Marina Milanović[1)]
Milan Stamenković[1)]

1) Faculty of Economics, University of Kragujevac, Serbia, {milanovicm, m.stamenkovic}@kg.ac.rs

CONTROL CHART AND DATA MINING METHODS IN THE ANALYSIS OF PROCESS DATA

Abstract: *The idea of finding patterns in process data is not new. Traditionally, in the pursuit focused on achieving the desired level of quality, which is primarily defined by the needs of customers, statistical process control is widely accepted methodological framework for monitoring, control and improvement of the quality characteristics of manufacturing and service processes. Control charts, as the most commonly used SPC tool for analysis, understanding and identification of non-standard process variations, are, on the other hand, unable to provide the proper identification of the causes of their occurrence. Therefore, the expert teams in search of the causes of quality problems engage the data mining process. In this Paper, the idea of integrated use of control chart and data mining methods in function of identification of not only the causes of (unusual) patterns detected by control charts, but also those of interesting patterns and their causes that are hidden in the huge repositories of process data, is presented and discussed.*

Keywords: *Control Chart Method, Data Mining, Pattern Analysis, Quality Control and Improvement*

1. INTRODUCTION

Changes and challenges in modern business environment continuously promote quality as a source of a long-term competitive advantage, and quality management as a paradigm of modern enterprise management. The need for constant monitoring and measuring of the level of quality of the process, resulted in the application of different statistical methods, which, in the field of quality management, have the special place (i.e. significant role). In the context of efforts aimed at achieving the desired level of quality (which is primarily defined by the customer needs), statistical process control (*SPC*) represents an important additional (supporting) tool in the management of manufacturing and service processes.

Traditionally, *SPC* methodology is effective in the control and improvement of quality characteristics of simple processes that generate relatively small amounts of data. However in conditions, where large amounts of data are generated within modern, complex organizational systems, and development of information technology enables their efficient storage, the limitations of traditional *SPC* methodology become more apparent and obvious. In fact, in situations when the decision makers are drowning in a sea of data while remaining thirsty for (useful) information due to the fact that through the implementation of data mining methods the efficient management and analysis of massive volumes of data can be ensured, their application in function of quality control and quality improvement of quality

characteristics of a process, becomes a necessity. Hence, in terms of methodology, the next generation in the evolution of quality includes the efforts and activities related to the use of highly sophisticated, algorithmic data mining methods.

Accordingly, the main purpose of this Paper is to emphasize the importance of data mining methods in the field of quality management as well as to present the possibilities of combined use of data mining tools and control charts in the context of monitoring and improvement of the quality of processes and their outputs. With the *Introduction*, the rest of the Paper is organized as follows. *Section 2* contains a discussion related to the relevant aspects of the *SPC*, including, the statistical stability of the process (*2.1*), the concept of control charts (*2.2*), and patterns analysis in the control charts (*2.3*). In *Section 3*, the basic determinations of data mining methodology are presented, and the role and necessity of its application in the detection of hidden patterns in huge amounts of process data, is emphasized (*3.1*; *3.2*). In addition, one of the possible approaches for the integration of control chart and data mining methods is proposed (*3.3*). The conclusions of this Paper are presented in the last, *Section 4*.

2. *SPC* AND CONTROL CHART METHOD

All processes exhibit variation. Control charts are primary *SPC* tools, used for the analysis and understanding of process variation. Hence, this *Section* focuses on using control charts to stabilize and improve a process.

2.1. Statistical stability of the process

Statistical process control is a proven and widely accepted methodological framework for monitoring, control, and improvement of the performances of manufacturing / service processes. The following are usually emphasized as the main objectives of *SPC*: *(a)* monitoring (usually by sampling) the main parameters of the process; *(b)* detecting process deviation, and *(c)* diagnosing and taking corrective actions, [8]. In addition, *SPC* should not be considered only in the context of outputs of the process, because the focus of *SPC* is the process as a whole and all its parts, from raw materials to the final products (outputs), from one activity to another.

The variability of quality is a universal feature of the processes, which occurs as a result of the effects caused by numerous common and special factors (causes/sources of variation) in the process [7]. Accordingly, process variations can be classified into following two categories: random and systematic. Random variations occur as a consequence of random factors, and they do not alter the essential characteristics of the process. The main characteristics of this type of variations are as follows: ▸ sudden occurrence; ▸ different intensities; ▸ the length of their duration cannot be determined; ▸ they do not affect the process permanently. Systematic factors are non-random, and they cause systematic variations of the quality of the process. These factors are responsible for disturbances in the production process and significantly affect the quality of all products that are produced during the period of their presence and activity. They occur from time to time, and affect the process until they are eliminated and removed from the process.

