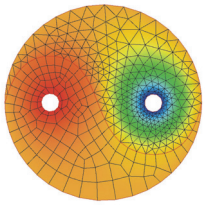
Publishing House
AKAPIT

COMPUTER METHODS IN MATERIALS SCIENCE

Informatyka w Technologii Materiałów

Vol. 9, 2009, No. 3



PRE-PROCESSING OF THE INDUSTRIAL DATA FOR DATA MINING AND MODELLING – APPLICATION TO THE COPPER FLASH SMELTING PROCESS

ANDRZEJ STANISŁAWCZYK, JAN KUSIAK

AGH - University of Science and Technology, Faculty of Metals Engineering and Industrial Computer Science, Al. Mickiewicza 30, 30-059 Kraków
Corresponding Author: astan@agh.edu.pl (A. Stanisławczyk)

Abstract

The paper presents the methodology of the pre-processing of the industrial data (measurements) collected from the copper flash smelting process (Kusiak, 2009). The application of data filtering and the cleaning method for the needs of the exploratory data analysis and modelling has been discussed. The influence of the appropriate data preparation on the quality of the developed model of the considered process has also been presented.

Key words: industrial data filtering, adaptive filtering, outlier detection

1. INTRODUCTION

In the data mining, the statement 'garbage in, garbage out' (GIGO) is widely known (Hand et al., 2001), which means that the use of 'garbage' data in analyses or modelling would result in obtaining poor models or erroneous analysis results. The industrial data is particularly exposed to disturbances, which may consist of human errors or failures of the measuring equipment. Therefore, the process of modelling by using the industrial data should be preceded by the analysis of the data quality, identification of disturbance types and data filtering and cleaning. Errors resulting from erroneous data that are caused by human activities are the most frequent source of disturbances (figure 1). Automatic data acquisition systems are, to a large extent, more resistant to gross error origination. In such systems, the possibility of difficult to detect maladjustment origination is a much more serious problem.

The paper presents an example of the pre-processing for the industrial data related to the analyses of the chemical composition of the copper concentrate as well as other parameters of the copper flash smelting process. The STATISTICA and MATLAB packages were used in the computations.

2. DATA QUALITY ASSESSMENT

The data quality assessment should be the first step after data reading and integration. This assessment may be performed visually based on the graphs drawn. The occurrence of gross errors may turn out to be especially dangerous. Such errors are visible on a box-whiskers graph (Turkey, 1977) in the form of extreme and outlier points.

Examples of the box-whiskers graphs for the chemical composition of copper concentrate are shown in figure 2. Extreme values exist for three variables (Fe, Pb and MgO), which may result from a gross error or accidental occurrence of a rather atypical composition of a concentrate. It would be the best to consult such measurements with process engineers, which have the required expertise to decide the correctness of the registered data.

 ISSN 1641-8581

a)

11.030000	1.990000	0.690000	32.270000	3.880000	12.330000
11.060000	1.960000	0.680000	32.099998	3.820000	12.170000
11.120000	1.970000	1.970000 ←	0.680000 ←	32.310001	12.190000
10.940000	1.950000	0.680000	32.310001	3.870000	12.210000
10.820000	1.960000	0.690000	32.750000	3.940000	12.360000

b)

2005-10-19 09:00:00	1309
2005-10-19 10:00:00	1317
2005-10-19 00:00:00	1314
2005-10-19 14:00:00	1314
2005-10-19 15:00:00	1311

Fig. 1. *Examples of disturbances caused by mistakes of operators (records with errors are highlighted): (a) error caused by wrong placement of the consecutive values in the record, (b) mistake in notation of the measurement time.*

