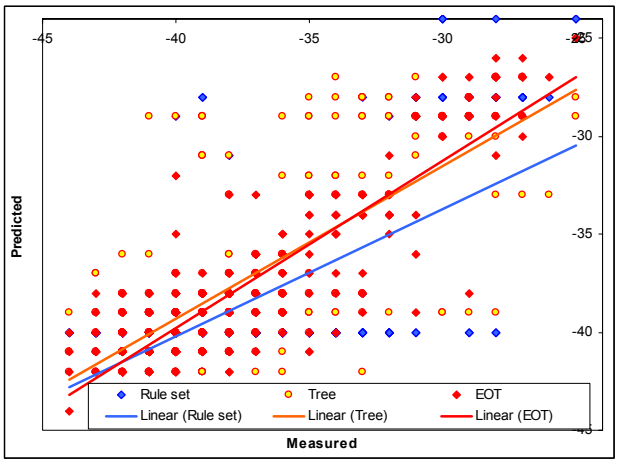


Fig. 13 - Predicted versus measured DIRM values in Trees, EOT and Rules Induction techniques (**Case 2B**)/ *Valori de predicție versus valori măsurate ale variabilei DIRM cu Metodele Arborilor, EOT şi Deducerea de reguli (****Cazul 2B****).*

More information, both theoretical and practical, related to Trees and Rules could be found in [4, 5]. A comparison between results of all three techniques is presented in figure 13. A simple visual analysis of the figure shows the poor correspondence between the sets in each predicted-measured pair, especially for Rules prediction. The complexity of some industrial systems – but not only – could impede, sometimes, on the accurate selection of the predictors that most influence the target(s) function(s). Sound correlations could be, therefore, difficult to extract in these cases, hence the need to predict a future evolution by analyzing only the past of the very target function. *Statistical Time Series* methods and *ANN Time Series* prediction were used in the paper. From the first category, an *Autoregressive Integrated Moving Average* (ARIMA) model has been selected. The best forecast model was selected by comparing the values of the *Akaike Information Criterion*. Results from **Cases 1** and **2D** are plotted against actual (measured) values in figures 14a and b.

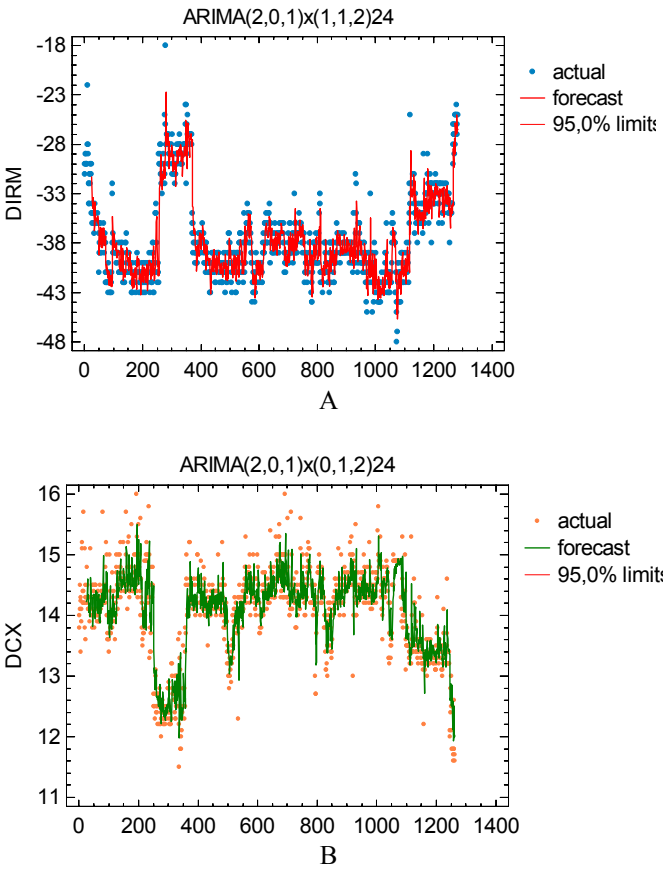


Fig. 14 - Time series predictions by ARIMA models: a) **Case 1**, b) **Case 2D**/ *Serii de timp – modele ARIMA: a)* ***Cazul 1****, b)* ***Cazul 2D.***

The corresponding ARIMA model summaries (see [6-9] for more information on these models about Time Series, in general) are:

a) Case 1, AR(1) = 1.40023, AR(2) = -0.404244, MA(1) = 0.747401, SAR(1) = 0.780023, SMA(1) = 1.84468, SMA(2) = -0.857715 and

b) Case 2D, AR(1)= 1.38738, AR(2)= -0.395685, MA(1)= 0.811392, SMA(1)= 1.03925, SMA(2)= -0.0806513

Seasonal differences of order 1 were taken.