The main goal and purpose (usefulness) of the *SPC* lies, primarily, in timely detection of the presence of variations in the quality of a process, caused mainly by the systematic factors, followed by a measurement of their intensity, and implementation of specific (proposed) corrective actions. In doing so, the emphasis is on determining whether a process is statistically stable or unstable.

The process is statistically stable (i.e. in a state of a statistical control) if there are only random variations present within it, while the process is statistically unstable (i.e. in a state out-of statistical control), if systematic variations are present along with random variations.

However, most of the processes are not, by themselves and their nature, in a state of a statistical control. Therefore, in order to achieve and ensure the statistical stability of the process, quality managers must dedicate (*"invest"*) considerable time and patience. Otherwise, decisions made on the basis of statistically unstable process can cause highly negative consequences on expected business results.

2.2. The concept of control charts

One of the basic quality improvement tools on which the *SPC* heavily relies are control charts. Control charts are widely accepted statistical tool used to analyze and understand the variations in a process, in order to stabilize and improve the process by detecting and reducing the variations in the behavior of any process quality characteristic. They were originally proposed by Walter Shewhart in the 1920s and, as an extremely useful method for efficient and effective quality management of the processes and products were strongly advocated by Deming. And still today, at the beginning of 21st century, fully justified, the use of control charts is widely present in virtually all situations where quality is (or needs to be) measured.

The methodology of control charts is based on statistical theory, i.e., the theory of distribution and sampling theory. In other words, a number of simple random samples are drawn from the observed process, in successive and strictly defined time intervals. Based on the sample (subgroup) statistics, the elements of control charts are determined and calculated. In fact, all control charts are characterized by a common structure consisting of the following three elements: centerline, upper control limit, and lower control limit. Generally, the centerline in the control chart represents the estimated average value of the monitored process quality characteristic, while the upper and lower control limits are determined by adding to / subtracting from the centerline the value of three standard deviations of an observed statistic (i.e. standard error), respectively. Hence, these control limits are also, often called *three-sigma limits*. In the constructed control chart, consisting of the described elements, data points representing the particular subgroup (sample) statistics of interest are entered, according to the order of performed sample extraction, and then, according to the distribution of presented data points, it can be determined whether the variability of the monitored quality characteristic is within or outside the control limits. In practice control charts are interpreted and used in the following way: if all plotted data points are within the control limits and/or if there are no unusual patterns present in their distribution, the process is under control, and *vice versa*.

According to the type of monitored process variable (quality characteristic) which is measured and then analyzed, control charts are divided into two fundamental groups:

- Control charts for numerical characteristics of the process (of which the most popular are $\overline{X}$, *R*, and *s* charts, including their appropriate combinations);
- Control charts for attribute characteristics of the process (of which the most popular are *p*, *np*, *c*, and *u* charts).

In both cases, the examination of either an attributive or a numeric process quality characteristics, based on the control chart method supported by the relevant expert opinions generally can be used *(1)* to evaluate the history of the process, *(2)* to evaluate the present state of the process, and/or, *(3)* to predict the near future state of the observed process [2].

2.3. Types of patterns in control charts

Analysis of patterns is an important aspect of the process quality analysis based on the control charts. In fact, control chart patterns are the result of the impact of different, non-random (special, assignable, or exogenous) and random (common or endogenous) causes of variations in the process. Their analysis enables identification of *in–control* and *out–of–control* situations in the process data, and provides valuable information for determination of possible potentials for the process improvement.

According to the nature of causes of variations, two basic groups of control chart patterns can be distinguished: *natural* (random) and *unnatural* (non-random), [2].

Natural pattern is the one that does not exhibit any points beyond the defined control limits, tendencies, or some other non-random behavior, and has most of the control chart points near the centerline (approximately, two-thirds of the points are located within a one sigma band of the centerline). Natural process, as a process characterized by a normal pattern, is therefore not affected by special causes of variation. Such process demonstrates functioning of a stable system variation.

The second group of patterns occurs as a result of the influence of special causes of variation. It consists of a series of different patterns and subgroups of patterns (e.g. *Western Electric Company* determined fifteen different types of patterns that can be encountered in practice). Basic determinations and graphical representations of some of the most common types of unnatural patterns can be presented as follows, [2].

- **Shift in Level patterns** – there are three types of shift in level patterns:

▸ *Sudden shift in level* – this type of pattern is characterized by a sudden rise or fall in the level of data regarding the monitored process quality characteristic presented on a control chart.

