Rantala, A.; Virtanen, H.; Saloheimo, Kari; Jämsä-Jounela, S-L

Using principal component analysis and self-organizing map to estimate the physical quality of cathode copper

Published in:
IFAC PROCEEDINGS VOLUMES

DOI:
10.1016/S1474-6670(17)37020-9

Published: 01/01/2000

Document Version
Early version, also known as pre-print

Please cite the original version:
Rantala, A., Virtanen, H., Saloheimo, K., & Jämsä-Jounela, S.-L. (2000). Using principal component analysis and self-organizing map to estimate the physical quality of cathode copper. *IFAC PROCEEDINGS VOLUMES*, *33*(22), 357-362. https://doi.org/10.1016/S1474-6670(17)37020-9

Author's accepted manuscript, published in Proceedings of IFAC Workshop on Future Trends in Automation in Mineral and Metal Processing, Helsinki, 2000

USING PRINCIPAL COMPONENT ANALYSIS AND SELF-ORGANIZING MAP TO ESTIMATE THE PHYSICAL QUALITY OF CATHODE COPPER

A. Rantala[1], H. Virtanen[2], K. Saloheimo[1] and S-L., Jämsä-Jounela[3]

[1] *Outokumpu Mintec Oy, P.O.Box 84, FIN-02201 Espoo, Finland*
[2] *Outokumpu Harjavalta Metals Oy, P.O Box 60, FIN-28101 Pori, Finland*
[3] *Helsinki University of Technology, Laboratory of Process Control and Automation, P.O.Box 5400, FIN-02015 HUT, Finland*

Abstract: The growing interest in utilising multivariable statistical dimension reduction techniques, PCA and PLS, and neural networks in process monitoring and analysis has resulted in a number of successful industrial applications. This paper describes a process study on the effects of the chemical quality of the anodes on the physical quality of produced cathodes at a copper electrorefining plant through PCA, SOM and a combination of these two techniques. The clustering of anode analysis data over time was compared with the physical quality data of the cathodes.

Keywords: Principal component analysis, Self-organizing map, Hybrid method, Process monitoring, Data mining, Refining, Copper.

1. INTRODUCTION

Process monitoring and analysis through statistical modelling techniques and neural networks have received considerable attention in recent years. The general objectives of process monitoring are to detect any abnormal events, reduce off-specification production and provide early warnings and identify important process disturbances, malfunctions or faults (Morris and Martin, 1997 and Kourti *et. al*, 1996). Process analysis by statististical or neural network methods promotes understanding of the process without the task of physical modelling, and ultimately improvement of the plant performance.

Statistical process monitoring appeared first as a traditional Statistical Process Control (SPC) concept. It provided control charts to determine the process state in a statistical manner. However, as discussed by MacGregor (1995), SPC was found to be inadequate for most processes because standard SPC methods examine less frequently collected quality variables, one at a time. At most industrial plants, the essential information is in the form of large history databases of process variables. It is reasonable to utilize all the collected data instead of extracting special quality information.

As reported by Hwang *et. al.* (1999), multivariable statistical monitoring methods have successfully replaced SPC. The multivariate projection techniques, Principal Component Analysis (PCA) and Partial Least Squares (PLS), especially have proved to be effective in a number of industrial applications (Wikström *et. al.*, 1998 and Taylor, 1998). According to Kourti *et. al.* (1996) and Kresta *et. al.* (1994), these methods address the traditional problems encountered in statistical analysis such as collinearity, missing data and large dimensionality.

Neural networks have been applied also in process monitoring and analysis, particularly to cases with nonlinearities and unknown mechanisms involved in the process. The use of a Self-Organizing Maps (SOM) has been successful in various industrial applications (Kohonen, 1990).

A number of attempts have been made to extend the properties of PCA and PLS to a nonlinear domain using various neural networks. Dong and McAvoy (1996) have described the concept of non-linear PCA (NLPCA). The combination of PCA and neural network was used in classification of iron ore by Cutmore *et al.* (1998). Holcomb and Morari (1992) discussed in depth how to combine PLS and

Feedforward Neural Networks. According to Wilson *et. al.* (1997), Radial Basis Function (RBF) neural networks can also be applied with PLS.