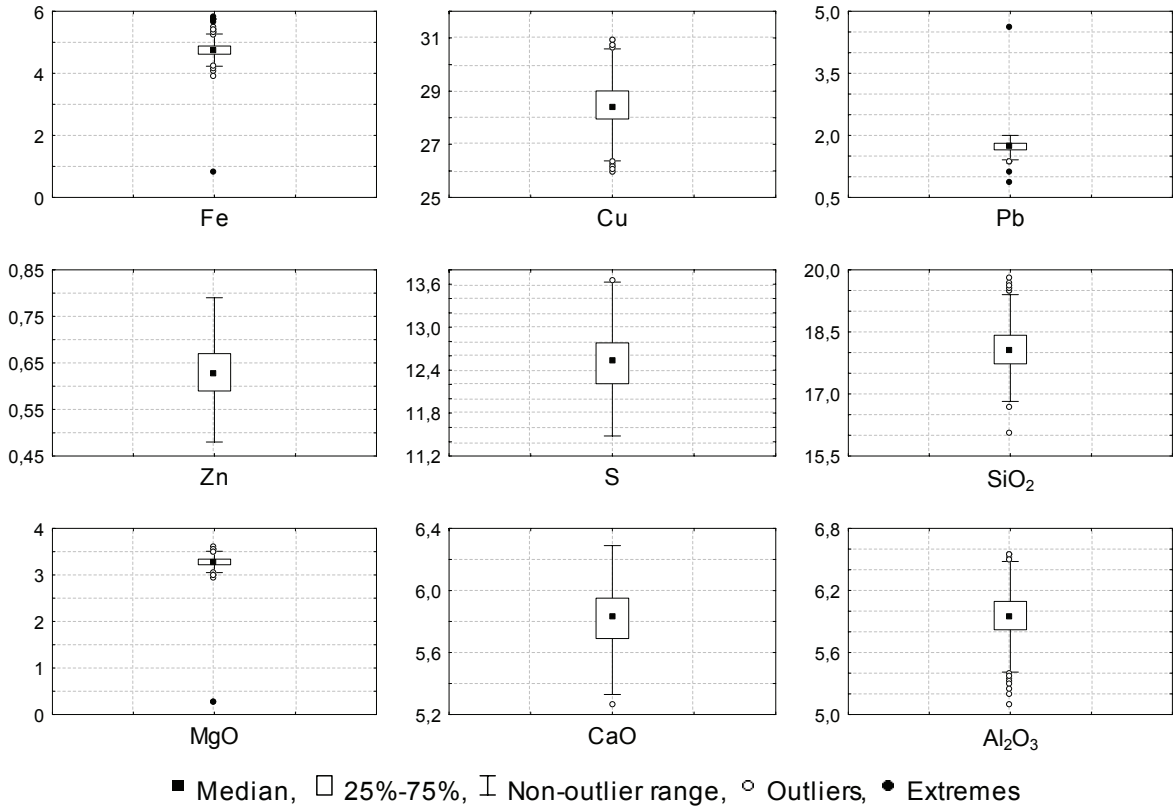


Fig. 2. *Box-whiskers graphs for the analysed variables.*

If the data is in the form of a time series, it also could be presented graphically in form of the line plot. Examples of such graphs, for which the occurrence of extreme values has been found, are shown in figure 3 in which arrows mark the extreme values, which are visible as peaks. This graph suggests that the observed extreme values result from gross error occurrence, which originated most likely as the mistake of typed data.

3. DATA FILTERING

The simplest type of the data filtering is a Min-Max method (Statsoft, 2006). The range of acceptable variation for a specific variable is determined in this type of filtering. If the value of the variable is from outside the permissible range, it may be removed and replaced with a code indicating the lack of the data in that point. The determination of the range of the permissible values of separate variables may be performed based on consultation with an expert or based on box-whiskers graphs' analysis.

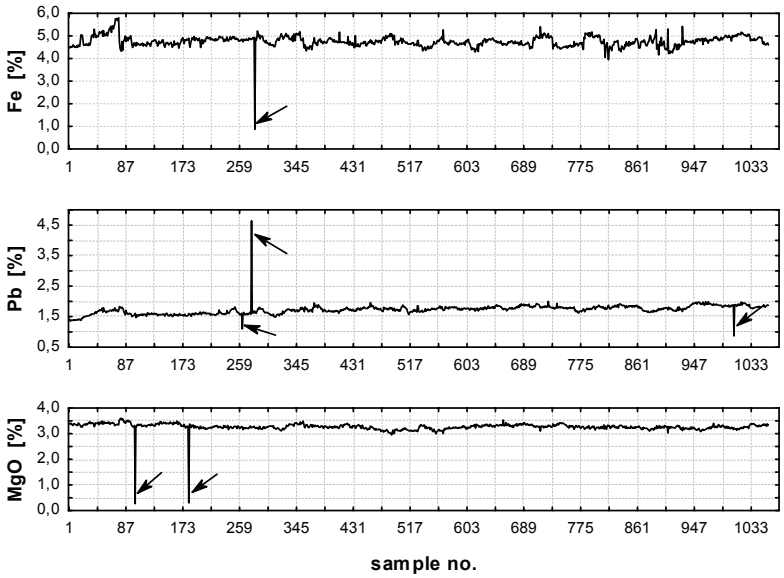


Fig. 3. Graphs of selected variables (arrows indicate extreme values).

