

**PROPOSAL OF USING SELECTED DATA MINING METHODS
IN SIX SIGMA METHODOLOGY**

Andrej Trnka¹

¹Department of applied informatics, University of ss. Cyril and Methodius in Trnava,
Nam. J. Herdu 2, Trnava 917 01, Slovak republic

Abstract

Data Mining methods give a lot of opportunities, how to change results in many sectors of manufacturing process. When manufacturing process is based on Six Sigma methodology, we can use some methods of Data Mining to improve results in DMAIC phases. Six Sigma is a business management strategy. Six Sigma seeks to improve the quality of process outputs by identifying and removing the causes of defects and minimizing variability in manufacturing and business processes. It consists of 5 phases. A lot of mathematical and statistical methods are used in these phases. With implementation Data Mining methods into these phases we can improve the level performance of Six Sigma methodology.

Some examples of implementation of Data Mining methods to Six Sigma methodology are nowadays tested.

Key words: *Six Sigma, DMAIC, Data Mining, Knowledge discovery in databases, KDD*

1. INTRODUCTION TO SIX SIGMA

During the early part of the 1980s, Total Quality Management (TQM), especially in the United States and Europe, concentrated on the application of quality management principles and SPC tools within all aspects of the organisation. The focus was upon their integration with key business processes. There were considerable successes in communication and culture but despite this something was missing in the rigour of project management principles. DMAIC (Define, Measure, Analyse, Improve, Control) was introduced by Motorola as the life cycle discipline for Six Sigma projects in the late 1980s. Since then, DMAIC has become the essential component of all Six Sigma initiatives and training programmes. In order to emphasise the sustainability of gains during the Control phase, some companies, (e.g. Dow Chemical) have added the process of Leverage (L) after Control and thus extended DMAIC to become DMAICL. While Demming's PDCA (Plan, Do, Check, Act) cycle has been extensively used in the development and deployment of quality policies, DMAIC has added the rigour of project life cycle to the implementation and closeout of Six Sigma projects. Figure 1 shows the relationship between DMAIC with PDCA and a typical project life cycle. (Basu, 2010)

A production process is, by definition, a chain of tasks, a sequence of events through which products are transformed from raw material to final goods. The tasks operate as customers and suppliers, with each task being a customer to the task upstream and a supplier to the task downstream. The quality of the product or service generated by each task depends on the variability at the current task and the combined variability of all the tasks upstream. Consequently, the quality of the final product depends on the combined variability that is exhibited by every step in the production process. To better control the quality level of the process yield and reduce operational expenses, it is necessary to control the potential errors at each step of the process and the overall variability of the process itself.