This paper presents a process analysis for a copper electrorefining plant using PCA, SOM and their combination. The object was to investigate the impact of anode impurities on the quality of produced cathode copper. Electrorefining is a good example of a process in which variables are strongly correlated and the time constants exceptionally large, both of which make the process very problematic to monitor and control. The chemical and physical quality of the anodes is recognised as the major disturbance source for the process, affecting the essential variables of the refinery. The physical influence mechanisms are very complex and mostly not completely understood. Investigating the problem through a physical model is thus laborious and uncertain

2. PROCESS DESCRIPTION

In the refining process, impure (appr. 99.0 w-%) copper anodes are electrically purified to high purity copper cathodes (over 99.99 w-%). As the electrolyte is continuously circulated through the electrolysis cells, its temperature and, additive agent composition can be controlled. The pure electrolyte is an aqueous copper sulphate and sulphuric acid solution. Purification of the soluble impurities (arsenic, nickel, etc.) is conducted in a separate process. Insoluble impurities, such as gold, silver and other noble metals, are handled in another process as well. E.g. Biswas and Davenport (1980) give a complete general description of the process.

The production rate is controlled by the electric current applied to the cells. The current density typically varies between 200 – 330 A/m^2. The most important variables include electrolyte composition, temperature, current density, additive agent concentrations and anode quality. The efficiency of the process is expressed as the overall quality (physical and chemical) of produced cathodes and the current efficiency of the refinery.

The annual capacity of the Outokumpu Harjavalta Metals Oy Pori Refinery is 125 000 tonnes of purified copper. There are 692 production cells, each containing 30 electrode pairs. The anodes are transported by train from the smelter at Harjavalta. The time needed to produce cathodes from starting sheets is eight days. Applying Periodically Reversed Current (PRC) allows higher current density (330 A/m^2) as compared with direct current feed.

3. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is a powerful dimensionality reduction technique (Jackson, 1991). PCA can be used to extract new latent variables, called principal components, from the original data without loss of any essential information. The principal components are linear combinations of the original variables that explain the maximum variability and covariance in the data. The first principal component describes the direction of greatest variability in the original data set. The next component is in the direction of the second greatest variability and orthogonal to the first principal component and so forth. The orthogonality of principal components means that they are independent from each other. PCA decomposes the original data matrix **X** (n x m) according to following model

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = \sum_{i=1}^{p} \mathbf{t}_{ni}\mathbf{p}_{mi}^T + \mathbf{E} \qquad (1)$$

where **T** is the scores matrix, t is the score vector, **P** is the loadings matrix, p is loading vector, **E** is residual matrix and p is the number of principal components.

The loading vectors $\mathbf{p}_p$ give the principal components in the original coordinates, and also define how the variables are related to each other in the original data space. The score vectors $\mathbf{t}_p$ contain the projection of the n observations in the reduced principal component subspace and therefore define the relationship between the observations in the original data matrix. Computationally PCA is based on the eigenvector decomposition of the covariance matrix of **X**. The corresponding eigenvalues of the covariance matrix are the variances of the principal components.

Usually only a few principal components are enough to explain the behaviour of the original data matrix because the greatest variability has been captured by the first components. There are several methods for determining a sufficient number of principal components, perhaps the most usable being crossvalidation (Jackson, 1991).

A PCA models of historical process variables data can be used in both process monitoring and analysis. However, a PCA model alone is not sufficient when determining differences between observations. The model explains the variation that is common to the data set. To idenfy a new type of variation, one has to determine the Squared Prediction Error (SPE) of each new observation (Kresta *et. al.*, 1991).

$$SPE_x = \sum_{i=1}^{p} (\mathbf{x}_{new,i} - \mathbf{P}_p \mathbf{t}_{p,new,i})^2 \qquad (2)$$

where $\mathbf{x}_{new}$ is the new observation vector. The control limits for the SPE are usually determined according to the chi-squared distribution.

A useful method in the analysis is to examine the formation of patterns or clusters in the score plots of the observations together with the loading plots in the principal component plane. Statistically deviating observations can be identified by drawing Hotelling T^2 ellipse which corresponds to the confidence interval at the given level. The Hotelling T^2 value is calculated according to the following equation

$$T^2 = \frac{(n^2 - 1)m}{n(n-m)} F_\alpha(m, n-m) \qquad (3)$$

where n corresponds to the number of observations, m is the number of variables and F_α is the α confidence value of the F-distribution.

In process monitoring, a statistically abnormal situation can be detected from the score plot when the projection of the observation passes outside the Hotelling T^2 ellipse or, equivalently, when the SPE increases above its control limit. The situation is illustrated in Figure 1.

