Management and Production Engineering Review
Volume 3 • Number 4 • December 2012 • pp. 47–53
DOI: 10.2478/v10270-012-0033-x

DATA MINING MODEL FOR QUALITY CONTROL OF PRIMARY ALUMINUM PRODUCTION PROCESS

Matus Horvath, Edita Vircikova

Technical University of Kosice, Faculty of Metallurgy, Department of Integrated Management, Slovakia

Corresponding author:
Matus Horvath
Technical University of Kosice
Faculty of Metallurgy, Department of Integrated Management
Letna 9 042 00 Kosice, Slovakia
phone: (+421) 905 377 293
e-mail: matus.horvath@tuke.sk

Received: 29 October 2012
Accepted: 28 November 2012

Abstract
Traditional statistical process control approaches are less effective in dealing with multivariate and autocorrelated processes. With the continual increase in process complexity, this inefficiency is becoming more apparent. A special type of multivariate and autocorrelated process is a process occurring within a heterogeneous production environment (a variety of types of machines, pots, etc. used for the same task). This makes the quality control of such processes more difficult. The approach presented in the paper utilizes time series fitting, cluster analysis and association mining in relation to a single data mining model for the analysis of complex multivariate autocorrelated processes. The aim is to divide the production cells (machines, pots, etc.) into groups exhibiting similar behaviors. This can then be used for more effective quality control of the entire process and afterwards to analyze the reasons for this behavior. This paper includes someof the results obtained from applying the model to an actual multivariate high autocorrelated process, the production of primary aluminum using the Hall-Heroult electrolysis process. The Hall-Heroult electrolysis process is a continual process that is ongoing in several pots simultaneously. The average plant operates 300 pots. Therefore, the quality control of such a complex process faces many issues concerning monitoring and problem diagnosis. The paper describes a method for dividing the pots into control groups exhibiting similar behaviors, which can then be used in the planning phase of the quality control analysis and to make improvements within these groups and thereby within the whole process.

Keywords
quality control analysis, data mining, multivariate autocorrelated process, quality improvement.

Introduction

Process monitoring and diagnosis have been widely recognized as key tools for detecting abnormal behavior and assessing quality improvement. Traditional statistical process control approaches are less effective in dealing with multivariate and autocorrelated processes. With the continual increase in process complexity, this inefficiency is becoming more apparent. A special type of multivariate process is a process occurring within a heterogeneous production environment (a variety of types of machines, pots, etc. used for the same task). Carrying out controls on this kind of process is more difficult if the process variables are autocorrelated. Both these conditions frequently occur in continual production processes, commonly found in the metallurgy industry, making the quality control of these processes more difficult. The purpose of this paper is to present some of our findings obtained during research into the use of the data mining approach in quality management systems. The approach discussed here is a new data mining model capable of promoting quality control and quality improvement in an organization with a heterogeneous production environment. The approach utilizes time series fitting, cluster analy-

sis and association mining in a single data mining model in order to analyze complex multivariate autocorrelated processes. The aim is to divide the production cells (machines, pots, etc.) into groups exhibiting similar behavior. These can then be used to make quality control of the entire process more effective and to analyze the causes of this behavior afterwards.

Data mining methods can be described as methods used to discover meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies as well as statistical and mathematical techniques [1]. Data mining methods include, for example, cluster analysis, association rules mining and classification. Data mining algorithms have the potential to aid process monitoring of complex processes because of their proven capability to manage and analyze large amounts of multivariate data. The data mining model describes the sequence and eventually the configuration of the particular data mining methods up to a specific algorithm. The data mining model described here consists of four steps. The inputs used in the model are daily means of monitored variables for each production cell. The model is not limited to a particular number of input variables. A higher number of variables can cause higher hardware performance demands. In this paper we have decided to apply the data mining model to a selected process in order to gain a better understanding of the model. A typical example of a multivariate and correlated process in a heterogeneous production environment is the primary production of aluminum. Section 2 briefly describes the production process, while Sec. 3 provides a step-by-step description of the data mining model and its use within the selected process.