If the data is a time series of a strong autocorrelation, the process of filtering may be more effective, when the analysis is carried out using the *one-time data differencing*. One-time data differencing is simply replacing the variable's value with the difference between its consecutive values: $x' \quad x_{i\ 1} \quad x_i$, where: x' – the value of variable x after one-time differencing; i – time step

Figure 4 presents a graph of the one-time differencing of the values of the Pb variable based on the graph presented in figure 3.

In a graph that is prepared in this way, the extreme values are more visible as well as more easily identifiable. The next step in the process of filtering consists of the determination of the threshold value for differenced variables. In general, a triple value of the variable's standard deviation is assumed as the threshold value (Hill & Lewicki, 2007). However, because of extreme points' occurrence for the differenced variables (figure 4), the quantile measures of the variables' variability were used in the present work. The following was assumed as the threshold value t_j of the j-th variable:

$$t_j = 3x_{0.95}\left(\left|z_{i+1}^j - z_i^j\right|\right) \qquad (1)$$

where: $x_{0.95}$ - quantile of the 0.95 order, z_i^j - i-th value of variable z of number j.

The variables' values that satisfy the condition:

$$\left|z_{i\ 1}^j \quad z_i^j\right| \quad t_j \qquad (2)$$

are replaced with the values of a linear approximation, calculated based on adjacent non-outstanding values.

To reduce a random error, filtering with a moving average was carried out in the next step, in which each value of the considered variable is replaced with the mean value calculated for a specified range (the range of a radius r=1 was used in the present paper).

Data graphs for the selected variables, after carrying out the filtering procedure described, are shown in figure 5.

The extreme values, which were the measurement errors with a high probability, were removed as a result of filtering. Data filtered in that way may then be used in the process of data exploratory analysis.

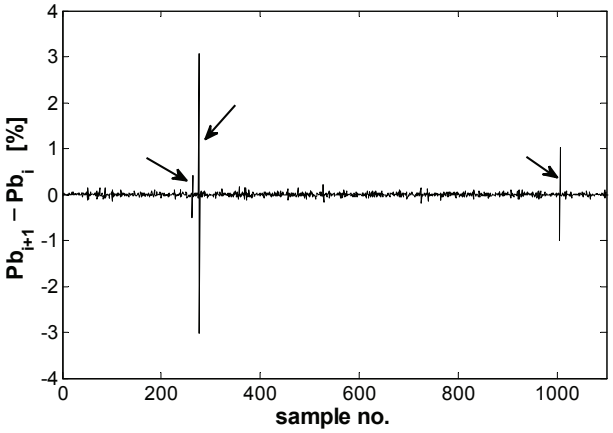


Fig. 4. Graph of the Pb variable (based on graph in figure 3) after one-time differencing (arrows indicate extreme values).

4. ADAPTIVE FILTERING

Another source of the data disturbance is the incompleteness of the measurements, in which one or more significant parameters are not available, e.g. due to the lack of their recording. In the case where an output parameter of a modelled object is subject to disturbance, the necessity of obtaining a model of the assumed accuracy may require the identification and removal of these disturbances from the training data set. The adaptive filtering technique works well in such a situation. The adaptive filtering procedure was described in the paper (Stanisławczyk et al., 2007). An example of a developed adaptive filter is shown in the figure 6.