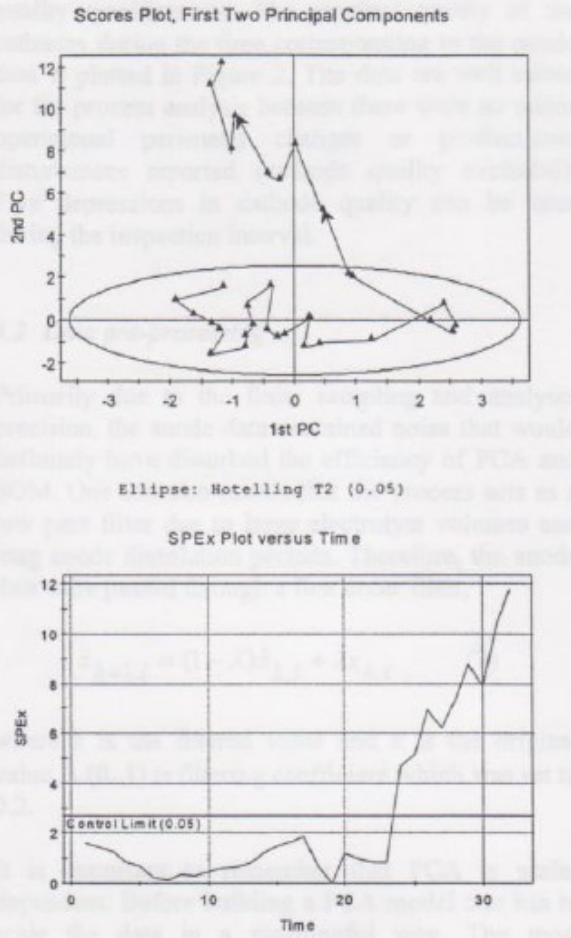

Fig. 1. Detection of process deviation from the scores plot and the SPE chart.

4. SELF-ORGANIZING MAP

Of the architectures and algorithms suggested for artificial neural networks, Kohonen's Self-Organizing Map has the special property of effectively creating spatially organised internal representations of various features of the input signals and their abstractions. Self-organization is based on competitive training that is able to find clusters from the learning data matrix **X**.

The SOM is composed of a set of elements, each of which represents a vector in the original data space. In training, the elements of the SOM compete for each input vector; the element that is closest to the input vector is the winner. The training algorithm then moves the winning element and its neighborhood closer to the presented input vector. In this way the elements of the network gradually learn to represent the training data. Since the neighborhood is taken into account, the properties of adjacent elements become similar. The map becomes ordered in such a way that clusters with similar properties are located near to each other (Kohonen, 1990).If the input vector is denoted by $\mathbf{x} = [x_0\ x_1\ \ldots\ x_{N-1}]$ and the location of a mapping element by $\mathbf{m}_i = [m_{i0}\ m_{i1}\ \ldots\ m_{iN-1}]^T$, the self-organizing algorithm is as follows,

1st step: Initiate the locations of the elements with random values

2nd step:

A: Find the SOM element $\mathbf{m}_c$ which best matches to the data vector $\mathbf{x}_i$ by searching all elements $\mathbf{m}_I$

$$\left\| \mathbf{x} - \mathbf{m}_C \right\| = \min_i \left\{ \left\| \mathbf{x} - \mathbf{m}_i \right\| \right\} \qquad (4)$$

B: Adjust the locations of these elements

$$\mathbf{m}_{i,k+1} = \begin{cases} \mathbf{m}_{i,k} + \alpha(t)\left[\mathbf{x}_k - \mathbf{m}_{i,k}\right] & , i \in N_{C,k} \\ \mathbf{m}_{i,k} & , i \notin N_{C,k} \end{cases} \qquad (5)$$

In equation (5), N_C denotes to the neighborhood area in the map. In equations (4) the Euclidian metric can be used as the distance measure.