Description of the example process

The main production process used in the commercial production of primary aluminum is the Hall-Heroult electrolytic process invented in 1886. The electrolytic production of primary aluminum is a process that has many variables and involves complicated systems such as mass and energy balance [2]. Production takes place in separate electrolytic pots with carbon based lining. Aluminum oxide (Al2O3) is dissolved in molten cryolite (Na3AlF6). A direct current is supplied to the electrolytic pot via graphite anodes and passes through the electrolyte at a low voltage but at high current rates (typically from 200 to 350 kA) towards the cathodes situated at the base of the electrolytic pot. Smelted aluminum is electrolytically deposited at the cathodes, in the base of the pot. The aluminum is periodically tapped. Aluminum oxide is constantly replenished from storage containers above the electrolytic pot. An electrolytic pot produces annually on average 150000 tons of aluminum.

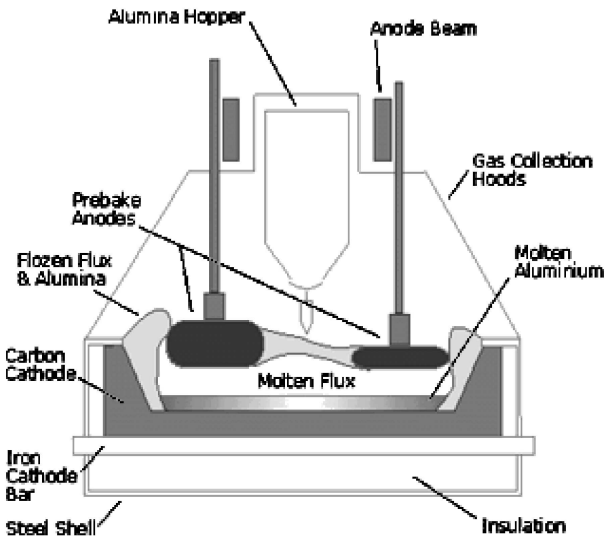

Fig. 1. Principle of Hall-Herout electrolytic process [3].

Studies on the nature of these physical and chemical actions indicate that this process is very unstable and affected by a number of factors. Performing controls on an electrolytic pot requires the continual monitoring of variables, rapid reactions to any changes and a very good knowledge of running reactions. Controlling the process is complicated by the fact that for efficiency reasons production occurs simultaneously in several electrolytic pots connected serially. Groups of electrolytic pots connected in this way form potlines. An average potline consists of 300 interconnected electrolytic pots. In order to account for the continual abrasion of the pot lining and cathodes, and to reduce downtimes for shutdown and repairs, production is begun on an individual pot basis in each potline (for example one pot per week). It is this that creates a heterogeneous production environment. It is very common to find that in one potline the electrolytic pots are running at various run times, that different kinds of pots are in use, and that the linings are worn to various degrees. In practice it is very common to find multiple potlines operating simultaneously in a single plant.

Fig. 2. Example of a potline [4].

Data mining model for production cell segregation

Monitoring and controlling complex production processes running simultaneously in several production cells is more challenging than monitoring and controlling a process involving a single production cell, given the large number of monitored and controlled variables. The scope of the data generated makes analysis more difficult. The heterogeneous production environment and the autocorrelation of the variables make effective analysis and quality improvement of the process extremely problematic. The data mining model discussed here can benefit the production process analysis by segregating production cells into control groups according to behavior. The control groups can then be used to further analyze reasons for the behavior and to more effectively control the process.

The data mining model has been used in relation to real production data from a primary aluminum production process operating in a heterogeneous production environment and with very strongly autocorrelated variables. The rules on the collaborative research mean that certain detailed information about the production process and the results of the analysis cannot be published. The input consisted of data from 226 electrolytic pots obtained over a calendar year. For each electrolytic pot 13 variables were monitored on a daily basis. Thus the input for the model was a database containing 1072370 data points. Figure 3 illustrates the variation of a single variable in January 2011. Each line represents an electrolytic pot. The Hall-Heroult electrolytic process produces high variability in production cell behavior and finding two production cells with similar behavior is a difficult task when analyzing just one variable.

Fig. 3. Variation in a single variable for each separate pot in January 2011.

Data preparation

It is apparent that we need to reduce the amount of input data. Given the scope of the data mining model, behavior analysis of the division of production cells can be conducted on a time basis (for example weeks, months). In order to make the analysis more effective, the next step is to summarize the input variables within the selected time frame. Using the usual method of summarizing and computing mean and standard deviation would mean losing some information on the input variable statistical distribution in the selected time frame.