Table 5

Comparison of the prediction efficiency for all prediction methods used in the paper
Comparație între eficiența predicției pentru toate metodele utilizate în articol

METHOD	CASE		MAE		STDEV	
			TS	VS	TS	VS
MULTIPLE REGRESSION	1		1.5503	1.5669	2.0234	2.3927
	2A		1.4590	1.4812	1.82937	1.9031
	2B		1.3627	1.4713	1.4719	1.4876
	2C (PLS)	DIRM	2.1132	2.4081	2.1558	2.1197
		DCX	0.5038	0.5314	0.5072	0.5383
	2		0.3401	0.3303	0.4398	0.4447
	3		1.8012	1.7574	2.3425	2.2213
ARIMA TIME SERIES	1		1.0737	1.3715	2.4359	2.2281
	2A		1.0289	1.2776	2.3979	2.2210
	2D		0.2565	0.2462	0.3717	0.3491
ANN FUNCTION APPROXIMATION	1		0.1591	0.4560	1.9987	1.9029
	2A		0.1678	0.4633	1.7630	1.7444
	2B		0.1836	0.5100	1.8517	1.7233
	2C	DIRM	0.1612	0.4837	1.8825	2.3122
		DCX	0.0110	0.0317	0.4991	0.4443
	2D		0.0043	0.0127	0.4865	0.3159
ANN TIME SERIES	1		0.0326	0.0794	0.4872	0.4019
TREE	1		-	1.6444	-	1.5065
	2B		-	1.6893	-	1.4510
	2D		-	0.3952	-	0.3857
RULES INDUCTION	1		-	1.9746	-	1.9416
	2B		-	3.2425	-	3.4944
	2D		-	0.4778	-	0.5304
ENSEMBLE OF TREES	1		-	1.4095	-	1.3427
	2B		-	1.4032	-	1.3431
	2D		-	0.2882	-	0.2959

Notations: MAE - Mean Absolute Error, STDEV - Standard Deviation, TS - Training Set, VS - Validation Set

Along with the graphical presentation of the prediction results and the R^2 values in table 1 for all regression type techniques all predictions were also assessed by their errors. *Mean Absolute Error* and the *Standard Deviation* of the errors were synthesized in table 5 for both training and validation sets, providing an overall, further way of evaluating the accuracy of the developed models. Consequently, *Rules Induction* used in **Case 2B** gave the worst response while the best was obtained for the *ANN Time Series* prediction. However, for some techniques, quantitative results should be corroborated with qualitative ones, meaning, for example, for all – *Trees, Rules* and *EOTs* – to analyze also figure 13 as well. Except for the worst case mentioned, the rest of the results could be compared, as values, with some others, take for example the ones coming from Regression methods. The difference is, in Regression predictions the STDEVs are higher while the errors distribution is normal. Trees, Rules and EOTs provide, along with comparable MAEs, a comforting STDEV; on the other hand, without the aid of the measured-predicted plots (or errors distribution study), one could hardly detect that these techniques provide a mainly constant prediction value for several subsets of data (i.e. a subset of consecutive – yet not only – different measured values correspond in many situations to only one prediction value).

By assessing all results within the paper, it could be stated without any doubt that ANN techniques (for function approximation, but also for time series prediction) gave the best outcomes for both target functions. Their training efficiency could be easily improved by monitoring the residuals over the number of training epochs and consequently by selecting the best network configuration, either manually or automatically via the aid of *Genetic Algorithms*.