In equation (5), the parameter $\alpha(t)$ is a coefficient that determines the movement of the winning element and its neighborhood in the direction of the data vector $\mathbf{x}_i$. It is equivalent to the learning rate

used in the back propagation algorithm for feedforward neural networks. It is recommended that $\alpha(t)$ decreases slowly with the progress of the iteration. Initially $\alpha(t)$ may be chosen as unity and in the final stages the recommended value is less than 0.01. One method for calculating $\alpha(t)$ is

$$\alpha(t) = \frac{\alpha_0 T}{T + 100t} \qquad (6)$$

where T is the total number of iterations and α_0 is the initial value for $\alpha(t)$. The neighborhood of the winning element C is defined by $N_C = \{i \mid d(i,C) < r(t)\}$, where d(i,C) is the distance between map elements i and C, and r(t) is the radius of the neighborhood. To ensure good global ordering, the radius r(t) should initially be more than half the diameter of the network. The radius should slowly decrease with time. In the final stage a small radius gives the map a good local spatial resolution. In addition to arranging the map topologically, the use of neighborhood equalises the number of input vectors classified in each cell.

A third training parameter is the number of iteration steps. To reach a good statistical accuracy the number of training steps should be at least five hundred times the number of elements in the SOM.

The topology of the trained SOM in the training data space R_T can be inspected graphically. Each SOM element carries a vector specifying its location in the R_T. In one-dimensional SOM, the value corresponding to the dimension is collected from the location vector of each element. The values are set into a matrix that can be presented graphically.

5. EXPERIMENTAL

5.1 Analysis data

The analyzed chemical quality history of the anodes supplied to Pori refinery consisted of 1052 chemical analyses during the past 1.5 years of operation. The sample is taken once per anode casting batch. There were nine variables representing the contents of the impurities occuring in the anode: antimony, arsenic, bismuth, lead, nickel, oxygen, selenium, silver and tellurium. There were some missing values for the lead (6.4 %), oxygen (1.3 %), silver (6.4 %) and selenium (2.9 %) contents mainly due to laboratory malfunctions.

The physical quality of the cathode is measured as the percentage of A-class cathodes in the total cathode production per day. The operators define the class, A or B, of the cathode to according to the quality specifications. The physical quality of the cathodes during the time corresponding to the anode data is plotted in Figure 2. The data are well suited for the process analysis because there were no major operational parameter changes or productional disturbances reported (cathode quality excluded). Five depressions in cathode quality can be seen during the inspection interval.

5.2 Data pre-processing

Primarily due to the finite sampling and analyser precision, the anode data contained noise that would definitely have disturbed the efficiency of PCA and SOM. One can also reason that the process acts as a low pass filter due to large electrolyte volumes and long anode dissolution periods. Therefore, the anode data were passed through a first order filter,

$$\hat{x}_{k+1,i} = (1-\lambda)\hat{x}_{k,i} + \lambda x_{k,i} \qquad (7)$$

where $\hat{x}$ is the filtered value and x is the original value. λ (0..1) is filtering coefficient which was set to 0.2.

It is important to remember that PCA is scale-dependent. Before building a PCA model one has to scale the data in a meaningful way. The most common method is standardization to unit variance,

$$Z_{k,i} = \frac{x_{k,i} - \bar{x}_{k,i}}{\sigma_i} \qquad (8)$$

where Z is the scaled variable, x is the original value, $\bar{x}$ is the mean and σ is the variance of the variable.

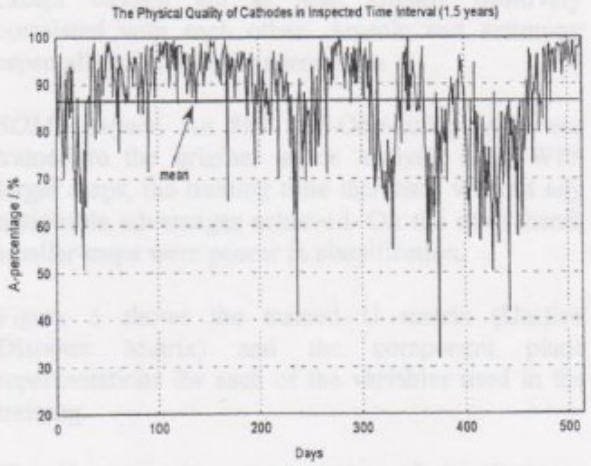

Fig. 2. The physical quality of the cathodes at the Pori Refinery during the inspected time interval.

5.3 Results and discussion

The estimation of the effect of the chemical composition of the anodes on the physical quality of the cathodes was done by three different methods: PCA models, SOM and a combination of PCA and SOM. The effects were judged by comparing the clusters formed among the observations (anode analysis) with quality data of the cathodes.