Therefore we decided to compute seven parameters of statistical distribution: mean, median, standard deviation, upper and lower quartile, min and max for each selected time frame and input variable. With the help of these parameters we can describe the statistical distribution of the summarized input variables more precisely. The benefit of this method is that it significantly reduces the amount of analyzed data with no major loss of accuracy. For example if we were to analyze segmentation of 100 production cells with 13 input variables for each production cell over one calendar year, we would be dealing with 474500 data points. By using this method to summarize these data points (within a time frame of a month) we can reduce the number of data points to 109200, which is more than a four-fold reduction in the amount of data.

Hierarchical cluster analysis of time frames

The second step in this data mining model is to conduct a hierarchical cluster analysis of the statistical distribution parameters of the input variables for each time frame. The hierarchical clustering method creates a nested sequence of clusters for one cluster to N clusters for a data matrix with N data points. The agglomerative hierarchical method starts with each data point as a separate cluster and merges them into successively larger clusters [6]. The results of hierarchical clustering are usually presented in a large

dendrogram. As we have stated, the similarity between the clusters is measured in terms of distance. In obtaining the results of the analysis it is very important to select the right distance metrics. For our type of data we chose correlation metrics. The mathematical formulation of our distance metrics for data matrix $X = \{X\_ij\}$ is formula (1).

$$d_r(x_j x_k) = 1 - \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2}}. \quad (1)$$

The second important setting of the hierarchical clustering algorithm is the linkage criterion. The linkage criterion determines the distance between sets of observations as a function of the pair wise distances between observations [5]. We have chosen Ward's method. Ward's method, also called Ward's minimum variance method, is a particular type of objective function method. Ward's minimum variance criterion minimizes the total within-cluster variance [6]. In each step a pair of clusters with minimum cluster distance is merged. The criterion for two clusters X_j and X_k, is mathematically defined as (2).

$$\Delta(x_j, x_k) = \frac{n_j n_k}{n_j + n_k} \left\| \overline{x}_j - \overline{x}_k \right\|. \quad (2)$$

Using the definition of Ward's criterion we can say that this will tend to create small clusters.

The allocation of individual production cells into clusters is denoted in the dendrogram on the basis of the estimated number of control groups and for each time frame. The estimated number of control groups can be obtained using input file analysis e.g. the evaluation graph or L method. More information on these methods can be found in [7]. Too many clusters would cause high segmentation and so the next step is to create control groups with a smaller number of production cells. Hierarchical cluster analysis was also selected since it provides a visual description of similarity. The dendrogram shows information on the similarity between the clusters and the production cells within the clusters for the selected time frame.

Applying the second step to the process

The second step of the data mining model is to perform a hierarchical cluster analysis of the statistical distribution parameters for each time frame. Given the number of production cells, it is only possible to show the results as a preview. The dendrogram preview for January 2011 using 226 electrolytic pots is shown in figure 4. The number of control groups was estimated to be 10 and the clusters were on the dendrogram indicated using rectangles (Fig. 4). The allocation of the electrolytic pots into the clusters was subsequently extracted. This procedure was repeated each month. In Fig. 4 we can see three large clusters. In the months that followed the dendrogram was similar in terms of cluster sizes.