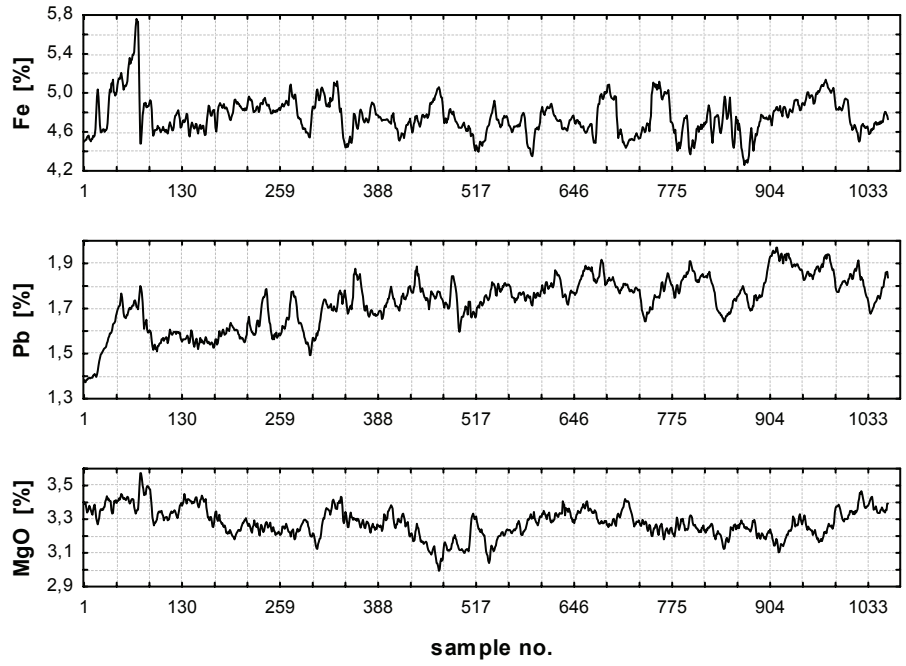


Fig. 5. Graphs of the filtered data of analysed variables.

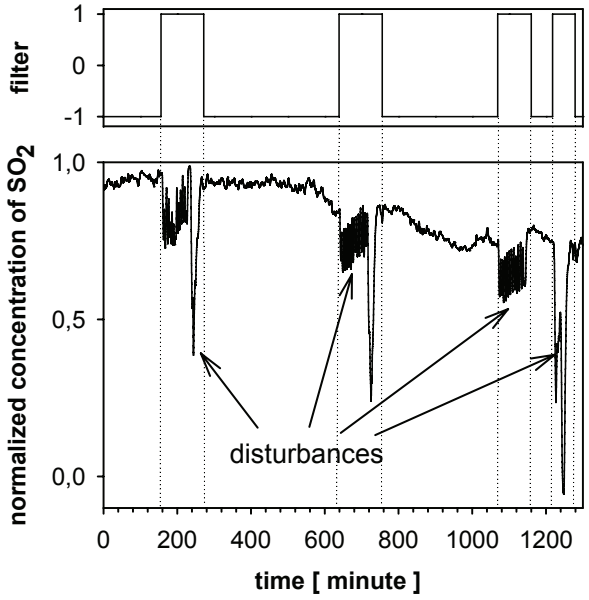


Fig. 6. Graph of SO_2 content in the exhaust gas registered with disturbances (lower graph) and the signal of the adaptive filter output (upper graph).

The adaptive filter identifies those changes in the output parameter that are not related to the changes in the measured input parameters. Most frequently, these changes result from the occurring of disturbances in the analysed output parameter of the object. In the next step, the records that are marked may be removed from the data set. The data cleaned in that way may be used in the modelling of analysed process.

5. DATA FILTERING INFLUENCE ON THE MODEL QUALITY

The influence of the data filtering on the quality of a model of the analysed process was examined by a comparison of two models that were developed based on artificial neural networks. The considered models allow the prediction of the SO_2 content in the exhaust gases of the copper flash smelting furnace. The first model, which was an MLP (Multi Layer Perception) type neural network (Tadeusiewicz, 1993) of the 27-10-1 structure, was developed based on a training set of unfiltered measurement data. The training set comprised 2,000 records, while the the validating data set consisted of 200 records.

The second neural network model (of the same structure) was developed on the base of the data filtered using the elaborated method of adaptive filtering. Disturbed records were removed from the training and validating data sets. As a result, 185 records were removed from the training set and 22 records from the validating set. The results of both models were compared on a test data set of 200 records (figure 7). The test set comprised a data from the later period of the furnace operation, in which the disturbances of the modelled parameter did not occur.

Mean-square error (equation 3) for the first model was equal 0.83, while for the second model, at which the development of the adaptive filtering was utilised, this error was more than four time smaller and was equal 0.20.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^o - y_i^p\right)^2} \qquad (3)$$

where: N – number of data records in a testing set, y_i^o, y_i^p – observed and predicted output values.