PCA modelling; Two PCA models were constructed. In the first one, the arsenic content was replaced by its ratio to certain impurity contents. In addition, a couple of other ratios were added largely due to metallurgical interest and knowledge. The selected variables are given in the loadings plot in Figure 3. The modelling resulted in three principal components with the explained variance (R2VX) of 72 %. The statistical key figures of the model are summarized in Table 1, and the scores and loadings plots of the first and second principal components are shown in Figure 3.

Table 1 The key figures of the first PCA model

	Eig.	R2VX / %	R2VX (cum) / %
1st PC	4.48	34	34
2nd PC	3.10	24	58
3rd PC	1.78	14	72

The arrows in Figure 3 point at certain clusters that could be distinguished from the data. The loadings plot revealed positive correlation among several impurities.

Only the oxygen content of the anode seemed to be negatively correlated with other basic impurities according to the first principal component.

The most interesting result was that the observations clearly clustered in separate areas in the principal component space when there was a fall in the cathode quality. This implies that the anode impurities have an effect on the physical quality of the cathodes. A classification could therefore be made to distinguish between problematic and non-problematic anode qualities. Physical explanations were also found for the problematic anode qualities.

The second PCA model was built using the nine original analysis variables. The model consisted of four principal components with 70 % explained variance. This model did not bring any improvements from the classification point of view compared with the first PCA model, since the clusters implied the same results. A minor setback was that a fourth principal component was needed to achieve a satisfactory explanation and classification.

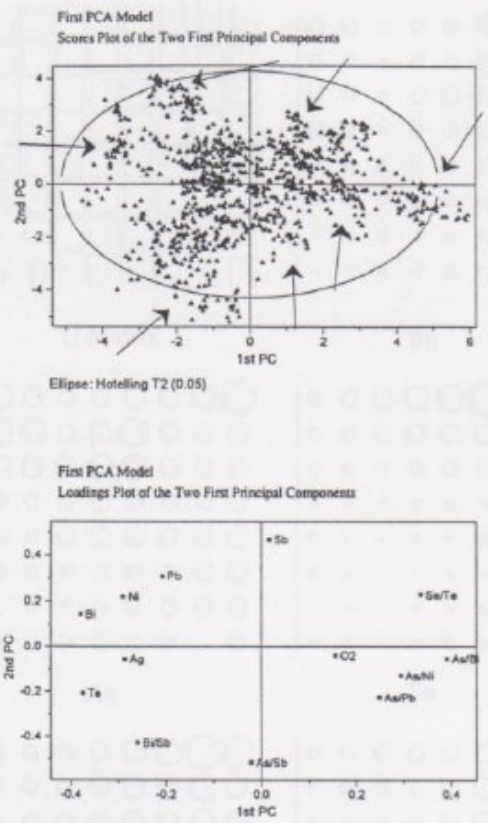

Fig. 3. The scores and loadings plot of the first PCA model.

This considerably complicated the visualization and analysis. However, this model gave a better view of the correlations between the basic anode impurity variables and there was not need to determine which ratios of impurities should be included in the model. The correlation structures are shown relative to the first two principal components in the loadings plot in Figure 4.

Figure 4 confirms the observation that all impurities except oxygen are at least slightly positively correlated with each other. Arsenic and antimony especially have a strong correlation.

SOM analysis; An 8x8 Self-Organizing Map was trained to the original anode analysis data. With larger maps, the training time increased without any noticeable advantages achieved. On the other hand, smaller maps were poorer in classification.

Figure 5 shows the trained U matrix (Unified Distance Matrix) and the component plane representations for each of the variables used in the training.

The U matrix is a presentation of the distances between the weight vectors of adjacent neurons in the two-dimensional map. A neuron located adjacent to other neurons indicate that the input vectors