▸ *Gradual shift in level* – this type of pattern indicates that some portion of the process has been changed, and the effect of this change is manifested as a gradual shift in the average level of the monitored quality characteristic.

▸ *Trend* – it is a slow, steady, continuous change, with increasing or decreasing tendency, of the average level of monitored quality characteristic. This pattern can also be defined as a gradual shift in level that does not settle down.

- **Cycles (Cyclical patterns)**

Cyclical patterns can be defined as a repeating series (waves) of periodic high and low values of the observed characteristic, relative to the centerline, caused by special disturbances that appear and disappear with some degree of regularity.

- **Systematic patterns (Oscillations)**

Unlike the natural pattern where the point-to-point fluctuations are, generally, unsystematic and unpredictable, oscillation (i.e. zigzag, or saw-tooth pattern) is a pattern of alternating high and low subgroup values of the observed quality characteristic (i.e. + – + – +, etc.) relative to the centerline.

- **Wild Patterns** - there are two types of wild patterns: ▸ *freaks* and ▸ *grouping / bunching patterns*. Usually, these patterns can be caused by the calculation errors or by the external disturbances that can dramatically affect one (*freak pattern*) or a "group" or a "bunch" (*grouping/bunching*) of control chart points that are close together. Affected point/s show up on a control chart as point/s significantly beyond the defined control limits or as very different from the rest of subgroups covered by the analysis.

- **Mixture patterns**

This group of patterns is characterized by the absence (or unusually high presence) of control chart points near the centerline (or near or beyond the upper and lower control limits), indicating the presence of large-scale fluctuations.

- **Instability patterns**

Instability is a pattern characterized by large, erratic (up and down) fluctuations (swings) in subgroup statistics mainly caused by one or more special causes of variation that can sporadically affect the average or variability of the observed process. As a result of this type of subgroup statistics distribution, defined control limits appear to be too narrow for the created control chart.

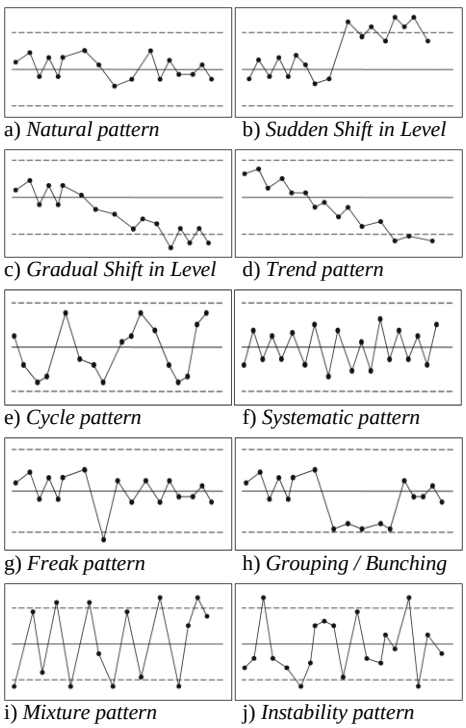

Figure 1.
Examples of some control chart patterns

Generally, according to the shape of its control chart, the particular process is classified into one of several categories, which indicate the presence of possible cause of deviation of this process from its natural status. This procedure, which is implemented within the *SPC*, is often called *Control Chart Pattern Recognition* (*CCPR*). Consequently, *CCPR* provides information to managers and employees that is relevant for the formulation and implementation of necessary corrective actions, aimed at solving the problems that are identified in the process. However, as problem–solving tools, control charts by themselves cannot determine the source of the existing problem. Therefore, expert teams, in search of the causes of the identified problem, must use and rely on some other problem–solving tools in combination with control chart method [9].

3. DATA MINING AND DISCOVERY OF PATTERNS IN CONTROL CHARTS

In recent years, the intensive application of data mining methods has begun in quality diagnosis and quality improvement in manufacturing and service processes. Hence, this *Section* focuses on using the data mining to overcome the problems related to the discovery of (unusual) patterns in the process data.

3.1. The necessity of application of data mining tools in process data

Identification of non–standard patterns, using the control charts, requires collection and organization of process data, determination of control limits, identification and analysis of (unusual) variations exhibited by the process. In modern organizational systems, data are collected electronically and stored in appropriate data repositories. Precisely, these data are potentially useful for the discovery of patterns and knowledge that is relevant in the context of quality improvement of a process and its outputs.