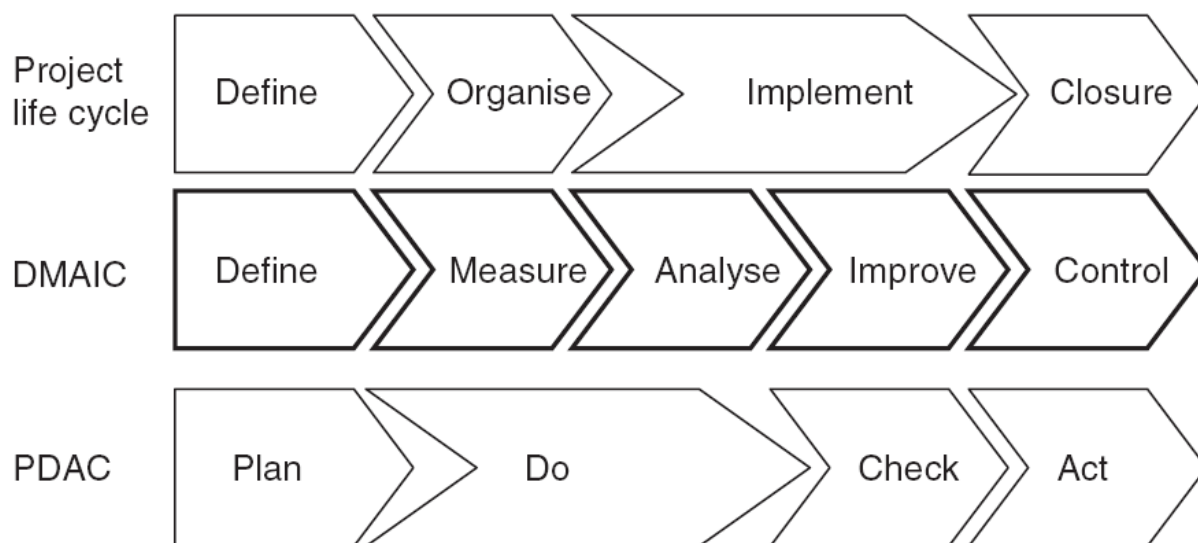


Fig. 1. Comparing DMAIC cycle with others approaches (Basu, 2010)

Quality targets are set for each step in a process and metrics are used to assess the performance of the steps and estimate the probability for a unit of product to go through every step free from defects. The metrics that are generally used in Six Sigma to assess quality are the total containment effectiveness (TCE), the defect per unit (DPU), the defect per opportunity (DPO), the defect per million opportunities (DPMO) and others. These metrics are calculated to determine the probability for a defect to occur at a given step of a production process. The Poisson and normal distributions are primarily used to determine the probability for defects to occur.

Total containment effectiveness (TCE). Since the objective of process improvement is to eliminate defects all across the production systems, the effectiveness of the defects tracking method should reveal the number of defects found at every step and the defects found after the final products have been released. The metric used to measure the effectiveness of the defects tracking method is the TCE. It compares the number of defects and errors prior to the product release to the total number of defects and errors after release.

The difference between errors and defects is that errors are found before the product leaves the processing step where they occurred, while defects are errors caught after the product leaves the step where they occurred.

$$TCE = \frac{\text{prerelease_defects} + \text{errors}}{\text{prerelease_defects} + \text{release_defects} + \text{errors}} * 100\%$$

The objective is to minimize the number of defects across the production system but in terms of a defects tracking process, the aim should be to have a high TCE because it would be a sign that more defects and errors are caught before the products are released.

Defect per unit (DPU). A defect is defined as a CTQ that does not meet the specified standard. If the measurement of an object is critical to quality and the specification for the CTQ is that it must fall within the interval [2, 4], any measurement that falls outside that interval will be considered a defect.

Most manufactured products have multiple CTQs; therefore, a defective product can have multiple defects.

The DPU is the most basic defects metric. It is the ratio of the number of defects found to the total units of products sent to customers:

$$DPU = \frac{\text{total_defects}}{\text{total_unit}} = \frac{D}{U}$$

Defects per opportunity (DPO). Every product or service exhibits multiple CTQs and has to be free of defects for those products and services to meet customers expectations. An opportunity is defined as any part of a product or a service where a defect can be found; in other words, opportunities are those CTQs that the customers expect to receive without any defects.

Total_opportunities is the total number of opportunities (O) that are found on a set of units (U):

$$\text{Total_opportunities} = O * U$$

DPO will be equal to the ratio of the defects found on all the units to the total number of opportunities:

$$DPO = \frac{\text{total_defects}}{O * U} = \frac{DPU}{O}$$

Defects per million opportunities (DPMO). The DPMO measures the number of defects that the process generates for every million opportunities (Bass, 2009):

$$DPMO = DPO * 10^6$$

The most important metric is DPMO. DPMO of 6th performance level of Six Sigma methodology is 3,4. Table 1 show others performance levels (DPMO) with the yield of the process.