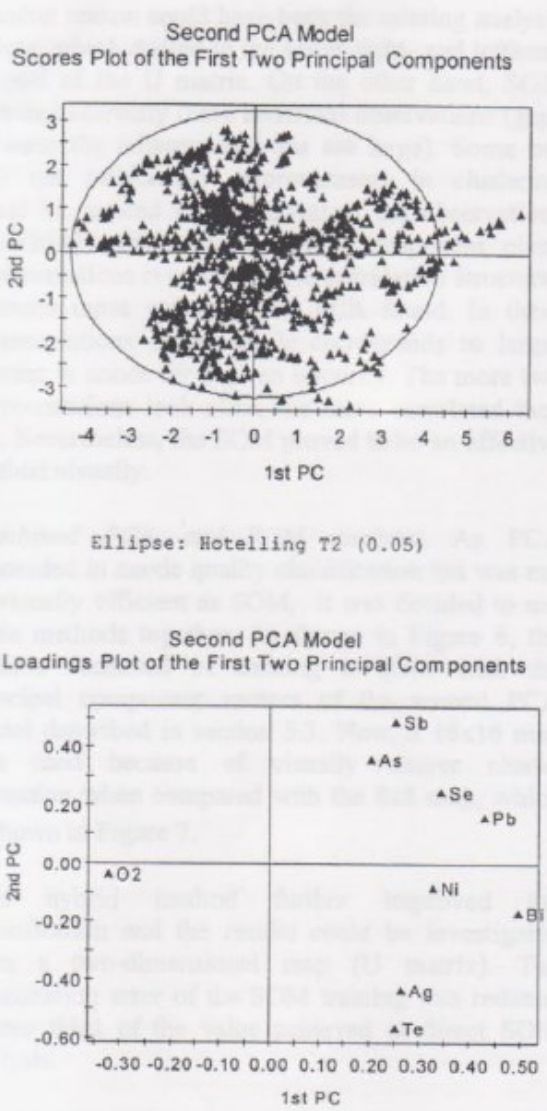

Fig. 4. The scores and loadings plot of the second PCA model.

mapped to these neurons are close to each other in input space and vice versa. The U matrix is used for detecting clusters and deviating input vectors in the original data.

The SOM was not as successful in classification as the PCA models since all the drops in cathode quality could not be traced to specific neurons. One reason for the poorer performance might have been the excessive dimension reduction - even the PCA models needed more than two principal components for the classification.

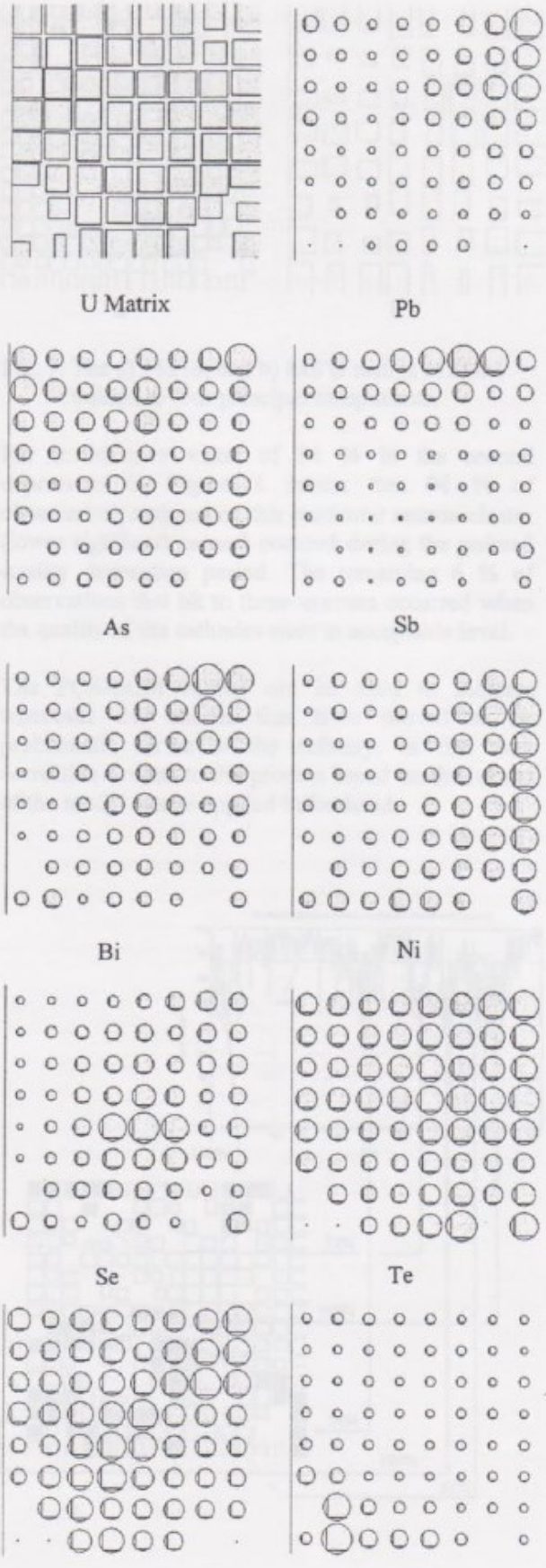

Fig. 5. The U matrix and component plane representations of the SOM trained with the anode analysis data