However, due to large amounts of available data, the discovery of patterns and knowledge hidden in the data, without the use of appropriate tools, can be very complicated, often impossible, and hardly feasible proposition (activity). In addition, identification of significant structures in data is further complicated by the fact that, besides data collected through the on-line monitoring of the process, data from

various internal and external sources, primarily collected for some other purposes, are also often used. Emphasized complexity has led to the creation of the need for the application of data mining techniques for extracting useful knowledge from huge amounts of raw data.

On the basis of these considerations, two key reasons that led to the need for integration of control charts and data mining tools can clearly be distinguished. The first reason refers to the limitations in the application of *SPC* tools for identifying the causes of the detected variations. In fact, *SPC* tools have already, for many years, been successfully used to detect non-standard variations in the process, but they, on the other hand, are unable to point to the possible causes of those variations. The second reason refers to the rapidly growing amount of raw process data. Due to the development of information–communication technologies (*ICT*), and storage of huge amounts of data, the detection of hidden patterns by manual analysis, or by using a standard statistical analysis, becomes practically impossible.

3.2. The concept of data mining

In modern business environment, managers make decisions based on the information obtained, among other ways, through the processing of huge amounts of data that result from performing daily, routine activities of the organizational entities, and which do not have, by default, high value and usefulness in decision-making process. Consequently, research and discovery of meaning in large (often highly heterogeneous) data collections through the extraction of (often very small parts of) useful information and knowledge, initiated directing of managerial attention to *data mining* as a methodology which in the decision-making process enables comparison of a large number of options, [6]. Therefore, data mining has become a significant area of interest of quality managers.

Data mining is a sub-(process) of a broader, interactive and iterative process of knowledge discovery from data, or *KDD*, which refers to "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [1]. This process and its constituent components are presented in *Figure 2*.

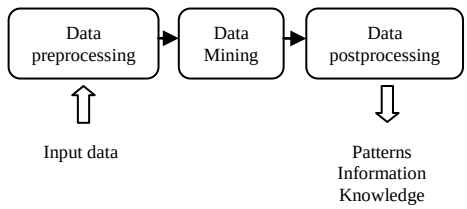

Figure 2. *Knowledge Discovery from Data*

Data mining component refers to the identification of significant patterns or models, hidden deep in the vast mass of data, through the application of highly sophisticated algorithmic methods. In addition, it is necessary to point out three main tasks, for whose implementation data mining methods and tools are usually used: exploration of data, discovery of patterns or models, and prediction. In general, all data mining tasks can be classified into following two categories: *descriptive tasks* (clustering, association rules, sequence discovery, summarization), and *predictive tasks* (classification, regression analysis, time series analysis, and prediction). (For details, see e.g. [3], [4].) For the realization of these tasks, a wide range of methods and tools can be used, generally adopted and combined by data mining from various research areas, such as, primarily, statistics, machine learning, and database management. Generally, in accordance with its interdisciplinary nature, data mining research is extensively and successfully used in many different areas.

It is important to emphasize that during the application of data mining procedures, researchers should be very careful, because data mining is easy to do badly [5].

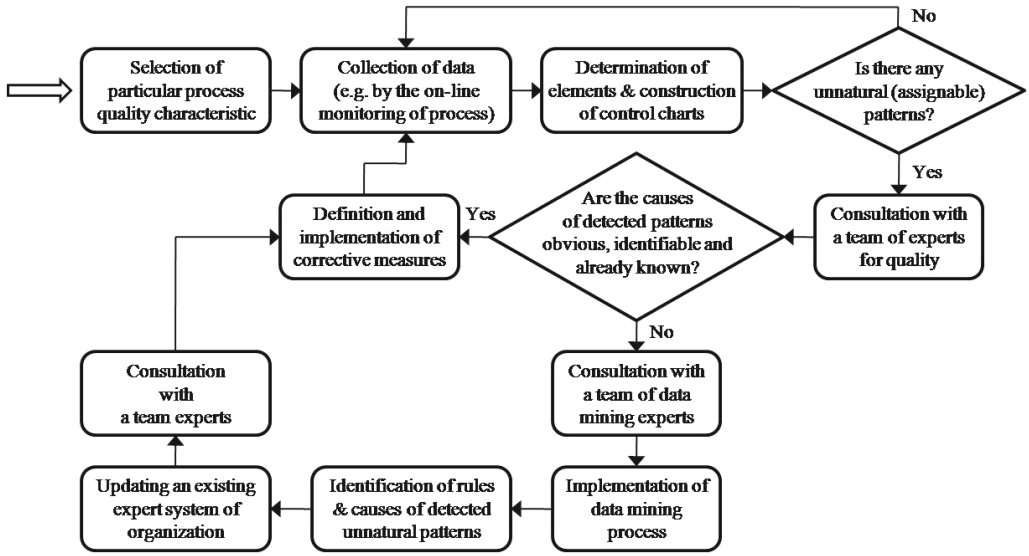

Figure 3. *Flowchart for the approach based on integration of control chart and data mining methods*