Table 1. Six Sigma performance level

Sigma	DPMO	Yield (%)
1	690 000	31
2	308 000	69,2
3	66 800	93,32
4	6 210	99,379
5	230	99,977
6	3,4	99,9997

It is necessary to obtain maximum performance level because of the yield. Normal production process reaches the Sigma performance level about 4-5.

1.1. DMAIC phases

When we are trying to control existing process, we have to use DMAIC cycle. We can also utilize DMADV cycle. This cycle is used to control a new process. In our research, we seek to improve the existing process, so we focus on DMAIC cycle.

DMAIC is an acronym for Define, Measure, Analyze, Improve and Control.

Define: Define is the first step in the process. In this step, it is important to define specific goals in achieving outcomes that are consistent with both your customer's demands and your own business's strategy. In essence, you are laying down a road map for accomplishment.

Measure: In order to determine whether or not defects have been reduced, you need a base measurement. In this step, accurate measurements must be made and relevant data must be collected so that future comparisons can be measured to determine whether or not defects have been reduced.

Analyze: Analysis is extremely important to determine relationships and the factors of causality. If you are trying to understand how to fix a problem, cause and effect is extremely necessary and must be considered.

Improve: Making improvements or optimizing your processes based on measurements and analysis can ensure that defects are lowered and processes are streamlined.

Control: This is the last step in the DMAIC methodology. Control ensures that any variances stand out and are corrected before they can influence a process negatively causing defects. Controls can be in the form of pilot runs to determine if the processes are capable and then once data is collected, a process can transition into standard production. However, continued measurement and analysis must ensue to keep processes on track and free of defects below the Six Sigma limit. (Tech-FAQ, 2009)

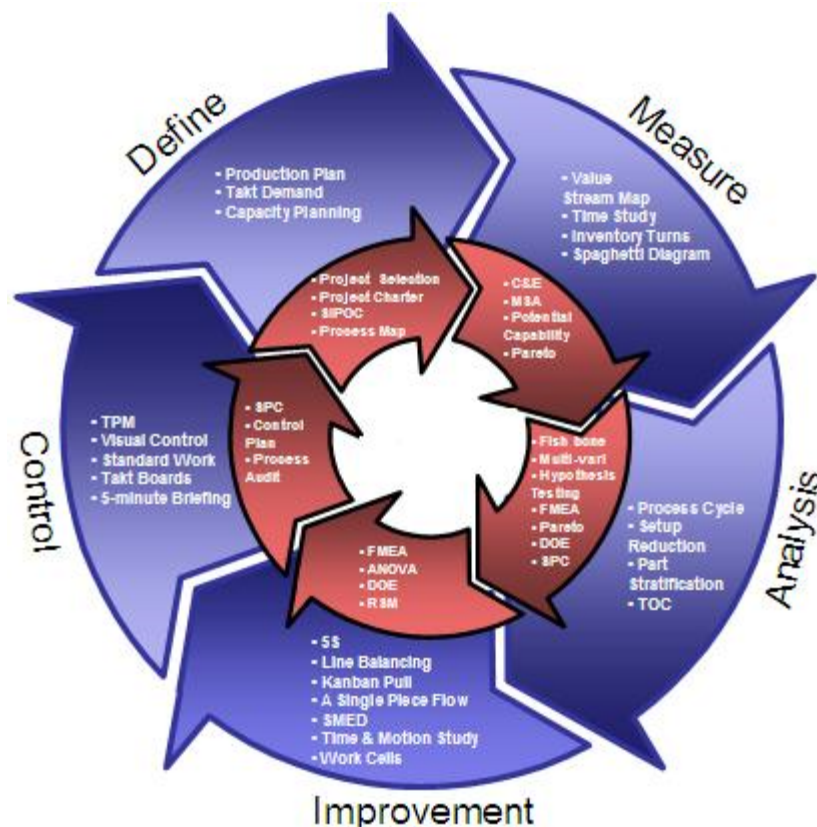


Fig. 2. DMAIC cycle with tools

As shown in fig. 2, these phases work and deal with a lot of statistical and mathematical tools or methods, so there is a possibility to use the Data Mining methods to improve results.