Another reason could have been the missing analysis values, which drifted to the lower right- and lefthand corners of the U matrix. On the other hand, SOM revealed correctly these abnormal observations (gaps between the adjacent neurons are large). Some but still not satisfactory improvements in clustering could be noticed after removal of the observations containing missing values. The component plane representations reveal the same correlation structures between input variables that PCA found. In these representations bigger circle corresponds to larger content in anode for a given impurity. The more two representations look alike, the more correlated they are. Nevertheless, the SOM proved to be an effective method visually.

Combined PCA and SOM analysis; As PCA succeeded in anode quality classification but was not as visually efficient as SOM, it was decided to use these methods together. As shown in Figure 6, the scheme consisted of training a SOM with the principal component vectors of the second PCA model described in section 5.3. Now, a 16x16 map was used because of visually clearer cluster formation when compared with the 8x8 map, which is shown in Figure 7.

This hybrid method further improved the classification and the results could be investigated from a two-dimensional map (U matrix). The quantization error of the SOM training was reduced to one third of the value achieved in direct SOM analysis.

As shown in Figure 8, the drops in the cathode physical quality could be traced to specific neurons and to clusters formed in the constructed map. The black coloured neurons indicate the areas in the map that anode analysis hit during the pointed quality depression periods. The reliability of the classification was studied by counting a hit percentage of each depression period hits, to total hits (whole time interval of inspection) in corresponding neuron cluster.

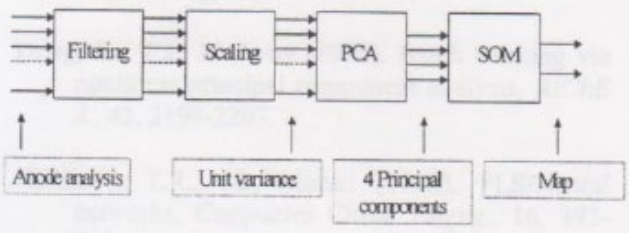

Fig. 6. The combined PCA and SOM method for classifying anode analysis.

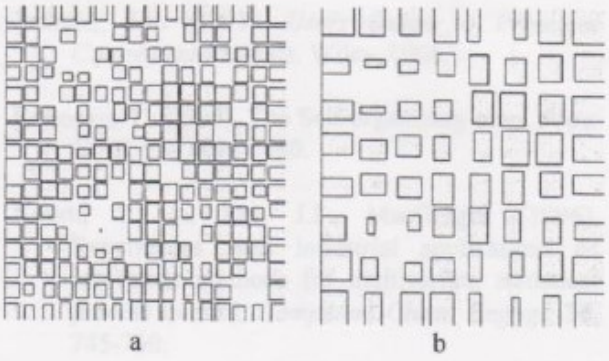

Fig. 7. The a) 16x16, and b) 8x8 U matrix of SOM trained to four principal components.

For instance, a value of 94 % in the second depression in Figure 8 means that 94 % of observations mapped on this particular neuron cluster (lower righthand corner) occured during the pointed quality depression period. The remaining 6 % of observations that hit to these neurons occurred when the quality of the cathodes were in acceptable level.

The PCA-SOM model can be used to indicate whenever the anodes that have proved to be problematic arrive at the refinery. In this way corrective actions to the process based on the results of the model can be applied beforehand.