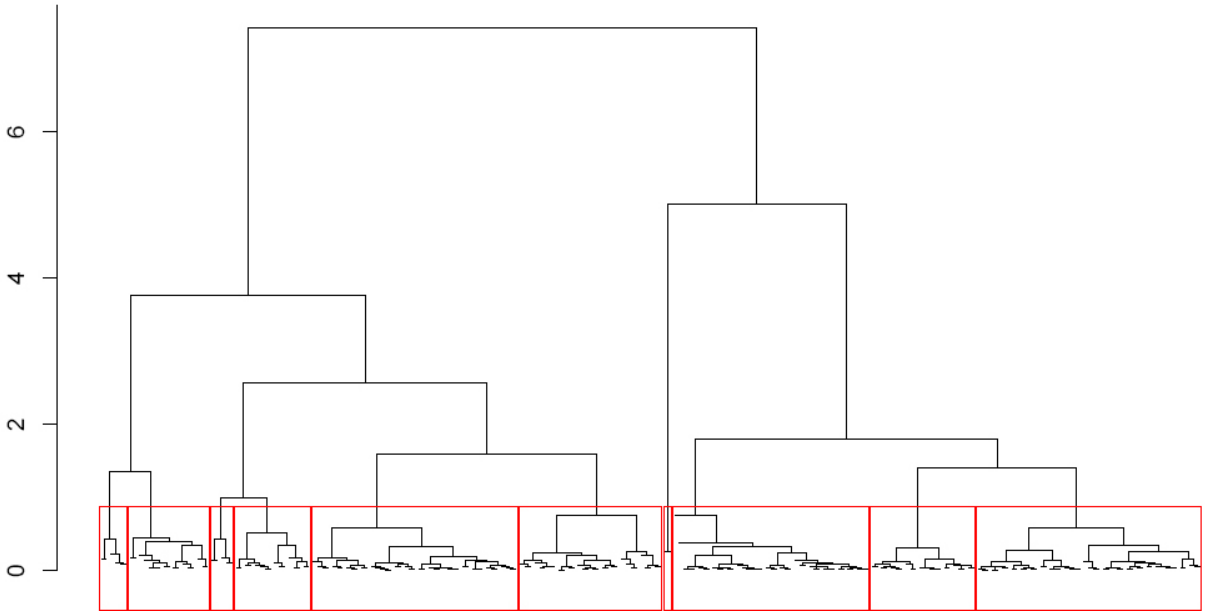

Fig. 4. Dendrogram for January using 226 electrolytic pots.

Generating control groups of production cells

In analyzing the individual time frame dendrograms, it is always possible to find movement in the production cells allocation into clusters. Where complex and autocorrelated processes operate in a heterogeneous production environment this instability in allocation is more visible. The instability in the behavior of the production cell can be explained by two factors. The first is the internal instability of the production cell, which may be caused by failure or incorrect controls etc. The second source of production cell instability could be changes in the operational process parameters that affect all production cells but because they have different construction or technical parameters they react differently. These changes may be result from changes to the chemical composition of the input materials, changes to work procedures or the weather (for example, the temperature). When the process involves several production cells and several monitored variables, some correlated or similar behavior may very easily be overlooked. Therefore we need to summarize the cluster allocation and find control groups of production cells that have been together in the same cluster from step two of the data mining model. To do this we can use another data mining method: association rules mining.

We have $I = \{I_1 ... I_m\}$ items. They contain the allocation of the individual production cell to the time frame and cluster. $X \subseteq I$ is the item set, D is the transaction set of T where each transaction T represents an item set, i.e. $T \subseteq I$. Set T in our case represent one cluster from a time frame. Set D represents the entire set of all clusters from all the time frames. Individual items in set X should be ordered in a predefined way. Set X consist of items from $x_1 ...$ to x_k i.e. $X = \{x_1 ..., x_k\}$. For the set of items $X \subseteq I$, support for set X in D is defined as a proportion of the transactions in D that contain X [8]. Therefore support for set X is the relative frequency of transactions in which the same production cells from set X were identified. An association rule is an implication in the form of $X \Rightarrow Y$ where X and Y are two disjunctive item sets i.e. $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$ [9]. The strength of an association rule is represented in its support $X \Rightarrow Y$ in D. It is the support of conjunction $X \cup Y$, which means that the relative frequency of joint occurrences of all items in the association rule [9] parameter in the association rules mining is confidence. Confidence in the association rule $X \Rightarrow Y$ in D is defined as a proportion of transactions containing set Y in the set of transactions from D that contains set X. From this it follows that the association rule $X \Rightarrow Y$ holds with confidence if the level of confidence is expressed as a percentage ratio of all clusters from D, which contains both X and Y.

If the value for support is correctly selected, it is possible to mine association rules that represent groups of production cells with similar behavior, during e.g. 8 months out of 12 or more. In this case, looking at 12 months with 10 clusters each, we would have a support value of 0.067 or more. A tendency to only look for groups of production cells that always behave similarly usually leads to the creation of small control groups or even failure. The value of the second parameter, confidence, is not usually utilized in this data mining model because of the nature of the input data. The input in this step is the cluster allocation for each time frame that implicates that the value of confidence always reaches the same value. This value depends only on the number of time frames in analyses where the production cell clusters are expected to behave similarly. The only exception occurs in situations where production cell composition changes between time frames.