2. KNOWLEDGE DISCOVERY IN DATABASES AND DATA MINING

Knowledge Discovery in Databases has been defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data. The KDD process (fig. 3) is iterative and interactive, consisting of nine steps. Note that the process is iterative at each step, meaning that moving back to previous steps may be required. The process starts with determining the KDD goals, and ends with the implementation of the discovered knowledge. (Bramer, 2007) (Maimon, 2005)

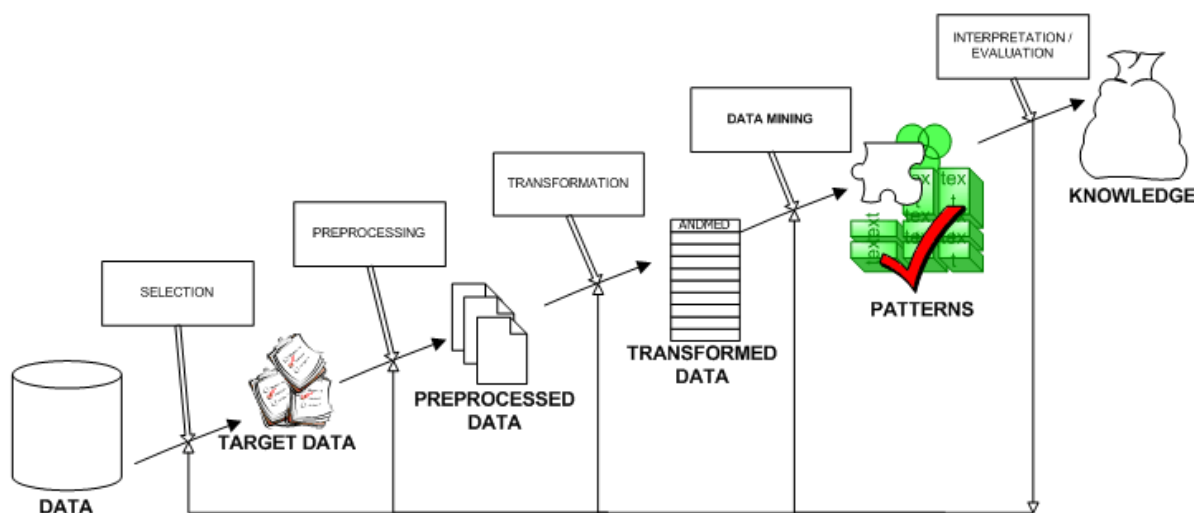


Fig. 3. KDD process

It is important to note, that Data Mining is only a part of KDD process.

Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. (Larose, 2005)

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities.

In data mining, the data is stored electronically and the search is automated – or at least augmented – by computer. Even this is not particularly new. Economists, statisticians, forecasters, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction. What is new is the staggering increase in opportunities for finding patterns in data. The unbridled growth of databases in recent years, databases on such everyday activities as customer choices, brings data mining to the forefront of new business technologies or methodologies. (Chakrabarti, 2009)

Six Sigma methodology seems to be one of those methodologies, where the Data Mining methods could be successfully and efficiently applied. This was the reason, why we decided to implement Data Mining methods to Six Sigma methodology.

All the new implementations to Six Sigma methodology have to be optimized and validated. (Schreiber, 2008)

3. MODELING TOOLS

All implementation is based on the IBM SPSS Modeler, which was chosen as a modelling tool because it is the most widely used tool for Data Mining.

IBM SPSS Modeler (formerly Clementine) makes it easy to discover insights in any data. Its simple graphical interface puts the power of data mining in the hands of business users and high-performance increases the analyst productivity. IBM SPSS Modeler has a wide range of users, from IT technology centric users to business users, requires no programming skill to reap the benefits of data mining, and supports the whole data mining process with comprehensive facilities. (SPSS, 2010)

The other tools we used in Six Sigma methodology for modelling or predicting:

- Crystal Ball
- SPC for Excel
- Six Sigma for Excel

The disadvantages of these three tools are in inability to make new potential patterns from data and obtained knowledge.