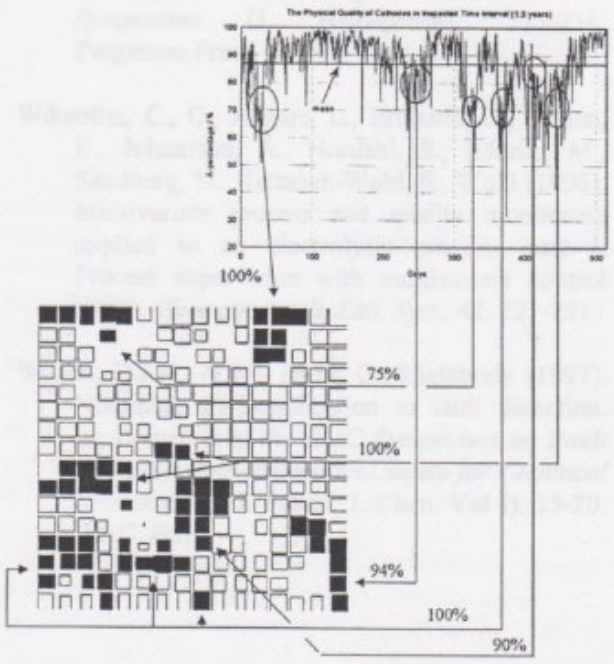

Fig. 8. The U-matrix representing the neurons and clusters where anode analysis hit during each drop in cathode quality. The percentages indicate the proportion of depression period hits to total hits counted in each pointed (black marked) neuron clusters.

6. CONCLUSIONS

The effects of anode impurities on the physical quality of cathode copper in the copper electrorefing process were analysed by PCA, SOM and a combination of the two techniques. The chemical composition of the anodes is widely recognized as the major disturbance source in the process. This study was focused on the clustering of the anode analysis over time and comparison of these formations to the physical quality data of the cathode copper.

PCA especially proved to be a powerful method for this kind of data mining application. The only drawback to this technique is the lack of visualization when there are more than two principal components involved. Direct SOM analysis was not as successful in classification, but was visually effective. Therefore, a SOM was trained to the principal component vectors in order to improve the visualization because the results can always be viewed from the two-dimensional map. In addition, classification was noticeably improved. This combination of PCA and SOM was found to give the best classifying results.

The classification clearly implied that the impurity concentration of the anodes had an effect on the physical quality of the cathodes. Based on the PCA models, the physical reasons for the cathode quality depressions could also be deduced. The adopted PCA-SOM model can be used to indicate whenever the anodes that have proved to be problematic arrive at the refinery. Early warnings of the upcoming disturbances can then be generated to the process control.

REFERENCES

Biswas, A.K., W.G., Davenport (1980). *Extractive Metallurgy of Copper*, 325-357. Pergamon, Oxford.

Cutmore, N.G., Y., Liu, A.G., Middleton (1998). On-line ore characterization and sorting, *Minerals Engineering*, **11**, 843-847.

Dong, D., T.J., McAvoy (1996). Batch tracking via nonlinear principal component analysis, *AIChE J.*, **42**, 2199-2207.

Holcomb, T.R., M., Morari (1992). PLS/Neural networks, *Computers Chem. Engng.*, **16**, 393-411.

Hwang, D.H., C., Han (1999). Real-time monitoring for a process with multiple modes, *Control Engineering Practice*, 7, 891-902.

Jackson, J.E. (1991). *Users Guide to Principal Component Analysis*. Wiley, USA.

Kohonen, T. (1990). The Self-organizing map, *Proc. IEEE*, **78**, 1464-1480.

Kourti, T., J., Lee, J.F., MacGregor (1996). Experiences with industrial applications of projection methods for multivariate statistical process control, *Computers Chem. Engng.*, **20**, 745-750.

Kresta, J., J.F., MacGregor, T.E., Marlin (1994). Multivariate statistical monitoring of process performace, *Can. J. Chem. Eng.*, **69**, 35-47.

MacGregor, J.F., T., Kourti (1995). Statistical process control of multivariate processes, *Control Engineering Practice*, **3**, 403-414.

Morris, A.J., E.B., Martin (1997). Process performance monitoring and fault detection through multivariate statistical process control. In: *Preprints of the IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes* (R. J. Patton, J. Chen. Vol I), 1-14. IFAC, Hull.

Taylor, A.G. (1998). The Application of principal component analysis for prediction blast furnace stability. In: *Preprints of the IFAC MMM'98 Symposium* (J. Heidepriem), 233-236. Pergamon Press, Oxford.

Wikström, C., C., Albano, L., Eriksson, H., Fridén, E., Johansson, Å., Nordahl, S., Rännar, M., Sandberg, N., Kettaneh-Wold, S., Wold (1998). Multivariate process and quality monitoring applied to an electrolysis process, part I. Process supervision with multivariate control charts, *Chemom. Intell. Lab. Syst.*, **42**, 221-231.

Wilson, D.J.H., G.W., Irwin, G., Lightbody (1997). Nonlinear PLS-application to fault detection. In: *Preprints of the IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes* (R. J. Patton, J. Chen. Vol I), 15-20. IFAC, Hull.