In other words, the application of data mining methods, without previously well–conducted preparatory activities, has in literature been marked as a dangerous activity that can easily lead to the discovery of irrelevant (spurious) patterns. In addition, improper use is often criticized and denoted with pejorative terms such as *data dredging*, *data snooping*, *data fishing*, and so on, [4]. It is also pointed out that long enough computer-based searching of large amounts of any data set (even randomly generated), will not only allow the identification of certain patterns that are generally trivial (i.e. irrelevant), but also present certain patterns as statistically significant, although they, realistically, are not that. In fact, data mining may generate thousands of patterns (or rules), but not all of them are significant and interesting. Only an interesting pattern represents knowledge, [3]. Therefore, in the process of evaluation and interpretation, discovered patterns must be analyzed by experts from specific (particular) areas, in order to determine which of them really represent (new) knowledge.

3.3. Integration of control chart and data mining methods

Starting from the considerations outlined above and the fact that through the integration of control charts and data mining analysis, elements can be provided to solve the problems related to the quality of the process, the authors propose one of the possible approaches for integration of these problem-solving tools. The proposed approach, for control and improvement of the quality of the process, based on this integration, is presented in *Figure 3*.

The key determinations of the proposed approach are:

- it is an interactive and iterative process that can be applied to every process and sub-process;
- control chart method provides preliminary information about the (in)stability of the process;
- the identification of causes of the detected unnatural patterns is based on the engagement of appropriate data mining techniques in the framework of the integral data mining process;
- for successful detection of hidden

patterns, understanding the causes of their occurrence, and generation of (new) knowledge, a creative team approach is necessary (essential) in consideration and examination of quality characteristic, which is based on the tight collaboration of domain experts, quality experts, and data mining experts;

- as the result of integration of knowledge extracted by data mining analysis and existing expert system of the organization, data mining is becoming an important support in the process of making optimal decisions regarding the realization of quality tasks.

4. CONCLUSION

Based on the presented considerations, the results of the Paper can be summarized in the form of the following concluding remarks. *SPC* approach has been, for many years, used extensively and widely for the analysis of the quality characteristics of a process, within which control charts are marked as a primary tool for understanding of process variation. Analysis of patterns, based on control charts, enables the identification of non-standard variations (unusual patterns) in process data, which contain valuable information in the context of determining the potential for process improvement. However, control charts do not provide the identification of the causes of quality problems, which are manifested in the form of unusual patterns. Due to the above stated, including the fact that, in the information era, as a result of performing daily activities, the amount of stored and available (process) data rapidly grows, the application of data mining methods for detection of, in a sea of raw process data, hidden, interesting patterns (useful knowledge), is necessary and unavoidable. Data mining is therefore a rapidly expanding scientific field with growing interests and importance in quality control and improvement, as an application area where it can provide significant sources of competitive advantage. Accordingly, in the Paper, the approach for recognizing unusual patterns and their causes (sources) based on the integration of control chart and data mining methods is proposed and described.

REFERENCES

[1] Fayyad, U., Shapiro, G. P., and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, AAAI Press, 17(3), pp 37–54.
[2] Gitlow, H., Oppenheim, A., and Oppenheim, R. (1995). *Quality Mnagement: Tools and Methods for Improvement*, 2nd edition, Irwin/McGraw-Hill.
[3] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*, 2nd edition, Morgan Kaufmann Publishers (Elsevier Inc.), USA.
[4] Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*, MIT Press, UK.
[5] Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley and Sons Ltd, New York, USA.
[6] Milanović, M. and Stamenković, M. (2011). "Data Mining in Time Series", *Economic Horizons*, Faculty of Economics, University of Kragujevac, Vol. 13, N° 1, pp 5-25.
[7] Milanović, M. and Stamenković, M. (2011). "The Power of Statistical Thinking in Quality Improvement", in the Proceedings of the *5th IQC*, Center for Quality, Faculty of Engineering, University of Kragujevac, May, 20 th 2011, pp 215-223.
[8] Montgomery, D. C. (1996). *Introduction to statistical quality control*, John Wiley, NY.
[9] Ramana, E.V. and Reddy, P.R. (2012). "Integration of Control Charts and Data Mining for Process Control and Quality Improvement", *IJAET*, Vol. 2, Issue 1, pp 640-648.