Where there is a large number of mined association rules, the production cell sets are found in multiple association rules. Therefore we need to extract the production cells that have a maximum number of members from this set of association rules. In order to perform this extraction we again decided to apply a hierarchical cluster analysis. This time, the aim of the hierarchical cluster analysis is to cluster the production cells from the association rules following the rule of the most frequent occurrence in the association rules. In this case we used Euclidean metrics (3) and single-link as the linkage criteria. The single link criterion merges clusters following the shortest distance between the data points of two clusters.

$$d_2(x_i, x_j) = \left(\sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \right)^{1/2}. \qquad (3)$$

Using hierarchical cluster analysis in this way generates the largest possible control groups of production cells, taking into account the selected level of support from the association rules mining phase. Since it is a hierarchical method we can identify the level at which the production cell belongs to the cluster and therefore to the control group from the dendrogram analysis. The results of the third step are clusters of production cells with minimal defined support, which means the number of time frames in which the production cells were in the same cluster. The order in which they appear in these clusters shows the level of similarity.

Applying association rules mining to the process

After analyzing the second step output we observed a high rate of inter time frame migration of electrolytic pots among clusters. This was due to the instability of this particular production process but also to the different behavior exhibited by the production cells in the heterogeneous production environment. The first step of the third phase of the data mining model is association rules mining. After discussions with the technicians we opted for mining association rules indicating control groups of electrolytic pots that had behaved similarly for 10 or more months during one year. The computed support value was 0.084 or above. The level of confidence was left unchanged at 0.1. The mined association rules were thereafter analyzed using hierarchical cluster analysis. The resulting dendrogram is shown in Fig. 5. Of the original 226 electrolytic pots, 209 electrolytic pots achieved the selected level of support. This phenomenon was caused by the fact that electrolytic pots were replaced during the time period under analysis. Thus, these electrolytic pots had not been in operation for long enough (the minimum was 10 months).

The analysis of data mining model results

The subsequent analysis of the data mining model results must focus on the reasons for segmentation. These may be due to differences in the technical construction of the production cells or to the use of different construction materials. Work procedures must also be considered as possible reasons. The authors recommend that discussions on segmentation should be held with process technicians and operators who may have additional information about the production cells. Following the basic analysis of the dendrogram in Fig.5 three blocks of control groups can be identified. The first block, indicated by a blue rectangle, includes 8 control groups with 11 to 6 electrolytic pots in each control group. These control groups represent production cells whose behavior in terms of probability cannot be random. Analyzing the reasons for this behavior may help ascertain how the control process and work procedures may be modified so as to achieve more effective and stable electrolytic pots from this control group. The second block of control groups is indicated by a green rectangle. The control groups from the second block have fewer members, between 4 and 2. Having a high number of these small control groups is undesirable and the analysis can therefore be focused on the reasons why this is the case and how they may be eliminated. The last block in the dendrogram is the third block of control groups indicated by a red rectangle.

These control groups contain just one member. These production cells were in operation for at least 10 months, but could not be grouped with other production cells from the analysis.