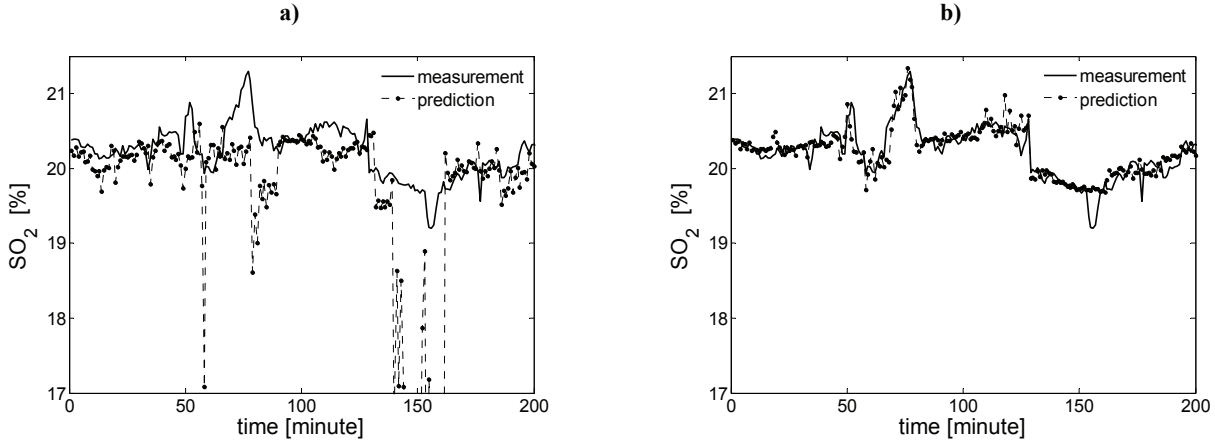


Fig. 7. Comparison of the results of developed models: (a) – model trained on the unfiltered measurement data and (b) – model developed based on the adaptively filtered training data. The solid line marks the measured values and the dotted line indicates the values predicted by the model.

6. SUMMARY

The crucial point in the industrial process modelling is adequate pre-processing and filtering of the measurement data. The first step of the data pre-processing should be the analysis of its completeness and identification of disturbances, especially gross errors. The next step is filtering of the considered data aiming on the elimination of identified errors.

An example of the pre-processing and filtering of the industrial data of the copper flash smelting process is presented in the work. The influence of the lack of the proper pre-processing and filtering on the efficiency of the process model was also presented. Presented example shows, how important is the proper pre-processing of the training data in a case of the process modelling based on the artificial neural network. The quality of obtained results of the developed model trained on filtered data is significantly higher than these obtained using the raw, disturbed data.

ACKNOWLEDGEMENTS

The financial support of the MNiSzW, project No 3 T08B 034 30 is acknowledged.

REFERENCES

1. Hand, D. J., Mannila, H., Smyth, P., 2001, *Principles of Data Mining*, MIT Press.
2. Hill, T., Lewicki, P., 2007, *STATISTICS Methods and Applications*, StatSoft, Tulsa.
3. Kusiak, J., 2009, Copper Flash Smelting process. Modelling and control, *Computer Methods in Materials Science*, 9, 3, 362-368.
4. Stanisławczyk, A., Talar, J., Jarosz, P., Kusiak, J., 2007, Filtering of industrial data using the artificial neural networks, *Computer Methods Material Science*, 7, 1, 311-316.
5. StatSoft, 2006, Elektroniczny Podręcznik Statystyki PL, Kraków. Available at: *http://www.statsoft.pl/textbook/stathome.html* (accessed: January 2009)
6. Tadeusiewicz, R., 1993, *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa.
7. Tukey, J. W., 1977, *Exploratory data analysis*, Addison-Wesley.

WSTĘPNE OPRACOWANIE DANYCH PRZEMYSŁOWYCH DO ANALIZY EKSPLORACYJNEJ I MODELOWANIA NA PRZYKŁADZIE ZAWIESINOWEGO PROCESU WYTOPU MIEDZI

Streszczenie

W pracy przedstawiono metodykę wstępnego opracowania danych pomiarowych otrzymanych w procesie zawiesinowego wytopu miedzi (Kusiak, 2009). Omówiono zastosowane metody filtrowania i czyszczenia danych na potrzeby eksploracyjnej analizy danych i modelowania. Przedstawiono również wpływ odpowiedniego przygotowania danych na jakość zbudowanego modelu procesu.

Received: May 11, 2009
Received in a revised form: May 31, 2009
Accepted: June 15, 2009