Conclusions

Analysis and prediction results depend on both dataset dimension and on the selection of the proper technique. A better insight that infers more

analysis techniques, corroboration of the results coming from several techniques (also confirmation of the analysis with prediction) and their interpretation in different presentation forms, errors analysis and playing *what-if* scenarios could provide more precise predictions. The major drawback when using these techniques - but not only - is the amount of results to be interpreted and, sometimes, but the most important, is the sharp difference between results obtained when little modifications were made over the dataset. Along with the vast number of combinations of variables and records in the original dataset, inaccurate judgments of how variables/records influence the variability of the dataset can be easily adopted, especially encouraging a direction that mostly corresponds to the expectations based on earlier experience. In the particular case of this paper, in an early phase of the work it was expected – say, it was imposed – the *Kiln Torque* as the target function, with very little reward. It is well known that *Kiln Torque* is highly correlated to the *free CaO*, thus constituting a good indicator of the clinker quality and an appealing target function. [In this particular case, *Kiln Torque* may be correlated with the other variables but with different time delays for each pair *Kiln Torque* – other variable, as gases and material needs time to pass through the clinker plant. Also, its sensitivity – that can be easily observed on site – indicates that, probably, here, a better approach could be to use instant values, not averaged ones and to accurately evaluate the time delays which are also dependant on particular conditions at a given moment of time – so being a very difficult to up to an impossible task to complete, for the entire plant]. Efforts progressed only when the target functions were isolated by means of analysis and not arbitrarily set. An increase in accuracy of some prediction techniques does not necessarily be recorded when using others (efficiency of the *Multiple Regression* and of the *ANN Function Approximation* varies inversely for **Cases 1** to **2A** and than to **2D**) that makes outliers removal an intricate procedure, depending not only on the selection of the variable(s) and/or record(s) but on the technique as well. In that light of very thin dependencies, knowledge discovery in datasets should pay respect also to the selection of the industrial site and to the seasonality. That makes the process highly dependant on situation, suggesting that it should be approached with attention to details. Uniqueness of the solution was not investigated, and will constitute the subject for a further research.

Knowledge discovery in an industrial dataset was targeting, here, both explicit and tacit information (see [10] for more details on these concepts) but with no respect to potential crucial knowledge identification, as this last one has to be related to the common/unique solution, yet to be proofed. However, the results could be used to improve human expertise, mainly in the areas where this one could be hardly formalized and capitalized.

Software packages used in the paper: Centurion Statgraphics for statistical Time Series prediction, XLStat for PLS regression on 1 response-variable, CAMO The Unscrambler for PLS regression on 2 response-variables, Peltarion Synapse for ANN Function Approximation and ANN Time Series prediction, Compumine Rule Discovery System for Tree, EOTs and Rule Induction.

REFERENCES

1. xxx, The Unscrambler Appendices: Method References, 2008
2. V. Lengard, M. Kermit, 3-way and 3-block PLS Regression in Consumer Preference Analysis, Food Quality and Preference, 2004, **17** (3-4), 234 .
3. Larsen O.V, Williams I, Lillelund A.C, Aastrup S, Byrne D.V, Practical application of multivariate statistical analysis for evaluation of sensory and process data from full-scale production, Technical quarterly - Master Brewers Association of the Americas ISSN 0743-9407, 2003, **40** (3), 193.
4. C. Apte, S.M. Weiss, Data Mining with Decision Trees and Decision Rules, Future Generation Computer Systems, November 1997.
5. F.P. Pach, J. Abonyi, Association Rule and Decision Tree based Methods, for Fuzzy Rule Base Generation, World Academy of Science, Engineering and Technology 2006, 13.
6. Ajoy K. Palit, Dobrivoje Popovic, Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications, Birkhäuser, 2005.
7. F. Hoppner, Learning Dependencies in Multivariate Time Series, Knowledge Discovery from Temporal- and Spatio-Temporal Data", European Conference on Artificial Intelligence (ECAI), Lyon, France, July 2002.
8. J. Kout, T. Vlcek, J. Klema, Predictive system for multivariate time series, AUTOMA International, 2005
9. Xiang Xuan, K. Murphy, Modeling changing dependency structure in multivariate time series, Proceedings of the 24th international conference on Machine learning, Corvalis, Oregon, 2007, 1055.
10. Ines Saad, Salem Chakhar, A decision support for identifying crucial knowledge requiring capitalizing operation, European Journal of Operational Research 2009, 195, 889.

****