4. SOME EXAMPLES OF IMPLEMENTATION

We tried to implement Data Mining methods to all the phases of DMAIC cycle. Our analysis showed that the best implementation is into the phases Define, Measure and Control. Statistical and mathematical methods are used in these phases. With the following Data Mining methods (or tasks) we can improve the results of DMAIC phases.

Define:

- Identifying promotional target customers (with Regency-Frequency-Monetary analysis)
- Making offers to customers (with Self-Learning Response model)

Measure:

- Preparing data for analysis
- Feature selection of fields
- Modelling for a binary target
- Modelling for a numeric range target

Analyze:

- Condition monitoring of machines

Improve:

- Market basket analysis

Control:

- Fraud detection
- Defects prediction

In example we focus to the Control phase and we explain defect prediction.

1.1. Defects prediction

To predict defects in manufacturing process we need historical manufacturing data (e.g. how many products were made, how many defects were made, when the products and defects were made ...). From these historical data we can plot a time series graph and use suitable model to predict defects from the process.

The watched process is about drill of the holes. The diameter measurements were executed every minute. The time series of this process is showed in fig. 4.

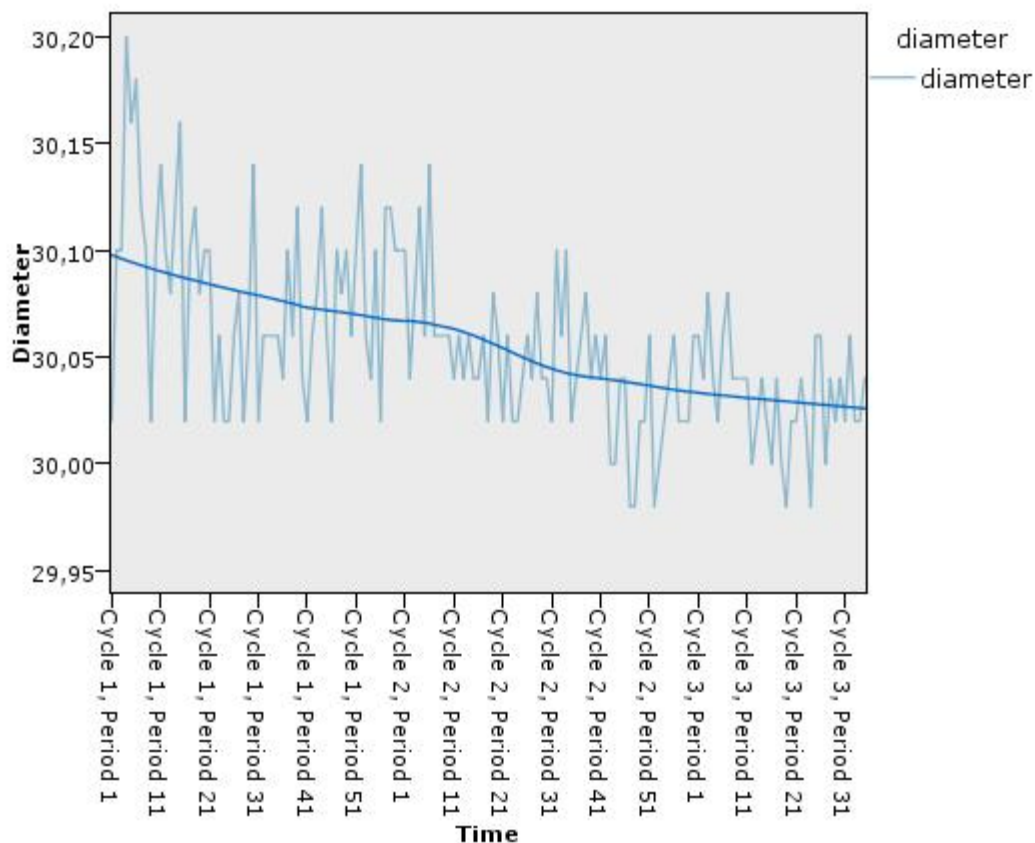


Fig. 4. Time series of the hole diameter

We have to predict, when will the process become poor. The Upper Specification Level is 30,2 mm and the Lower Specification Level is 29,96 mm. We used exponential smoothing to choose the best predict model.

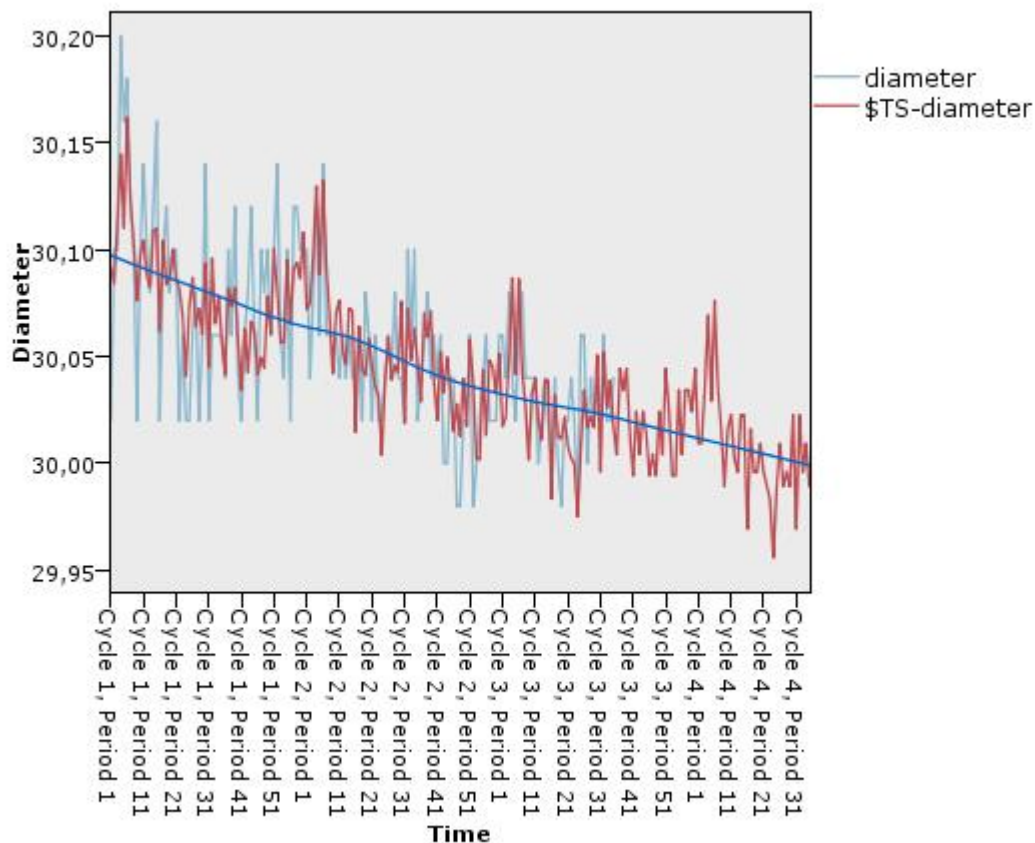


Fig. 5. Time series with prediction

Fig. 5 shows prediction of diameter from Cycle 3. The Cycle corresponds with hour and the Period corresponds with minute. In cycle 4 is clear that the defect will happen. So, it is the time to make intervention to the process. This prediction improves the process with reduction of defects.