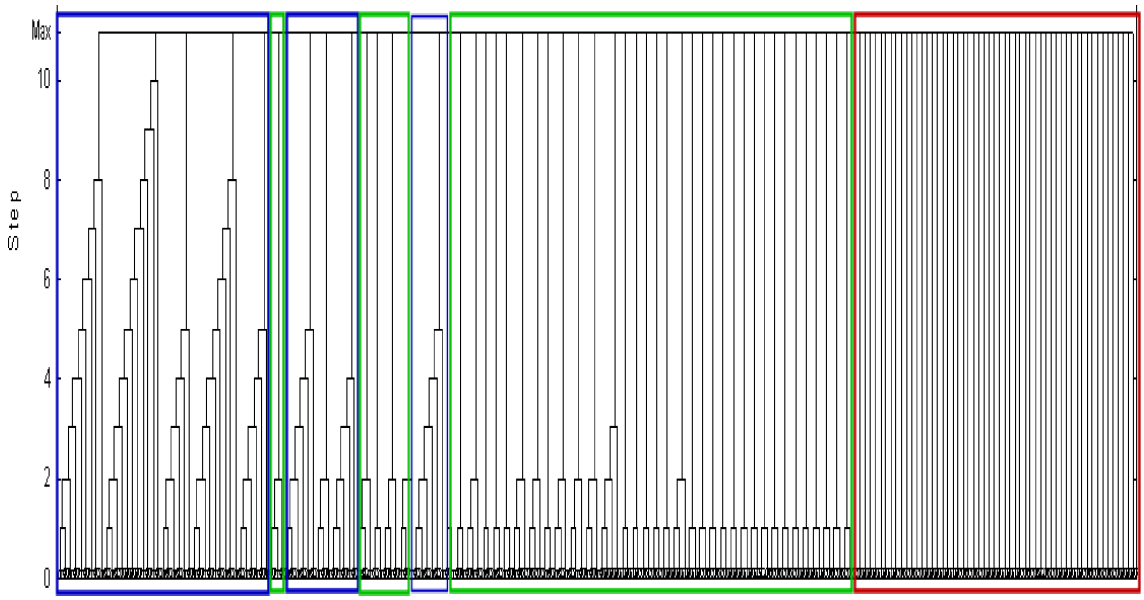

Fig. 5. Dendrogram for extracted electrolytic pots control groups.

Analyzing the reasons for this behavior may help improve the process so that the number of electrolytic pots in the third block of control groups can be reduced.

Conclusions and future work

The aim of this paper was to introduce a new data mining model for process control and process improvement support for special types of production processes. Analyzing behavior in complex production processes with autocorrelated variables within a heterogeneous production environment often presents difficulties in practice. With the help of our data mining model the set of all production cells can be segmented into control groups exhibiting similar behaviors. Analyzing the reasons for these similar behaviors can reveal the impact technical parameters or different operational controls have on long-term behavior. This information can be used to improve selection of the technical parameters of the production cells or to improve the control and work procedures for production cells in particular control groups. The results of the analysis are also revealing in terms of highlighting unstable behavior or atypical production cells that are not desirable in the process. All this new information can be used to improve the quality and efficiency of the production process under analysis. Repeated use of the data mining model may indicate the impact new technical parameters or changes to work procedures have on the behavior of production cells in the long term.

Acknowledgments

This work was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic. KEGA Project 009TnUAD-4/2011: Creative Laboratory Education at Technical Faculties (CRELABTE).

References

[1] The Gartner Group, www.gartner.com (accessed: 21.03.2010).

[2] Chen J.J.J., Taylor M.P., *Control of Temperature and Aluminum Fluoride in Aluminum Reduction, in Aluminum*, International Journal of Industry, Research and Applications, 81, 678–682, 2005.

[3] The International Aluminum Institute, *Technologies*, http://www.world-aluminum.org/About+Aluminum/Production/Smelting/Technologies (accessed: 01.06.2012).

[4] Štrauch M., *How it looks in aluminum plant Slovalco 2011* (*Ako to vyzerá v hlinikárni Slovalco 2011*), http://www.etrend.sk/galeria/ako-to-vyzera-v-hlinikarni-slovalco.html, (accessed: 01.06.2012).

[5] Pham D.T., Afify A.A., *Clustering techniques and their applications in engineering*, in Proc. IMechE Vol. 221 Part C: J. Mechanical Engineering Science, 221, 1445–1459, 2007.

[6] *The CLUSTER Procedure: Clustering Methods: SAS/STAT 9.2 Users Guide*, SAS Institute, 2009.

[7] Salvador S., Chan P., *Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms*, In proc. Tools with Artificial Intelligence, 2004, ICTAI 2004, 16th IEEE International Conference, pp. 576–584.

[8] Paralic J., *Knowledge discovery in databases* (*Objavovanie znalostí v databázach*), FEI TUKE, Kosice, 2003, ISBN: 80-89066-60-7.

[9] Agrawal R., Imielinski T., Swami A., *Mining Association Rules between Sets of Items in Large Databases*, In proc. 1993 ACM SIGMOD Conference Washington DC, USA, ACM, New York, pp. 207–216.

[10] Ward J.H., *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 58, 301, 236–244, 1963.