Fig. 6 shows the proposed time series model in IBM SPSS Modeler.

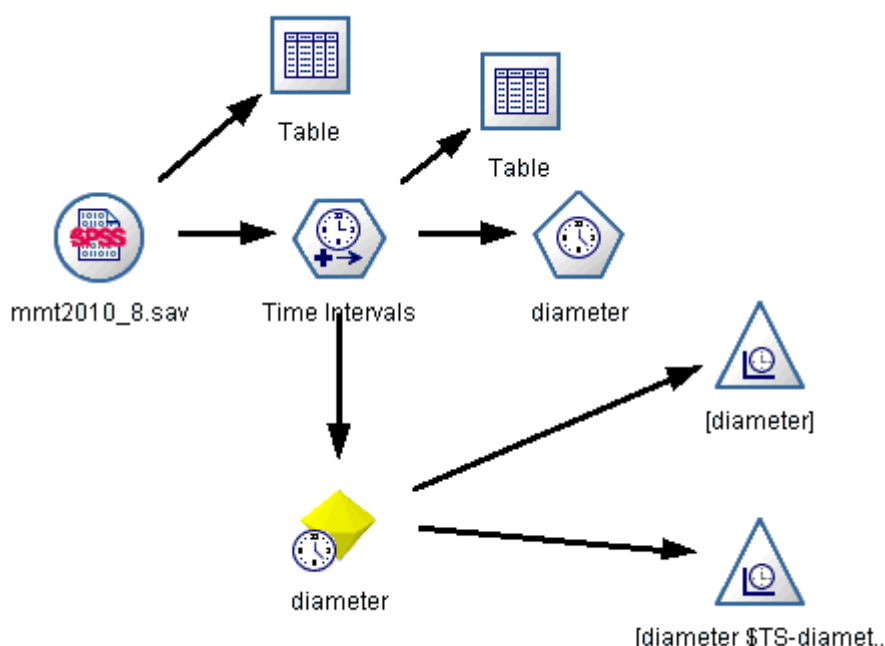


Fig. 6. Created model in IBM PASW MODELER

REFERENCES

1. Basu, R. (2010) Implementing Six Sigma and Lean: A Practical Guide to Tools and Techniques. Elsevier Ltd., 349 p., ISBN 978-1-8561-7520-3
2. Bass, I., Lawton, B. (2009) Lean Six Sigma: Using SigmaXL and Minitab. McGraw-Hill Companies, 362 p., ISBN 978-0-07-162130-4
3. Tech-FAQ (2009). What is DMAIC. [online] [cited 11.1.2010]. Available at <www.tech-faq.com/dmaic.shtml>
4. Chakrabarti, S., et. al (2009) Data Mining: Know it all. Elsevier Inc., 460 p., ISBN 978-0-12-374629-0
5. Larose, D. (2005) Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley, 222 p., ISBN 0-471-66657-2
6. Bramer, M. (2007) Principles of Data Mining. Springer, 343 p., ISBN 1-84628-765-0
7. Maimon, O., Rokach, L. (2005) Introduction to Knowledge Discovery in Databases. Data mining and Knowledge Discovery Handbook. Edited by Maimon O., Rokach L., p. 1-17, ISBN 0-387-24435-2
8. IBM SPSS Modeler - Data mining, text mining, predictive analysis (2010), [online] [cited 13.1.2010]. Available at <<http://spss.com/software/modeling/modeler>>
9. Schreiber, P., Vazan, P., Tanuska, P. (2008) Production optimization by using genetic algorithms and simulation model of production system. In: Annals of Daaam for 2008 & Proceedings of the 19th International Daaam Symposium - Intelligent Manufacturing & Automation: Focus on Next Generation of Intelligent Systems and Solutions; 2008 October 22-25; Trnava, Slovakia. Daaam Int Vienna, Vienna Univ Technology, p. 1229-1230, ISBN 978-3-901509-68-1.